universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Contributions to physicochemical approaches in bioinformatics“

verfasst von / submitted by

Lukas Michael Bartonek BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 419
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Chemie
Betreut von / Supervisor:	Univ.-Prof. Dr. Bojan Zagrovic, BA

Lukas Bartonek: *Contributions to physicochemical approaches in bioinformatics,* Exploration of one-dimensional physicochemical property profiles as a means to study biomolecular sequence attributes, © September 2015 - November 2019

SUPERVISOR:
Univ.-Prof. Dr. Bojan Zagrovic

LOCATION:
Vienna, Austria

TIME FRAME:
September 2015 - November 2019

Part of the inhumanity of the computer is that,
once it is competently programmed and working smoothly,
it is completely honest.

— Isaac Asimov

ACKNOWLEDGMENTS

This thesis would not have been possible in this form without the direct and indirect involvement of many people.

First and foremost, I wish to express my sincere gratitude to my advisor Prof. Bojan Zagrovic for the continued support throughout all stages of work on my doctoral thesis - you have been an outstanding mentor for me. With his persistent and continuous guidance, he not only encouraged me in all stages of this thesis but also helped me grow as a scientist.

Furthermore, I would like to pay my special regards to my coauthors - Bojan Zagrovic, Anton Polyansky and Daniel Braun - I thoroughly enjoyed researching and writing in collaboration with you. Your comments and criticism were always welcome and proved to be of the highest quality.

I want to especially thank my fellow lab-colleagues Matea, Markus, Gerald, Mathias, David, Clemens, Daniel, Yuliia, Sarah, and Amy for the spirited discussions, the long hours spent working together and generally being supportive whenever I was in need of help.

Last but not least I want to extend a special thanks to my family. I cannot express how grateful I am for your continuous support allowing me to be in this position right now. Additionally, I would also like to thank my friends who supported me throughout the time of working on this thesis. Finally, I would like to express my deepest gratitude to Karo - thank you for being always supportive and cheerful whenever I was in need of that.

PUBLICATIONS

Ideas, figures and text presented in this work have appeared previously in the following publications which were (co)authored by the author of this piece of work. The cited publications are included in their full length, including supplementary text, within this piece of word.

[1] **Lukas Bartonek** and Bojan Zagrovic. "mRNA/protein sequence complementarity and its determinants: The impact of affinity scales." In: *PLoS computational biology* 13.7 (2017), e1005648.

[2] Bojan Zagrovic, **Lukas Bartonek**, and Anton A. Polyansky. "RNA-protein interactions in an unstructured context." In: *FEBS Letters* 592.17 (2018). ISSN: 18733468. DOI: 10.1002/1873-3468.13116.

[3] **Lukas Bartonek** and Bojan Zagrovic. "VOLPES: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences." In: *Nucleic Acids Research* 47.W1 (2019), W632–W635. ISSN: 0305-1048. DOI: 10.1093/nar/gkz407. URL: https://academic.oup.com/nar/article/47/W1/W632/5494750.

[4] **Lukas Bartonek**, Daniel Braun, and Bojan Zagrovic. "Invariants of Frameshifted Variants." In: *bioRxiv* (2019), p. 684076. DOI: 10.1101/684076. URL: https://www.biorxiv.org/content/10.1101/684076v1.full.

ABSTRACT

Most computational methods for treating physicochemical properties of biopolymers, such as proteins and nucleic acids, are computationally expensive and are usually applied to a limited number of systems only. Most high-throughput analyses of large datasets are, therefore, performed at the primary structure level only and use highly symbolic, lexical representation of individual monomers. By merging the concept of bioinformatic sequence analysis with weights describing measurable physicochemical properties of individual monomers, a connection between sequence information and general polymeric properties can be established. This results in fast, parallelizable workflows with a wide range of applications, which can provide deeper insight into the information present at the sequence level. This work explores different aspects of this idea, including an analysis of the space the monomer weights reside in, development of an easy-to-use web application for the analysis of physicochemical sequence profiles and a novel result, obtained by using this method, concerning the impact of frameshifts on protein sequence properties.

ZUSAMMENFASSUNG

Physikalische und chemische Eigenschaften von Biopolymeren, wie Proteine oder Nukleinsäuren, werden bei computergestützten Analysen von großen Datensätzen meist nicht in Betracht gezogen. Da moderne Methoden, die physikochemische Eigenschaften berechnen können, einen sehr hohen Rechenaufwand benötigen, werden diese Methoden üblicherweise nur auf eine geringe Anzahl an Systemen angewandt. Analysen von großen Datensätzen werden meist auf der Primärstruktur - der biologischen Sequenz an Monomeren - durchgeführt. Durch eine Erweiterung dieser Methode um experimentell bestimmte Gewichte, die eine messbare Eigenschaft repräsentieren, kann eine Verbindung zwischen der Sequenzinformation und den Eigenschaften des Polymers hergestellt werden. Dies führt zu sehr schnellen, parallelisierbaren Arbeitsabläufen mit einem breiten Spektrum an Anwendungen, die einen Einblick in bereits auf Sequenzebene verfügbare Informationen ermöglichen. Diese Arbeit erforscht verschiedene Aspekte dieser Idee, zunächst wird der mathematische Raum in dem sich die experimentelle Gewichte befinden analysiert, eine einfach zu verwendende Webanwendung vorgestellt, die es jedem erlaubt solche Analysen durchzuführen und letztendlich wird ein völlig neues Ergebnis präsentiert, das durch Anwendung dieser Methode auf Sequenzen mit Frameshift-Fehlern erzielt wurde.

CONTENTS

LIST OF FIGURES

Part I

INTRODUCTION

Readers are introduced to the basic concepts dealt with in the thesis and a general overview of the thesis is given.

1 INTRODUCTION

The analysis of physicochemical properties of biopolymers, including RNA, DNA or proteins, has a long history. From concrete applications, such as drug development, to abstract ones, such as elucidating the tree of life, it has touched upon every imaginable aspect of life science. A short summary of 'state of the art' of the theoretical and computational methods to study biopolymers is presented here, ranging from the most complex, physically realistic models to the most simplified ones. This is followed by a definition of the concept of *physicochemical sequence analysis*, the central topic of this thesis and a discussion of technical details related to this concept. Specifically, we focus on the particulars of the actual approach of representing and comparing sequences of biopolymers as one-dimensional physicochemical profiles, including a discussion of the physicochemical property scales and the scoring functions used. Finally, we discuss the hypothesis of the stereochemical origin of the universal genetic code since it represents a recurring theme behind several applications of physicochemical sequence analysis presented herein.

1.1 METHODS TO INVESTIGATE PROPERTIES OF BIOPOLYMERS

A wide array of theoretical and computational methods to analyze properties and interactions of biological polymers have been established in the literature. While some of these methods are widely accepted and applied, each comes with different specific benefits and drawbacks. We start here with a short overview of such methods.

Quantum Mechanical Simulations

Computer simulations of systems relevant in chemistry and biology based on the application of quantum mechanical models have a veritable record and can be broadly separated into two categories, Hartee-Fock method based [1, 2] and density functionl theory based approaches [3–5]. While there are serious differences in computational efficiency and accuracy between the two approaches depending on the system of interest, both are computationally considerably more demanding, but also more physically realistic and arguably more precise than all the other categories discussed here. The long calculation times involved, however, severely limit the applicability of these methods to biological systems. Complete proteins or nucleic acids are usually impossible to simulate, especially in their natural, aqueous environment, and state-of-the-art efforts do not go beyond treatments

of individual monomer building blocks or short oligomers - often in conjunction with molecular dynamics treatment for the rest of the structure[6–9].

1.1.1 *Molecular Dynamics*

Unlike quantum mechanical methods, classical molecular dynamics simulations typically completely avoid explicit treatment of electrons [10, 11]. Interactions between simulated atoms with typically fixed partial charges are modelled by defining a semi-empirical interaction potential. Bonded interactions - those involving one or more chemical bonds - are usually defined using harmonic potentials (bonds and angles) or trigonometric functions (torsions), while non-bonded interactions are modeled by the Lennard-Jones potential for van-der-Waals interactions and the Coulomb potential for electrostatic interactions. An example of such a potential, corresponding to the widely used Amber force field[12], is shown in equation (1).

$$\begin{aligned} V = & \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \\ & \sum_{dihedral} \frac{V_n}{2}[1+cos(n\phi)] \\ & \sum_{nonbonded} 4\varepsilon_{ij}\left(\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \left[\frac{q_i q_j}{\varepsilon R_{ij}}\right]\right) \end{aligned} \tag{1}$$

Here, r is a bond length, θ an angle with the subscript 0 indicating the equilibrium value in both cases and K the respective force constants of the harmonic oscillator. ϕ denotes a dihedral angle, q a charge and ε and σ the Lennard-Jones parameters.

The benefit of such a simple description of the system is that the computational effort is severely reduced. All-atom simulations involving thousands of explicit water molecules as solvent and full-length proteins or nucleic acids are nowadays accessible even on standard desktop computers. Highly optimized hardware on distributed computing systems allows for simulations on long timescales and/or involving relatively large molecular systems [13, 14]. While reasonably good at sampling the underlying ensemble, the results are strongly dependent on the *forcefield* parameters used. In exchange for relatively fast computation times – ranging, for example, from days to months for simulations of realistic proteins on the timescale of microseconds or longer - any processes or features dependent on the outer-shell electrons remain inaccessible. This includes chemical reactions, excited states - including states induced by light – and, in the general case, even polarization of atoms. These deficiencies can be addressed

by using advanced modern approaches, such as QM/MM methods, but this usually comes with a serious performance cost[15].

1.1.2 *Simplified atomistic methods*

While molecular dynamics simulations are highly efficient as compared to methods that deal with actual wavefunctions, molecular dynamics simulations of relevant systems on biologically meaningful timescales still may require months rather than days to be performed. In applied research - especially related to development of pharmaceuticals - *in silico* models are often used as predictive filters[16–18]. To achieve such a goal, multiple models need to be considered and tested in quick succession: for example, *in silico* libraries of small molecules can reach sizes in excess of 100 million compounds listed. While most of the filtering is usually performed without taking structures of the biopolymers in question into the account, a final set still needs to be evaluated against the actual structure. This is where modern methods of docking come into play[19, 20]. The already severely simplified potentials in MD simulations experience a further reduction in complexity with the main goal of reducing the computational effort. All-atom models are often exchanged for pharmacophore models in which local properties are represented as sets of steric and electronic features, while dynamic behavior is at best crudely modeled and static description is preferred.

1.1.3 *Bioinformatics*

Finally, when it comes to large-scale analysis, there is no way around introducing further simplifications into the model. For example, in the case of large multiple-sequence alignments, as required in modern applications of genome, transcriptome or proteome sequencing or phylogenetic analyses, the information is reduced simply to biological sequences of the molecules in question. In other words, these applications demand fast, high-throughput calculations, a challenge which is solved by disregarding most of the structural, physical or chemical properties of biopolymers and investigating their sequences only. While some of these aspects can be dealt with by introducing ensemble-averaged information like substitution matrices or performing actual in-silico folding, as routinely performed for RNA, the resulting information is far more coarse-grained as compared to what is obtained by all the previously mentioned methods, with the upside being the ability to analyze genomic-scale information on a reasonable time-scale. Finally, the rapid decline in price for full genome sequencing is currently resulting in a major discrepancy between the amount of the raw data and the available computational power to

draw conclusions from the data, something only fast bioinformatics methods can currently deal with[21–23].

1.1.4 *Physicochemical bioinformatics*

In this thesis, an intermediate approach was chosen to deal with the large quantity of sequence information required for analysis, but also keep a connection to physical and chemical properties of the biomolecules involved, admittedly on a rather coarse-grained level. In this approach, sequence information is coupled with residue-type-specific information concerning a wide range of physicochemical properties, while sequences are analyzed and compared in the form of one-dimensional physicochemical property profiles. With the ability to compare sequences in this high-dimensional property space, high-throughput analysis becomes possible without being limited to purely "text-based" analysis characteristic of classical bioinformatics. One should emphasize that the idea of treating biological sequences as 1D physicochemical profiles, as explored here, is by no means novel in-and-of-itself, but it is also true that the use of such profiles as an analytical tool has not been widely promulgated. As attempted to be shown in this thesis, such an approach could indeed yield significant novel advances in our understanding of basic biology and merits a wider and more detailed scrutiny. The basic ideas of such an analysis can be found in the classical attempts to analyze the hydropathy of protein sequences [24, 25], but have also been applied to analysis of charge or structural disorder and development of complex prediction tools[26–31].

1.2 PHYSICOCHEMICAL APPROACH TO BIOINFORMATICS

The concept of a physicochemical approach to bioinformatics is based on several technical principles. All the essential ideas are presented here ranging from the basic concept of converting a sequence to a property profile to the choice of the most relevant property scales and scoring functions to compare different profiles and, finally, to the different approaches and implications of smoothing the profiles.

1.2.1 *Central Approach*

The central assumption behind the physicochemical approach to analyzing biopolymers is that the physicochemical properties of individual monomers in a given sequence can be used to characterize the properties of longer sequences, whereby different *property scales* - lists of numeric properties of individual monomers - are applied to a given sequence as weights. Importantly, such a procedure converts the lexical sequence into a numerical one, which now serves as

a representation of a specific property of the polymer along its sequence. Depending on the property of interest, further steps may be applied. In some cases, it may be advantageous to average over certain domains to obtain a more global, coarse-grained view. In other cases, it may be sufficient to only include the local environment in the model. If multiple representations have been obtained using compatible models, they may be compared using an appropriate scoring function.

1.2.2 *Scoring Function*

While 1D physicochemical property profiles can be easily compared visually, a similarity measure is necessary if multiple profiles are to be compared or a high-throughput approach is chosen. In principle, multiple options are available, each with their individual benefits and downsides. Four very common similarity measures used herein are Pearson correlation coefficient R, coefficient of determination R^2, RMSD and Spearman rank correlation coefficient ρ.

1.2.2.1 *Pearson R and R^2*

Initially, in all the publications exploring the similarity between mRNA nucleobase-density profiles and nucleobase-affinity profiles of their cognate proteins [32–34], which served as a theoretical basis for the present thesis, Pearson R (see equation 2) was used as a distance measure. This was due to some very practical reasons.

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

First, the Pearson correlation coefficient has a beneficial property that it is normalized, always resulting in a value between -1 and 1. This allows for an easy comparison between different analyses. Furthermore, this normalization also allows one to compare profiles capturing different physicochemical property profiles, like the mRNA nucleobase-density and the protein nucleobase-affinity profiles. What this effectively captures is a tendency of higher, or lower, values occurring simultaneously within the same position of a sequence. Since the application of the sample Pearson correlation coefficient effectively disregards all information concerning the specific position of a given residue in the sequence, the coefficient of determination R^2 can simply be calculated by squaring the Pearson R and assuming a linear, least-squares regression with a single explanatory variable and an intercept term.

$$\rho_{xy} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{3}$$

The major downside of the Pearson R as well as the coefficient of determination R^2 is that the physical meaning of a given level of correlation is hard to judge due to the normalization that is implicitly included in the calculation. Profiles which claim to reasonably represent a given physicochemical property may have very different levels of intrinsic variance in corresponding profiles or even an overpowering intercept term, resulting in high levels of correlation that cannot be attributed to actual similarities between the two biopolymers that are being compared.

1.2.2.2 *RMSD*

When comparing two profiles capturing the same physicochemical property, a more natural measure of similarity is the root mean square deviation between the two (see equation 4). By having direct knowledge regarding the values in an applied scale, an experienced researcher can easily judge the level of similarity as captured by RMSD to acquire a deeper insight. On the flipside, users with less experience regarding a specific property in question and system might have a difficulty interpreting these values, due to a lack of normalization. Depending on the applied scale, *good* fits can have values ranging over multiple orders of magnitude.

$$RMSD_{xy} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{4}$$

While there have been attempts made to normalize the RMSD, and such methods are described in the literature[35, 36], none of them improve significantly upon the normalization already included in the calculation of Pearson R and were, therefore, not included in this work.

1.2.2.3 *Spearman rank correlation coefficient*

As an alternative to the Pearson sample correlation coefficient, Spearman rank correlation coefficient (see equation 5) can also be used. The Spearman correlation coefficient can be seen as a Pearson correlation coefficient of the ranks of the data points. This leads to a metric that is not as strongly affected by outliers as Pearson R, but still shows a low correlation when applied on elliptically distributed data as one would expect from Pearson R.

$$\begin{aligned} r_{xy} &= R(rg(x), rg(y)) \\ r_{xy} &= 1 - \frac{6\sum_{i=1}^{n}(rg(x_i) - rg(y_i))^2}{n(n^2-1)} \end{aligned} \tag{5}$$

1.2.3 *Amino-acid physicochemical property scales*

The essential ingredient for the physicochemical analysis of biomolecular sequences are the scales describing the properties and/or identity of individual monomers in a given biopolymer (e.g. amino-acid hydrophobicity scales or nucleobase-identity scales). They represent a connection between the purely mathematical model and physicochemical reality by reporting, in most cases, on the actual experimentally determined properties of individual monomers. A scale is usually given as a list of property values for each individual amino acid or nucleobase. Here some widely used scales, relevant for the present thesis, are introduced and discussed.

1.2.3.1 *Polar requirement and nucleobase affinity scales*

The polar requirement amino-acid scale was reported by Woese in 1966 [37–39] as a result of chromatographic experiments. Specifically, the experimental readout that was used to define the scale was based on the change in the chromatographic retardation of amino acids when changing the polarity of the mobile phase by altering its molar ratio of 2,4-dimethylpyridine to water. The polar requirement value is the slope in a log-log plot of the chromatographic factor R_m against the mole fraction of water in the mobile phase χ_w as given in equation 6.

$$\begin{aligned} R_m &= (1 - R_f)/R_f \\ PR &= -\frac{\Delta log R_m}{\Delta log \chi_w} \end{aligned} \tag{6}$$

The chemical similarity between 2,4-dimethylpyridine and pyrimidine nucleobases, and subsequent usage of the PR scale as a proxy of the amino-acid affinity for pyrimidines, have been noted in several publications [32, 37, 40–42]. Nevertheless, the lack of backbone substituent and the substantially different substitution pattern on the pyridine skeleton results in a drastic difference between this 2,4-dimethylpyridine/water mixture and real RNA polymers in water. It can also be argued that any change in polarity of the solvent, in a similar magnitude, would result in comparable values.

For this reason, efforts have been directed in the host group towards deriving nucleobase/amino-acid affinity scales using different computational approaches. In particular, by exploiting the statistics of nucleobase/amino-acid contacts in high-resolution structures of RNA-protein complexes, relative nucleobase/amino-acid affinities were derived by using a knowledge-based formalism[33]. This has, in turn, allowed one to the analyze the results obtained by using the PR scale, while employing scale values extracted from actual biopoly-

mers and not monomeric units, resulting arguably in a biologically more realistic treatment.

1.2.3.2 *Consensus scales*

Both Kidera et.al. [43] as well as Atchley et.al. [44] realized the complications of having a large number of scales that capture in different ways one and the same property, such as amino-acid hydrophobicity. Kidera et.al. presented an attempt to classify and organize the large number of different amino-acid property scales using multivariate data-analysis already in 1985. 20 years later, Atchley et.al. tackled the issue again and used multivariate statistical analysis on a set of approximately 500 different amino-acid scales, resulting in a set of 5 consensus factor scales. Importantly, each of the 5 factor scales was attributed a physical property based on expert knowledge including, in the order of variance explained, hydrophobicity, secondary structure, molecular size, amino-acid composition and charge.

The Factor I scale, in particular, gained some popularity because it solved two problems at once. First, it can be considered a consensus hydropathy scale, capturing in a balanced manner the properties of different individual hydropathy scales. Secondly, it removed the confusion stemming from two definitions of hydropathy i.e. hydrophobicity and hydrophilicity, by providing a standardized convention.

For most of the work regarding hydropathy of proteins presented herein, the Atchley Factor I scale is used as it is based on a more up-to-date set of scales. Individual hydropathy scales, such as Levitt [45], Eisenberg [46], Engelman [47] and Kyte & Doolitle [24] scales, are only considered because of their popularity and as a point of comparison with other results.

1.2.4 *Contemporary application of profiles*

Physicochemical property profiles have historically been used in different biological contexts to investigate the global properties of biomolecular sequences. Initially applied mostly as a visualization aide, such profiles have recently found application in the area of predictive modeling as well [27, 33, 48]. In most of predictive applications, however, the focus has been on individual properties, while large-scale comparative analyses have been performed only in the context of property scales themselves (see Consensus scales) [43, 44]. Nevertheless, several important developments could be observed in recent years. In the following section, we present some of these developments in order to give the reader an overview of the tools and methods currently in use.

1.2.4.1 *Prediction of properties*

One of the most popular applications of profiles as predictive and sometimes even quantitative tools is as predictors of transmembrane protein domains. Tools like TM Finder [48] analyze proteins using both hydrophobicity scales and helicity scales and combine the results in a metric predicting whether a given protein segment is part of a membrane bound domain. Similar examples focusing on membrane bound proteins abound [29, 49, 50], but it should be emphasized that other properties have been investigated using similar approaches as well. Examples include analyses of charge patterns in proteins [51, 52], rationalization of their cellular localization [53, 54] and even elucidation of processes behind the formation of phase-separated granules in cells [55, 56]. Finally, analysis of physicochemical properties encoded in the linear sequences of RNA and proteins, including secondary structure predictions, hydrogen bonding and van der Waals' interactions, has been used to predict interactions between these biopolymers [30, 57].

1.2.4.2 *Sequence alignment*

Classical sequence alignment of biological sequences is essentially performed by defining transition matrices defining the cost for every possible combination of monomers. Usually an additional term is introduced defining the cost of a gap in an alignment. Once defined, a set of high-performance algorithms can be applied to sample all the possible alignments and return the one with the lowest total cost [58, 59]. The implication of these methods and the transition matrices used is that such an alignment is a good representation of evolutionary distance. If, on the other hand, one is interested in an alignment according to a specific property instead of the more global 'relatedness', novel approaches may be necessary. Recently, several advances have been made in this direction showing that such alignment is possible [60, 61], publishing tools [62, 63] and showing results only obtainable using such approaches [64, 65].

1.2.5 *Window averaging*

Window averaging, usually employed as a means to remove local noise in one-dimensional profiles and expose features on different length-scales of organization, can have a profound impact on the results of the analysis. In particular, not only can multiple window sizes be used, but also different functions can be applied in order to average the values in a given window. Each of these choices has implications on the resulting model. For hydropathy plots, a smaller window, which only takes into the account the most local environment, is often used. If one is interested in domain structure, especially when

it comes to membrane proteins, wider windows of 21 residues or more have been used successfully in the past[33, 54]. On the side of functions used for averaging, most tools employ a boxcar, or moving-average, filter. Since this type of function weighs every residue inside a given window equally, the window size has an important effect. Other methods like Savitzky-Golay [66], on the other hand, reweight every residue within a given window, making the window size itself have less impact on the final result. The final choice of parameters needs to be determined on a case-by-case basis, but as long as these parameters are seen as a part of the final model, a wide range of values may be considered.

1.3 COMPLEMENTARY ENCODING

Another relevant topic that plays a role in most publications in this thesis is a generalization of the stereochemical hypothesis[37, 67, 68]. The stereochemical hypothesis, popularized early after the first results of the Miller-Urey experiments[69], states that the assignment of amino acids to their codons in the universal genetic code is not random i.e. a result of a 'frozen accident', as suggested by others[70], but rather follows defined physical and chemical principles. In particular, the hypothesis states that the primordial assignment of amino acids to specific codons was based on favorable direct interaction between the two[37, 68]. Furthermore, substantial progress has been made when it comes to devising a plausible scenario for the transition from a simple primordial translation mechanisms, which is based on direct interactions, to a modern-day, tRNA-based system of translation [71–73]. Importantly, the recently proposed complementarity hypothesis echoes one of the central principles for explaining biological function at a molecular level regarding non-covalent interactions between "complementary" molecules [37, 74]. Specifically, by practically expanding the stereochemical hypothesis [37] to longer sequences, it has been suggested that proteins and their cognate mRNA molecules may be complementary to each and interact in a co-aligned fashion, especially in the unstructured state [32, 33, 75]. Support for this idea was obtained by applying the idea of profile analysis to cognate protein/mRNA pairs, that is, when comparing protein profiles of amino-acid affinity for nucleobases with the cognate mRNA's nucleobase composition profiles [32, 33, 75]. Importantly, while high correlations can be observed in these cases, this does not suggest actual interaction over the full length of the two biomolecules. In fact, the coding sequence of a mature mRNA is approximately five times longer than the protein product encoded by it, a discrepancy that can get even more severe when folding is accounted for. On the other hand, a more dynamic type of interaction cannot be disregarded just

based on the length difference [75] and should be seen as a fruitful direction for further research.

1.4 SUMMARY OF THE THESIS

The present thesis explores the idea of one-dimensional physicochemical property profiles, capturing biologically relevant features of biopolymers, from multiple angles. In Chapter 2, we explore the impact of nucleobase/amino-acid affinity scales and their internal structure on the degree of matching between mRNA nucleobase-density profiles and their cognate protein nucleobase-affinity profiles. The work presented in this chapter serves as a formal foundation for understanding the cognate mRNA/protein complementarity hypothesis and its limitations, but also extends significantly beyond. In Chapter 3, we present a review dealing with the usage of property profiles to study interactions between RNA and proteins, primarily in an unstructured context. In Chapter 4, we present a novel web-based tool for visualizing, analyzing and comparing physicochemical property profiles of proteins and nucleic acids. The key aim behind the work in this chapter is to enable the general biology community to perform facile, multifaceted analysis of physicochemical property profiles without a need for advanced programming knowledge. In Chapter 5, we explore a novel finding of a potentially wider biological significance, dealing with the invariance of certain physicochemical property profiles of proteins with respect to frameshifts in their coding mRNA sequences. The results in this chapter illustrate the power of analyzing 1D property profiles to study fundamental principles of biological organization at the molecular level, in this case relating to the evolutionary mechanisms for generating novel protein sequences.

