8874–8882 Nucleic Acids Research, 2012, Vol. 40, No. 18 Published online 25 July 2012
doi:10.1093/nar/gks679

Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels

Mario Hlevnjak, Anton A. Polyansky and Bojan Zagrovic*

Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Vienna 1030, Austria

Received April 24, 2012; Revised June 11, 2012; Accepted June 20, 2012

ABSTRACT

A potential connection between physico-chemical properties of mRNAs and cognate proteins, with implications concerning both the origin of the genetic code and mRNA–protein interactions, is unexplored. We compare pyrimidine content of naturally occurring mRNA coding sequences with the propensity of cognate protein sequences to interact with pyrimidines. The latter is captured by polar requirement, a measure of solubility of amino acids in aqueous solutions of pyridines, heterocycles closely related to pyrimidines. We find that the higher the pyrimidine content of an mRNA, the stronger the average propensity of its cognate protein's amino acids to interact with pyridines. Moreover, window-averaged pyrimidine profiles of individual mRNAs strongly mirror polar-requirement profiles of cognate protein sequences. For example, 4953 human proteins exhibit a correlation between the two with $|R| > 0.8$. In other words, pyrimidine-rich mRNA regions quantitatively correspond to regions in cognate proteins containing residues soluble in pyrimidine mimetics and *vice versa*. Finally, by studying randomized genetic code variants we show that the universal genetic code is highly optimized to preserve these correlations. Overall, our findings redefine the stereo-chemical hypothesis concerning code's origin and provide evidence of direct complementary interactions between mRNAs and cognate proteins before development of ribosomal decoding, but also presently, especially if both are unstructured.

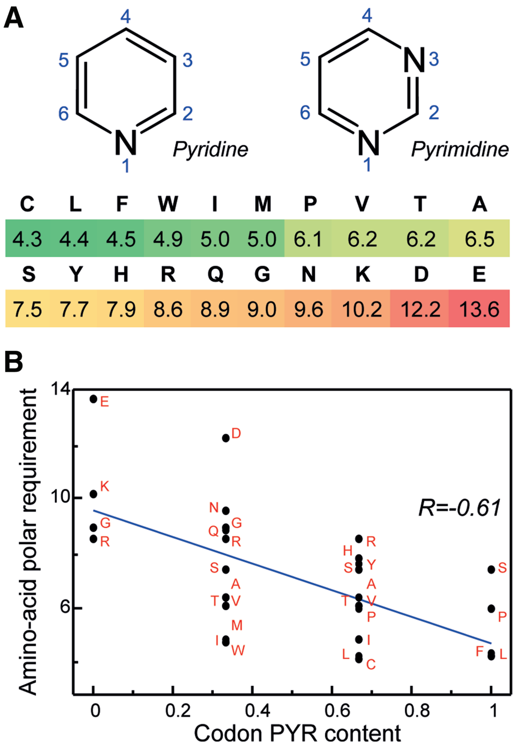
INTRODUCTION

The universal genetic code is the central building block at the foundation of all of life as we currently know it, but the problem of how it evolved is still completely open (1,2). In fact, it has been argued that the origin of the genetic code, i.e. of translation is one of the central foundational questions in molecular biology that are still unresolved (3). Obviously, the origin of the code must be related to its overall structure and function. Early on it was noticed that the structure of the genetic code is highly non-random: codons coding for the same amino acid resemble each other, and so do the codons for different, but chemically similar amino acids (1,4). This suggests that the code somehow embodies features that go beyond a simple mapping between codons and amino acids, and that it might actually also be in service of other, higher-level functions. For example, Lesnik and Reiss (5) showed that hydropathy profiles of membrane proteins are related to sequence profiles of the thymine/adenine ratio in their genes, i.e. uracil/adenine ratio in their mRNAs. More recently, Prilusky and Bibi (6) showed the same when it comes to mRNA uracil-density profiles, and hypothesized that there might exist cellular factors which recognize uracil-rich regions and are responsible for membrane targeting of membrane proteins' mRNA. Moreover, it has recently become apparent that mRNA, in addition to its coding function, plays a multitude of other roles in the cell, such as ensuring proper protein localization (7), determining the rate of translation (8,9) or even affecting the ways nascent peptides get post-translationally modified (10).

Concerning the evolutionary development of the genetic code, there exist several main classes of hypotheses. The 'frozen-accident hypothesis' suggests that the code is essentially inherited from the last universal common

*To whom correspondence should be addressed. Tel: +43 1 4277 52271; Fax: +43 1 4277 9522; Email: bojan.zagrovic@univie.ac.at

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



C	L	F	W	I	M	P	V	T	A
4.3	4.4	4.5	4.9	5.0	5.0	6.1	6.2	6.2	6.5
S	**Y**	**H**	**R**	**Q**	**G**	**N**	**K**	**D**	**E**
7.5	7.7	7.9	8.6	8.9	9.0	9.6	10.2	12.2	13.6

Figure 1. Pyrimidine content of individual codons versus PR of cognate amino acids. (**A**) chemical structure of pyridines and pyrimidines with the PR scale (23) for the 20 natural amino acids. (**B**) correlation between the pyrimidine content of individual codons, as based on the universal genetic code, and the PR of cognate amino acids.

ancestor and could simply not evolve further as any change would be detrimental to a large number of proteins (2,11). Second, the 'coevolution hypothesis' proposes that the code coevolved together with amino acid biosynthetic networks and that similar codons code for metabolically connected amino acids (2,12). On the other hand, the 'error minimization hypothesis' is based on computational analyses which have shown that the code is robust to point mutations and translational errors, such that certain properties of the original and the mutated amino acids, like hydrophobicity, differ to a minimal degree (2,13,14). Finally, the 'stereo-chemical hypothesis' suggests that the principal feature of translation before the development of ribosomal decoding machinery was a direct interaction between codons and amino acids they code for (2,15–19). Although binding of individual codons and cognate amino acids has never been observed, analysis of amino acid-binding RNA aptamers and different RNA–protein complexes has revealed that some codons are present at a higher-than-random frequency in the binding sites of cognate amino acids (e.g. isoleucine and arginine) (18–20).

Importantly, the original support for the stereo-chemical hypothesis came from the analysis of polar requirement (PR), an experimental measure of the way amino acids partition in aqueous solutions of substituted pyridines, nitrogenous bases closely related to pyrimidines (Figure 1A) (21–23). Depending on type, amino acids were shown to exhibit different, clearly defined preferences for interacting with pyridines and this was then used not only as evidence supporting the stereo-chemical hypothesis, but also as an explanation for the general structure of the genetic code itself (21,22). Namely, when grouped according to the PR of their cognate amino acids, compositionally similar codons automatically cluster together (21). More recently, Mathew and Luthey-Schulten employed molecular dynamics simulations to provide a microscopic picture behind the PR scale and also to improve it, albeit marginally, by avoiding some experimental artifacts (Pearson $R = 0.95$ between the two scales) (23). Their PR scale, derived using 2,6-dimethylpyridine and used in all of our analysis (if not indicated otherwise), is given in Figure 1A (the higher the PR, the lower the affinity for pyridines and *vice versa*). A very similar ordering is obtained for pure pyridines and 2-methylpyridines as well (21).

While it has been recognized that individual amino acids encoded by pyrimidine-rich codons generally have low PR, i.e. a relatively high affinity for pyridines, and *vice versa* (21,22), this connection has never been quantitatively explored. Moreover, thanks to the modern large-scale sequencing efforts, one can now also examine a potential connection between the pyrimidine content of complete mRNA-coding regions and the PR of cognate protein sequences on the whole proteome level. There are several reasons why this might be important. First, regardless of how the genetic code evolved, its essence has always been to encode sequence properties of one large polymer (protein) in sequence properties of another polymer (mRNA). Second, if the direct binding between amino acids and their codons exists, but is weak, the cooperative effect of combining multiple such interactions in the context of long polymers may yield a stronger net effect. Finally, depending on codon usage bias and specific amino acid sequences in question, the strength of correlation seen just on the level of the genetic code table might significantly change in either direction. Motivated by these rationales, as the central objective in this study we explore the relationship between the pyrimidine content of mRNA-coding sequences and the PR (23) of cognate protein sequences for complete proteomes of 15 different organisms, five from each domain of life (Supplementary Table S1).

MATERIALS AND METHODS

Datasets

Complete proteomes for 15 organisms (five Archaea: *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum*, *Methanocaldococcus jannaschii*, *Methanosarcina acetivorans and Pyrococcus horikoshii*; five Bacteria: *Bacillus subtilis, Escherichia coli, Mycobacterium tuberculosis, Salmonella typhimurium and Synechocystis sp.*, and five Eukaryota: *Arabidopsis thaliana, Drosophila melanogaster, Homo sapiens, Musmusculus and Saccharomyces cerevisiae*) were extracted from UniProtKB (24) database (October 2011 release), with maximal-protein-evidence-level set at 4 (i.e. proteins annotated as 'uncertain' were excluded), and only the reviewed Swiss-Prot entries used for further analysis. The coding sequences of their corresponding mRNAs were extracted using the 'Cross-references' section of each UniProtKB entry, where out of several possible translated RNA or DNA sequences the first one satisfying the length criterion (RNA length = 3 × protein length + 3) was selected and its sequence downloaded from The European Nucleotide Archive (25). The protein as well as RNA sequences with only canonical amino acids or nucleotides were chosen for analysis. Proteins were sorted into mutually exclusive cytosolic and membrane groups using controlled vocabulary within the 'Subcellular location' subsection of each UniProtKB entry, employing the following criteria: membrane proteins are those labeled with any of the 'Membrane', 'Multi-pass membrane protein', 'Single-pass membrane protein', 'Single-pass type I membrane protein', 'Single-pass type II membrane protein', 'Single-pass type III membrane protein' or 'Single-pass type IV membrane protein' identifiers, but are not labeled with the 'Cytoplasm' identifier, while the opposite was used for the cytosolic proteins. Proteins that did not fall into either category were designated as 'other'. Homology filtering was performed using CD-HIT web server (26) by multiple runs with default settings and sequence identity cutoff values for first, second and third run set to 90, 60 and 30%, respectively.

Window-averaging procedure for profile comparison

Sequence profiles were generated using a sliding-window averaging procedure, whereby each position in a given sequence is associated with the average value of the property in question calculated over a window centered at that position. Initially, we tested window sizes in the range of 1–41 residues/codons and calculated the average value of the Pearson correlation coefficient between the mRNA pyrimidine content and protein PR profiles over the entire human proteome. This value attains approximately constant value for all windows greater than or equal to 21 residues (Figure 3A inset), and therefore the window size of 21 residues/codons was used for all further calculations. Note that for all profile comparisons and windows of size $N+1$, the first $N/2$ and the last $N/2$ positions in the compared sequences were not included because for these positions a window of $N+1$ residues is not defined. Similarity between individual mRNA and protein sequence profiles was estimated using the Pearson correlation coefficient between them. Superposition of mRNA and protein profiles was carried out by aligning the average values of the two profiles and rescaling each profile by its standard deviation.

Randomization of the genetic code

Randomized genetic codes were generated by randomly shuffling the 64 codons in the natural genetic code. In this way, each natural codon is mapped to another codon in the randomized code, allowing straightforward rewriting of the mRNA for a given protein. Here, the number of codons for each amino acid remains the same as in the natural code. In total, 10^6 randomized genetic codes were generated for each organism, and compared with the natural code. The reported *P*-values correspond to the fraction of randomized codes for which the absolute value of the Pearson correlation coefficient $|R| > |R_{\mathrm{natural}}|$ for sequence-average or $|<R>| > |R_{\mathrm{natural}}|$ for sequence-profile comparisons.