Part II

SHOWCASE OF PUBLISHED WORK

As a central part of this cumulative thesis, we present here four articles dealing with the analysis of biomolecular sequences from a fundamentally physicochemical perspective. Three out of the four presented articles have been published in peer-reviewed scientific journals, while the fourth one is currently under review (please, see below for details). The articles are presented in the original layout corresponding to the respective journals they were published in, with the only addition being continuous page numbering. The layout design should be considered intellectual property of the respective journals.

A collection of all research articles published and/or in peer-review. Each publication is introduced by a short summary and a statement of personal contribution and is accompanied by all supporting figures, tables and text. Supporting datasets that could not be reproduced here, can be found in the respective journal's online version of each article, as indicated below.

2

MONTE CARLO OPTIMIZATION OF AFFINITY SCALES

Lukas Bartonek and Bojan Zagrovic. "mRNA/protein sequence complementarity and its determinants: The impact of affinity scales." In: *PLoS computational biology* 13.7 (2017), e1005648 [34]

As demonstrated in several publications before [32, 33, 40, 42], the nucleobase-density profiles of mRNA coding sequences correlate with the respective nucleobase-affinity profiles of their cognate proteins, as determined by several independently derived nucleobase/amino-acid affinity scales. It has been suggested that this behavior may reflect the propensity of the two biopolymers to actually physically interact, but experimental proof for this has not been provided so far. One way to gain deeper insight into the previous results is to investigate the dependence between the properties of nucleobase/amino-acid scales and the degree of similarity between cognate mRNA and protein profiles. While randomization approaches have been used elsewhere[33] to calculate the significance of a given degree of similarity under a particular null distribution, here we primarily studied the determinants of such similarity at the level of the scales used. Specifically, we explored the space of scales by zero-temperature Monte Carlo sampling, using the degree of profile similarity as captured by Pearson correlation coefficient R as a fitness function. By analyzing the diversity of scales at different levels of profile similarity, we could gain insight into which amino acids contribute more and which less to the observed similarity.

The applied method, which in this case specifically targeted the question of similarity in the sequence properties of cognate mRNA and protein pairs, could in principle be used for the analysis of determinants of similarity between any two profiles that are derived from sequence information in combination with different property scales.

PERSONAL CONTRIBUTION

The author of this thesis was involved in the conceptualization, data curation, analysis, development of methodology, software development, validation, visualization as well as writing of the presented article.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

mRNA/protein sequence complementarity and its determinants: The impact of affinity scales

Lukas Bartonek, Bojan Zagrovic*

Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5, Vienna, Austria

* bojan.zagrovic@univie.ac.at

OPEN ACCESS

Citation: Bartonek L, Zagrovic B (2017) mRNA/protein sequence complementarity and its determinants: The impact of affinity scales. PLoS Comput Biol 13(7): e1005648. https://doi.org/10.1371/journal.pcbi.1005648

Editor: Nikolay V. Dokholyan, University of North Carolina at Chapel Hill, UNITED STATES

Received: March 24, 2017

Accepted: June 26, 2017

Published: July 27, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the European Research Council to BZ (Starting Independent grant PROTINT 279408), https://erc.europa.eu/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

It has recently been demonstrated that the nucleobase-density profiles of mRNA coding sequences are related in a complementary manner to the nucleobase-affinity profiles of their cognate protein sequences. Based on this, it has been proposed that cognate mRNA/protein pairs may bind in a co-aligned manner, especially if unstructured. Here, we study the dependence of mRNA/protein sequence complementarity on the properties of the nucleobase/amino-acid affinity scales used. Specifically, we sample the space of randomly generated scales by employing a Monte Carlo strategy with a fitness function that depends directly on the level of complementarity. For model organisms representing all three domains of life, we show that even short searches reproducibly converge upon highly optimized scales, implying that the topology of the underlying fitness landscape is decidedly funnel-like. Furthermore, the optimized scales, generated without any consideration of the physicochemical attributes of nucleobases or amino acids, resemble closely the nucleobase/amino-acid binding affinity scales obtained from experimental structures of RNA-protein complexes. This provides support for the claim that mRNA/protein sequence complementarity may indeed be related to binding between the two. Finally, we characterize suboptimal scales and show that intermediate-to-high complementarity can be reached by substantially diverse scales, but with select amino acids contributing disproportionally. Our results expose the dependence of cognate mRNA/protein sequence complementarity on the properties of the underlying nucleobase/amino-acid affinity scales and provide quantitative constraints that any physical scales need to satisfy for the complementarity to hold.

Author summary

Messenger RNAs and proteins, two essential types of biopolymers, have recently been shown to exhibit closely related, complementary physicochemical properties. Specifically, density profiles of certain groups in messenger RNA sequences directly match the affinity profiles for precisely those groups in protein sequences they encode. Based on this, it has been suggested that these molecules may interact with each other specifically and in a co-

PLOS COMPUTATIONAL BIOLOGY

aligned fashion, especially when unstructured. Here, we explore different amino-acid scales used in the above analysis to assess which of their properties dictate the observed matching. Specifically, we define the constraints that need to be satisfied by physical scales for the complementarity to hold and show that the previously derived nucleobase/amino-acid affinity scales indeed satisfy these constraints. As a whole, our work provides a quantitative foundation for understanding the putative messenger RNA/protein complementarity with implications in different areas of RNA/protein biology including transcription, translation, splicing and viral assembly.

Introduction

The relationship between mRNAs and the proteins they encode is one of the key defining characteristics of life at the molecular level [1–3]. While this relationship has primarily been studied in the context of biological information transfer, less attention has been paid to a potential link between the physicochemical properties of the two biopolymers. Recently, however, we have reported an unexpectedly strong connection between the nucleobase-density profiles of mRNA coding sequences and the nucleobase-affinity profiles of their cognate proteins [4–6]. For example, purine (PUR) density profiles of *E. coli* mRNA coding sequences match their cognate protein's guanine (GUA) affinity profiles with an average Pearson correlation coefficient of -0.76 (note the negative values for R indicate matching as a result of the standard definition of binding affinities) [4,5]. As illustrated in Fig 1A, the protein GUA-affinity profiles in this analysis were calculated by weighting their sequences with the GUA-affinity values for individual amino acids, which in turn were derived from known 3D structures of RNA/protein complexes by using a knowledge-based formalism [4]. In addition, we have also studied other affinity scales derived by diverse experimental and theoretical approaches: 1) a chromatographically determined scale of amino-acid affinity for pyrimidine (PYR) mimetics pyridines [7], 2) a computationally derived variant of the same scale [8], 3) absolute binding free energy scales between nucleobases and amino-acid sidechain analogs in different solvents [9], and 4) affinity scales obtained from simulated partitioning experiments using realistic RNA nucleobases [10,11]. The consensus of these studies has been that the mRNA regions rich in a particular nucleobase or a type of nucleobases (PUR or PYR) tend to encode the protein regions with a pronounced affinity for precisely those or similar bases. This novel finding is well illustrated by the mRNA PUR-density profile and its cognate protein's GUA-affinity profile of a typical, representative mRNA/protein pair in *E. coli* whose Pearson correlation coefficient (-0.76) equals that of the mean of the entire *E. coli* distribution (Fig 1A). Importantly, the only exception to the above rule was seen in the case of adenine (ADE). Namely, protein regions with a high affinity for the purine base ADE tend to be encoded by mRNA regions rich in PYR bases [5,6].

Although robust and of potentially far-reaching implications, the above findings still lack a definitive explanation. We have suggested two possibilities: one, concerning the reasons for the observed complementarity, and another one, concerning its implications. First, the above observations are consistent with a possibility that genetic encoding may have arisen from binding, as postulated by the stereochemical hypothesis of the origin of the genetic code [2,12–13]. This hypothesis dates back to 1960s, but we believe our results provide the strongest-yet evidence for it. Importantly, however, our results shift the focus towards analyzing the binding in the context of longer biopolymers and not just isolated codons and amino acids [4–6,14]. Second, the compositional mirroring between mRNAs and their cognate proteins at the level of

PLOS COMPUTATIONAL BIOLOGY

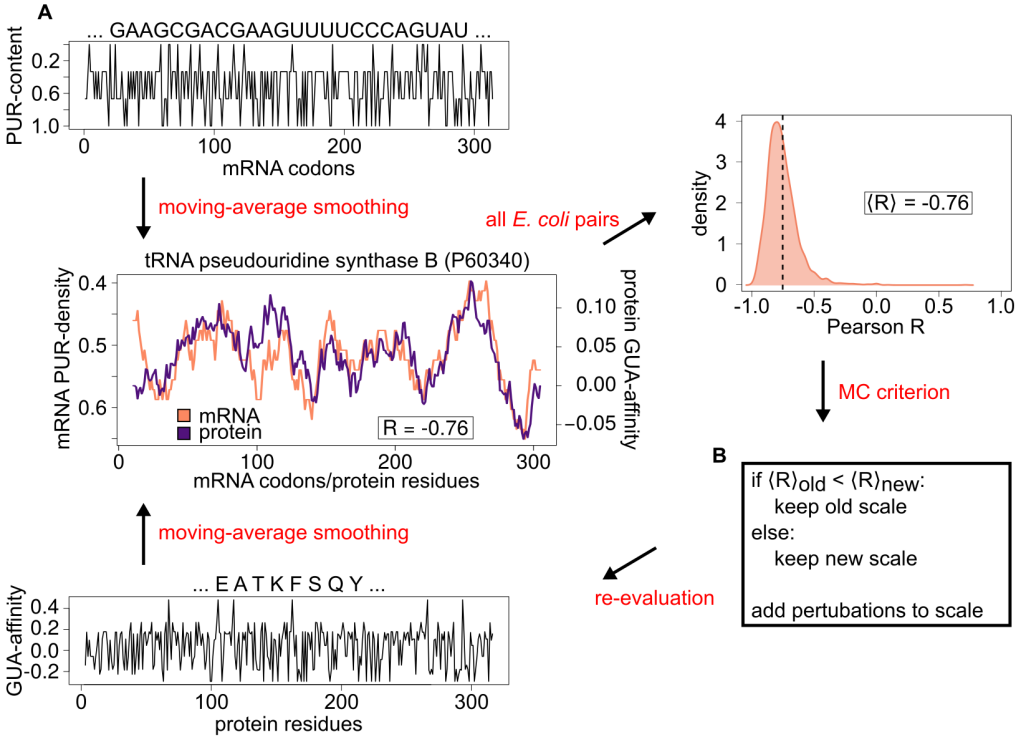


Fig 1. Cognate mRNA/protein sequence profile comparison and scale optimization. A) *top and bottom*: mRNA nucleobase-density profiles are obtained by window-averaged smoothing of their nucleobase composition strings, while nucleobase-affinity profiles of their cognate proteins are obtained by weighting their sequences with the aid of nucleobase-affinity scales and subsequent window-averaged smoothing; *center*: comparison between an mRNA PUR-density profile and the cognate protein's GUA-affinity profile for a typical, representative mRNA/protein pair in *E. coli*. Matching between the two profiles (Pearson R = -0.76) corresponds to the mean Pearson correlation among all cognate mRNA/protein pairs in *E. coli*; *right*: distribution of Pearson Rs for all E. coli cognate mRNA/protein pairs. Note that the scale used for these profiles is the knowledge-based scale derived from the structures of RNA-protein complexes. In our simulation, we have explored the full space of such scales as shown in the text; B) pseudo code of the Monte Carlo strategy employed in the simulation.

https://doi.org/10.1371/journal.pcbi.1005648.g001

primary sequences supports the possibility of complementary, co-aligned binding between the two even in modern systems, especially if they are unstructured [4–6,14]. On the side of proteins, this pertains to both intrinsically disordered proteins as well as the unfolded states of otherwise folded proteins, such as during translation or upon thermal/chemical stress. While we do not exclude the possibility of interactions in the structured states of either partner [15], this also more likely involves mRNA stretches that are not base-paired. Overall, the potential implications of this interpretation concern different facets of RNA/protein biology including translation control, structure of ribonucleoprotein particles, the behavior of non-membrane-bound cellular compartments, viral capsid assembly and others [5]. More than two decades ago, Kyrpides and Ouzounis proposed that cognate mRNA/protein interactions may be an ancient mechanism for auto-regulation of mRNA stability [16]. Our findings now provide mechanistic details behind such a possibility. Here it should also be mentioned that the opposite behavior of ADE suggests that there may have been at least two major phases in the code's development, the more recent of which introduced ADE to modulate in a negative manner the complementarity engendered by other bases [5].

The statistical significance of the cognate mRNA/protein sequence complementarity has already been probed by different tests involving randomizations of the genetic code table,

shuffling of both the affinity scales and the primary mRNA/protein sequences and analyzing the behavior of other physically realistic amino-acid property scales in the same context [4,6,14]. However, these approaches, although valid and necessary, have thus far not included a systematic exploration of the whole space of nucleobase/amino-acid affinity scales. On the other hand, several important questions can only be answered with an in-depth knowledge of the properties of the space of affinity scales and their influence on the observed complementarity. A key open problem in this regard concerns the uniqueness of the affinity scales that yield a high degree of cognate sequence-profile matching. Is there an optimal scale that results in maximized matching between mRNA nucleobase-density profiles and their cognate protein sequence profiles or are there multiple scales that produce the same or similar levels of matching over complete proteomes? How many different scales are there that show intermediate-to-high levels of matching? Finally, are there specific amino acids whose affinities for different nucleobases dominate the matching? It is easily imaginable that there are amino acids that exhibit similar affinities to different nucleobases, while the specificity i.e. the sequence complementarity, is dictated by a select few. To address these questions, one needs to consider not only the already published, physically realistic scales, but also those that do not produce strongly matching profiles. In other words, one would like to sample the space of randomly-generated scales using a fitness function that is related to the degree of cognate mRNA/protein sequence complementarity engendered by those scales.

Here, we present a Monte Carlo (MC) search method that satisfies the above criteria and explores the space of nucleobase/amino-acid affinity scales, while at the same time assessing their impact on the degree of cognate complementarity (Fig 1A and 1B). More specifically, our MC searches start from a uniform scale with identical weights for all 20 amino acids and evolve through a succession of random perturbations, which are accepted or rejected according to a fitness function. The latter, in turn, is related to the degree of complementarity as captured by the proteome-average Pearson correlation coefficient $<R>$ between the cognate sequence profiles (see Materials and Methods for details). In other words, we search for amino-acid scales (i.e. 20-element linear arrays of amino-acid weights), which result in a given level of proteome-wide average matching between mRNA nucleobase-density profiles and the cognate protein profiles generated by weighting their sequences using these scales. Notably, our procedure is completely computational and does not impose any physicochemical constraints on the sampled scales. As a consequence, it provides an unbiased, detailed characterization of the space of amino-acid scales and their effect on the cognate mRNA/protein sequence complementarity. Finally, there are ongoing efforts in our and other laboratories to test the hypothesis that compositional sequence complementarity between mRNAs and their cognate proteins implies binding between them. The primary aim of the present work, however, is to assess the impact of affinity scales on such sequence complementarity, which remains a fact even in the absence of experimental verification of the hypothesis that it implies binding [4–6,9,11].

Materials and methods

Data sets

Complete annotated proteomes of *Escherichia coli*, *Methanocaldococcus jannaschii* and *Saccharomyces cerevisiae* along with the corresponding mRNA coding sequences were analyzed. The protein sequence data was extracted from the UniProtKB database with the maximal-protein-evidence-level set to 4, including only reviewed Swiss-Prot entries for the analysis [17–19]. Coding sequences for each protein were extracted using the 'Cross-references' section of each entry in the UniProtKB. Of all the entries, the first one satisfying the length criterion of mRNA

PLOS COMPUTATIONAL BIOLOGY

length = 3 x protein length + 3 was selected and the sequence downloaded from the European Nucleotide Archive Database. All sequences containing non-canonical amino acids or nucleobases were disregarded in the analysis. The complete sets of mRNA/protein data used in this study are included in S1 Dataset. Note that in the present study our analysis is reserved for primary sequences of complete mRNA coding sequences and their cognate proteins only, without consideration of higher-order structural organization. This not only enables a direct 1-to-1 mapping between mRNA and protein sequences, but also allows for a full exploration of the complementarity hypothesis in the unstructured context. Analysis of the influence of structure on the side of protein has been published elsewhere [15], while an analogous analysis on the side mRNAs will be the topic of our future work.

Correlation calculation

The mRNA nucleobase-density profiles and the corresponding protein nucleobase-affinity profiles were compared by calculating the linear Pearson correlation coefficients R between them. Prior to calculation, the sequences were smoothed via a window-averaging procedure with a 63-nucleotide window for mRNAs and a 21-residue window for proteins as used before [4–6,20]. Importantly, sequence profile comparison is largely insensitive to the size of the averaging window, as shown previously [4]. Furthermore, the scale values obtained during the simulation may result in any arbitrary value including negative numbers. Once the simulation is finished, the resulting scales are rescaled so that the values are in the range of [0, 1] i.e. the lowest value is set to 0, the highest to 1 and the rest are rescaled accordingly. This is done in order to distinguish between truly different scales and the rescaled versions of the same scale. Importantly, this procedure has no impact on our analysis since both window-averaging and calculation of Pearson correlation coefficients are invariant with respect to linear rescaling.

Monte Carlo simulations

The simulations were carried out using a combination of a C++ program for calculating the proteome-average correlation coefficients between mRNA and protein sequence profiles [6] and a Python script for implementing the Monte Carlo (MC) search. At each step in a given MC simulation, anywhere between 1 and 4 randomly chosen scale values were perturbed by a randomly chosen offset. A simulated annealing procedure was implemented in order to vary the size of the offsets from a randomly chosen value between [-0.1, 0.1] in the beginning of the simulation to a value between [-0.01, 0.01] in the end, with a linear ramp between the two. A given MC move i.e. a given scale, is accepted according to a zero-temperature, downhill Metropolis criterion: if a new scale results in a lower average Pearson R across the proteome (<R>) as compared to the scale it was derived from, it is accepted, and it is rejected otherwise. Affinity scales were generated individually for each of the four RNA nucleobase (uracil—URA, cytosine—CYT, adenine—ADE and guanine—GUA) as well as PUR (preference for both ADE and GUA). Given that the mRNA PYR fraction in a given stretch is directly related to the PUR fraction (%PYR = 1—%PUR), the PYR scales are by definition the inverses of the PUR scales and were for this reason not explicitly included in our analysis. Finally, the number of steps and the speed of simulated annealing both influence the final result. The number of steps has been chosen to be at least 3 times the number of moves necessary to reach a stable minimum.

Selection of scales for landscape generation

The MC approach does not produce interaction scales for a given level of matching but the other way around–only once a scale has been created, its level of matching is calculated. In

PLOS COMPUTATIONAL BIOLOGY

order to generate the full landscape, we chose those scales that showed the closest value of matching to the target value of the reported Pearson R. In the construction of the landscape, we include individual scales that are within +/-0.01 in Pearson R from each reported value of <R>. Considering the very fast evolution of interaction scales, for some low values of <R> we did not obtain a full set of 1000 scales that would match this criterion, which were therefore not included.

Analysis and software used

Data analysis was performed using the R statistical programming language. Calculations were performed using custom software written in C++ and Python. Plotting and data visualization was performed in R and Python, while figures were generated in Gimp and Inkscape.

Statistical significance tests

The MC-optimized scales were compared with the corresponding physically realistic knowledge-based scale by calculating the Pearson correlation coefficient R between them. The significance of the obtained correlation coefficients was ascertained by a randomization procedure whereby the reported p-values correspond to the fraction of a set of 10^6 scales with randomly chosen values exhibiting a more negative Pearson R than the MC-generated scales (a one-tailed significance test). Two-tailed p-values were calculated by multiplying the initial p-value by 2 if below 0.5, or as 1—p-value otherwise. Combined p-values were calculated according to Fisher's method based on two-tailed p-values [21,22]. The p-values were calculated from their respective X distributions utilizing Microsoft Excel's function CHIDIST().

Results

Optimized scales, which lead to a close matching between mRNA nucleobase-density profiles and sequence-weighted cognate protein profiles, could be identified in an extremely low number of MC moves. For example, the number of MC moves required to reach a proteome-wide average matching of mRNA PUR-density profiles with $<R> \leq -0.86$ in *E. coli* is only 322 ± 66.5 (standard deviation) as evaluated over 1000 independent MC simulations initiated with the system time as a random seed for each run (S1 Fig). Importantly, for all the combinations tested, the optimal scales appear to be extremely well defined and emerge robustly at the end of all independent MC simulations (scales with average weights are given in S1 Table). Also, the equivalent optimized scales for the three organisms studied are highly similar to each other with pairwise Pearson Rs in all cases exceeding 0.86 (S1 Table). The minor deviation between the scales of different organisms can be explained by the codon usage bias as well as different amino-acid composition of the respective organisms. For example, the difference between $CYT_{E.\ Coli}$ and $CYT_{M.\ Jannaschii}$ can be traced back to the different choice of codons for arginine.

In Fig 2A and 2B, we trace the evolution of the average and the standard deviation of the normalized values of individual amino-acid weights corresponding to a scale that was optimized for matching the mRNA PUR-density profiles in *E. coli*. The mean weights corresponding to the majority of amino acids, as obtained from 1000 independent repetitions, exhibit well-defined ranks starting already with low levels of matching and attain their definitive ranks already at approximately $<R> = -0.5$ (Fig 2A). In this particular case, the extreme weights correspond reproducibly to Phe on the high side and Glu and Lys on the low side. The reproducibility of optimal scales is attested by the extremely low standard deviations of amino-acid weights at the extremely high values of <R> (Fig 2B). Importantly, although a sharp drop in standard deviation towards high levels of matching is observed, a clear trend for the specific

PLOS COMPUTATIONAL BIOLOGY

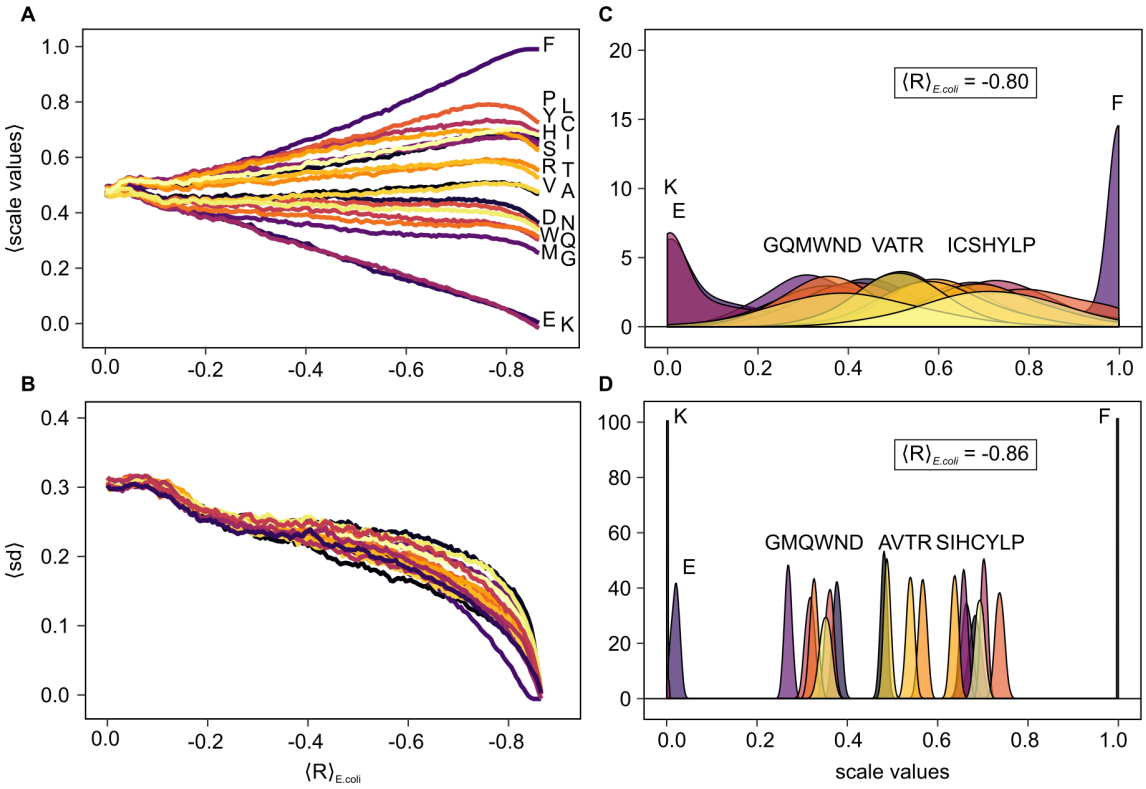


Fig 2. Dependence of profile matching on scale properties. The average level of cognate matching between mRNA PUR-density profiles and cognate protein PUR-affinity profiles in *E. coli* in relation to A) the average and B) the standard deviation of the rescaled PUR-affinity values for each amino acid yielding the given level of matching. The color code in B is the same as in A; C) distribution of PUR-affinity scale values at a matching level of <R> = -0.80; D) distributions of values of the best-matching PUR-affinity scales generated in 1000 different MC simulations (<R> = -0.86). Panels A,B and C,D share the same x-axis.

https://doi.org/10.1371/journal.pcbi.1005648.g002

values is achieved only very late in the simulations i.e. for the most extreme values of <R> only. Expectedly, the highest diversity of scale values is obtained for <R> ~ 0, but a substantial variability is retained even for intermediate-to-high values of matching: e.g. for <R> between -0.4 and -0.6, the standard deviations remain close to 0.25 in normalized units (Fig 2B). For comparison, a uniform random distribution between 0 and 1 results in a standard deviation of $1/\sqrt{12} \sim 0.29$.

The sequences of mRNAs and their cognate proteins are, of course, linked by the universal genetic code. Therefore, suitable amino-acid scales for weighting protein sequences to match their cognate mRNA's density profiles correspond, for a particular nucleobase, to the relative fractions of that nucleobase in the respective codons, weighted by the codon usage bias. For example, the scale derived in such a way results in an average matching of <R> = -0.86 for the mRNA PUR-density profiles in *E. coli* and the cognate protein sequences weighted by the

Table 1. Minimum <R> values achieved for each organism for the MC-optimized scales.