Amino acid scales and calculation of average properties

A total of 531 amino acid scales, describing different physico-chemical or biological properties of the 20 natural amino acids, were extracted from the AAindex database (27), with additional nine extracted from the literature (23,28–31). The hydrophobicity-related scales were separated from the rest by searching for hydrophobicity-related keywords (hydrophobicity, hydrophilicity, transfer/solvation free energy, polarity, membrane preference/composition, exposure, buriability, accessibility, surface area, partitioning, retention coefficients and variations thereof) in the AAindex database description, as well as by consulting the original literature. The average protein sequence property¸ as defined by a given amino acid scale ($x_1 \ldots x_{20}$), was calculated as a weighted sum where each x_i is weighted by the fraction of residue type i in the sequence. The average protein sequence disorder was calculated by averaging the disorder scores of its residues as predicted by IUPred web server (32). Proteins with average disorder scores >0.5 were considered to be disordered.

RESULTS

We first focus on the relationship between the pyrimidine content of individual codons and the PR of cognate amino acids. To address this, in Figure 1B we plot these two properties against each other using all 61 coding triplets from the universal genetic code and all 20 naturally occurring amino acids. Several patterns become immediately obvious. First, the general inverse dependence between the two properties is apparent with the Pearson correlation coefficient $R = -0.61$, i.e. pyrimidine-rich codons do preferentially code for amino acids having higher affinity for pyrimidine mimetics as captured by PR. However, just on the level of the genetic code table, this relationship is not overly quantitative with a large level of scatter around the central trend line (Figure 1B). Second, some amino acids such as arginine are coded for by multiple codons with significantly different pyrimidine content having an adverse impact on the observed correlation. Finally, there are amino acids, such as tryptophan and methionine, which clearly deviate from the trend followed by other amino acids in that they exhibit low PR, while being coded for by purine-rich codons. However, it should be mentioned that tryptophan in particular is thought to have appeared later in evolution (33) and the definition of its codon may have been influenced by different factors than for other amino acids.

How does the correlation seen just on the level of the genetic code table change if one looks at the average pyrimidine content of realistic mRNA coding sequences and

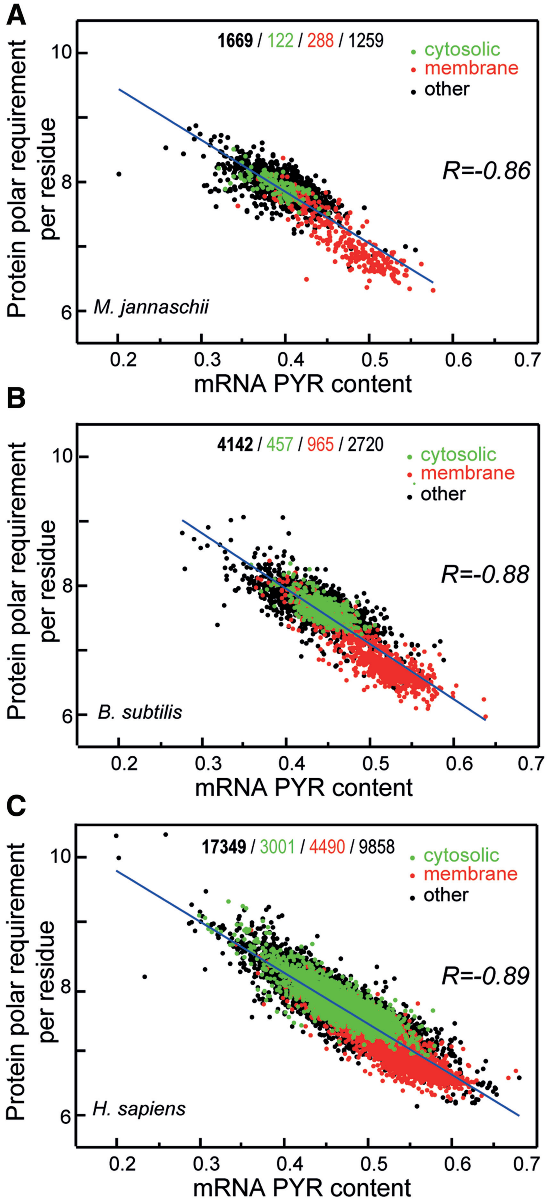


Figure 2. Relationship between the average pyrimidine content of mRNAs and the average PR of cognate proteins. Correlation between the mRNA coding-sequence pyrimidine content and the average PR of cognate protein sequences for the complete proteomes of: (**A**) *M. jannaschii*, (**B**) *B. subtilis* and (**C**) *H. sapiens*. We indicate the total number of proteins (all/cytosolic/membrane/other) and Pearson correlation coefficients over all proteins for each organism. Annotated cytosolic and membrane proteins are depicted in green and red, respectively, while all other proteins are in black.

compares it with the average PR of cognate protein sequences? In principle, depending on codon usage bias and specific sequence composition, one could expect the above level of correlation to change in either direction. To address this question, we have analyzed complete proteomes of 15 different organisms, and as representative examples, in Figure 2A–C we show the results for the proteomes of *M. jannaschii, B. subtilis* and *H. sapiens,* respectively. Similar results hold for all of the proteomes examined (Supplementary Table S2). Remarkably, regardless of the organism studied, the average pyrimidine content of mRNA coding sequences exhibits an extremely strong inverse correlation with the average PR of cognate protein sequences over complete proteomes. For example, the two exhibit a correlation with $R = -0.89$ over 17 349 annotated human proteins (Figure 2C), and similar results are seen for all other organisms. What is more, the results do not change if the protein homology level within a given proteome is reduced to 30%. For example, in such homology-filtered human proteome, including now 10 113 proteins instead of the original 17 349, the correlation changes to $R = -0.88$. Overall, such strong, universally found correlations attest to a curious, close-to-quantitative correspondence: the higher the propensity of a given protein's amino acids to interact with pyrimidine mimetics in water mixtures, the more pyrimidine-rich is its cognate mRNA coding sequence and *vice versa*. What is more, moderate tendencies embodied in the genetic code (Figure 1B) get significantly amplified if one analyzes the average features of realistic mRNA and protein sequences on the whole-proteome scale, thus accounting for both sequence composition and codon usage bias. In the context of the stereo-chemical hypothesis, these findings can be thought of as a coarse-grained, generalized analog of Chargaff's complementarity rules for DNA. Namely, the adenine or guanine level in DNA is quantitatively predictive of the level of thymines and cytosines, i.e. precisely those bases which have affinity for adenines and guanines, respectively. Similarly, here we show that the pyrimidine level in mRNAs is quantitatively predictive of the PR-weighted affinity for pyrimidine mimetics of cognate protein sequences, hinting at a possibility of complementarity between the two.