	ADE	CYT	GUA	URA	PUR
E. coli	-0.89	-0.75	-0.84	-0.80	-0.86
M. jannaschii	-0.90	-0.85	-0.89	-0.87	-0.92
S. cervisiae	-0.87	-0.78	-0.86	-0.84	-0.89

https://doi.org/10.1371/journal.pcbi.1005648.t001

average, usage-bias-weighted PUR content of the respective codons. Equivalent levels of matching were obtained for all nucleobases and organisms (Table 1). Interestingly, the scale composed of the average MC-optimized weights derived from *E. coli* mRNA PUR-density profiles correlates with the scale derived from codon PUR fractions with a Pearson R = 0.997 (S1 Table). Similar results were obtained for all other nucleobases and in all other organisms, albeit with slightly lower levels of correlation in some cases (S1 Table).

The main advantage of the MC approach, besides its efficiency, is that it also provides thorough sampling of suboptimal scales. This has allowed us to explore the development of scale properties in relation to the degree of mRNA/protein sequence complementarity. When looking at the distribution of specific values for a given level of <R> over the whole set of mRNAs and proteins, strong differences between the final optimized values and the intermediate values are identified. In Fig 2C and 2D, we show the complete distributions of weights for different amino acids at two different levels of average matching in *E. coli* (<R> = -0.80 and <R> = -0.86) for the PUR-density scales. Here, from each independent simulation, the scale resulting in <R> closest to the target value was selected. Note that the <R> values of all selected scales round to the reported target value at the second decimal place. As can be seen, only the most extremely optimized scales i.e. those with <R> = -0.86, exhibit well-defined weights for the majority of amino acids. For example, although an average Pearson <R> of -0.80 can be considered a high level of matching, most scale weights are still broadly distributed at that level (Fig 2C). Importantly, different amino acids exhibit distributions of different widths at a given value of <R>, with some converging to tighter distributions earlier than others. For example, the weights for Phe, Lys and Glu attain their final values early on and exhibit standard deviations that are lower than for any other amino acids at most values of <R>. In general, the distinct behavior of the optimized weights corresponding to individual amino acids is also seen for other scales and in other organisms (S2 Fig).

How many distinctly different scales are able to perform similarly well when it comes to the sequence profile matching of mRNAs with their cognate proteins? To separate scales at a given level of matching into several subsets, we have applied a hierarchical clustering algorithm, which has allowed us to build dendrograms for each level of <R>. As a natural distance measure between scales in this clustering approach, we have used 1—R, with lower numbers signifying higher similarity between two given scales and *vice versa*. In Fig 3A, we show two such dendrograms capturing the diversity of scales obtained by matching mRNA PUR-density profiles in *E. coli* at <R> = -0.8 and <R> = -0.86. To build a landscape of affinity scales, we have cut these dendrograms at specified distance cutoffs and have reported the number of clusters at a given cutoff as exemplified in Fig 3A for *E. coli*. The landscape is presented in Fig 3B with the values given representing the upper cutoff for the distance (1—R) of two scales in one cluster. From this landscape, it is clear that only at the highest levels of correlation between mRNA nucleobase profiles and nucleobase-affinity profiles of their cognate proteins do the affinity scales show a very similar structure. At lower levels, a very diverse set of scales is able to perform comparably well i.e. the lower values of matching can be attributed to a wider class of interaction scales.

Previously, several complete interaction scales for nucleobase/amino-acid interactions have been reported from different groups [23–26]. While some of these scales focus on the general affinity of amino acids for RNAs, but do not differentiate between specific nucleobases, other scales report on the specific propensities of the 20 amino acids for each of the four RNA nucleobases. The latter, for example, include scales derived in different ways including chromatographic experiments [7], absolute binding free energy calculations [9], classical and quantum-mechanical interaction enthalpy calculations [26], knowledge-based analysis of nucleobase/amino-acid contacts derived from X-ray and NMR structures [6] and simulated

PLOS COMPUTATIONAL BIOLOGY

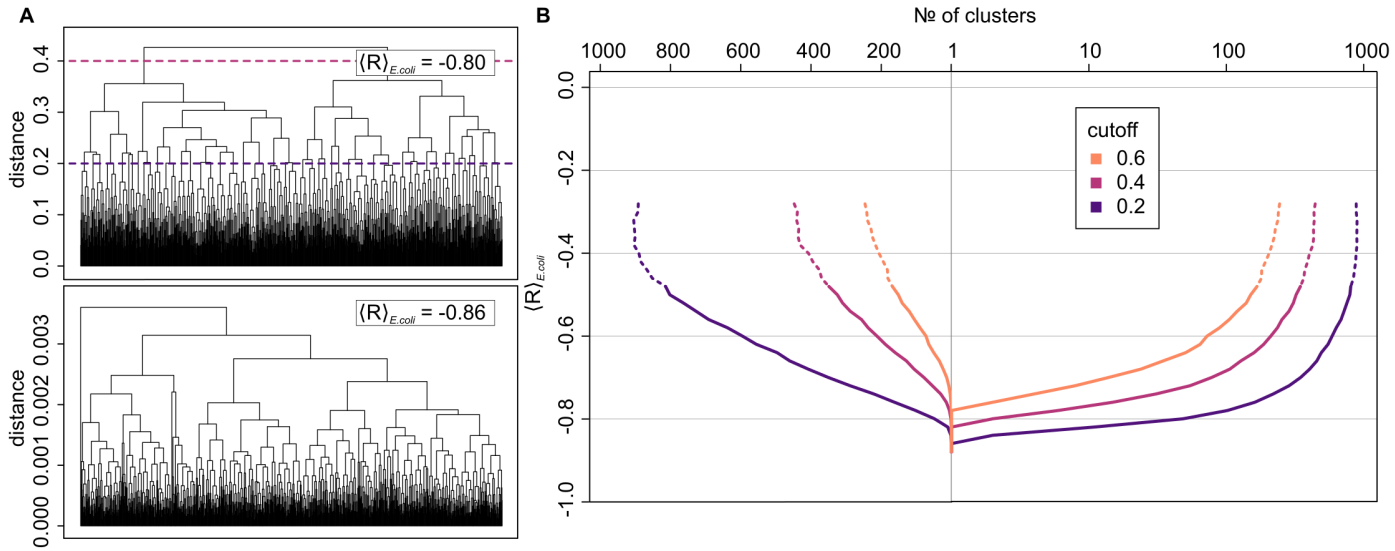


Fig 3. Landscape of affinity scales. A) two dendrograms exemplifying the process of clustering of scales for landscape generation (top, <R> = -0.80; bottom, best level of matching with <R> = -0.86). The height of the barrier between two clusters (left axis) is given as distance = 1—R between two scales. The presented results are obtained for mRNA PUR-density profiles and cognate protein PUR-affinity profiles in *E. coli*. The dotted lines capture the cutoffs of 0.2 and 0.4 used in Fig 3B; B) the number of clusters of scales at a given average level of matching as obtained by Pearson R-based clustering (left: linear scale; right: logarithmic scale) at different values of the upper cutoff for the distance (1—R) of two scales in one cluster.

https://doi.org/10.1371/journal.pcbi.1005648.g003

partitioning experiments [10,11]. Here, we focus on the knowledge-based scales as they are arguably the most relevant proxies for the nucleobase/amino-acid affinities at the realistic RNA/protein interfaces. In Fig 4A, we show the Pearson Rs between each of MC-derived scale

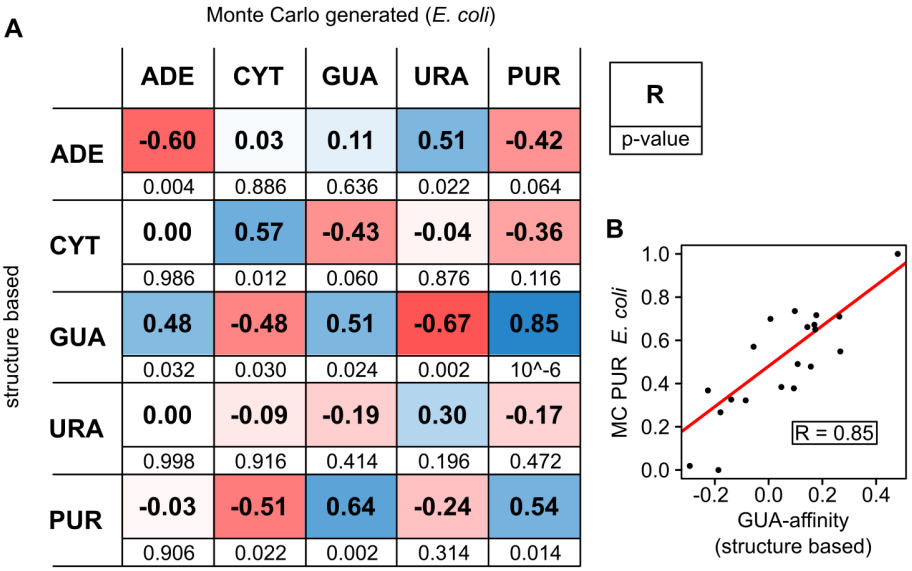


Fig 4. Comparison between MC-optimized scales and physical affinity scales. A) Pearson Rs (bold, top) and the associated p-values (regular, bottom) between all the MC-optimized scales derived from *E. coli* and the published knowledge-based nucleobase/amino-acid affinity scales [6]; B) comparison between the optimal scale derived from matching the mRNA PUR-density profiles in *E. coli* and the knowledge-based GUA-affinity scale (Pearson R = 0.85 between the two).

https://doi.org/10.1371/journal.pcbi.1005648.g004

for the *E. coli* dataset with each of the physical, knowledge-based scales derived by Polyansky and Zagrovic [6]. Below the correlation coefficients, we list the p-value capturing the statistical significance for the specific correlations. Importantly, three out of five pairs of the corresponding scales (GUA, CYT, PUR) exhibit Pearson Rs > 0.5 and p-values ≤ 0.024 each, suggesting high statistical significance. Moreover, the URA scales also exhibit a positive correlation coefficient (R = 0.30, p-value = 0.098), albeit not as strong as the other three pairs. Interestingly, the knowledge-based ADE-affinity scale displays a strong anti-correlation with its MC-generated counterpart. This anti-correlating behavior is also seen if one focuses on PURs only. Namely, although ADE is a purine base, its affinity scale correlates inversely with the values for MC-generated PUR scale. On the other hand, the GUA-affinity knowledge-based scale shows by far the strongest correlation with the generated PUR scales: the relative preferences of amino acids when it comes to interaction with GUA in RNA-protein structures show values very similar to those obtained by our MC procedure, which only considers the matching between mRNA PUR-density profiles and appropriately weighted cognate protein profiles. The two correlate with a Pearson R of 0.85 and no major outliers (Fig 4B).

We have also calculated the combined p-values from the two-tailed values given in Fig 4 for two different combinations of entries in the table. In the case of the Fisher method for combining p-values, the individual tests need to be independent from each other [21,22]. Arguably, the closest subset to this requirement is the combination of the four individual diagonal elements, entailing a comparison between the corresponding scales for individual bases. This set results in a highly significant combined p-value of 1.7×10^{-4}. Moreover, when including all combinations of URA, CYT, ADE GUA and PUR scales, a significance level of 8.3×10^{-11} is reached. Here, it should also be noted that there exist other suboptimal scales, which correlate better with the knowledge-based scales than do the optimal scales. For example, in the case of the *E.coli* PUR scale vs. the knowledge-based GUA scale, the highest correlation achieved between the two is R = 0.94, which is obtained for a suboptimal scale that itself results in an average proteome-wide correlation of $\langle R \rangle = -0.79$. A complete analysis of suboptimal scales and their relationships with knowledge-based scales is presented in S2 Table. In general, the fact that our optimization scheme produces scales that are similar to the physically realistic nucleobase/amino-acid binding affinity scales shows that our sampling is thorough and, more importantly, suggests that compositional matching of mRNA and protein sequence profiles may indeed be related to binding between them.

In addition, a total of 544 different one-dimensional scales [27,28], capturing different physicochemical properties of amino acids including size, hydrophobicity or interaction propensities, have been compared with the scales derived in this work by calculating the Pearson R between them. The result is visualized in Fig 5 in the case of the *E. coli* PUR scale. Here, it should be noted that the sign of correlation in this comparison depends only on the definition of a given scale and does not carry additional significance: for example, a hydrophilicity scale may be defined as a hydrophobicity scale and result in the same absolute result, but with an inverted sign. In general, the majority of the scales do not correlate closely with the MC-derived scales, but most of those that do are indeed in some way related to RNA/protein interactions or, interestingly, protein structural disorder. For example, the strongest correlation (R = 0.85) is obtained for the knowledge-based GUA-affinity scale [6], while the Woese pyridine affinity scale [29] ranks among the top 3% of all scales and exhibits an R = -0.63 with the MC-derived PUR-based scale. Parenthetically, a potential explanation for the observed lower density of Pearson Rs around 0 (Fig 5) may be that approximately 1/3 of all physical scales examined are related to amino-acid hydrophobicity [4]. Since our optimized PUR scale in general correlates negatively with hydrophobicity scales i.e. positively with hydrophilicity scales, this could create a somewhat lower density of Pearson Rs around 0.

PLOS COMPUTATIONAL BIOLOGY

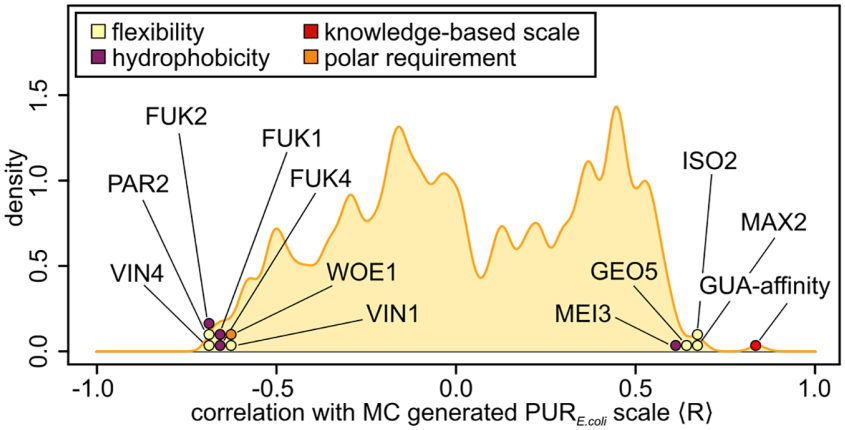


Fig 5. Comparison with other amino-acid properties. Distribution of Pearson Rs between all the AAindex [27,28] scales and the optimal scale derived from matching the mRNA PUR-density profiles in *E. coli*, with all scales showing a correlation |R|>0.60 explicitly marked: GUA-affinity scale by Polyansky et al. [6]; WOE1, a pyridine affinity scale derived by Woese et al. [29]; GEO5, a scale capturing the properties of linker regions in proteins [41]; ISO2, a flexibility scale characterizing bends in proteins [40]; MAX2, a predictor for protein backbone topology [39]; PAR2, a scale related to X-ray B-factor values of residues [38]; VIN1 & VIN4: scales assessing the correct flexibility prediction in proteins [37]; MEI3, a scale describing the effect of protein size on residue hydrophobicity [42]; FUK1, FUK2 & FUK4, scales capturing the protein surface composition of several bacterial proteins [43].

https://doi.org/10.1371/journal.pcbi.1005648.g005

Discussion

In the present work, we have sampled the space of amino-acid scales using as a fitness function the proteome-average matching between the mRNA nucleobase density profiles and the scale-weighted sequence profiles of cognate proteins. Importantly, our framework disregards all available biological information except the sequences of the two biopolymers and approaches the amino-acid scales from an abstract perspective as arrays of 20 numerical weights that can be chosen arbitrarily. This provides the benefit that the physicochemical interpretation of the scales does not need to be given *a priori*. Rather, the features of the fitness landscape of scales provide the constraints that any physical scales need to fulfill in order to be consistent with cognate mRNA/protein sequence complementarity. At the same time, effects like codon-usage bias are included by default. Utilizing this method, we have shown that it is possible to identify scales, which are highly optimized to lead to pronounced complementarity, and that our method is highly efficient at doing so.

Notably, it was not *a priori* clear that a simple MC search would result in unique optimal solutions for the matching problem. A remarkable result that virtually one single scale is the sole result of numerous independent simulations suggests that the intrinsic features of the underlying fitness landscape guide the development of these affinity scales towards a narrow range of values. Against our expectations, especially considering the vast combinatorial space the scales reside in, evolution of highly optimized scales could be performed in less than 200 MC moves (S1 Fig). This result on its own provides some pertinent information about the space the scales reside in. Namely, such a rapid and reproducible convergence can only be explained by a landscape that is shaped like a funnel. In this picture, the continuous downhill slope guides the search reproducibly towards the final optimized scale. Although simulated annealing was included in the MC approach, this would not suffice to reach the exact same minimum in every MC run: if there existed another local minimum separated by a significant barrier, we would also have sampled it. A potential criticism of this interpretation could be

that we start all our MC runs from the same scale (all weights equal to 0), which in principle could bias the final outcome. However, the MC runs initiated with different random number seeds decorrelate rapidly from each other in a few steps (as seen in the standard deviations in Fig 2B), only to converge again at the end of the runs. This, in effect, suggests that regardless of the starting point on the landscape, the MC searches reproducibly end up in the same minimum.

Our MC procedure has also resulted in a large number of suboptimal scales at all values of the average Pearson R between ~0.3 and -1. We have applied a hierarchical clustering algorithm to this data set and identified clusters of mutually similar scales. As reported in Fig 3, the development of the number of clusters as a function of the degree of matching exhibits a funnel-like shape. The important point to make is that the funnel is shallow and wide even up to the <R> values of -0.7 or higher. This means that there exist many rather different scales, which could lead to suboptimal, yet still relatively high levels of proteome-average complementary matching. This carries significant implications for our previous investigations of the complementarity hypothesis. As a case in point, our first report on the hypothesis involved a computationally derived scale of amino-acid affinity for PYR mimetics, which had led to a value of <R> of -0.74 across the human proteome [4]. This was interpreted as a strong signal of putative complementary interactions between mRNAs and their cognate proteins. On the other hand, our present analysis shows that there exist over 200 clusters of scales, whereby members of different clusters exhibit a correlation of at most 0.8 or less, all leading to a proteome-average correlation better than -0.74. In other words, the PYR-mimetic affinity scale used in our original study is by no means unique in its ability to lead to relatively high matching. While these findings do call for caution, it should be emphasized that they in no way contradict our previous interpretations. Namely, our present study only enumerates the list of possible scales that could lead to high complementarity. What is more, this list represents only a minor fraction of the space of all possible scales as indicated by our present and previous randomization studies [4,6,14]. The matching, in other words, is not a consequence of just any scale, but rather it can be achieved by only a select few, however mutually different from each other they may be.

The centerpiece of the present study is the comparison of the computationally-derived optimal scales with the published knowledge-based scales derived from structural data [6]. On the one hand, the computational scales are derived from a singular requirement that, when protein sequences are weighted by them, they yield profiles that resemble the cognate mRNA nucleobase-density profiles. Conversely, the knowledge-based scales are derived from the contact statistics of nucleobases and amino-acid side chains at the RNA-protein interfaces. They, therefore, report on the intrinsic binding preferences of the two sets of monomers. It is remarkable that Pearson correlation coefficients of up to 0.85 with highly significant p-values can be achieved between the two (Fig 4). The same can be said for the high overall p-values resulting from combining multiple scales. In other words, we start here from a simple computational exercise in which we search for scales that when applied to protein sequences yield profiles that match their cognate mRNA's nucleobase density profiles. The fact that the scales obtained in this way resemble the nucleobase/amino-acid binding affinity scales strongly suggests that profile matching and binding indeed may be related, as put forth by the complementarity hypothesis. However, the rather weak correlation with some other affinity scales shall not be ignored. A question remains as to why protein affinity profiles for GUA strongly match the mRNA PUR-density profiles, but not as well the mRNA GUA-density profiles. In line with this, why are the protein ADE-affinity profiles inversely related to the mRNA PUR-density profiles? It has been suggested that the ADE and URA nucleobases may be newer additions to biological systems, while the GUA and CYT may have been the very first nucleobases adopted [5]. If both the matching behavior as postulated

by the stereochemical hypothesis and the assumed timeline of RNA evolution hold true, this would suggest that the usage of ADE as an anti-matching base could have served a biologically important purpose. Namely, the presence of ADE has the potential to negatively regulate the level of complementarity and, therefore, the strength of binding between cognate partners, as previously suggested [5].

The present work accounts only for sequence data on both the mRNA and the protein side, with no secondary or higher-order structure elements being considered. This, of course, does not rule out structured mRNAs and proteins as interaction partners, but certainly limits the generality of the current work. It should, however, be noted that the binding between unstructured RNAs and intrinsically disordered proteins belongs to an important, large class of RNA-protein interactions and, moreover, provides a relevant context in which to look for cognate interactions [30–36]. In this regard, it may be potentially informative that some of the closest physically realistic scales to the optimal MC scales derived herein are linked with protein disorder (Fig 5) [37–43]. Finally, our results suggest that the degree of cognate mRNA-protein complementarity is heavily determined by the intrinsic binding affinities of just a handful of nucleobase/amino-acid pairs. For example, the opposite behavior of Phe and Glu/Lys define a large fraction of the PUR-density/PUR-affinity matching. At the same time, the nucleobase-binding preferences of other amino acids are much more ambiguous and diverse, even at relatively high levels of matching. The main biological implication of this is that it defines constraints on the mechanism by which genetic encoding could have evolved from binding, as proposed by the stereochemical hypothesis and our generalization of it. Specifically, the establishment of a coding relationship between codons and amino acids, which would be a consequence of complementary binding between cognate mRNA/protein pairs, is possible only for a narrow set of nucleobase-affinity values for several key residues, as defined by our study. Future research should provide information about this and other related open questions.

Supporting information

S1 Fig. Development of <R> as a function of the number of MC steps in the case of PUR-based scales.
(TIF)

S2 Fig. Scale value distributions for all scales and organisms investigated. Distributions of scale values for the most optimized scale of each individual MC simulation. The ordering of amino-acid symbols corresponds to the ordering of the means of the respective distributions. Results for ADE, CYT, GUA, URA and PUR are shown for each of the three organisms investigated–*E. coli*, *M. jannaschii* and *S. cervisiae*. The average level of matching over all scales contributing to a distribution is given in the plot.
(PDF)

S1 Table. Comparison between the MC-optimized scales and the codon-based scales. Simple scales obtained from the average nucleobase-content of the codons of individual amino acids as weighted by their codon usage bias. Pearson correlation coefficients R between the scales are given on the right. Correlations of scales obtained for different organisms are reported at the end.
(PDF)

S2 Table. Comparison of suboptimal scales which best match the knowledge-based scales. For each type of scale, a specific MC scale was selected that resembles the corresponding knowledge-based scale most closely. For these scales, the Pearson correlation coefficients R with the knowledge-based scales and the average, proteome-wide correlation coefficients that

result by their application, are reported.
(PDF)

S1 Dataset. Datasets employed in this study. Each of the original whole proteome mRNA/protein sequence datasets of *E. coli*, *M. jannaschii* and *S. cervisiae* used in this study in a space-separated text format.
(ZIP)

Author Contributions

Conceptualization: Lukas Bartonek, Bojan Zagrovic.

Data curation: Lukas Bartonek.

Formal analysis: Lukas Bartonek, Bojan Zagrovic.

Funding acquisition: Bojan Zagrovic.

Investigation: Lukas Bartonek, Bojan Zagrovic.

Methodology: Lukas Bartonek, Bojan Zagrovic.

Project administration: Bojan Zagrovic.

Resources: Bojan Zagrovic.

Software: Lukas Bartonek.

Supervision: Bojan Zagrovic.

Validation: Lukas Bartonek, Bojan Zagrovic.

Visualization: Lukas Bartonek, Bojan Zagrovic.

Writing – original draft: Lukas Bartonek, Bojan Zagrovic.

Writing – review & editing: Lukas Bartonek, Bojan Zagrovic.