To explore this idea more closely, we have evaluated correlation coefficients between window-averaged pyrimidine profiles of individual mRNA coding sequences and PR profiles of cognate protein sequences for all 15 proteomes. We use the averaging window of 21 amino acids/codons, but similar results are obtained for all other windows above 16 (Figure 3A, inset). Remarkably, the distributions of the thus-obtained correlations coefficients for all 15 species show great similarity, with the average correlation coefficient typically in excess of 0.7 in absolute value (Figure 3A). For example, human mRNA pyrimidine and protein PR profiles exhibit the average correlation coefficient of -0.73 (median of -0.74), with 28.5% of sequences (4953 out of 17 349 proteins) having a correlation of $|R| > 0.8$! Again, these results are not affected if the protein homology level within a given proteome is lowered to 30%. For example, for the human proteome the average correlation coefficient remains at -0.73 and the median changes to -0.75 after homology filtering.

To illustrate what these correlation coefficients mean when it comes to actual sequences of typical length, we show the pyrimidine mRNA profile overlaid with the respective protein PR profile for serine/threonine protein kinase VRK1, a human cytosolic protein whose correlation coefficient between the two profiles ($R = -0.74$) is the same as the median over all proteins, i.e. the most representative protein (Figure 3B, upper panel). Second,

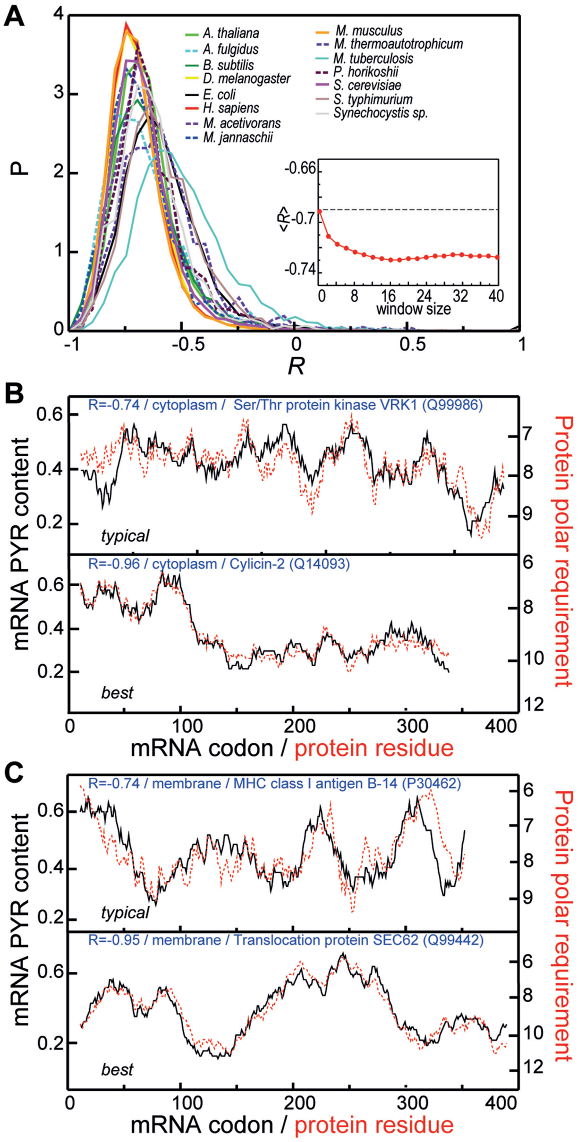


Figure 3. mRNA coding-sequence pyrimidine profiles mirror protein sequence PR profiles. (**A**) distributions of correlation coefficients R between window-averaged pyrimidine-content profiles of individual mRNA-coding sequences and window-averaged PR sequence profiles of the respective proteins for all 15 proteomes (window size = 21). In the inset, we show the dependence of the average R for the human proteome on the size of the averaging window. (**B**) Typical and best-matching pairs of mRNA pyrimidine-content and protein PR profiles for human cytosolic proteins. (**C**) Same as in 3B, but for human membrane proteins. All proteins in B and C were chosen to be of similar, representative length (300–400 residues).

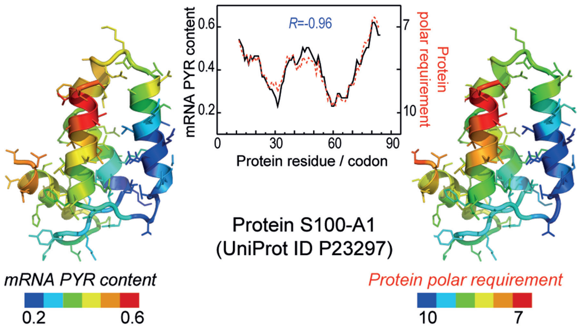


Figure 4. Structure-mapped sequence profiles. NMR structure of protein S100-A1 (PDB code: 2L0P) colored according to the window-averaged mRNA coding-sequence pyrimidine content (left) or window-averaged protein-sequence PR (right), together with the superimposed sequence profiles of the two variables.

we show the same superposition for the best cytosolic protein sequence in this regard, cylicin-2, with $R = -0.96$ (Figure 3B, lower panel). Likewise, in Figure 3C we do the same for the most representative (MHC class I antigen B-14, $R = -0.74$, upper panel) and the best-performing (translocation protein SEC62, $R = -0.96$, lower panel) human membrane protein, respectively. Clearly, both types of profiles exhibit strong characteristic features with peaks and valleys on the scale of tens to hundreds of residues/codons. More importantly, there is a remarkable degree of matching between mRNA coding-sequence pyrimidine profiles and the respective protein PR profiles even for average human proteins. Despite the completely different chemical nature of mRNAs and proteins, these two key biopolymers exhibit a striking complementarity when it comes to their physico-chemical properties: sequence profiles of pyrimidine density in mRNAs strongly mirror the sequence profiles of cognate proteins capturing their affinity for pyrimidine mimetics. Similar results are obtained for all the species studied, except to a smaller degree for *M. tuberculosis* where the average correlation coefficient is reduced by ~0.2 (Figure 3A). As *M. tuberculosis* utilizes the same universal genetic code as other organisms, this difference attests to an important feature of the above findings. Namely, it is a combination of the genetic code together with codon usage bias and specific sequence composition that all together determine the level of matching seen in different cases. Finally, to further illustrate the level of matching observed, in Figure 4 we show the experimental NMR structure of the cytosolic human protein S100-A1 (PDB code: 2L0P(34)) where we color its residues according to either window-averaged pyrimidine content of its mRNA or window-averaged amino acid PR: the similarity between the two is evident (Figure 4). A detailed analysis of the relationship between the level of matching seen in different proteins and their functional and structural characteristics or evolutionary age will be presented elsewhere.

How optimized is the universal genetic code to preserve the above relationships between mRNA pyrimidine content and protein sequence PR? To study this, for each proteome we have generated 10^6 randomized genetic codes, and evaluated the above correlations for each one of them, both on the level of sequence-averages and complete sequence profiles (Figure 5A and B). Remarkably, for human proteome, not one out of 10^6 randomized genetic codes results in higher correlations than the natural genetic code (i.e. $P\text{-value} < 10^{-6}$) in either of the two cases. In Figure 5A (inset), we show for *H. sapiens* the distribution of correlation coefficients for sequence-averages obtained for all randomized genetic

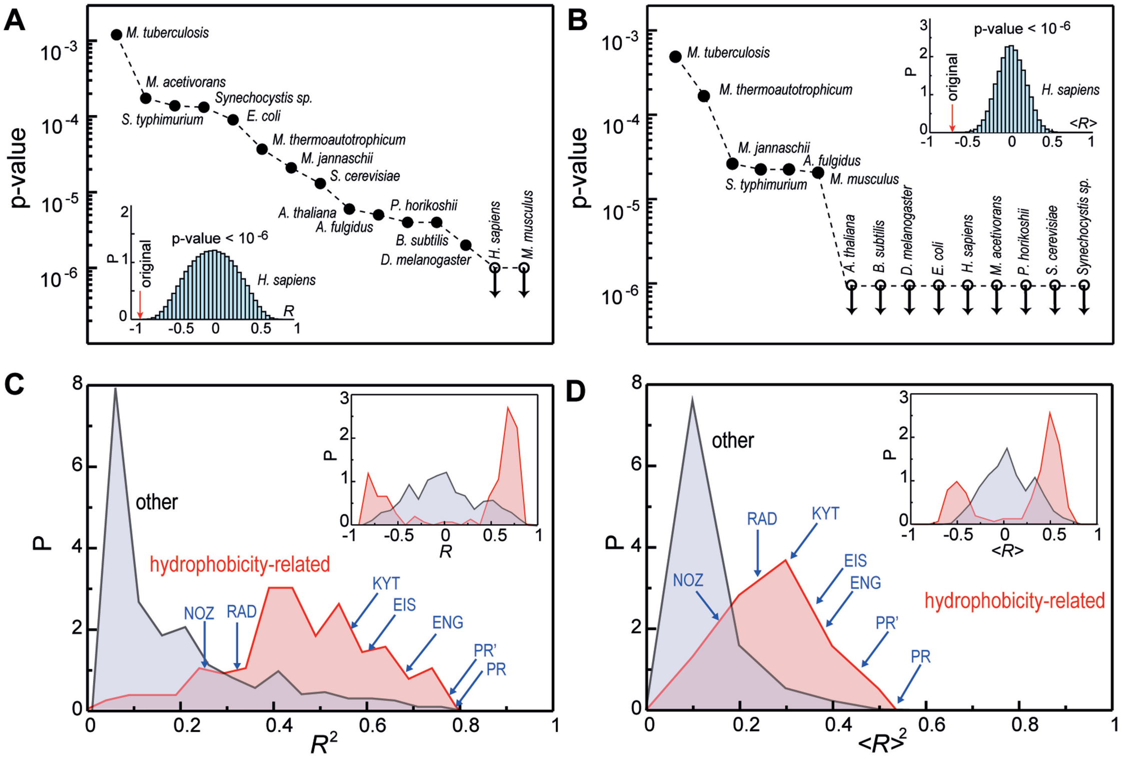


Figure 5. Randomized genetic codes and analysis of different amino acid scales. (**A**) probability that a randomized genetic code gives stronger correlation in terms of $|R|$ between average mRNA pyrimidine content and average protein PR for different proteomes (*P*-values) with empty circles with arrows denoting *P*-value $< 10^{-6}$. Inset: distribution of *R* for randomized genetic codes, with the same value for the natural genetic code indicated with a red arrow (*P*-value $< 10^{-6}$) for *H. sapiens*. (**B**) same as in 5A, but calculated for the means of distributions of sequence-profile correlation coefficients $<R>$. (**C**) Probability density distributions of R^2 for 152 hydrophobicity-related (red) and 388 non-hydrophobicity-related (blue) amino acid properties for *H. sapiens* (17 349 sequences) for sequence-average analysis. Inset: the same for correlation coefficients *R*. Values of R^2 for some notable scales and their correlation coefficients are indicated explicitly: PR—Mathew *et al.* PR scale (23); PR'—Woese PR scale (22); ENG—Engelman *et al.* hydrophobicity scale (35); EIS—Eisenberg *et al.* hydrophobicity scale (36); KYT—Kyte-Doolittle hydrophobicity scale (37); RAD—Radzicka-Wolfenden scale (33); NOZ—Nozaki-Tanford scale (34). (**D**) same as in C, but for sequence-profile analysis.

codes, with the value obtained for the natural genetic code indicated with an arrow. Similar results are found for all other species, with the weakest level of significance obtained for *M. tuberculosis*, where still the natural genetic code is in the top 99.9% of all codes (*P*-value of 1.2×10^{-3}) (Figure 5A). Finally, an even more dramatic picture is seen if one looks at average correlation coefficients between mRNA and protein sequence profiles for natural and randomized genetic codes (Figure 5B): the natural code is highly optimized with respect to maximizing this correlation, with more than half of the organisms exhibiting *P*-values $< 10^{-6}$.