References

1. Pederson T. RNA: Life's Indispensable Molecule, by James E. Darnell. 2011. Cold Spring Harbor Laboratory Press. ISBN: 978-1-936113-19-4. RNA. Cold Spring Harbor Laboratory Press; 2011;17: 1771–1774.
2. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. IUBMB Life. 2009; 61: 99–111. https://doi.org/10.1002/iub.146 PMID: 19117371
3. Yansheng Liu, Andreas Beyer, Ruedi Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016; 165: 535–550. https://doi.org/10.1016/j.cell.2016.03.014 PMID: 27104977
4. Hlevnjak M, Polyansky AA, Zagrovic B. Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. Nucleic Acids Res. 2012; 40: 8874–8882. https://doi.org/10.1093/nar/gks679 PMID: 22844092
5. Polyansky AA, Hlevnjak M, Zagrovic B. Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic. RNA Biol. 2013; 10: 1248–1254. https://doi.org/10.4161/rna.25977 PMID: 23945356
6. Polyansky AA, Zagrovic B. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. Nucleic Acids Res. 2013; 41: 8434–8443. https://doi.org/10.1093/nar/gkt618 PMID: 23868089
7. Woese CR. Evolution of the genetic code. Sci Nat. Springer Berlin Heidelberg; 1973; 60: 447–459.
8. Mathew Damien C, Luthey-Schulten Zaida. On the physical basis of the amino acid polar requirement. J Mol Evol. 2008; 66: 519–528. https://doi.org/10.1007/s00239-008-9073-9 PMID: 18443736

9. de Ruiter Anita, Zagrovic Bojan. Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. Nucleic Acids Res. 2015; 43: 708–18. https://doi.org/10.1093/nar/gku1344 PMID: 25550435
10. Matea Hajnic, Osorio Juan I, Zagrovic Bojan. Interaction preferences between nucleobase mimetics and amino acids in aqueous solutions. Phys Chem Chem Phys. Royal Society of Chemistry; 2015; 17: 21414–21422. https://doi.org/10.1039/c5cp01486g PMID: 26219945
11. Matea Hajnic, Osorio Juan Iregui, Zagrovic Bojan. Computational analysis of amino acids and their side-chain analogs in crowded solutions of RNA nucleobases with implications for the mRNA-protein complementarity hypothesis. Nucleic Acids Res. 2014; 42: 12984–12994. https://doi.org/10.1093/nar/gku1035 PMID: 25361976
12. Woese CR, Dugre DH, Saxinger WC, Dugre SA. The molecular basis for the genetic code. Proc Natl Acad Sci U S A. National Academy of Sciences; 1966; 55: 966–974. PMID: 5219702
13. Michael Yarus, Widmann Jeremy Joseph, Knight Rob. RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code. J Mol Evol. Springer-Verlag; 2009; 69: 406–429. https://doi.org/10.1007/s00239-009-9270-1 PMID: 19795157
14. Polyansky Anton A, Hlevnjak Mario, Zagrovic Bojan. Analogue encoding of physicochemical properties of proteins in their cognate messenger RNAs. Nat Commun. Nature Publishing Group; 2013; 4: 2784. https://doi.org/10.1038/ncomms3784 PMID: 24253588
15. Andreas Beier, Bojan Zagrovic, Polyansky Anton A. On the Contribution of Protein Spatial Organization to the Physicochemical Interconnection between Proteins and Their Cognate mRNAs. Life (Basel, Switzerland). Multidisciplinary Digital Publishing Institute; 2014; 4: 788–99.
16. Kyrpides Nikos C, Ouzounis Christos A. Mechanisms of Specificity in mRNA Degradation: Autoregulation and Cognate Interactions. J Theor Biol. 1993; 163: 373–392. https://doi.org/10.1006/jtbi.1993.1126 PMID: 8246508
17. UniProt: a hub for protein information. Nucleic Acids Res. Oxford University Press; 2015; 43: D204–D212. https://doi.org/10.1093/nar/gku989 PMID: 25348405
18. Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Bridge Alan J, et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods in molecular biology (Clifton, NJ). 2016. pp. 23–54.
19. Rolf Apweiler, Amos Bairoch, Wu Cathy H. Protein sequence databases. Curr Opin Chem Biol. 2004; 8: 76–80. https://doi.org/10.1016/j.cbpa.2003.12.004 PMID: 15036160
20. Mario Hlevnjak, Bojan Zagrovic. Malleable nature of mRNA-protein compositional complementarity and its functional significance. Nucleic Acids Res. 2015; 43: 3012–3021. https://doi.org/10.1093/nar/gkv166 PMID: 25753660
21. Fisher Ronald A. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1925.
22. Frederick Mosteller, Fisher RA. Questions and Answers. Am Stat. 1948; 2: 30–31.
23. Suxin Zheng, Robertson Timothy A Varani Gabriele. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J. 2007; 274: 6378–9391. https://doi.org/10.1111/j.1742-4658.2007.06155.x PMID: 18005254
24. Michael Fernandez, Yutaro Kumagai, Standley Daron M, Sarai Akinori, Kenji Mizuguchi, Shandar Ahmad, et al. Prediction of dinucleotide-specific RNA-binding sites in proteins. BMC Bioinformatics. BioMed Central; 2011; 12: S5.
25. Treger Michèle Westhof Eric. Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recognit. John Wiley & Sons, Ltd.; 2001; 14: 199–214. https://doi.org/10.1002/jmr.534 PMID: 11500966
26. Jakubec Dávid, Hostaš Jiří, Laskowski Roman A, Hobza Pavel, Vondrášek Jiří. Large-Scale Quantitative Assessment of Binding Preferences in Protein–Nucleic Acid Complexes. J Chem Theory Comput. American Chemical Society; 2015; 11: 1939–1948. PMID: 26894243
27. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000; 28: 374. PMID: 10592278
28. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2007; 36: D202–D205. https://doi.org/10.1093/nar/gkm998 PMID: 17998252
29. Woese CR. On the evolution of the genetic code. Proc Natl Acad Sci U S A. National Academy of Sciences; 1965; 54: 1546–1552. PMID: 5218910
30. Rinn John L, Ule Jernej, Attar N, Borsenberger V, Crowe MA, Lehbauer J, et al. 'Oming in on RNA–protein interactions. Genome Biol. BioMed Central; 2014; 15: 401. https://doi.org/10.1186/gb4158 PMID: 24485348

31. Baltz Alexander G, Munschauer Mathias, Schwanhäusser Björn, Vasile Alexandra, Murakawa Yasuhiro, Schueler Markus, et al. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. Mol Cell. 2012; 46: 674–690. https://doi.org/10.1016/j.molcel.2012.05.021 PMID: 22681889

32. Wimberly Brian T, Guymon Rebecca, McCutcheon John P, White Stephen W, Ramakrishnan V. A Detailed View of a Ribosomal Active Site: The Structure of the L11–RNA Complex. Cell. 1999; 97: 491–502. PMID: 10338213

33. Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Beckmann Benedikt M Strein Claudia, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 2012; 149: 1393–1406. https://doi.org/10.1016/j.cell.2012.04.031 PMID: 22658674

34. Beckmann Benedikt M, Horos Rastislav, Fischer Bernd, Castello Alfredo, Eichelbaum Katrin, Alleaume Anne-Marie, et al. The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat Commun. Nature Publishing Group; 2015; 6: 10127. https://doi.org/10.1038/ncomms10127 PMID: 26632259

35. Mihaly Varadi, Fruzsina Zsolyomi, Mainak Guharoy, Peter Tompa, Moore PB, Ravindranathan S, et al. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. PLoS One. Public Library of Science; 2015; 10: e0139731. https://doi.org/10.1371/journal.pone.0139731 PMID: 26439842

36. Alfredo Castello, Bernd Fischer, Frese Christian K Horos Rastislav, Alleaume Anne-Marie Foehr Sophia, et al. Comprehensive Identification of RNA-Binding Domains in Human Cells. Mol Cell. 2016; 63: 696–710. https://doi.org/10.1016/j.molcel.2016.06.029 PMID: 27453046

37. Mauno Vihinen, Esa Torkkila, Pentti Riikonen. Accuracy of protein flexibility predictions. Proteins Struct Funct Genet. Wiley Subscription Services, Inc., A Wiley Company; 1994; 19: 141–149. https://doi.org/10.1002/prot.340190207 PMID: 8090708

38. Parthasarathy S, Murthy MR. Protein thermal stability: insights from atomic displacement parameters (B values). Protein Eng. Oxford University Press; 2000; 13: 9–13. PMID: 10679524

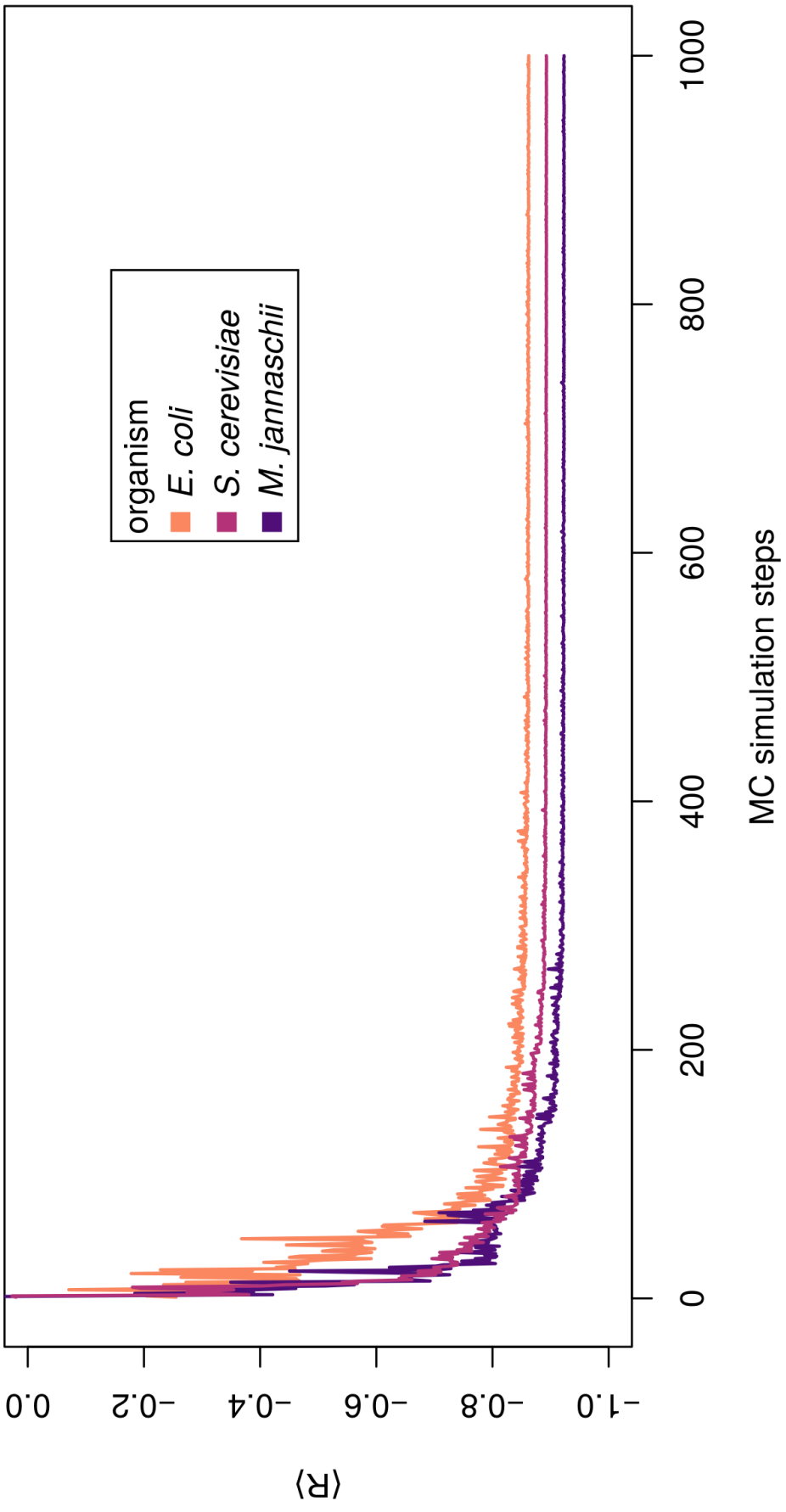
39. Maxfield FR, Scheraga HA. Status of empirical methods for the prediction of protein backbone topography. Biochemistry. 1976; 15: 5138–5153. PMID: 990270

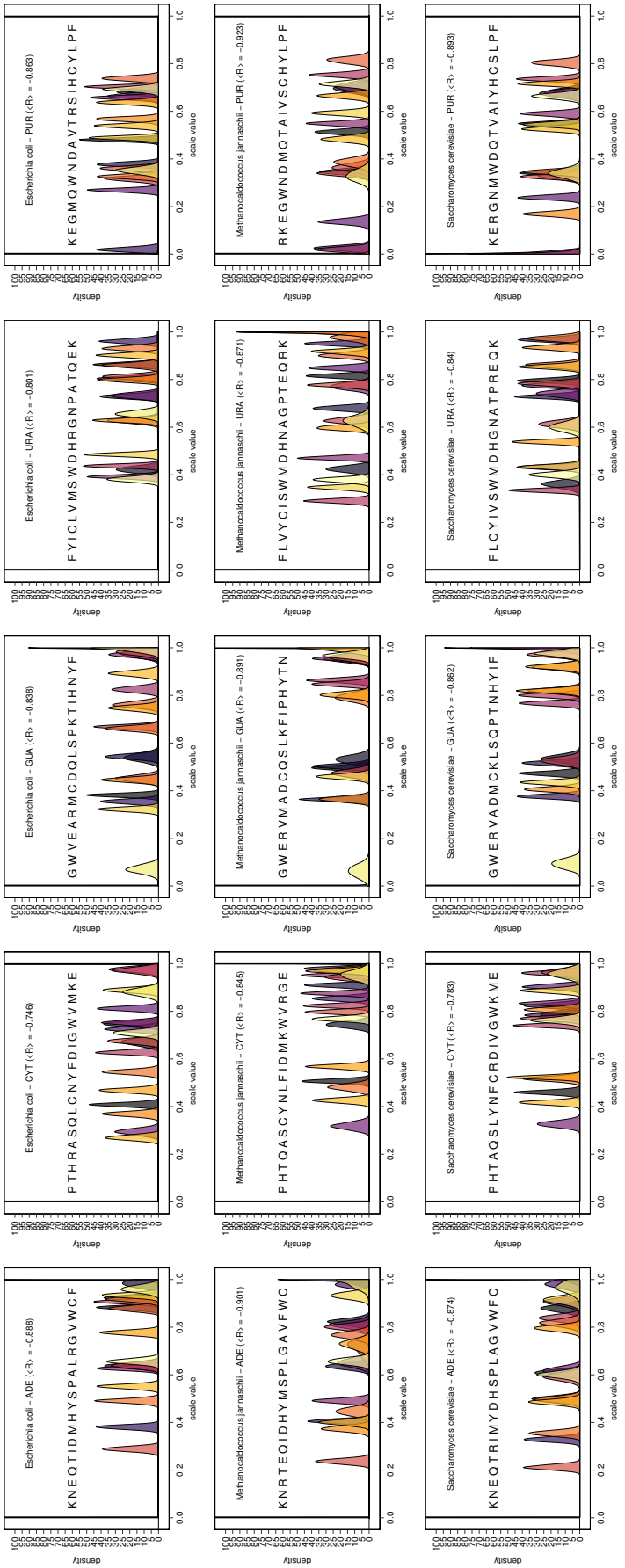
40. Isogai Y, Némethy G, Rackovsky S, Leach SJ, Scheraga HA. Characterization of multiple bends in proteins. Biopolymers. Wiley Subscription Services, Inc., A Wiley Company; 1980; 19: 1183–1210. https://doi.org/10.1002/bip.1980.360190607 PMID: 7378550

41. George Richard A, Heringa Jaap. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng. Oxford University Press; 2002; 15: 871–879. PMID: 12538906

42. Meirovitch H, Rackovsky S, Scheraga HA. Empirical Studies of Hydrophobicity. 1. Effect of Protein Size on the Hydrophobic Behavior of Amino Acids. Macromolecules. American Chemical Society; 1980; 13: 1398–1405.

43. Satoshi Fukuchi, Ken Nishikawa. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol. 2001; 309: 835–843. https://doi.org/10.1006/jmbi.2001.4718 PMID: 11399062


organism
E. coli
S. cerevisiae
M. jannaschii
⟨R⟩
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0
200
400
600
800
1000
MC simulation steps


Escherichia coli – ADE (<R> = –0.888)
KNEQTIDMHYSPALRGVWCF
Methanocaldococcus jannaschii – ADE (<R> = –0.901)
KNRTEQIDHYMSPLGAVFWC
Saccharomyces cerevisiae – ADE (<R> = –0.874)
KNEQTRIMYDHSPLAGVWFC
Escherichia coli – CYT (<R> = –0.746)
PTHRASQLCNYFDIGWVMKE
Methanocaldococcus jannaschii – CYT (<R> = –0.845)
PHTQASCYNLFIDMKWVRGE
Saccharomyces cerevisiae – CYT (<R> = –0.783)
PHTAQSLYNFCRDIVGWKME
Escherichia coli – GUA (<R> = –0.838)
GWVEARMCDQLSPKTIHNYF
Methanocaldococcus jannaschii – GUA (<R> = –0.891)
GWERVMADCQSLKFIPHYTN
Saccharomyces cerevisiae – GUA (<R> = –0.862)
GWERVADMCKLSQPTNHYIF
Escherichia coli – URA (<R> = –0.801)
FYICLVMSWDHRGNPATQEK
Methanocaldococcus jannaschii – URA (<R> = –0.871)
FLVYCISWMDHNAGPTEQRK
Saccharomyces cerevisiae – URA (<R> = –0.84)
FLCYIVSWMDHGNATPREQK
Escherichia coli – PUR (<R> = –0.883)
KEGMQWNDAVTRSIHCYLPF
Methanocaldococcus jannaschii – PUR (<R> = –0.923)
RKEGWNDMQTAIVSCHYLPF
Saccharomyces cerevisiae – PUR (<R> = –0.893)
KERGNMWDQTVAIYHCSLPF
density
scale value

E.coli

	Scale	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	correlation (R)
Monte Carlo	ADE	0.886	0.988	0.640	0.385	1.000	0.935	0.649	0.642	0.000	0.916	0.644	0.298	0.885	0.503	0.922	0.782	0.560	0.937	0.958	0.663	0.998
Simple		0.924	1.000	0.639	0.389	1.000	0.962	0.639	0.615	0.000	0.941	0.639	0.278	0.931	0.513	0.946	0.802	0.593	0.945	1.000	0.639	
Monte Carlo	CYT	0.406	0.665	0.749	1.000	0.720	0.821	0.295	0.749	0.981	0.622	0.973	0.671	0.000	0.543	0.361	0.464	0.262	0.883	0.880	0.699	0.989
Simple		0.401	0.736	0.824	1.000	0.799	0.808	0.326	0.801	1.000	0.600	1.000	0.739	0.000	0.529	0.363	0.528	0.321	0.898	1.000	0.796	
Monte Carlo	GUA	0.385	0.542	0.557	0.361	0.996	0.000	0.990	0.982	0.834	0.668	0.454	0.991	0.769	0.675	0.447	0.750	0.894	0.317	0.062	1.000	0.996
Simple		0.368	0.535	0.535	0.391	1.000	0.000	1.000	1.000	0.892	0.707	0.535	1.000	0.753	0.696	0.482	0.728	0.875	0.361	0.070	1.000	
Monte Carlo	URA	0.866	0.429	0.736	0.964	0.000	0.810	0.735	0.398	1.000	0.437	0.639	0.823	0.872	0.937	0.804	0.632	0.903	0.494	0.661	0.380	0.994
Simple		0.938	0.440	0.757	1.000	0.000	0.869	0.779	0.413	1.000	0.472	0.611	0.827	0.939	1.000	0.850	0.664	0.936	0.511	0.611	0.391	
Monte Carlo	PUR	0.481	0.681	0.379	0.019	1.000	0.272	0.665	0.668	0.000	0.698	0.315	0.355	0.736	0.331	0.566	0.631	0.534	0.490	0.363	0.697	0.997
Simple		0.477	0.667	0.333	0.000	1.000	0.248	0.667	0.644	0.000	0.736	0.333	0.333	0.759	0.333	0.579	0.622	0.535	0.491	0.333	0.667	

M.janaschii

	Scale	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	correlation (R)
Monte Carlo	ADE	0.825	0.983	0.635	0.410	0.987	0.822	0.678	0.495	0.000	0.809	0.730	0.238	0.764	0.451	0.366	0.730	0.411	0.933	1.000	0.656	0.992
Simple		0.848	1.000	0.630	0.407	1.000	0.792	0.630	0.470	0.000	0.762	0.630	0.260	0.754	0.394	0.370	0.748	0.449	0.919	1.000	0.630	
Monte Carlo	CYT	0.495	0.732	0.904	1.000	0.852	0.977	0.317	0.870	0.947	0.818	0.940	0.791	0.000	0.471	0.977	0.559	0.422	0.966	0.949	0.765	0.990
Simple		0.461	0.843	0.914	1.000	0.898	0.967	0.371	0.950	1.000	0.861	1.000	0.856	0.000	0.510	0.988	0.598	0.458	0.967	1.000	0.888	
Monte Carlo	GUA	0.506	0.539	0.502	0.373	0.961	0.000	0.973	0.955	0.862	0.844	0.487	1.000	0.954	0.801	0.365	0.813	0.993	0.465	0.047	0.975	0.998
Simple		0.518	0.538	0.538	0.354	1.000	0.000	1.000	1.000	0.863	0.897	0.538	1.000	0.981	0.833	0.416	0.825	0.980	0.499	0.075	1.000	
Monte Carlo	URA	0.822	0.435	0.690	0.961	0.000	0.851	0.767	0.475	0.995	0.294	0.628	0.786	0.905	0.978	1.000	0.610	0.931	0.351	0.642	0.382	0.993
Simple		0.841	0.399	0.705	1.000	0.000	0.926	0.743	0.476	1.000	0.337	0.642	0.747	0.909	1.000	0.997	0.599	0.870	0.417	0.642	0.366	
Monte Carlo	PUR	0.516	0.682	0.340	0.020	1.000	0.124	0.695	0.549	0.013	0.754	0.353	0.336	0.809	0.371	0.000	0.661	0.479	0.591	0.333	0.711	0.998
Simple		0.515	0.667	0.333	0.000	1.000	0.091	0.667	0.522	0.000	0.711	0.333	0.333	0.765	0.333	0.011	0.646	0.489	0.565	0.333	0.667	

S.cerevisiae

	Scale	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	correlation (R)
Monte Carlo	ADE	0.890	0.998	0.617	0.337	1.000	0.918	0.616	0.506	0.000	0.846	0.607	0.218	0.825	0.366	0.507	0.805	0.505	0.928	0.978	0.615	0.999
Simple		0.885	1.000	0.613	0.342	1.000	0.913	0.613	0.505	0.000	0.837	0.613	0.226	0.843	0.348	0.523	0.812	0.494	0.915	1.000	0.613	
Monte Carlo	CYT	0.462	0.791	0.834	1.000	0.781	0.912	0.328	0.834	0.965	0.744	0.961	0.775	0.000	0.522	0.811	0.525	0.416	0.887	0.969	0.767	0.998
Simple		0.434	0.826	0.839	1.000	0.812	0.909	0.371	0.880	1.000	0.768	1.000	0.814	0.000	0.537	0.828	0.540	0.439	0.907	1.000	0.800	
Monte Carlo	GUA	0.475	0.540	0.516	0.372	1.000	0.000	0.973	1.000	0.766	0.803	0.532	0.973	0.931	0.821	0.396	0.826	0.923	0.433	0.095	0.979	0.999
Simple		0.476	0.529	0.529	0.388	1.000	0.000	1.000	1.000	0.804	0.816	0.529	1.000	0.942	0.852	0.409	0.825	0.935	0.438	0.058	1.000	
Monte Carlo	URA	0.796	0.364	0.730	0.961	0.000	0.766	0.727	0.426	1.000	0.340	0.620	0.785	0.855	0.973	0.939	0.540	0.856	0.436	0.584	0.402	0.997
Simple		0.858	0.374	0.749	1.000	0.000	0.824	0.752	0.437	1.000	0.350	0.615	0.769	0.880	1.000	0.946	0.557	0.868	0.466	0.615	0.396	
Monte Carlo	PUR	0.548	0.668	0.342	0.005	1.000	0.234	0.672	0.583	0.000	0.734	0.325	0.325	0.798	0.338	0.171	0.714	0.525	0.537	0.337	0.664	0.999
Simple		0.530	0.667	0.333	0.000	1.000	0.218	0.667	0.573	0.000	0.729	0.333	0.333	0.823	0.333	0.170	0.714	0.518	0.529	0.333	0.667	

Correlation between organisms

	ADE	CYT	GUA	URA	PUR
E.coli vs *M.janaschii*	0.878	0.864	0.968	0.968	0.869
E.coli vs *S.cerevisiae*	0.940	0.927	0.976	0.979	0.930
M.janaschii vs *S.cerevisiae*	0.971	0.981	0.993	0.987	0.982

E. coli

compared scales	Pearson R	scale fitness
ADE_{KB} vs ADE_{MC}	0.719	-0.048
CYT_{KB} vs CYT_{MC}	0.868	-0.556
GUA_{KB} vs GUA_{MC}	0.803	-0.540
URA_{KB} vs URA_{MC}	0.818	-0.491
PUR_{KB} vs PUR_{MC}	0.862	-0.736
GUA_{KB} vs PUR_{MC}	0.943	-0.789

M. jannaschii

compared scales	Pearson R	scale fitness
ADE_{KB} vs ADE_{MC}	0.699	0.314
CYT_{KB} vs CYT_{MC}	0.912	-0.472
GUA_{KB} vs GUA_{MC}	0.823	-0.322
URA_{KB} vs URA_{MC}	0.787	-0.517
PUR_{KB} vs PUR_{MC}	0.864	-0.767
GUA_{KB} vs PUR_{MC}	0.918	-0.865

S. cerevisiae

compared scales	Pearson R	scale fitness
ADE_{KB} vs ADE_{MC}	0.676	-0.200
CYT_{KB} vs CYT_{MC}	0.866	-0.542
GUA_{KB} vs GUA_{MC}	0.772	-0.347
URA_{KB} vs URA_{MC}	0.807	-0.346
PUR_{KB} vs PUR_{MC}	0.839	-0.729
GUA_{KB} vs PUR_{MC}	0.947	-0.843

KB … knowledge based (refers to scales published in Polyansky AA, Zagrovic B. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. Nucleic Acids Res. 2013;41: 8434–8443. as Set 2+)

MC … Monte Carlo generated – closest matching scales found in the total set of scales generated

scale fitness … mean whole proteome correlation resulting from the given scales applied on the respective proteome. Negative values show better interaction according to the definition of affinities. For calculation details please refer to the Materials and Methods section.

3

A REVIEW OF CURRENT METHODS TO ASSESS RNA/PROTEIN INTERACTIONS

Bojan Zagrovic, Lukas Bartonek, and Anton A Polyansky. "RNA-protein interactions in an unstructured context." In: *FEBS letters* 592.17 (2018), pp. 2901–2916 [75]

In this review article, we summarized the current knowledge on RNA/protein interactions with a strong focus on interactions in an unstructured context and the potential implications of the *stereochemical hypothesis* on our understanding of such interactions. The central idea discussed and defended in the article concerns the usage of one-dimensional physicochemical property profiles as an analytical tool for studying RNA/protein interactions, especially in an unstructured context.

PERSONAL CONTRIBUTION

The author of this thesis contributed the derivation of the original knowledge-based affinity scales, first presented in this review article (Figure 3A) and referenced as *NDB2017* throughout the main text. Furthermore, the author was involved in writing the parts of the original draft related to these scales as well as the applications of machine learning in RNA/protein interaction prediction, in addition to reviewing and editing the complete draft.

FEBS Letters

FEBS PRESS science publishing by scientists

REVIEW ARTICLE

RNA-protein interactions in an unstructured context

Bojan Zagrovic[1], Lukas Bartonek[1] and Anton A. Polyansky[1,2]

1 Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Austria
2 MM Shemyakin and Yu A Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia

Correspondence
B. Zagrovic, Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5, A-1030 Vienna, Austria
Fax: +43 1 4277 9522
Tel: +43 1 4277 52271
E-mail: bojan.zagrovic@univie.ac.at

(Received 23 April 2018, revised 12 May 2018, accepted 13 May 2018, available online 21 June 2018)

doi:10.1002/1873-3468.13116

Edited by Wilhelm Just

Despite their importance, our understanding of noncovalent RNA–protein interactions is incomplete. This especially concerns the binding between RNA and unstructured protein regions, a widespread class of such interactions. Here, we review the recent experimental and computational work on RNA–protein interactions in an unstructured context with a particular focus on how such interactions may be shaped by the intrinsic interaction affinities between individual nucleobases and protein side chains. Specifically, we articulate the claim that the universal genetic code reflects the binding specificity between nucleobases and protein side chains and that, in turn, the code may be seen as the Rosetta stone for understanding RNA–protein interactions in general.

Keywords: intrinsically disordered proteins; long noncoding RNAs; nucleobase/amino acid interaction affinity scales; RNA–protein granules; RNA–protein interactions

Our understanding of biology at the molecular level is transforming rapidly and central to this is a radical reappraisal of the importance and ubiquity of RNA–protein interactions [1–6]. From gene expression regulation to RNA processing and decay to protein localization, it is now clear that many cellular processes are unimaginable without direct, specific interactions between RNA and RNA-binding proteins (RBPs) [1–7]. A good example in this regard is the case of mRNAs and their interaction networks. Namely, recent proteome-wide experiments utilizing UV-cross-linking and mass spectrometry have led to hundreds of newly identified proteins, which interact directly with mRNAs, but do not contain any known RNA-binding domains (RBDs) [2,5,6,8–12]. For example, 40% of the 570 identified yeast mRNA-binding proteins lack well-defined RBDs, are not associated with any presently known functions in RNA biology and even include different metabolic enzymes and transcription factors [10]. Importantly, many such 'enigmRBPs' contain repetitive, low-complexity sequence regions and are intrinsically unstructured i.e., disordered [2,5,6,8–12].[1] A systematic Gene Ontology (GO) analysis of the dependence between the calculated degree of structural disorder and functional enrichment clearly shows that 'poly(A) RNA binding' is the most enriched function among highly disordered proteins (>80%) in human (Fig. 1A) [13]. In general, RNA/DNA binding functions feature strongly among the highly disordered proteins (Fig. 1A). The reverse is also true: ~30% of all mRNA-binding RBPs in HeLa cells, for example, have one half or more of their residues in an unstructured state [5,10] (Fig. 1B). Traditionally, structural disorder in proteins has not only been linked with

[1]Throughout this text, we use the terms "unstructured" and "disordered" interchangeably when referring to a lack of well-defined, persistent secondary or tertiary structure in macromolecules.

Abbreviations
ADE, adenine; CYT, cytosine; FMRP, Fragile X Mental Retardation Protein; FUS, Fused-In-Sarcoma; GO, Gene Ontology; GUA, guanine; IDPs, unstructured/disordered proteins; lncRNAs, long noncoding RNAs; MD, molecular dynamics; PUR, purine; PYR, pyrimidine; RBDs, RNA-binding domains; RBPs, RNA-binding proteins; RBR, RNA-binding regions; RF, random forest; RRM, RNA recognition motif; SVM, support vector machine; URA, uracil; UTRs, untranslated regions.

transcription, chromatin modification, and signaling [14,15] but is also key for the assembly of phase-separated organelles such as P-bodies and stress granules, major sites of RNA processing and storage [16–20]. Although conserved flexibility and disorder are important for RNA–protein binding, the fundamental principles behind such interactions remain largely unexplored. As a consequence, the computational tools to predict binding between RNA and unstructured proteins are scarce [21]. The present review focuses on the recent work on RNA–protein interactions in an unstructured context with a particular focus on how the specificity in such interactions may be determined.

Cellular context of RNA interactions with unstructured RBPs

Unstructured RBPs are closely involved at all stages of the life cycle of different RNAs in the cell. In the case of mRNAs and long noncoding RNAs (lncRNAs), this includes transcription, processing by the spliceosome and the exon junction complex, nuclear export, translation at the ribosome, and decay (reviewed in [5] and [22]). Importantly, a series of recent studies have reported instances of phase separation in the cytoplasm and nucleoplasm, a process similar to lipid domain (e.g., so-called 'rafts') formation in the membrane, which in aqueous environment results in the formation of liquid droplets (reviewed in [20], [23], and [24]). These droplets are typically rich in proteins and RNA and define nonmembrane-bound cellular compartments such as nucleoli, P-bodies, stress granules, and Cajal bodies. They are also known to be the sites of mRNA storage, processing, and decay. What is critical here is that it has been shown that both multivalency (i.e., existence of many low-affinity binding sites) and the presence of low-complexity, disordered regions in proteins may be required for the formation of such phase-separated compartments [19,23,25,26]. Moreover, such compartments are known to preferentially attract single-stranded nucleic acids and are stabilized in their presence [25,27]. For these reasons, it is likely that understanding of RNA interactions with disordered proteins and their contextual dependence may contribute directly to our understanding of the structure, assembly, and function of phase-separated cellular compartments as well. Conversely, this also suggests that it is critical to study such interactions in the context of crowded, dehydrated, low-dielectric environments similar to those present in P-bodies or stress granules, where RNA molecules may also be in a single-stranded, unstructured form.

Lack of 3D organization motivates sequence-based analysis

Due to diminished structural constraints, the properties of intrinsically unstructured/disordered proteins (IDPs) depend much more on their linear sequence features than in the case of folded proteins [14,15,28].

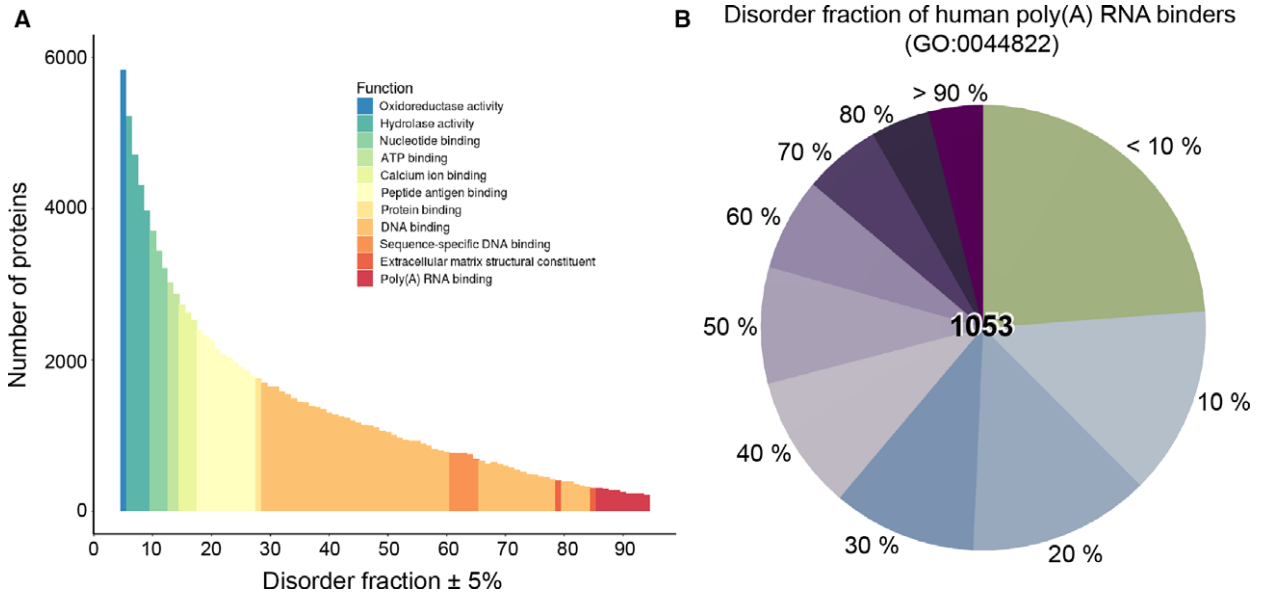


Fig. 1. RNA binding and protein disorder. (A) Distribution of the degree of structural disorder among human proteins according to IUPRED [42], with the most enriched GO functional term for each bin indicated in color; (B) Structural disorder among poly(A) RNA-binding proteins in human according to IUPRED [42]. Figure 1A was reproduced with permission (Oxford University Press) from Ref. [13].

For example, accurate prediction of protein disorder using sequence information is almost routine nowadays [28,29]. In general, the structural, dynamical, and functional characteristics of IDPs can in many cases be successfully related to the linear distribution of different physicochemical properties of individual residues and/or short fragments along their sequence, that is, 1D physicochemical profiles. As an illustration, we present in Fig. 2 several such 1D profiles obtained for Fused-In-Sarcoma (FUS) protein [26,30–32] including its disorder probability, charge density, hydrophobicity, and GUA-affinity profiles. FUS is an abundant nuclear protein involved in mRNA transcription, processing, and transport [30,31,33–37] and implicated in the pathophysiology of amyotrophic lateral sclerosis and frontotemporal lobar degeneration [37,38]. Importantly, FUS contains four highly disordered RNA-binding regions (RBR) including RGG/RG domains [5], whose arginine residues are known to be the targets of arginine methyl transferases [39]. This post-translational modification has been shown to modulate the nuclear transport of FUS [40] as well as RNA-binding activity of another RNA-binding IDP FMRP [41].

A distribution of disorder probability along the FUS sequence, that is, its disorder profile as calculated in this case by IUPRED [42], gives one a possibility to identify unstructured regions within the protein and match them with the known RBRs. For instance, repetitive linear sequence elements required for RNA binding, like RS and RGG/RG domains, reside in highly unstructured regions and/or the flanking regions of structured RNA-binding domains (e.g., RRM) as was also shown for a number of RNA-binding IDPs (see Järvelin *et al.* [5] for detailed review of disorder organization in 46 different well-characterized RNA-binding IDPs). The FUS RRM in particular is surrounded by two RGG/RG domains (RBR 2 and 3, Fig. 2). Such an organization enables a synergy between disordered and structured RNA-binding domains and can increase the RNA affinity of the RRM in question as compared to an isolated one. For example, it has been shown recently that in FUS the flanking RGG/RG domains actually enable the RNA-binding activity of its RRM [32]. While RRMs typically exhibit highly conserved sequence features, the flanking regions can also display sequence conservation and more importantly—disorder conservation [29]. Using DisCons, a server for the analysis of disorder score conservation in multiple sequence alignment [43], Varadi and coworkers have shown that disorder conservation is especially pronounced for the residues at RNA-binding interfaces.

A 1D distribution of charged residues along protein sequence gives one the possibility to identify prominent negatively or positively charged regions. Interestingly, the FUS RRM together with the flanking RBRs displays a prominent alternation in the net charge (Fig. 2). Many IDPs are polyampholytes and exhibit simultaneously positively and negatively charged regions, whereby their distribution within the sequence shapes protein conformational behavior [44] and can also modulate phosphorylation patterns within the disordered regions [45]. The alternating sequence of net charge has also been shown to be an important determinant in phase-separated droplet formation involving DDX4 RNA-helicase [25]. A distribution of hydrophobic residues along an IDP sequence is another important property which affects liquid phase separation, due largely to the contribution of aromatic residues to multivalent interactions with RNA and other proteins (e.g., π–π and cation–π), as shown for a number of RNA-binding IDPs including FUS [19,23,25,26,46–48]. Interestingly, while the FUS RRM is prominently hydrophobic, its RBRs display a different level of hydrophobicity depending on the neighboring sequence context (Fig. 2; the FUS hydrophobicity profile was determined by using a consensus hydrophobicity scale Factor I [49]). This could contribute to a slightly different sequence specificity of FUS RBRs and the preference of its RRM to interact with the relatively more hydrophobic ADE-rich sequences [32]. Finally, in order to map directly the RNA-binding specificity along the FUS sequence, we plot its 1D guanine (GUA) affinity profile as derived by using a knowledge-based nucleobase/amino acid side-chain affinity scale (see below) and the formalism described elsewhere [50]. Here, one can see that the disordered RBR regions perfectly match the peaks of GUA-affinity, which is in line with an experimental observation of the general preference of these regions for GUA-rich sequences [5,32] (N.B. following the standard thermodynamic convention, the low ΔG values correspond to high affinity and *vice versa*). This also agrees with the analysis of the distribution of GUA-preferring amino acids in structured proteins, which shows that they are mostly enriched in unstructured loop regions [51].

Sequence-based prediction of RNA–protein interactions

Computational prediction of RNA–protein interactions in an unstructured context is still relatively underdeveloped due in part to our lack of understanding of the basic physicochemical principles at play and a general lack of high-resolution information when it comes to

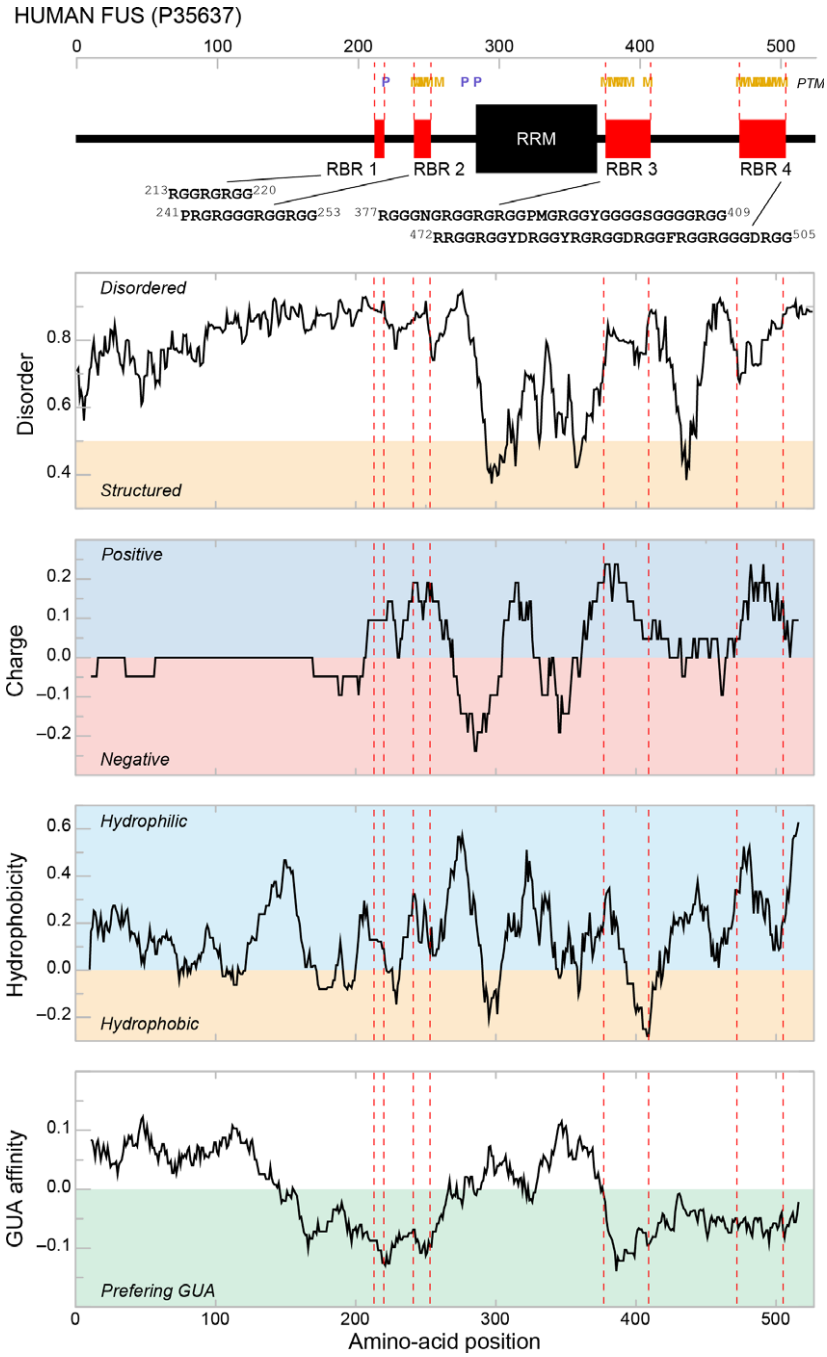


Fig. 2. 1D physicochemical profiles of FUS. Disorder, charge, hydrophobicity and GUA-affinity profiles of human FUS in relation to its domain structure. The charge, hydrophobicity and GUA-affinity profiles were determined by using a running-average window of 21-residues.

unstructured partners. This situation, however, will likely change in the near future and the lessons learned and the methods developed when it comes to RNA–protein interactions in general will likely be important in the case of structural disorder as well. Most modern methods for predicting RNA–protein binding from sequence information, physicochemical properties of individual RNA and protein building blocks, or global RNA–protein characteristics are based on machine learning strategies. In particular, predictive models are trained on the known interactors by using features such as composition, hydrophobicity, and evolutionary information [52–55]. An early such approach, developed by Pancaldi and Bahler, is based on support vector machine (SVM) and random forest (RF) formalisms and uses the known RNA and protein features as predictors, including *inter alia* GO terms, protein localization, and chromosome position information [56]. A comparable approach developed by

Dobbs and coworkers, also using SVM and RF approaches, showed that using just sequence information as descriptors may result in a significantly better prediction [57]. Chen and coworkers employed an alternative approach and used an extended naïve Bayes approach on sequence composition data [58]. In contrast, Tartaglia and coworkers put a major focus on the physicochemical properties and trained their catRAPID model on features such as secondary structure propensity, hydrogen bonding, and van der Waals interactions [59,60]. Neural networks have also been applied to identifying RNA–protein interactions, following their overwhelming success in areas such as image or speech recognition and natural language processing. DeepBind, for example, uses a neural network trained on several experimental datasets including RNAcompete, ChIP-seq, and HT-SELEX experiments and performs well over a wide range of metrics [61]. Qu and coworkers have shown the benefits of a deeper network in the rather similar case of DNA binding: for an extensive realistic dataset, an accuracy of 94.2% could be achieved [62]. RCK, on the other hand, uses k-mers to evaluate the binding propensity and slightly outperforms DeepBind, at least on the RNAcompete dataset [63,64]. Finally, RNAcontext, although somewhat outdated, still performs well in comparison to RCK and DeepBind, sometimes even outperforming the two. In this approach, both RNA sequence and secondary structure information are used to accurately predict binding to RBPs [65].

Although they tackle an important and difficult problem, the above methods still frequently suffer from a limited accuracy, a lack of general applicability, and the fact that only a few of them [21,50,53,59] are fundamentally steeped in basic physicochemical principles. The latter criticism is probably the most important downside of the machine learning approaches: they frequently do not allow for a deeper insight into the physicochemical underpinnings of RNA–protein interactions. Another limitation concerns the availability of experimental data to train these models on. For instance, the only existing method for predicting RNA binding to IDPs, DisoRDPbind [21], was trained on a set of only 14 annotated RNA-binding proteins. Here, we would like to argue that it may in many cases be more advantageous to approach this problem purely from a physicochemical perspective and exploit the intrinsic affinities between nucleobases and amino acids as a foundation for predicting the interaction between longer biopolymers. This approach seems promising especially in the context of single-stranded RNA interacting with unstructured proteins, given that in those cases the dependence on the properties of basic building blocks is likely most pronounced.

Nucleobase/amino acid affinities as a basis for understanding RNA–IDP interactions

The recent findings about the mode of interaction between RNA and IDPs highlight the relevance of predicting and understanding such interactions from the perspective of first principles. For instance, it is known that short linear motifs with <10 residues are one of the primary ways of how IDPs interact with partners [22,28]. Moreover, even for IDPs that fold upon binding, the participating regions comprise mostly local, 10–70 residue segments [22,28]. On the other hand, RNA-binding sites for proteins also typically include just a few nucleotides organized in single-stranded, linear stretches [63,66]. For example, RNAcompete studies have shown that most RBPs bind single-stranded RNAs with <10 nucleotides, and none absolutely requires a defined RNA secondary structure [63]. Hence, one may expect that the principles of RNA interactions with IDPs or the unfolded states of otherwise folded proteins can be deduced by examining structural and thermodynamic aspects of interactions between individual nucleobases and amino acids. With diminished structural constraints, the behavior of long polymers can more easily be related to their constituent building blocks, a possibility that has remained unexplored until recently.

Significant progress in understanding nucleobase–amino acid interactions has over the years been made using computational approaches [50,67–80]. For example, analysis of 3D structures of RNA– or DNA–protein complexes has yielded the relative binding preferences of nucleobases and amino acids together with a geometric and energetic characterization of their interactions [50,67,68,72,75–79]. Despite a limited amount of statistics that could be extracted from the analysis of 3D complexes [50], the obtained amino acid preferences for GUA and adenine (ADE) are significantly robust and reproducible when it comes to the scale values and a moderate anticorrelation between the GUA and ADE scales as shown for different sets of RNA–protein complexes (Fig. 3A). Also, *ab initio* methods have been used to study the quantum-mechanical aspects of such binding, including hydrogen bonding [71], π–π [73,74] and cation–π interactions [70]. Finally, the binding free energy maps for several amino acids and DNA base pairs have also been reported [69]. While these early studies have typically either focused on a few bases and amino acids only or have simply been insufficiently quantitative, recently there appeared several computational studies with a more comprehensive outlook. For example, our recent

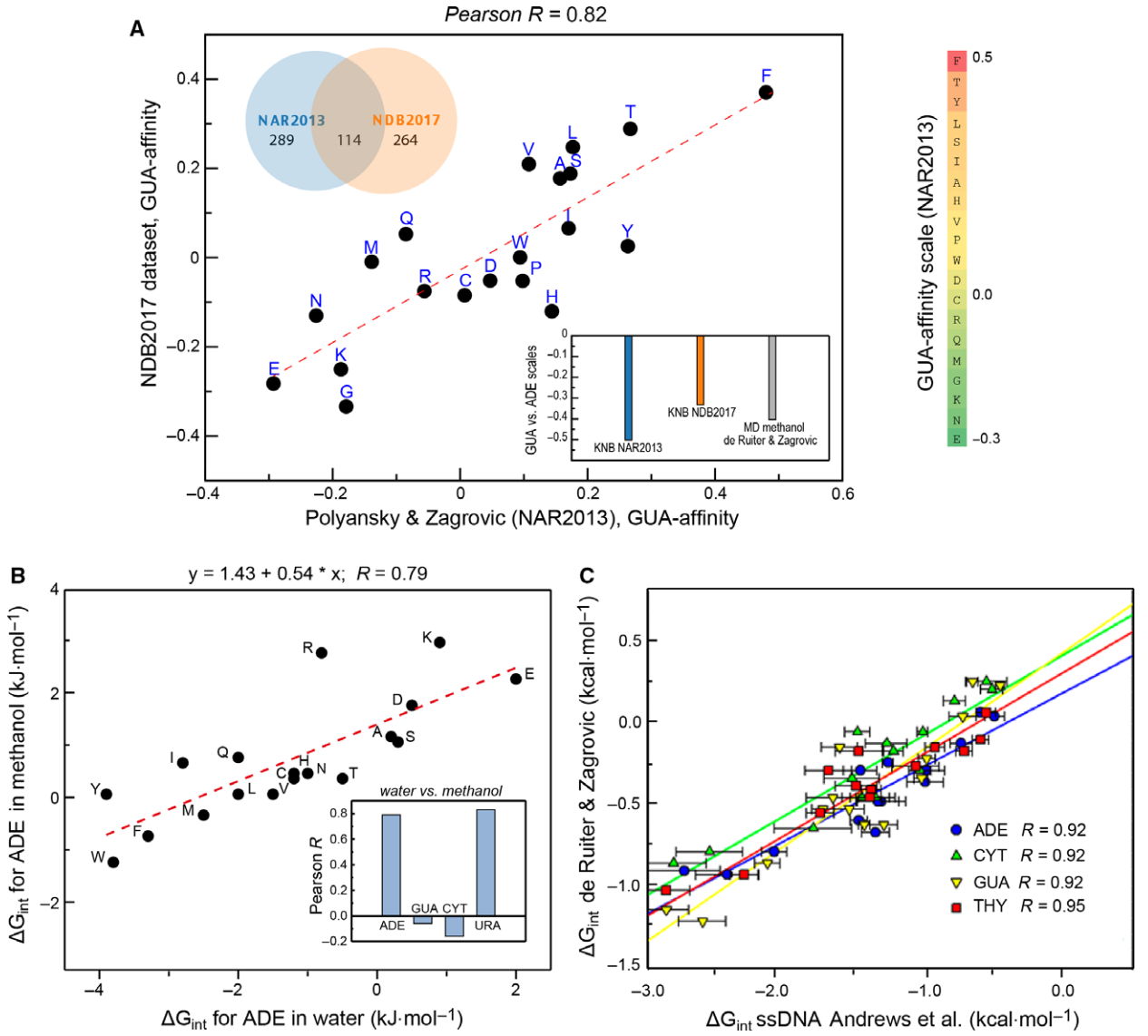


Fig. 3. Robustness of affinities between nucleobases and amino acid side chains. (A) A close correlation between the knowledge-based (KNB) scales of GUA/side-chain affinity derived from two largely independent sets of structures of RNA–protein complexes (NAR2013 [50] and NDB2017). The NDB2017 set was generated using a representative dataset from the Nucleic Acid Database (http://ndbserver.rutgers.edu) [110] with the resolution cutoff of 2.5 A using the identical method as described by Polyansky *et al.* [50]. The overlap between the two sets is given by the Venn diagram. Inset: anticorrelation between the GUA and ADE side-chain affinity scales is observed for two sets of knowledge-based scales (NAR2013 [50] and NDB2017) and the MD-based scales of nucleobase–amino acid affinity in methanol [81]. The bars indicate the Pearson R coefficients between the GUA and ADE scales. (B) ADE and URA amino acid-binding free energy scales are up to a constant largely insensitive to local dielectric constant, while those for GUA and CYT strongly depend on it (Pearson Rs given in the inset) [81]. (C) Andrews *et al.* [80] nucleobase–amino acid-binding ΔG scales (*x*-axis) correlate closely with de Ruiter *et al.* scales [81] (*y*-axis) Figure 3C was reproduced with permission (American Chemical Society) from Ref. [80].

analysis of the absolute binding free energies between all standard RNA–DNA nucleobases and amino acid side-chain analogs in different solvents [81,82] and their dependence on the local dielectric properties is, to the best of our knowledge, the first example where this key property has been determined within a single, self-consistent framework. Interestingly, the dielectric properties of the environment tune the affinity of amino acid side chains for GUA and cytosine (CYT), while having little or no effect on ADE and uracil (URA) scales, as shown by using umbrella-sampling simulations of individual nucleobases and amino acid side-chain analogs in water and methanol (Fig. 3B) [81,82]. These results are especially important when it

comes to understanding the basic principles of RNA–IDP interactions in the context of liquid RNA–protein granules characterized by a reduced dielectric constant as compared to bulk water. Importantly, the obtained affinities in water were found to be in an excellent agreement with the results of a subsequent independent analysis by Elcock and coworkers who have performed explicit-solvent MD simulations of long single- and double-stranded DNA molecules in heterogeneous aqueous mixtures of amino acids using a different MD force field [80] (Fig. 3C). The latter analysis also provided a comprehensive dissection of the salt dependence of nucleobase–amino acid interactions and the contribution of DNA sugar and phosphate groups to binding. Moreover, the nucleobase–amino acid affinity scales were also derived based on a simulated partitioning of amino acids between nucleobase-rich phases and water [83,84]. Finally, Vondrasek and coworkers have used the known PDB structures of DNA–protein complexes together with molecular mechanics and DFT-D *ab initio* calculations to estimate the binding preferences between all 20 natural amino acids and the four DNA bases [79].