In addition to PR, are there other protein sequence properties that are equally well predicted by the cognate mRNA's pyrimidine content? We have repeated the above analyses using 538 additional amino acid property scales describing *inter alia* various geometric features, secondary structure propensities and, as arguably the most studied amino acid property, hydrophobicity (23,27–31). In Figure 5C, we present the distribution of the thus-obtained correlation coefficients for sequence-average analysis, while in Figure 5D we show the same for the mean correlation coefficients for sequence profile comparison, both for *H. sapiens,* including in the analysis also both PR scales (for a total of 540 scales). We divide amino acid properties into a set related to hydrophobicity (152 scales, see 'Materials and Methods' section for definition), and a set containing all the remaining scales (388 scales). Note that the exact partitioning depends on how one defines hydrophobicity: in Supplementary Table S2 we give the sequence-average correlation coefficients for all 540 scales. Significantly, mRNA pyrimidine content appears to be predictive of the hydrophobic properties of cognate protein sequences as captured by a number of different hydrophobicity scales, as was noticed before (1,4–6,35). However, the optimized PR scale (23) in both types of analyses exhibits higher correlation with the mRNA pyrimidine content than any of the other 538 amino acid property scales, with the experimental PR scale (22) closely following (Figure 5C and D and Supplementary Tables S2 and S3). In other words, it is the affinity for pyrimidine mimetics of cognate protein sequences that appears to be best encoded by mRNA pyrimidine content, rather than just hydrophobicity in general. In fact, there is a number of widely-used hydrophobicity scales that result in little correlation with the mRNA pyrimidine content, such as the Radzicka-Wolfenden scale (36) or the Nozaki-Tanford scale (37)

(Figure 5C and D). On the other hand, as an important part of amino acid/pyridine interactions is of hydrophobic nature, it is not surprising that one also sees correlations with a number of other hydrophobicity-related scales, but this is arguably just one aspect of the issue.

We have also zeroed in on individual mRNA nitrogenous bases and combinations thereof to analyze which of them are most responsible for the sequence-average and sequence-profile correlations. As demonstrated in Supplementary Table S4 for *H. sapiens*, the content of individual bases (e.g. A-content for all proteins or U-content for membrane proteins in some cases) already alone gives sizable correlations in some cases, but overall, pyrimidine content significantly outperforms all individual mRNA nitrogenous bases and combinations thereof when it comes to sequence-average and sequence-profile correlations with protein PR, both for cytosolic and membrane proteins (Supplementary Tables S2 and S4).

DISCUSSION

If one assumes that the interaction of amino acids with pyridines is analogous to their interaction with pyrimidines (21), the strong correlation between mRNA coding-sequence pyrimidine content and protein-sequence PR (Figures 2 and 3) gives support to an experimentally-testable hypothesis that mRNA-coding regions may be physically complementary to cognate protein regions, especially if both are unstructured. Here, direct interaction between the two is facilitated through direct pairing between pyrimidine-rich mRNA regions and amino acid stretches, encoded by them, which at the same time exhibit high propensity to interact with pyrimidines (Figure 6). Consequently, our results give strong support to the idea that the universal genetic code reached most of its present-day features in the era before the development of tRNA-based ribosomal-decoding machinery (15,22,41). In this framework, ancient proteins were directly templated off of cognate mRNAs, but the code was ‘fuzzy’, i.e. the exact nature of bases and amino acids was not fully defined, but their general physico-chemical characteristics were. A stretch of pyrimidine-rich bases on mRNA would code for different protein sequences, but all of them would have low PR in common, and *vice versa* as suggested in Carl Woese’s ‘translation error’ model for the evolution of the code (15,22).

In all previous formulations of the stereo-chemical hypothesis concerning the origin of the genetic code, the focus has been exclusively on the interactions between *individual* amino acids and *individual* codons or anti-codons (2,15–19). Contrary to this, our results support a more coarse-grained picture of these interactions, whereby the direct complementarity between mRNAs and proteins may exists predominately on the level of: (i) longer sequences and (ii) general physico-chemical characteristics of participating groups, i.e. nitrogenous bases and amino acids. Although our results leave room for well-defined, rigid, stereospecific interactions between amino acids and individual codons, as proposed before (2,15–19,42), we suggest that they could also be more general, dynamic and liquid-like. In this model, pyrimidine-rich regions in mRNAs, which are physically significantly larger (the contour length of an mRNA coding region is approximately 4.5 times longer than that of a cognate protein), solubilize the complementary stretches of pyrimidine-soluble protein sequence in a partially non-specific way. The very key element in our model is that amino acids with similar properties (e.g. low PR) tend to come in longer, contiguous stretches (such as in block copolymers) (Figure 3B and C), which could then interact with the corresponding mRNA stretches, themselves built of similar codons when it comes to pyrimidine content. It is the cooperative interaction between these longer stretches that is the essence of the present model (Figure 6). Note also that our hypothesis may be generalized to interactions between proteins and DNA-coding strands as their pyrimidine profiles are the same as those of cognate mRNAs. What is more, similar complementarity may exist for any protein and nucleic acid whose PR and pyrimidine profiles match, and not just cognate protein and mRNA or DNA pairs. This makes the present model potentially generalizable to the level of all protein–nucleic acid interactions including, for instance, those between proteins and long non-coding RNAs. Finally, we see strong matching when it comes to primary-sequence profiles of mRNAs and cognate proteins, suggesting that the proposed model may be particularly relevant in those situations where both biopolymers are unstructured. Consistent with this idea, we see a statistically significant enrichment of disordered proteins in the top 10% of human proteins when it comes to the level of matching ($P < 0.0001$, Pearson’s chi-squared test, Supplementary Table S5) as well as their slight depletion in the bottom 10% ($P < 0.02$, Pearson’s chi-squared test, Supplementary Table S5). However, the model may apply even to situations where otherwise natively structured mRNAs and proteins are largely unstructured, such as during translation or upon thermal stress, and we also do not exclude the possibility of similar interactions even in the folded state.

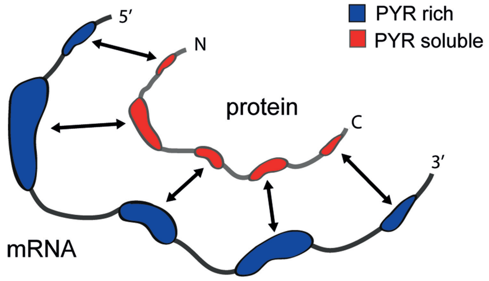


Figure 6. Model of mRNA–protein complementarity. Pyrimidine-rich regions in mRNA correspond to pyrimidine-soluble regions in protein, facilitating their complementary interactions. Hypothetically, an analogous effect might be observed for purine-rich regions. Note: polymer sizes not drawn to scale.

It is intriguing to speculate that a similar correspondence to what we have observed might be obtained when it comes to purines on the side of mRNA and purine-affinity when it comes to amino acids. It is already known that arginines (which exhibit high PR, i.e. have low pyridine affinity) interact directly with their codons and especially purine-only AGG (18,19), lending support to this possibility. Although at present we do not see any biologically relevant reason for an asymmetry between pyrimidines and purines in this context, it is also possible that pyrimidine-based interactions might alone be sufficient to stabilize binding between mRNAs and cognate proteins. Furthermore, our analysis is based on the assumption that amino acid interactions with pyridines are a quality proxy for their interactions with pyrimidines, as has been suggested before (21). Given the close chemical similarity between the two species and the large success of PR in explaining the structure of the genetic code (21,22) and its robustness towards random mutations (2,13,14,43), this assumption appears reasonable. In fact, it would be difficult to explain the strength of the correlations observed herein (Figures 1–3) if pyridines and pyrimidines interacted very differently with amino acids. However, future work should examine the validity of this assumption in more detail, as well as analyze PR-like scales for specific pyrimidines (U, T or C) and purines (A or G).

Previous computational analyses have shown that the genetic code is robust to random point mutations and translational errors, such that hydrophobicity of the original and the mutated amino acids differs to a minimal degree ('error-minimization' hypothesis) (2,13,43). Given the fact that the PR scale is related to hydrophobicity, one might argue that our present findings could be explained as a consequence of the error-minimization idea. However, one must recognize that PR is first and foremost a measure of amino acid affinity for pyrimidine-like nitrogenous bases, and for all of our principal conclusions it is secondary what the physical basis of this interaction is. In fact, we would like to suggest that error-minimization could naturally arise as a consequence of evolutionary optimization of mRNA–protein-binding interactions. Namely, a genetic code, which assigns similar codons to amino acids with similar PR, i.e. with similar affinity for pyrimidines, will also be error-minimizing in terms of hydrophobicity change of amino acids upon random mutation. What is more, the fact that of all the 152 hydrophobicity-related scales examined, the PR scales result in the best matching with mRNA pyrimidine content (Figure 5C and D), strongly suggests that the underlying physical causes are more specific than what is just encompassed by the general term 'hydrophobicity'.

Finally, it should be emphasized that the fact that all of our analysis was performed on *present-day* sequences suggests that the complementarity between mRNAs and cognate proteins, particularly under destabilizing conditions, may still have a vital biological function. This potentially concerns all facets of mRNA and protein biology including transcriptional and translational control, splicing, cellular localization, structure and function of ribonucleoprotein complexes and others, and will be explored elsewhere. It is already known that binding of several specific proteins to their cognate mRNAs (44–46) has functional importance. While the modes of binding could and have been shown to be variable, our mechanism opens up the possibility that this might be a much more widespread phenomenon, especially under conditions where both proteins and mRNAs are unstructured. Alternatively, it is possible that the correlations we observe are entirely a remnant of an era preceding the development of ribosomal-decoding machinery. In either case, it appears reasonable to suggest that our findings must somehow be related to mRNA–protein complementarity, be it exclusively ancient or both ancient and present-day. As a whole, our results are consistent with the RNA world hypothesis (47) and provide a framework for connecting it with the protein-dominated biology of today. Future work should elucidate the full biological significance of this connection, when it comes to both its evolutionary as well as present-day aspects.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5.

ACKNOWLEDGEMENTS

The authors thank A. Barta, S. Dorner, I. Moll, A. Vaziri, U. Blaesi, G. Warren, J. Davis, D. Herschlag, R. Schroeder, J. D. Dunitz and members of the Laboratory of Computational Biophysics at MFPL for useful advice and critical reading of the manuscript. The authors gratefully acknowledge M. Hajnic for help with protein disorder prediction.

FUNDING

The Austrian Science Fund FWF [START Y 514-B11 to B.Z.]; European Research Council [ERC Starting Independent to B.Z.]. Funding for open access charge: The Austrian Science Fund FWF.

Conflict of interest statement. None declared.

REFERENCES

1. Nirenberg,M.W., Jones,O.W., Leder,P., Clark,F.C., Sly,S. and Petska,S. (1963) On the coding of genetic information. *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 549–557.
2. Koonin,E.V. and Novozhilov,A.S. (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, **61**, 99–111.
3. Woese,C.R. (2001) Translation: in retrospect and prospect. *RNA-a Publicat. RNA Soc.*, **7**, 1055–1067.
4. Woese,C.R. (1965) Order in genetic code. *Proc. Natl Acad. Sci. USA*, **54**, 71–75.
5. Lesnik,T. and Reiss,C. (1998) Detection of transmembrane helical segments at the nucleotide level in eukaryotic membrane protein genes. *Biochem. Mol. Biol. Int.*, **44**, 471–479.
6. Prilusky,J. and Bibi,E. (2009) Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. *Proc. Natl Acad. Sci. USA*, **106**, 6662–6666.
7. Lecuyer,E., Yoshida,H., Parthasarathy,N., Alm,C., Babak,T., Cerovina,T., Hughes,T.R., Tomancak,P. and Krause,H.M. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**, 174–187.