When it comes to experimental determination of nucleobase–amino acid-binding preferences, only limited progress has been made over the years. Akinrimisi *et al*. [85] and Thomas *et al*. [86] have studied spectroscopically the solubility in water of several amino acids in the presence of either purines or different nucleosides, respectively. In the same way, Thomas *et al*. [86] have determined binding constants for several nucleoside and amino acid pairs. On the other hand, Woese *et al*. have evaluated chromatographically the interaction propensities of amino acids and different pyridine derivatives in water [87,88]. Finally, several groups have studied interactions between different nucleotides and polyamino acids, focusing typically on polylysine or polyarginine peptides [89–91]. The scarcity of the experimental studies in this context can in part be attributed to the low solubility of bases and some side chains as well as the weak interaction strengths involved. For example, our recent work suggests that only a handful of nucleobase side-chain affinities exceed 1 k_BT [81,82]. As discussed below, however, we see a systematic experimental determination of such affinities as one of the major open challenges for the future.

Nucleobase–amino acid affinities reflect the genetic code organization

As discussed above, RNA–protein interactions in an unstructured context are a ubiquitous feature of modern biological systems. However, they also provide an important perspective for understanding the establishment of the RNA–protein relationship in primordial systems in which structural disorder is likely to have been even more pervasive [92]. This, in particular, concerns the central aspect of the whole RNA–protein relationship: the process of translation and the genetic code [93,94]. Namely, with minor variations, the genetic code is universally conserved and it, without exaggeration, represents the very point where biological phenotype and genotype meet. However, despite 50 years of effort, the nature of the driving forces behind its establishment has remained largely unknown. Over the years, multiple theories have been proposed in this regard with varying levels of evidence [94–97]. Of relevance here, the 'stereochemical hypothesis' suggests that the key feature of ancient translation was a direct interaction between codons and amino acids they code for [87,88,94,98]. Although specific binding of isolated codons and their cognate amino acids has never been observed, analysis of amino acid-binding RNA aptamers [98] and RNA–protein interactions in the ribosome [99] has revealed that not only some codons but also anticodons, preferentially colocalize with their cognate amino acids. Importantly, early support for the hypothesis came from Woese and coworkers who analyzed the interaction preferences between amino acids and pyrimidine mimetics pyridines [87,88,94,100] (see also above). They showed that amino acids with a similar propensity to interact with pyridines also have similar codons.

A common feature of most studies of the code's origin has been their focus on individual codons and amino acids only [93,94,101] with little attention paid to the properties of longer biopolymers. However, any biases present at the level of individual groups may get cooperatively amplified in such cases, facilitating their detection. Moreover, if the stereochemical hypothesis is indeed true, then the genetic code could also be seen as a key for understanding RNA–protein interactions in general. Following this paradigm, we have recently explored the link between the physicochemical properties of mRNAs and the proteins they encode [102,103] and have made a surprising discovery. First, using both experimental and computational nucleobase–amino acid affinity scales, we could show that the nucleobase content of a given codon is directly related to the affinities of its cognate amino acids for the respective nucleobases (Fig. 4A). For example, the codon PYR content correlates with the PYR-mimetic affinity of the cognate amino acids with a Pearson R of −0.61, while the codon PUR content correlates with the knowledge-based GUA-affinities of the cognate amino acids with a Pearson R of −0.68 (Fig. 4A). What is more, this

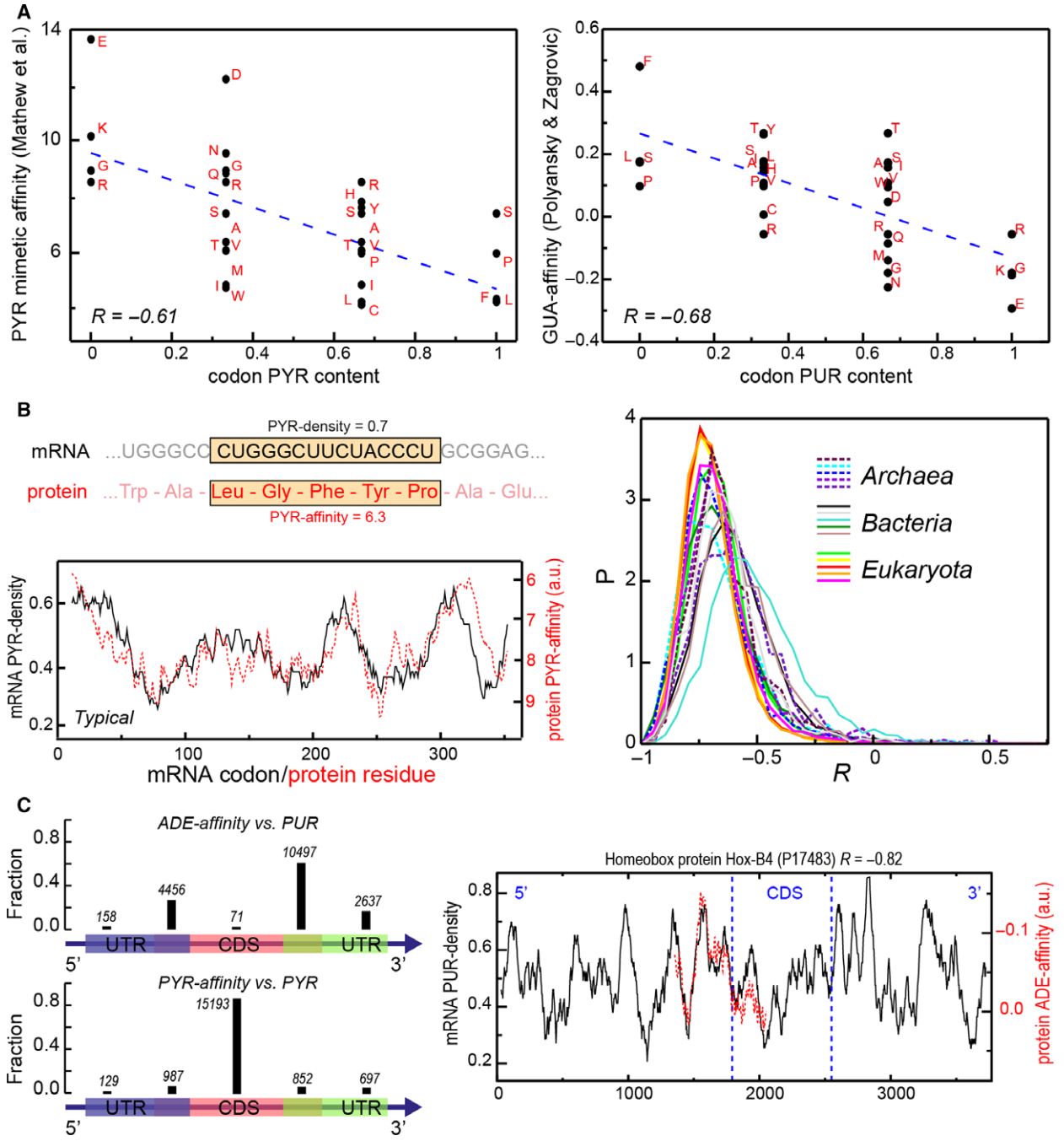


Fig. 4. The complementarity hypothesis. (A) Codon PYR content correlates with the cognate amino acid affinity for PYR mimetics [100], while codon PUR content correlates with the cognate amino acid affinity for GUA [50]. (B) Right: profile calculation method together with a typical pair of mRNA PYR density and protein PYR-mimetic affinity profiles in human; left: Pearson Rs for mRNA PYR density/protein PYR-mimetic affinity profiles in 15 species. (C) Left: Location of top matches for human mRNAs and cognate proteins, including UTRs and transition regions (violet/olive) for mRNA PUR density/protein ADE affinity, and mRNA PYR density/protein PYR affinity cases. Numbers of top matches are given above bars. Right: An example of a top match between mRNA PUR density and cognate protein ADE affinity. Figures 4A and 4B were reproduced with permission (Oxford University Press) from Ref. [102].

FEBS Letters **592** (2018) 2901–2916 © 2018 The Authors. *FEBS Letters* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

relationship was even further magnified in the context of complete biopolymers. A pyrimidine (PYR) density profile of an mRNA coding sequence can be obtained as a running average of the PYR content of its codons (Fig. 4B). Similarly, a PYR-affinity profile of the cognate protein can be estimated by weighting its sequence by the amino acid propensities to interact with PYR mimetics as captured in Woese's experiments or subsequent simulations [88,100]. The remarkable finding is that for most mRNA–protein pairs the two profiles match closely when aligned [102,103]. For example, the median Pearson correlation coefficient R between the two profiles over all human pairs is $R = -0.74$ (note that all affinity scales are defined such that negative Rs mean matching [102,103]).

This is unexpected and potentially far-reaching: although mRNAs and their cognate proteins are completely different biopolymers, they exhibit a strong, quantitative complementarity in that the density of a particular type of groups in one of them can be accurately predicted from the affinity profile for similar groups in the other one. Importantly, we could show that this finding is statistically extremely robust and holds equally well for organisms from all three domains of life (Fig. 4B) [102,103]. Moreover, we have confirmed these findings for biologically relevant PYRs as well as extended them to purines (PURs) by using knowledge-based nucleobase–amino acid affinities derived from structures of RNA–protein complexes [50,103]. For example, mRNA PUR density profiles quantitatively match GUA-affinity profiles of cognate proteins, with a median of $R = -0.80$ in human [50,103]. Notably, protein ADE affinity profiles exhibit a reverse property in that they match the PYR density of cognate mRNAs. We have shown that this stems from biosynthetically more complex amino acids that are thought to have entered biology later [50,103]. Finally, we have fully corroborated the above findings by using affinities derived by orthogonal approaches including umbrella-sampling MD simulations [81,82] or modeling of partitioning experiments [83,84].

The above results provide support for the stereochemical hypothesis of the origin of the genetic code, but they emphasize the importance of an extended, polymeric context in which the relatively weak affinities of individual building blocks can be amplified. Moreover, these findings support a novel hypothesis that in the unstructured state, mRNAs and the proteins they encode may be complementary to each other and bind in a coaligned manner, whereby the complementarity level is negatively regulated by mRNA ADE content [50,51,81–84,103–105]. Since compositional matching is seen for primary sequence profiles, we expect that the strongest interactions will occur if the partners are unstructured, yielding dynamic, multivalent, liquid-like complexes: in addition to IDPs, the hypothesis applies equally well to the unfolded states of otherwise folded proteins [51]. Importantly, the coarse-grained nature of the complementarity hypothesis allows one to extend it to interactions between unstructured proteins and nucleic acids other than their cognate mRNA coding regions [102,103]. The key point is that a physicochemical view of biomolecular sequences provides a measure of both binding propensity in an unstructured context and evolutionary relatedness for different RNA and protein molecules. In this sense, the above findings could be interpreted as general rules for understanding RNA–protein interactions: amino acids whose codons are enriched in PYR also display proteome-wide tendencies to be specific for PYR or ADE, while polar amino acids, encoded by PUR, predominantly interact with GUA (Fig. 4A). Moreover, these rules could be applied in noncoding situations as well. This can be illustrated in the case of noncoding 5′ and 3′ untranslated regions (UTRs) of mRNAs. For example, for hundreds of human mRNA sequences, the top match between the mRNA nucleobase density profiles and their cognate proteins' nucleobase-affinity profiles is observed in the UTRs or in the transitional regions including a UTR and a part of the coding sequence (Fig. 4C). In fact, for mRNA ADE density profiles, the majority of mRNAs and their cognate proteins fall into these categories. This is well illustrated in the case of HoxB4 mRNA and its cognate protein, where the top matching region includes a significant part of the 5′ UTR (Fig. 4C). As a whole, the above findings suggest that the structure of the universal genetic code reflects, in part, the binding specificity between nucleobases and amino acids and, conversely, support an exciting possibility that the universal genetic code may be seen as a key for understanding RNA–protein interactions in the unstructured context and beyond (Fig. 5).

Significance and outlook

An improvement in our understanding of the RNA–protein interactions in the unstructured context, including the ability to predict binding sites and relative affinities, would represent a major step forward from both fundamental and practical perspectives. Currently, most of our knowledge of RNA–protein interactions concerns structured RNA-binding motifs and, consequently, computational methods for predicting RNA-protein interactions in a physicochemically and structurally realistic manner are necessarily limited and biased by such knowledge [21,53–55,59,106]. Similarly

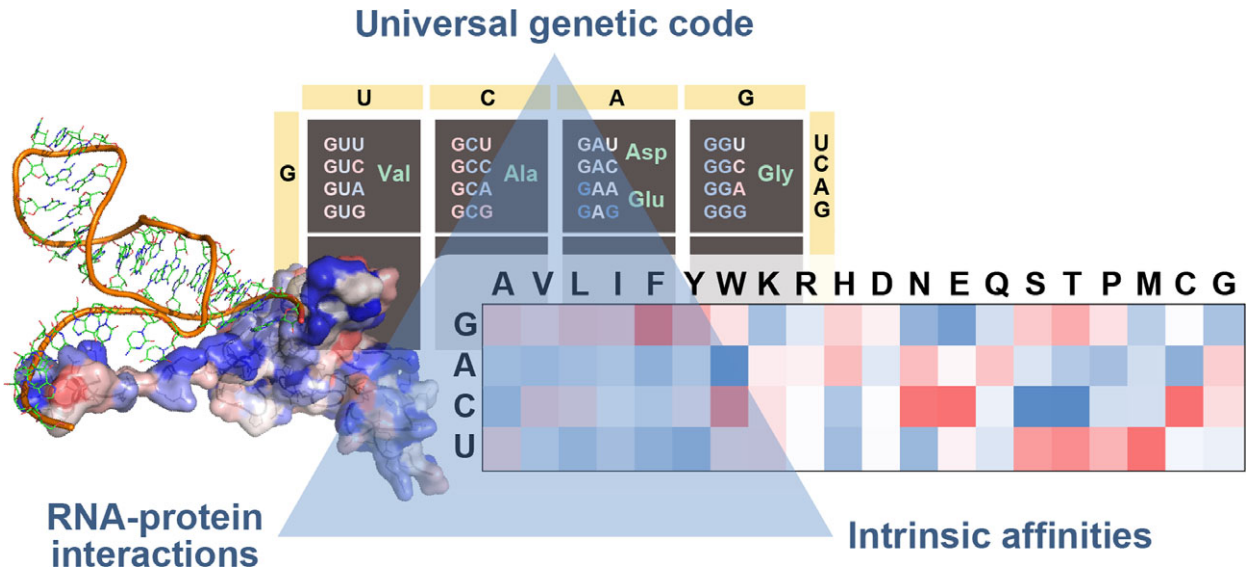


Fig. 5. The universal genetic code as the Rosetta stone for understanding RNA–protein interactions.

problematic, the sequence-based computational approaches for predicting RNA–protein interactions typically involve machine learning strategies and only indirectly rely on or help further understand the microscopic physicochemical principles behind such interactions [53–55]. While experimental approaches toward enumerating and characterizing RNA interactions with IDPs are making significant advances [2,5,8–10,29,63], they are limited by the fact that the microscopic, high-resolution features of such complexes are largely beyond the reach of modern structural biology methods. It is, therefore, imperative to explore the existence of general, novel physicochemical principles behind such interactions. Our 'complementarity hypothesis' may provide one such principle. In general, physicochemical complementarity is one of the most powerful paradigms used to explain biological function at the molecular level. Complementarity between DNA strands is the key element behind gene duplication, antibody–antigen complementarity guides immune response, while enzymes cannot be understood without the complementarity between active sites and reaction transition states. Therefore, it is potentially extremely far-reaching that RNA and protein sequences, including both cognate mRNA–protein pairs, as shown in our recent work [50,51,81,83,84,103–105] and those not connected by coding, as illustrated above, would exhibit such a robust compositional complementarity. This finding simply demands detailed exploration and full explanation and we see it as a major challenge for the future. Having said this, the range of validity and general applicability of the 'complementarity hypothesis' remain unclear: for example, the fact that the contour length of a typical mRNA coding region is approximately five times longer than the contour length of its cognate protein suggests that any physical realization of putative complementary binding must negotiate a significant spatial challenge. However, the hypothesis provides a well-defined, testable framework for relating the fundamental physicochemical properties of nucleobases and amino acids with the interactions of complete RNAs and proteins in an unstructured context. Moreover, the hypothesis promotes a way of looking at RNA–protein interactions that could be of practical importance regardless of whether the hypothesis itself turns out to be true or not. In particular, viewing RNA and protein sequences as physicochemical entities, with specific properties and local interaction propensities, presents a powerful paradigm for linking the speed and power of standard bioinformatic techniques with the atomistically realistic rigor. The consequences of this line of research potentially concern all canonical areas of bioinformatics and computational biology ranging from sequence comparison to multiple sequence alignment to building of phylogenetic trees.

There are a number of important open challenges regarding the RNA–protein interactions in an unstructured context. Here, we outline what we believe are the three most relevant and fundamental such challenges that have thus far, somewhat surprisingly, escaped wider attention. We firmly believe that these issues hold a key to a deeper understanding of RNA–protein interactions in an unstructured context, but could also prove to be essential on a much wider scale.

CHALLENGE 1: experimental determination of nucleotide–amino acid affinities

Nucleotides and amino acids are arguably the most fundamental building blocks in all living matter. Moreover, the specificity in interactions between nucleic acids and proteins is directly shaped by the specificity in interactions between these individual building blocks. Considering this, it is remarkable that the pioneering, yet incomplete experimental studies aimed at determining the affinities between individual nucleotides and amino acids, that is, their fragments [85–91], have not been extended since the 1960s and 1970s when they were first performed. There currently exists no self-consistent, experimentally determined table of absolute or, for that matter, relative binding ΔGs between the 5 standard RNA–DNA nucleotides (i.e., nucleosides/nucleobases) and the 20 standard amino acids (i.e., amino acid side chains). As discussed above, there is a growing body of computational/theoretical work in this regard [50,77–80], but the experimental contributions remain surprisingly incomplete. A part of the reason for this are the low solubilities of some of the groups and/or the low affinities involved, but we are firmly convinced that these difficulties can be successfully addressed with the aid of clever experimental strategies using, for example, affinity chromatography, NMR, microscale thermophoresis, or high-precision ITC. A natural extension of such studies would be a careful dissection of the binding contributions of different fragments of individual nucleotides and amino acids, such as the sugar or the phosphate groups, as well as an analysis of the impact of post-transcriptional nucleotide modifications and post-translational amino acids modifications on the individual binding affinities. For example, by using a purely computational analysis, we have recently shown that deamination of adenine, one of the most biologically important post-transcriptional nucleobase modifications, changes its interaction pattern with amino acids to that of guanine [107]. It would be critical that such and similar studies be extended from an experimental side as well. One criticism may be that nucleotide–amino acid interactions will be severely context dependent and susceptible to local environmental influences (pH, ionic strength, dielectric constant, temperature etc.) and, therefore, too difficult to accurately pin down. While we share this apprehension, we are convinced that a systematic analysis of such influences is in order. Moreover, our computational results strongly suggest that certain robust patterns of behavior remain even under changing environmental conditions [81,82]. Such robustness, after all, must be there simply for the biological systems to be able to function in a stable way.

CHALLENGE 2: analysis of sequence specificity in the formation of phase-separated granules

Phase-separated granules represent arguably the best, biologically relevant system for studying the interactions between RNA and proteins in an unstructured context [16–20]. On the one hand, single-stranded RNA and disordered proteins are ubiquitous constituents of such granules. On the other hand, the weak, multivalent, dynamic complexes formed by disordered RNA and proteins naturally lead to liquid–liquid phase separation that can also be well studied *in vitro*. In other words, phase separation and RNA–protein interactions in an unstructured context go hand in hand and should therefore be studied from a joint perspective: understanding of RNA–protein interactions in the unstructured state will provide critical information for the understanding of granule formation and *vice versa*. Importantly, a growing body of work has shown that the presence of RNA, and in particular single-stranded RNA, markedly reduces the critical concentration required for the formation of phase-separated protein granules [20,25,27]. However, a major open question concerns the sequence specificity behind such effects. In most cases, the authors have used random RNA sequences and/or select homooligonucleotides, but there have been no fully systematic attempts at understanding how different RNA, that is, protein sequences influence each other in this regard. We predict that the impact of different RNAs on the formation of protein granules will follow the rules given by the universal genetic code [50,102,103]. For example, we predict that GUA-rich sequences will have a stronger effect on the phase separation of protein sequences containing mostly polar residues, while ADE/PYR-rich sequences will have a stronger impact on the more hydrophobic protein sequences. In the context of the complementarity hypothesis discussed above, it would also be particularly interesting to systematically study the impact of cognate mRNA–protein pairs when it comes to granule formation. These studies should go hand in hand with the determination of the individual binding affinities between nucleotides and amino acids and their environmental dependence as outlined in the first challenge above. For example, phase-separated granules are known to be partially dehydrated, low-dielectric environments in which the relative nucleotide–amino acid affinities may follow different rules than in the more aqueous environments. Our umbrella-sampling

calculations have shown that in methanol, a solvent with a significantly lower dielectric constant as compared to water, the affinity of GUA for the negatively charged Asp and Glu side chains increases multiple-fold as compared to that in water [81,82]. Such and similar analyses will be necessary for the full understanding of microscopic driving forces behind the formation of phase-separated granules. Conversely, phase-separated granules will provide the proper biological context for studying the intricacies of the binding preferences between individual nucleotides and amino acids and RNA–IDP interactions in general.

CHALLENGE 3: development of computational tools for the prediction/analysis of RNA–protein binding in an unstructured context

The third open challenge concerns the development of computational frameworks for predicting the sites and the strength of interaction between RNA and unstructured protein regions that would be based on fundamental physicochemical principles. While the top-down machine learning-based approaches definitely have significant merit, the more physicochemically motivated, bottom-up strategies could provide a deeper mechanistic insight and have a greater predictive power. When it comes to the interaction between single-stranded RNAs and largely unstructured proteins, successful strategies could be based on the knowledge of the intrinsic interaction affinities between the individual building blocks of the two polymers, as discussed above. Presently, we do not have a clear prescription for how this could be implemented practically, but are motivated by a simple analogy. Namely, hybridization of two strands of DNA or folding of an RNA molecule can be well predicted from a simple thermodynamic quantification of Chargaff pairing rules and local stacking propensities [108,109]. Local affinities of nucleotides for each other, together with some understanding of the effect of local neighboring sequences and structures are often sufficient to predict, for example, the melting temperatures of duplexes or the folds of individual RNA molecules [108,109]. The point here is that the global structural and thermodynamic behavior of large nucleic acid molecules can be related to their linear sequence features and local interaction preferences. It is our hope that the rich world of RNA–protein interactions in an unstructured context could also in part be understood in such simple terms, which in turn would open up a myriad of different fundamental and applied possibilities. We are particularly excited by the possibility that the rules behind such interactions may actually be embedded in an ancient, already familiar codebook: the universal genetic code.

Acknowledgements

The funding by the European Research Council (Starting Independent Grant 279408 to BZ) and the Austrian Science Fund FWF (Grant 30680–B21 to BZ) is gratefully acknowledged.