8. Robinson,M., Lilley,R., Little,S., Emtage,J.S., Yarranton,G., Stephens,P., Millican,A., Eaton,M. and Humphreys,G. (1984) Codon usage can affect efficiency of translation of genes in Escherichia coli. *Nucleic Acids Res.*, **12**, 6663–6671.
9. Qu,X.H., Wen,J.D., Lancaster,L., Noller,H.F., Bustamante,C. and Tinoco,I. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature*, **475**, 118–121.
10. Zhang,F.L., Saha,S., Shabalina,S.A. and Kashina,A. (2010) Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science*, **329**, 1534–1537.
11. Crick,F.H.C. (1968) The origin of the genetic code. *J. Mol. Biol.*, **38**, 367–379.
12. Wong,J.T. (2005) Coevolution theory of the genetic code at age thirty. *Bioessays*, **27**, 416–425.
13. Haig,D. and Hurst,L.D. (1991) A quantitative measure of error minimization in the genetic-code. *J. Mol. Evol.*, **33**, 412–417.
14. Freeland,S.J., Knight,R.D., Landweber,L.F. and Hurst,L.D. (2000) Early fixation of an optimal genetic code. *Mol Biol Evol.*, **17**, 511–518.
15. Woese,C.R. (1965) On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA*, **54**, 1546–1552.
16. Woese,C.R. (1968) Fundamental nature of genetic code: prebiotic interactions between polynucleotides and polyamino acids or their derivatives. *Proc. Natl Acad. Sci. USA*, **59**, 110–117.
17. Woese,C.R. (1969) Models for the evolution of codon assignments. *J. Mol. Biol.*, **43**, 235–240.
18. Yarus,M. (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J. Mol. Evol.*, **47**, 109–117.
19. Yarus,M., Widmann,J.J. and Knight,R. (2009) RNA-amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.*, **69**, 406–429.
20. Johnson,D.B. and Wang,L. (2010) Imprints of the genetic code in the ribosome. *Proc. Natl Acad. Sci. USA*, **107**, 8298–8303.
21. Woese,C.R., Dugre,D.H., Saxinger,W.C. and Dugre,S.A. (1966) The molecular basis for the genetic code. *Proc. Natl Acad. Sci. USA*, **55**, 966–974.
22. Woese,C.R. (1973) Evolution of the genetic code. *Naturwissenschaften*, **60**, 447–459.
23. Mathew,D.C. and Luthey-Schulten,Z. (2008) On the physical basis of the amino acid polar requirement. *J. Mol. Evol.*, **66**, 519–528.
24. The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **40**, D71–D75.
25. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Research*, **39**, D28–D31.
26. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
27. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.
28. Moelbert,S., Emberly,E. and Tang,C. (2004) Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci.*, **13**, 752–762.
29. Atchley,W.R., Zhao,J.P., Fernandes,A.D. and Druke,T. (2005) Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA*, **102**, 6395–6400.
30. Efremov,R.G., Chugunov,A.O., Pyrkov,T.V., Priestle,J.P., Arseniev,A.S. and Jacoby,E. (2007) Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.*, **14**, 393–415.
31. Zhao,G. and London,E. (2009) Strong correlation between statistical transmembrane tendency and experimental hydrophobicity scales for identification of transmembrane helices. *J. Membrane. Biol.*, **229**, 165–168.
32. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
33. Trifonov,E.N. (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene*, **261**, 139–151.
34. Nowakowski,M., Jaremko,L., Jaremko,M., Zhukov,I., Belczyk,A., Bierzynski,A. and Ejchart,A. (2011) Solution NMR structure and dynamics of human apo-S100A1 protein. *J. Biol. Chem.*, **174**, 391–399.
35. Wolfenden,R., Andersson,L., Cullis,P.M. and Southgate,C.C.B. (1981) Affinities of amino-acid side-chains for solvent water. *Biochemistry*, **20**, 849–855.
36. Radzicka,A. and Wolfenden,R. (1988) Comparing the polarities of the amino-acids - side-chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous-solution. *Biochemistry*, **27**, 1664–1670.
37. Nozaki,Y. and Tanford,C. (1971) Solubility of amino acids and 2 glycine peptides in aqueous ethanol and dioxane solutions - establishment of a hydrophobicity scale. *J. Biol. Chem.*, **246**, 2211–2217.
38. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino-acid-sequences of membrane-proteins. *Annu. Rev. Biophys. Bio.*, **15**, 321–353.
39. Eisenberg,D. and Mclachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
40. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
41. Noller,H.F. (2012) Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol.*, **4**, 1–U20.
42. Yarus,M., Caporaso,J.G. and Knight,R. (2005) Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.*, **74**, 179–198.
43. Butler,T., Goldenfeld,N., Mathew,D. and Luthey-Schulten,Z. (2009) Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement. *Phys. Rev. E*, **79**, 060901.
44. Mosner,J., Mummenbrauer,T., Bauer,T., Sczakiel,G., Grosse,F. and Deppert,W. (1995) Negative feeback regulation of wild-type p53 biosynthesis. *EMBO J.*, **14**, 4442–4449.
45. Tai,N., Schmitz,J.C., Liu,J., Lin,X., Bailly,M., Chen,T. and Chu,E. (2004) Translational autoregulation of thymidylate synthase and dihydrofolate reductase. *Front. Biosci.*, **9**, 2521–2526.
46. Schuttpelz,M., Schoning,J.C., Doose,S., Neuweiler,H., Peters,E., Staiger,D. and Sauer,M. (2008) Changes in conformational dynamics of mRNA upon AtGRP7 binding studied by fluorescence correlation spectroscopy. *J. Am. Chem. Soc.*, **130**, 9507–9513.
47. Cech,T.R. (2011) The RNA worlds in context. *Cold Spring Harb Perspect Biol.*, **4**, a006742.