References

1 Licatalosi DD and Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* **11**, 75–87.
2 Rinn JL and Ule J (2014) 'Oming in on RNA-protein interactions. *Genome Biol* **15**, 401.
3 Gerstberger S, Hafner M and Tuschl T (2014) A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829–845.
4 Jankowsky E and Harris ME (2015) Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol* **16**, 533–544.
5 Jarvelin AI, Noerenberg M, Davis I & Castello A (2016) The new (dis)order in RNA regulation. *Cell Commun Signal* **14**, 9.
6 Koster T, Marondedze C, Meyer K and Staiger D (2017) RNA-binding proteins revisited - The Emerging Arabidopsis mRNA Interactome. *Trends Plant Sci* **22**, 512–526.
7 Haberman N, Huppertz I, Attig J, Konig J, Wang Z, Hauer C, Hentze MW, Kulozik AE, Le Hir H, Curk T *et al.* (2017) Insights into the design and interpretation of iCLIP experiments. *Genome Biol* **18**, 7.
8 Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* **46**, 674–690.
9 Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406.
10 Beckmann BM, Horos R, Fischer B, Castello A, Eichelbaum K, Alleaume AM, Schwarzl T, Curk T, Foehr S, Huber W *et al.* (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **6**, 10127.
11 Castello A, Fischer B, Frese CK, Horos R, Alleaume AM, Foehr S, Curk T, Krijgsveld J and Hentze MW (2016) Comprehensive identification of RNA-binding domains in human cells. *Mol Cell* **63**, 696–710.

12 He CS, Sidoli S, Warneford-Thomson R, Tatomer DC, Wilusz JE, Garcia BA and Bonasio R (2016) High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells. *Mol Cell* **64**, 416–430.
13 Weichselbaum D, Zagrovic B and Polyansky AA (2017) Fuento: functional enrichment for bioinformatics. *Bioinformatics* **33**, 2604–2606.
14 Babu MM, van der Lee R, de Groot NS and Gsponer J (2011) Intrinsically disordered proteins: regulation and disease. *Curr Opin Struc Biol* **21**, 432–440.
15 Wright PE and Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Bio* **16**, 18–29.
16 Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoege C, Gharakhani J, Julicher F and Hyman AA (2009) Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732.
17 Hyman AA, Weber CA and Julicher F (2014) Liquid-liquid phase separation in biology. *Annu Rev Cell Dev Biol* **30**, 39–58.
18 Jain S and Parker R (2013) The discovery and analysis of P Bodies. *Adv Exp Med Biol* **768**, 23–43.
19 Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, Mittag T and Taylor JP (2015) Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133.
20 Alberti S (2017) The wisdom of crowds: regulating cell function through condensed states of living matter. *J Cell Sci* **130**, 2789–2796.
21 Peng ZL & Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* **43**, e121.
22 Calabretta S and Richard S (2015) Emerging roles of disordered sequences in RNA-binding proteins. *Trends Biochem Sci* **40**, 662–672.
23 Mitrea DM and Kriwacki RW (2016) Phase separation in biology; functional organization of a higher order. *Cell Commun Signal* **14**, 1.
24 Boeynaems S, Tompa P and Van Den Bosch L (2018) Phasing in on the cell cycle. *Cell Div* **13**, 1.
25 Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD *et al.* (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell* **57**, 936–947.
26 Lin Y, Currie SL and Rosen MK (2017) Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J Biol Chem* **292**, 19110–19120.
27 Nott TJ, Craggs TD and Baldwin AJ (2016) Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nat Chem* **8**, 569–575.
28 van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589–6631.
29 Varadi M, Zsolyomi F, Guharoy M and Tompa P (2015) Functional advantages of conserved intrinsic disorder in RNA-binding proteins. *PLoS ONE* **10**, e0139731.
30 Wang XY, Schwartz JC and Cech TR (2015) Nucleic acid-binding specificity of human FUS protein. *Nucleic Acids Res* **43**, 7535–7543.
31 Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynov S, Mahamid J, Saha S, Franzmann TM *et al.* (2015) A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077.
32 Ozdilek BA, Thompson VF, Ahmed NS, White CI, Batey RT and Schwartz JC (2017) Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Res* **45**, 7984–7996.
33 Stolow DT and Haynes SR (1995) Cabeza, a Drosophila gene encoding a novel RNA-binding protein, shares homology with EWS and TLS, 2 genes involved in human sarcoma formation. *Nucleic Acids Res* **23**, 835–843.
34 Bertolotti A, Lutz Y, Heard DJ, Chambon P and Tora L (1996) hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *EMBO J* **15**, 5022–5031.
35 Mastrocola AS, Kim SH, Trinh AT, Rodenkirch LA and Tibbetts RS (2013) The RNA-binding protein Fused in Sarcoma (FUS) functions downstream of poly(ADP-ribose) polymerase (PARP) in response to DNA damage. *J Biol Chem* **288**, 24731–24741.
36 Burke KA, Janke AM, Rhine CL and Fawzi NL (2015) Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Mol Cell* **60**, 231–241.
37 Alberti S and Hyman AA (2016) Are aberrant phase transitions a driver of cellular aging? *BioEssays* **38**, 959–968.
38 Deng H, Gao K and Jankovic J (2014) The role of FUS gene variants in neurodegenerative diseases. *Nat Rev Neurol* **10**, 337–348.
39 Blanc RS and Richard S (2017) Arginine methylation: the coming of age. *Mol Cell* **65**, 8–24.
40 Dormann D, Madl T, Valori CF, Bentmann E, Tahirovic S, Abou-Ajram C, Kremmer E, Ansorge O, Mackenzie IR, Neumann M *et al.* (2012) Arginine methylation next to the PY-NLS modulates

Transportin binding and nuclear import of FUS. *EMBO J* **31**, 4258–4275.
41 Blackwell E, Zhang X and Ceman S (2010) Arginines of the RGG box regulate FMRP association with polyribosomes and mRNA. *Hum Mol Genet* **19**, 1314–1323.
42 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.
43 Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, Boone C, Bader GD, Myers CL and Kim PM (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* **12**, R14.
44 Das RK and Pappu RV (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci USA* **110**, 13392–13397.
45 Das RK, Huang Y, Phillips AH, Kriwacki RW and Pappu RV (2016) Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc Natl Acad Sci USA* **113**, 5616–5621.
46 Quiroz FG and Chilkoti A (2015) Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat Mater* **14**, 1164–1171.
47 Brangwynne Clifford P, Tompa P and Pappu Rohit V (2015) Polymer physics of intracellular phase transitions. *Nat Phys* **11**, 899.
48 Basu S and Bahadur RP (2016) A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell Mol Life Sci* **73**, 4075–4084.
49 Atchley WR, Zhao J,Fernandes AD and Druke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci USA* **102**, 6395–6400.
50 Polyansky AA and Zagrovic B (2013) Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res* **41**, 8434–8443.
51 Beier A, Zagrovic B and Polyansky AA (2014) On the contribution of protein spatial organization to the physicochemical interconnection between proteins and their cognate mRNAs. *Life (Basel)* **4**, 788–799.
52 Puton T, Kozlowski L, Tuszynska I, Rother K and Bujnicki JM (2012) Computational methods for prediction of protein-RNA interactions. *J Struct Biol* **179**, 261–268.
53 Livi CM and Blanzieri E (2014) Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. *BMC Bioinformatics* **15**, 123.
54 Si J, Cui J, Cheng J and Wu R (2015) Computational prediction of RNA-binding proteins and binding sites. *Int J Mol Sci* **16**, 26303–26317.
55 Yan J, Friedrich S and Kurgan L (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* **17**, 88–105.
56 Pancaldi V and Bahler J (2011) In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res* **39**, 5826–5836.
57 Muppirala UK, Honavar VG and Dobbs D (2011) Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* **12**, 489.
58 Wang Y, Chen X, Liu ZP, Huang Q, Wang Y, Xu D, Zhang XS, Chen R and Chen L (2013) De novo prediction of RNA-protein interactions from sequence information. *Mol BioSyst* **9**, 133–142.
59 Bellucci M, Agostini F, Masin M and Tartaglia GG (2011) Predicting protein associations with long noncoding RNAs. *Nat Methods* **8**, 444–445.
60 Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D and Tartaglia GG (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* **29**, 2928–2930.
61 Alipanahi B, Delong A, Weirauch MT and Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831–838.
62 Qu YH, Yu H, Gong XJ, Xu JH and Lee HS (2017) On the prediction of DNA-binding proteins only from primary sequences: a deep learning approach. *PLoS ONE* **12**, e0188129.
63 Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177.
64 Orenstein Y, Wang Y and Berger B (2016) RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. *Bioinformatics* **32**, i351–i359.
65 Kazan H, Ray D, Chan ET, Hughes TR and Morris Q (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol* **6**, e1000832.
66 Auweter SD, Oberstrass FC and Allain FHT (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* **34**, 4943–4959.
67 Luscombe NM, Laskowski RA and Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* **29**, 2860–2874.
68 Allers J and Shamoo Y (2001) Structure-based analysis of Protein-RNA interactions using the program ENTANGLE. *J Mol Biol* **311**, 75–86.
69 Yoshida T, Nishimura T, Aida M, Pichierri F, Gromiha MM and Sarai A (2001) Evaluation of free

energy landscape for base-amino acid interactions using ab initio force field and extensive sampling. *Biopolymers* **61**, 84–95.

70 Biot C, Buisine E and Rooman M (2003) Free-energy calculations of protein-ligand cation-pi and amino-pi interactions: From vacuum to proteinlike environments. *J Am Chem Soc* **125**, 13988–13994.

71 Cheng AC, Chen WW, Fuhrmann CN and Frankel AD (2003) Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J Mol Biol* **327**, 781–796.

72 Hoffman MM, Khrapov MA, Cox JC, Yao J, Tong L and Ellington AD (2004) AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res* **32**, D174–D181.

73 Ebrahimi A, Habibi-Khorassani M, Gholipour AR and Masoodi HR (2009) Interaction between uracil nucleobase and phenylalanine amino acid: the role of sodium cation in stacking. *Theor Chem Acc* **124**, 115–122.

74 Rutledge LR, Navarro-Whyte L, Peterson TL and Wetmore SD (2011) Effects of extending the computational model on DNA-protein T-shaped interactions: the case of adenine-histidine dimers. *J Phys Chem A* **115**, 12646–12658.

75 Kondo J and Westhof E (2011) Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res* **39**, 8628–8637.

76 Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D and Altman RB (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15**, 189–199.

77 Perez-Cano L, Solernou A, Pons C and Fernandez-Recio J (2010) Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pac Symp Biocomp* **15**, 269–280.

78 Tuszynska I & Bujnicki JM (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC Bioinform* **12**, 348.

79 Jakubec D, Hostas J, Laskowski RA, Hobza P and Vondrasek J (2015) Large-scale quantitative assessment of binding preferences in protein-nucleic acid complexes. *J Chem Theory Comput* **11**, 1939–1948.

80 Andrews CT, Campbell BA & Elcock AH (2017) Direct comparison of amino acid and salt interactions with double-stranded and single-stranded DNA from explicit-solvent molecular dynamics simulations. *J Chem Theory Comput* **13**, 1794–1811.

81 de Ruiter A and Zagrovic B (2015) Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. *Nucleic Acids Res* **43**, 708–718.

82 de Ruiter A, Polyansky AA & Zagrovic B (2017) Dependence of binding free energies between RNA nucleobases and protein side chains on local dielectric properties. *J Chem Theory Comput* **13**, 4504–4513.

83 Hajnic M, Osorio JI and Zagrovic B (2014) Computational analysis of amino acids and their sidechain analogs in crowded solutions of RNA nucleobases with implications for the mRNA-protein complementarity hypothesis. *Nucleic Acids Res* **42**, 12984–12994.

84 Hajnic M, Osorio JI and Zagrovic B (2015) Interaction preferences between nucleobase mimetics and amino acids in aqueous solutions. *Phys Chem Chem Phys* **17**, 21414–21422.

85 Akinrimisi EO & Tso PO (1964) Interactions of purine with proteins and amino acids. *Biochemistry* **3**, 619–626.

86 Thomas PD and Podder SK (1978) Specificity in protein-nucleic acid interaction – solubility study on amino acid nucleoside interaction. *FEBS Lett* **96**, 90–94.

87 Woese CR, Dugre DH, Saxinger WC and Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci USA* **55**, 966–974.

88 Woese CR (1973) Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459.

89 Lacey JC & Pruitt KM (1969) Origin of genetic code. *Nature* **223**, 799.

90 Rifkind JM and Eichhorn GL (1970) Specificity for interaction of nucleotides with basic polypeptides. *Biochemistry* **9**, 1753.

91 Wagner KG and Arfmann HA (1974) Properties of basic-amino-acid residues – nucleotide-poly(amino acid) interaction. *Eur J Biochem* **46**, 27–34.

92 Noller HF (2012) Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol*. **4**, 1-U20.

93 Woese CR (2001) Translation: In retrospect and prospect. *RNA* **7**, 1055–1067.

94 Koonin EV and Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111.

95 Crick FHC (1968) The origin of the genetic code. *J Mol Biol* **38**, 367–379.

96 Freeland SJ, Knight RD, Landweber LF and Hurst LD (2000) Early fixation of an optimal genetic code. *Mol Biol Evol* **17**, 511–518.

97 Wong JT (2005) Coevolution theory of the genetic code at age thirty. *BioEssays* **27**, 416–425.

98 Yarus M, Widmann JJ and Knight R (2009) RNA-amino acid binding: a stereochemical era for the genetic code. *J Mol Evol* **69**, 406–429.

99 Johnson DB and Wang L (2010) Imprints of the genetic code in the ribosome. *Proc Natl Acad Sci U S A*. **107**, 8298–8303.

100 Mathew DC and Luthey-Schulten Z (2008) On the physical basis of the amino acid polar requirement. *J Mol Evol* **66**, 519–528.
101 Di Giulio M (2005) The origin of the genetic code: theories and their relationships, a review. *Biosystems*. **80**, 175–184.
102 Hlevnjak M, Polyansky AA and Zagrovic B (2012) Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res* **40**, 8874–8882.
103 Polyansky AA, Hlevnjak M and Zagrovic B (2013) Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol* **10**, 1248–1254.
104 Hlevnjak M and Zagrovic B (2015) Malleable nature of mRNA-protein compositional complementarity and its functional significance. *Nucleic Acids Res* **43**, 3012–3021.
105 Bartonek L and Zagrovic B (2017) mRNA/protein sequence complementarity and its determinants: The impact of affinity scales. *PLoS Comput Biol* **13**, e1005648.
106 Polyansky AA, Hlevnjak M & Zagrovic B (2013) Analogue encoding of physicochemical properties of proteins in their cognate messenger RNAs. *Nat Commun* **4**, 2784.
107 Hajnic M, Ruiter A, Polyansky AA and Zagrovic B (2016) Inosine nucleobase acts as guanine in interactions with protein side chains. *J Am Chem Soc* **138**, 5519–5522.
108 Mathews DH, Sabina J, Zuker M and Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–940.
109 Lorenz R, Bernhart SH, Siederdissen CHZ, Tafer H, Flamm C, Stadler PF & Hofacker IL (2011) ViennaRNA package 2.0. *Algorithm Mol Biol* **6**, 26.
110 Narayanan BC, Westbrook J, Ghosh S, Petrov AI, Sweeney B, Zirbel CL, Leontis NB and Berman HM (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res* **42**, D114–D122.

4

VOLPES – INTERACTIVE ANALYSIS OF PHYSICOCHEMICAL SEQUENCE PROPERTIES

Lukas Bartonek and Bojan Zagrovic. "VOLPES: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences." In: *Nucleic acids research* 47.W1 (2019), W632–W635 [31]

The development of VOLPES followed a realization that the physicochemical view of biological sequence properties is severely underrepresented in the current literature and, more importantly, among the available webtools. By allowing users to quickly, easily and without any need of programming knowledge investigate and compare a multitude of biomolecular sequence properties, VOLPES targets a wide audience and promotes the view of sequences as actual physical objects with defined, biologically meaningful properties. The design as an interactive webtool and not a standalone program to be installed on users' machines removes significant barriers that could impede the usability of such a tool. The decision, on the other hand, to implement most of the logic in a front-end facing JavaScript library enabled a fully interactive experience for the end-user. Such a direct way of interacting with the underlying data allows one to get an in-depth understanding of both the data and the analytical model applied on it.

PERSONAL CONTRIBUTION

The author of this thesis was involved in the conceptualization, data curation, analysis, development of methodology, software development, validation, visualization as well as writing of the presented article.

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz407

VOLPES: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences

Lukas Bartonek and Bojan Zagrovic*

Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, 1030 Vienna, Austria

Received March 07, 2019; Editorial Decision May 02, 2019; Accepted May 02, 2019

ABSTRACT

The structure, dynamics and, ultimately, biological function of proteins and nucleic acids are determined by the physicochemical properties of their primary sequences. Such properties are frequently captured via one-dimensional profile plots depicting a given physicochemical variable as a function of sequence position. Hydrophobicity, charge or structural disorder in proteins or nucleobase-density in nucleic acids are routinely visualized in this manner to analyze sequences at a glance. Such visualizations, however, are typically created case-by-case in a purely static manner, employ fixed visualization parameters only and do not enable a quantitative comparison between different sequences. Here, we present VOLPES (volpes.univie.ac.at), a user-friendly web server and the corresponding JavaScript library that enable a fully interactive, multifunctional visualization, analysis and comparison of the physicochemical properties of protein and nucleic-acid sequences, allowing unprecedented insight into biological sequence data and creating a starting point for further in-depth exploration.

INTRODUCTION

A foundational principle of bioinformatics is that biological function is encoded in and can ultimately be related to the primary sequences of nucleic acids and proteins. While sequences can be analyzed in different abstract ways, they first and foremost refer to the composition of real, physical objects i.e. linear-chain biopolymers. Importantly, the physicochemical properties of biopolymers are frequently visualized via one-dimensional profile plots where each sequence position is associated with a quantitative measure of the property of interest. For example, hydrophobicity profiles are often employed to learn about the structural features and domain boundaries in proteins. In particular, such profiles are routinely used to ascertain the location of transmembrane regions in membrane proteins (1–4). Another common application of profile plots is to visualize the distribution of charge in protein sequences (5–7). Such analyses, for example, have been instrumental in understanding the fundamental principles behind the formation of phase-separated granules in the cell (8,9). Finally, nucleobase-density profiles in DNA and RNA have been used to study different biological phenomena ranging from nucleosome positioning (10–12) to RNA–protein interactions (13–15) to protein localization (16,17).

While many programs or servers output results in the form of profile plots, these tools are usually highly specialized and/or feature static visualization only (2,18,19). Despite the fact that the usage of such tools is mostly exploratory in nature, users are typically unable to change visualization parameters on-the-fly or opt for a different property without extensive recalculation or even having to switch to a different tool altogether. Furthermore, most tools do not enable a comparison between multiple profiles, although such a feature may be critical in a wide range of applications including mutational analyses and domain comparisons or simply to highlight the relationship between different properties of interest. On the other hand, tools featuring more advanced features are not available as web servers but rather as downloadable standalone packages (3,20).

In response to these challenges, we present here VOLPES, a multifunctional web server and a JavaScript library for highly customizable, interactive visualization, analysis and comparison of the physicochemical properties of protein and RNA sequences. As its integral part, the VOLPES server provides direct access to hundreds of published amino-acid and nucleobase physicochemical property scales, many of which were sourced from the AAindex (21,22), facilitating a rapid and multifaceted exploration of the sequences of interest. By treating biomolecular sequences as physicochemical objects, which they invariably are, VOLPES enables detection of biologically meaningful patterns and similarities that may be inaccessible to stan-

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

dard methods of primary sequence analysis and comparison.

MATERIALS AND METHODS

Input

The VOLPES server requires only protein or RNA primary sequence data as the initial user input. All further details required for the analysis are already provided by the server itself. The most basic way of inputting sequence information is either in the form of a string of single-letter-encoded monomers (amino acids or nucleobases) without further specifications or in the FASTA format, which is automatically detected and handled appropriately by the server, including support for gap characters. However, the most convenient way to import sequence data is via an integrated connection to public databases whereby the relevant sequence data and sequence names, if available, are fetched automatically. In the current release, UniProt (23) and ENA (European Nucleotide Archive) (24) identifiers are supported.

Savefiles. Input data can also be provided by uploading a previous savefile generated by the server itself or a compatible external program (please see below for details).

Output

A multi-tiered approach allows users to generate output files with different levels of complexity.

Image output. Conveniently available from the export menu, the active visualization state can be exported as a rasterized PNG image. The saved state is exactly the same as that defined by the user while exploring the data and includes all set visualization parameters as well as the zoom and the shift state of the plotting window. Vector data are supported in the format of SVG files, which can be used as a basis for making publication quality figures in post-processing. This allows figures to be easily combined to make more advanced illustrations, adjust font properties and modify colors afterward. Furthermore, this format allows figures to be rendered to rasterized formats at arbitrary resolution.

Data output. VOLPES can also be employed to compute physicochemical profile data for uses other than visualization. To accommodate this, output of raw data in the form of CSV files is supported. Users can choose any combination of staged sequence data, including a subset of select profiles, to be downloaded for further use.

Savefiles. Any given visualization state can also be exported as a *.viz* savefile to be shared with collaborators or saved for later analysis. These files specify all details about the current visualization visible to the user and, therefore, allow anyone with access to this file to continue working at the same point at a later time. This file standard is open, easy to read/write and allows developers of other tools to access the stored information with ease.

Interactivity

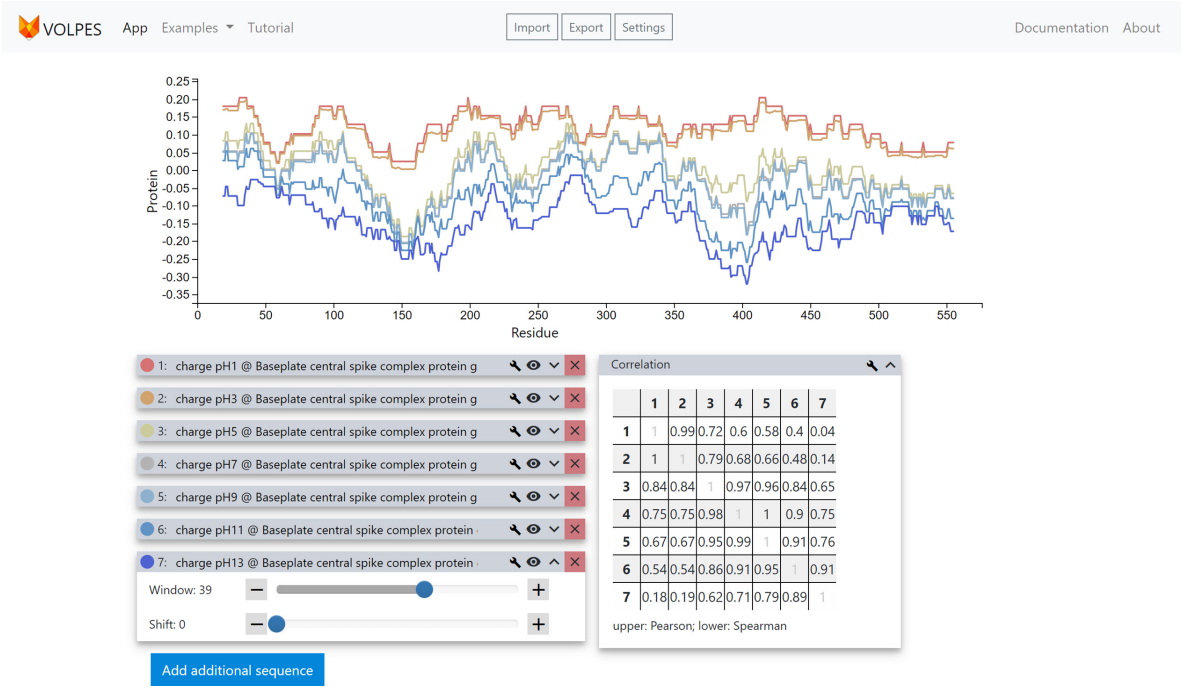
The principal advantage of VOLPES as compared to other tools for visualizing sequence profiles is its inherent interactivity and ease of accessing the underlying data. This is achieved by performing most of the visualization-related calculations client-side in the users' browser. In other words, VOLPES enables real-time interactive visualizations by removing any delays introduced by internet connectivity, while all of its modifiable parameters are accessible via an user-friendly interface, as illustrated in Figure 1.

Smoothing. Virtually all visualizations of physicochemical properties in the form of profiles employ smoothing of the raw data to make key features stand out, while reducing the inherent noise in the data. The size of the window used for such smoothing depends strongly on the feature of interest. The VOLPES interface allows users to change the smoothing window size by using a slider and observing the impact in real time. This is essential for users to develop an intimate feel for the underlying data.

Shifting and zooming. Unlike other related tools, VOLPES allows multiple sequence profiles to be visualized at the same time as well as shifted against each other along the residue axis. This is implemented on an individual sequence basis such that each sequence can be shifted against all others independently. To keep the drawing canvas to a reasonable size, the implementation ensures that shifts are limited in a way that all possible relative superpositions are possible, but not more. Importantly, users can visualize each profile in relative units of the standard deviation of the corresponding data series, thus enabling superposition of profiles that capture different physicochemical variables. At last, users can zoom in or zoom out of any region in a given profile at will.

Visual appearance. Multiple visual properties are modifiable in real-time. While the color of each profile can be changed directly, other properties have to be accessed in the menu for individual settings. Line-widths and line styles can be changed freely from within the menu. The number of line-styles has been limited to four on purpose to allow users a wide enough selection of parameters without overwhelming them.

Property scales. At the heart of VOLPES is a database of over 600 different physicochemical property scales for the 20 standard amino acids and 4 standard RNA nucleotides or components thereof (amino-acid side chains, nucleosides, nucleobases etc.). The scales were sourced from the AAindex (21,22) and complemented by additional scales obtained either from the recent literature (cited within the tool) or derived to be used in VOLPES (e.g. amino-acid charge at different pH). To catalogue this vast dataset efficiently, we have categorized the scales by using the approach originally developed by Tomii *et al.*, (25). In the case of amino-acid scales, this has resulted in the following categories: α-propensity, β-propensity, charge, composition, hydrophobicity, general physico-chemical properties, RNA-affinity and other. RNA scales were grouped by expert



	1	2	3	4	5	6	7
1	1	0.99	0.72	0.6	0.58	0.4	0.04
2	1	1	0.79	0.68	0.66	0.48	0.14
3	0.84	0.84	1	0.97	0.96	0.84	0.65
4	0.75	0.75	0.98	1	1	0.9	0.75
5	0.67	0.67	0.95	0.99	1	0.91	0.76
6	0.54	0.54	0.86	0.91	0.95	1	0.91
7	0.18	0.19	0.62	0.71	0.79	0.89	1

Figure 1. Example of a typical analysis performed via the VOLPES server: charge profiles calculated at different values of pH for a specified protein. A quantitative analysis of the similarity between profiles is given in the correlation matrix.

knowledge into three groups corresponding to composition, energy-related properties or general physico-chemical properties. Upon selection of a scale of interest in VOLPES, a detailed view is shown providing a short name for the scale, a link to the AAindex, if present in the database, the explicit values of the scale, a longer description of the scale, a literature reference and a link to the Pubmed reference database, if available. This allows users to scroll through a large number of scales and identify those that are relevant for their purposes without a need for extensive prior knowledge. Once a scale is selected, it is immediately applied to the active sequence and the plot is updated accordingly.

Comparison. The level of similarity between different physicochemical profiles can be quantified in several ways. A pairwise matrix of correlation coefficients between all uploaded sequences is displayed automatically once two or more profiles are present on the canvas. As default parameters, the Pearson correlation coefficient R is displayed in the upper triangle of the matrix and the Spearman rank correlation coefficient in the lower triangle. Both can be exchanged for each other or the coefficient of determination R^2 or the root-mean-square deviation (RMSD) by selecting the respective measures in the associated settings menu. The values in the matrix are dynamically updated whenever a change that affects the correlation level is implemented.

Library

In addition to the implementation in the form of a freely accessible web server, the JavaScript library, used to power the server frontend, is freely available as a git repository (found at github.com/BarLuk/volpini). A detailed documentation can be found in the repository as well as on the server itself. The library is built upon the popular d3.js library (26), enabling comfortable access and modifications to SVG nodes as DOM elements with a wide range of custom functionality built in to accommodate the analysis of biological sequences as a special application.

Easy to use. In its most basic functionality, the library requires a single function call together with an already provided dataset to produce an interactive figure. To allow for additional interactivity aimed at sampling the parameter space, the drawing function just needs to be called whenever a parameter is updated to redraw the figure.

Wide range of applications. The library can be used in any environment in which JavaScript can be executed and HTML can be rendered. Due to the widespread use of modern web browsers, such an environment is present on virtually any computer. Besides the obvious use as the frontend of a web server, the library has already been employed to produce interactive conference posters as well as presentation slides for scientific talks. Examples can be found in the online documentation. At last, the library can also in principle be used for publishing interactive figures in the scientific literature, enabling the readers to explore the presented data themselves. While the authors would still be able to present their data using the visualization parameters they prefer, the readers could easily vary the parameters themselves for additional insight.

Data security and access

Except for IP access logs, no other user data are stored server-side. This explicitly also refers to the specific sequences or combinations of sequences and scales that have been processed by the tool. Access to the server is free without a need to register. The server can be accessed at volpes.univie.ac.at.

CONCLUSION

VOLPES is a freely available online tool for visualizing the physicochemical properties of biopolymers that only requires an up-to-date web browser for access. Both protein and RNA sequences are supported, with database connectivity enabling user-friendly data access. Over 600 different physicochemical property scales of amino acids and nucleobases are available in VOLPES, with detailed descriptions for each scale, allowing the user to quickly browse through. Enabling users to interact with visualization parameters in real-time makes it possible to develop an intimate understanding of the underlying dataset. The speed of interaction with the dataset, on the other hand, opens up the possibility of a data-based, exploration-driven approach to research that puts the raw data first, allowing one to develop hypotheses based on observations made on this level.

VOLPES was launched in August 2018 and it is expected to have high visibility among bioinformaticians and experimentalists alike. Since the original launch, several thousands of pageviews have been recorded, but due to the nature of the service, the exact number of analyses performed cannot be reported. The service has been tested thoroughly by internal and external testers and remains under active support.

ACKNOWLEDGEMENTS

The authors thank the members of the Laboratory of Molecular Biophysics at Max Perutz Labs of the University of Vienna for useful and critical input when developing the server.

FUNDING

European Research Council Starting Independent Grant [279408 to B.Z.]; Austrian Science Fund FWF Standalone Grants [P 30550 and P 30680-B21 to B.Z.]. Funding for open access charge: Austrian Science Fund FWF Standalone Grant [P 30680-B21 to B.Z.].
Conflict of interest statement. None declared.

REFERENCES

1. Cuthbertson,J.M., Doyle,D.A. and Sansom,M.S. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.*, **18**, 295–308.
2. Snider,C., Jayasinghe,S., Hristova,K. and White,S.H. (2009) MPEx: a tool for exploring membrane proteins. *Protein Sci.*, **18**, 2624–2628.
3. Deber,C.M., Wang,C., Liu,L., Prior,A.S., Agrawal,S., Muskat,B.L. and Cuticchia,A.J. (2001) TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.*, **10**, 212–219.
4. Zhao,G. and London,E. (2006) An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Protein Sci.*, **15**, 1987–2001.
5. Das,R.K. and Pappu,R.V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 13392–13397.
6. Das,R.K., Huang,Y., Phillips,A.H., Kriwacki,R.W. and Pappu,R.V. (2016) Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5616–5621.
7. Zagrovic,B., Bartonek,L. and Polyansky,A. (2018) RNA-protein interactions in an unstructured context. *FEBS Lett.*, **592**, 2901–2916.
8. Wang,J., Choi,J.-M., Holehouse,A.S., Lee,H.O., Zhang,X., Jahnel,M., Maharana,S., Lemaitre,R., Pozniakovsky,A., Drechsel,D. *et al.* (2018) A molecular grammar governing the driving forces for phase separation of Prion-like RNA binding proteins. *Cell*, **174**, 688–699.
9. Nott,T.J., Petsalaki,E., Farber,P., Jervis,D., Fussner,E., Plochowietz,A., Craggs,T.D., Bazett-Jones,D.P., Pawson,T., Forman-Kay,J.D. *et al.* (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell*, **57**, 936–947.
10. Hebert,C. and Roest Crollius,H. (2010) Nucleosome rotational setting is associated with transcriptional regulation in promoters of tissue-specific human genes. *Genome Biol.*, **11**, 1–13.
11. Wang,J.-P.Z. and Widom,J. (2005) Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res.*, **33**, 6743–6755.
12. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I.K., Wang,J.-P.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
13. Hlevnjak,M., Polyansky,A. and Zagrovic,B. (2012) Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res.*, **40**, 8874–8882.
14. Polyansky,A. and Zagrovic,B. (2013) Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res.*, **41**, 8434–8443.
15. Bartonek,L. and Zagrovic,B. (2017) mRNA/protein sequence complementarity and its determinants: the impact of affinity scales. *PLoS Comput. Biol.*, **13**, 1–16.
16. Lesnik,T. and Reiss,C. (1998) Detection of transmembrane helical segments at the nucleotide level in eukarayotic membrane protein genes. *IUBMB Life*, **44**, 471–479.
17. 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. U.S.A.*, **106**, 6662–6666.
18. Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S., Wilkins,M.R., Appel,R.D. and Bairoch,A. (2005) Protein identification and analysis tools on the ExPASy server. In: Walker,JM (ed). *The Proteomics Protocols Handbook*. Humana Press, Totowa, pp. 571–607.
19. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
20. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
21. Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374–374.
22. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2007) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.
23. UniProt Consortium,T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699–2699.
24. Harrison,P.W., Alako,B., Amid,C., Cerdeño-Tárraga,A., Cleland,I., Holt,S., Hussein,A., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2019) The European Nucleotide Archive in 2018. *Nucleic Acids Res.*, **47**, D84–D88.
25. Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.
26. Bostock,M., Ogievetsky,V. and Heer,J. (2011) D^3 Data-Driven Documents. *IEEE Trans Vis. Comput. Graph.*, **17**, 2301–2309.

5

INFLUENCE OF FRAMESHIFTS ON MUTANTS AS SEEN BY AFFINITY SCALES

Lukas Bartonek, Daniel Braun, and Bojan Zagrovic. "Invariants of Frameshifted Variants." In: *bioRxiv* (2019), p. 684076 [76]

While exploring the physicochemical perspective on biomolecular sequence analysis, we made a potentially far-reaching observation, namely that protein sequence profiles corresponding to certain specific physicochemical properties appear to be significantly invariant to frameshift mutations in their cognate mRNA coding sequences. This observation was investigated in depth, including a detailed analysis of statistical significance and limitations of the observed effect.

What we observed was that this invariance is primarily applicable to protein hydrophobicity profiles, with several other additional specific properties following suit. In particular, we could show that both the matching between mRNA nucleobase-density profiles and their cognate protein's nucleobase-affinity profiles as well protein disorder profiles are also largely unaffected by mRNA frameshift mutations.

Finally, we proposed a hypothesis that frameshift mutations may be an important evolutionary strategy for generating a high diversity of protein sequences, while retaining some of their basic properties, such as hydrophobicity. We suggested that this effect would most likely be relevant in protein classes that strongly depend on the hydrophobicity pattern of their underlying sequence, as in the case of membrane proteins.

At the time of writing the manuscript is under peer-review, the submitted version is attached. The initial manuscript [76] has been submitted on 29.6.2019 for review, the attached revised version has been resubmitted on 31.10.2019.

PERSONAL CONTRIBUTION

The author of this thesis was involved in the conceptualization, data curation, analysis, development of methodology, software development, validation, visualization as well as writing of the presented article.

Frameshifting preserves key physicochemical properties of proteins

Lukas Bartonek¶, Daniel Braun¶ & Bojan Zagrovic*

Department of Structural and Computational Biology, Max Perutz Labs, University of Vienna, Campus Vienna Biocenter 5, Vienna A-1030, Austria

October 2019

¶ These authors contributed equally to this manuscript.

***Correspondence:**

Email: bojan.zagrovic@univie.ac.at

Tel: +43 2 4277 52271

Fax: +43 1 4277 9522

Abstract

Frameshifts in protein coding sequences are widely perceived as resulting in either non-functional or even deleterious protein products. Indeed, frameshifts typically lead to markedly altered protein sequences and premature stop codons. By analyzing complete proteomes from all three domains of life, we demonstrate that, in contrast, several key physicochemical properties of protein sequences exhibit significant robustness against +1 and -1 frameshifts. In particular, we show that hydrophobicity profiles of many protein sequences remain largely invariant upon frameshifting. For example, over 2900 human proteins exhibit a Pearson correlation coefficient between the hydrophobicity profiles of the original and the +1-frameshifted variants greater than 0.7, despite a median sequence identity between the two of only 6.5% in this group. We observe a similar effect for protein sequence profiles of affinity for certain nucleobases as well as protein sequence profiles of intrinsic disorder. Finally, we show that frameshift stability is directly embedded in the structure of the universal genetic code and may have contributed to shaping it. Our results suggest that frameshifting may be a powerful evolutionary mechanism for creating new proteins with vastly different sequences, yet similar physicochemical properties to the proteins they originate from.

Significance Statement

Genetic information stored in DNA is transcribed to messenger RNAs and then read in the process of translation to produce proteins. A frameshift in the reading frame at any stage of the process typically results in a significantly different protein sequence being produced. Here, we show that, nevertheless, several essential properties of many protein sequences, such as their hydrophobicity profiles, remain largely unchanged upon frameshifts. This finding suggests that frameshifting could be an effective evolutionary strategy for generating novel protein sequences, which retain the functionally relevant physicochemical properties of the sequences they derive from.

Introduction

Frameshifts in the mRNA coding sequences of proteins are typically considered to be unproductive events, which, if unchecked, could result in non-functional and sometimes even deleterious protein products (1–4). This notion is mainly based on the dramatic difference between the sequences of wildtype proteins and their frameshifted counterparts. For example, the average sequence identity between wildtype human proteins and proteins obtained by +1 frameshifting their mRNAs is only 6.2% (Figure S1). Following results like this, it has been widely assumed that frameshifting produces polypeptides that are essentially unrelated to wildtype proteins in terms of their physicochemical properties and suitability to carry out biological function (5–7). Equally importantly, frameshifted mRNAs frequently contain premature stop codons and in eukaryotes are rapidly degraded by the nonsense-mediated decay machinery (8). It has even been suggested that the genetic code has been optimized such that the hidden stop codons would prevent extensive out-of-frame gene

reading (6). This is supported by the fact that frameshifted mRNAs feature stop codons at a frequency that is typically higher than expected at random (9, 10). In a more practical context, an introduction of frameshifts coupled to nonsense-mediated decay has become a standard strategy for disabling gene expression via CRISPR/CAS9 (11).

On the other hand, it is also known that changes in the reading frame do not necessarily lead to unwanted consequences. For example, there exist several known genes that include frameshifts as compared to related genes in other species (12). Moreover, it has been suggested that frameshifts may result in proteins with completely novel functions (13, 14). Finally, instances of overlapping genes are well described not only in viruses, but even in human (15, 16). In a related context, it has long been known that codons with similar composition encode amino acids with related physicochemical properties (17–19). Although the impact of this feature of the genetic code has been well appreciated in the case of point mutations, its influence in the case of frameshifts has only recently been addressed (20–23). In particular, Wang et al. showed that wildtype sequences and their frameshifted variants exhibit higher than expected similarity as captured by several classic substitution matrices and sequence alignment with gaps (20, 23). On the other hand, Geyer and Mamlouk analyzed a particular physicochemical property of amino acids i.e. a measure of their partitioning in pyridine/water mixtures, and demonstrated a weak, but significant correlation between frameshifted counterparts at the genetic code level (21). Finally, Wnetrzak and coworkers showed that the code may have been optimized in part to lessen the impact of frameshift mutations (22). Importantly, however, all these studies analyzed a limited set of amino-acid properties or substitution frequencies only and focused primarily on the genetic code.

The structure, dynamics and, ultimately, biological function of proteins are determined by the physicochemical properties of their sequences. For example, sequence hydrophobicity profiles of membrane proteins allow one to accurately identify transmembrane segments (24). Even in the case of cytosolic proteins, the hydrophobic/polar alterations in the sequences are thought to be important determinants of their tertiary folds (25). Moreover, nucleobase-affinity sequence profiles of proteins have been suggested to provide relevant information about their propensity to interact with RNA (26–29). Finally, the lack of well-defined tertiary structure in the case of intrinsically disordered proteins is directly encoded in the sequences and their physicochemical properties (30). A fundamental question, therefore, is how frameshifting affects different physicochemical properties of biologically realistic protein sequences. To address this question, we have analyzed the complete sets of protein sequences in multiple representative organisms and compared them against their +1 and -1 frameshifted counterparts using over 600 different physicochemical properties of amino acids. Our results show that several such properties are significantly robust against frameshifting, a finding with potentially far-reaching biological implications.

Results

Effect of frameshifting at the level of the genetic code

We have first evaluated the impact of frameshifting on different physicochemical properties of individual amino acids (AAs) as encoded in the universal genetic code (UGC). A frameshift of the genetic code table produces a total of 232 pairs involving original and all possible respective frameshifted AAs (64 x 4 - 24 pairs involving stop codons). As a measure of the impact of frameshifting, we have calculated the Pearson R over this set of pairs for each of the 604 different amino-acid property scales studied (Figure 1A). A distribution of Pearson Rs over all properties corresponds approximately to a background distribution calculated from 10^7 scales with randomly chosen values (see Methods for details). Specifically, the distribution is centered around zero and the best-performing scales reach approximately 0.4. In comparison, the scales, which were computationally optimized for frameshift robustness of the UGC (see Methods for details), exhibit Pearson Rs of ~0.56 (Figure 1A, arrow). Importantly, a p-value analysis demonstrates that a subset of scales, belonging mostly to the hydrophobicity category (Figure 1, green bars), performs outstandingly well when compared to the random background (Figure 1B).

As representative examples, we highlight two consensus hydrophobicity scales: the Factor 1 scale (p-value = 0.0014), derived by Atchley et al. (31) via factor analysis of more than 500 different amino-acid scales including over 100 hydrophobicity scales, and its predecessor (p-value = 0.0005), derived by Kidera et al. (32) using similar means. A high degree of significance is also reached by several individual scales in other categories, including the knowledge-based scale of amino-acid affinity for the RNA/DNA nucleobase guanine (27) and an amino-acid β-propensity scale obtained by a neural network model (33). On the other hand, separating the AA scales according to physicochemical categories (Figure 1B) reveals the superiority of the hydrophobicity category, with over 100 scales exhibiting p-values < 0.05 and some reaching p-values < 10^{-4} (Figure 1B and Datasets S1, S2). This is supported by an enrichment analysis via Fisher's exact test showing that hydrophobicity is the only significantly enriched category among scales with p-values < 0.05. Similar results are also obtained with different p-value cutoffs or a somewhat different assignment of categories (26). Importantly, randomizing the genetic code with its block structure retained yields practically identical p-values for all scales as the random background used above (see Figures S2, S3). Moreover, alternative, less conservative methods of randomizing the genetic code also give qualitatively similar results (see Figures S2, S3). Finally, our findings are in a general agreement with the previous results obtained for a few select scales only (21, 22).

A set of scales that exhibit optimal frameshift stability in the context of the UGC could be derived in a consistent manner via different local minimization algorithms (see Methods for details, Figure S4). This set is

practically entirely described by the first two dimensions of a principal component analysis (PCA, Figure S4). The contribution of individual amino acids to these two principal components is visualized by the arrows in Figure 1C. Transforming the 604 studied physicochemical scales to this PCA space (Figure 1C) reveals strong clustering of hydrophobicity scales in close proximity to the optimal scales. At the same time, scales that themselves exhibit significant frameshift stability, yet belong to other categories, colocalize with the hydrophobicity clusters. This indicates that these scales share certain characteristics that are relevant for frameshift stability with hydrophobicity scales. Please note that the two main clusters shown in Figure 1C are congruent in the sense that inverting a scale moves it to the opposite side. In other words, in the case of the hydrophobicity category, the two clusters correspond to hydrophobicity and its inverse, hydrophilicity.

Effect of frameshifting at the level of protein sequences

In a biological setting, frameshifts always appear in the context of protein sequences. There, the reading frame can be shifted one nucleotide either towards or away from the 5' end of the mRNA, resulting in two different frameshifted protein variants (+1 shift and -1 shift, respectively). In Figure 2A, we show the hydrophobicity sequence profile of an exemplary wildtype protein (UniProt ID Q6NUP7) overlaid with the hydrophobicity profiles of its +1 and -1 frameshifted variants in the case of the Factor 1 consensus hydrophobicity scale (31). In order to quantify the similarity between the profiles before and after frameshifting, we use the Pearson R, which in this case is approximately 0.7 for both the +1 and the -1 shift as compared to the wildtype (Figure 2A). Here, it should be emphasized that over 2900 human protein sequences exhibit a Pearson R between their wildtype and +1 frameshifted Factor 1 profiles of 0.7 or greater. This is particularly striking considering the fact that the wildtype sequences in this subset exhibit a sequence identity with their +1 frameshifted variants of only 6.5% on average (see also Figure S1). In order to illustrate the impact of frameshifts at a proteomic scale, we show in Figure 2B the distributions of Pearson R as obtained by comparing wildtype Factor 1 profiles against the respective +1 (R_{median} = 0.55) or -1 shifted (R_{median} = 0.45) profiles over the whole human proteome.

Extending the proteomic analysis to all scales, in Figure 2C we report the distribution over all 604 studied scales of median Pearson Rs in human (wildtype vs. +1 shift) together with the appropriate background calculated from 10^6 scales with randomly chosen values (see Methods for details). As expected from the results obtained at the level of the genetic code, scales of the hydrophobicity category dominate among the scales with the highest median Pearson Rs and the lowest corresponding p-values in *H. sapiens* (Figures 2C, 2D) as well as *E. coli* and *M. jannaschii* (Figures S5, S6). Remarkably, the Factor 1 (31) and Kidera et al. (32) consensus hydrophobicity scales rank among the top four scales in this regard in human, where they are also joined by the Levitt hydrophobicity scale (34) and the knowledge-based scale of amino-acid affinity for guanine (27) (Figure 2D). In general, scales which are significantly resistant to frameshifts at the level of the

genetic code also exhibit significant invariance in a sequence context (Figure 2D, S6). In many cases, in fact, the p-values tend to be lower in the sequence context than in the case of the genetic code (Figures 2D, S6 and Datasets S1, S2).

As a next step, we have analyzed the enrichment of different Gene Ontology (GO) terms for human proteins with the highest frameshift invariance using the Factor 1 hydrophobicity scale. In the case of +1 frameshifts, there is a significant enrichment of integral membrane proteins with approximately one third of all proteins in the top quartile of the human proteome having this GO annotation, while for -1 frameshifts, one observes an enrichment of RNA-binding functions and nucleolus localizations (Figure 2E and Dataset S3). Finally, a comparison of Factor 1 frameshift stability, as captured by the Pearson R between a wildtype profile and its frameshifted counterpart, for more than 12000 orthologous proteins in *H. sapiens* and *M. musculus* reveals a high degree of similarity (Figure 2F). On the other hand, no such similarity is found for orthologous proteins in organisms belonging to different domains of life (Table S1). Nevertheless, GO analysis in *E. coli* along the same lines as in *H. sapiens* also shows significant enrichment of integral membrane proteins (p-value: +1 shift $9.7*10^{-3}$).

Frameshift stability of a membrane protein's hydrophobicity profile

The strong frameshift stability in the hydrophobicity profiles of many membrane proteins is exemplified in Figure 4 in the case of sodium/potassium/calcium exchanger 1 (Uniprot ID O60721). The wildtype Factor 1 hydrophobicity profile of this transmembrane protein differs only slightly from its +1 frameshifted variant (R = 0.90), despite their markedly different sequences (sequence identity = 5.4%). Importantly, different domains of the protein can easily be identified by analyzing its Factor 1 hydrophobicity profile with the transmembrane helices adopting extremely low values and the intervening soluble linkers and domains adopting significantly higher values (Figures 3A-C). Remarkably, the +1 frameshifted variant exhibits a pronounced similarity in its alteration of hydrophobic and hydrophilic regions, with the first transmembrane stretch (Figure 3B) showing a somewhat closer matching than the second (Figure 3C). In Figure 3D, we highlight an N-terminal stretch in which only two out of 19 residues align with each other and, in Figure 3E, a C-terminal stretch where not a single residue out of 31 is identical. Despite this, the high similarity in the hydrophobicity profiles of wildtype and frameshifted sequences in these regions is obvious. On the other hand, while the absolute value of electrostatic charge is largely retained upon +1 frameshifting, the shift results in an inversion of the sign of the charge (Figure 3E). Specifically, the wildtype Glu residues are frameshifted to mostly Arg and Lys residues. This suggests that, while hydrophobicity is mostly unaffected, the charge profile of frameshifted variants show major differences with respect to wildtype profiles. In fact, at a whole proteome level, the +1 frameshifting produces sequences whose net charge is negatively

correlated with the wildtype net charge (R = -0.45), although charge density profiles show a weaker relationship (median R = -0.15).

Frameshift stability of nucleobase-affinity and intrinsic disorder

So far, we have focused mainly on frameshift stability of properties in the hydrophobicity category. As an example of a scale from other categories that also shows significant behavior, we present analogous results for the knowledge-based scale of amino-acid affinity for the nucleobase guanine (Figures 4A-C). Importantly, this scale provides context for analyzing frameshift stability from a different perspective i.e. comparing both wildtype and frameshifted protein profiles with the nucleobase density profile of the wildtype coding mRNA. Namely, we have recently demonstrated a strong degree of matching between nucleobase-density profiles of mRNA coding sequences and the nucleobase-affinity profiles of the proteins they encode (26–29). For example, purine density profiles of human mRNA coding sequences match their autologous proteins' guanine-affinity profiles with a median Pearson |R| of 0.80 (wildtype in Figure 4B). We have used this to hypothesize that mRNAs and the proteins encoded by them could interact in a complementary, co-aligned fashion, especially if unstructured (26–29). As indicated above, some nucleobase-affinity profiles of proteins, such as GUA-affinity profiles (Figure 4A) in all studied organisms and ADE-affinity profile in *M. jannaschi* (Datasets S1, S2), exhibit significant robustness against frameshifting. Moreover, the previously observed matching of mRNA purine density profiles and their autologous proteins' GUA-affinity profiles is also retained for frameshifted proteins, as shown in Figure 4B. Here we show the distributions of Pearson R between the wildtype mRNA purine density profiles and either wildtype, +1 or -1 frameshifted protein GUA-affinity profiles. As a specific example, the wildtype mRNA purine density profile shown in Figure 4C matches the GUA-affinity profile of its -1 frameshifted protein with a Pearson |R| of 0.76. This value corresponds to the median of the proteomic distribution (-1 shift in Figure 4B), thus reflecting the high level of profile similarity even in a typical case. This is, in part, a consequence of the robustness in the GUA-affinity profiles (Figure 4A), but even more so a natural corollary of the fact that mRNA nucleobase-density profiles are unaffected by frameshifts combined with the original observation that protein GUA-affinity profiles match their mRNA purine density profiles.

Finally, we have also focused on intrinsic disorder (Figures 4D, 4E) as a more complex property, which does not only depend on the nature of individual AAs in a given polypeptide stretch, but is also context dependent. The average disorder in wildtype human proteins, as assessed using IUPred (35), correlates with the average disorder of their +1 and -1 frameshifted counterparts with median Pearson Rs of 0.49 and 0.41, respectively (Figures 4D, S7). Similar results are also seen for sequence profiles of intrinsic disorder. In Figure 4E, we show an example with a Pearson R of 0.40 between wildtype and +1 frameshift profiles, corresponding approximately to the median of the respective distribution in human. Despite this relatively modest level of

correlation in the median case, there still exist thousands of proteins with an undeniably strong frameshift stability. For example, there are close to 2800 proteins in human for which the wildtype disorder profiles correlate with their +1 frameshifted counterparts with a Pearson R > 0.7 (Dataset S4), with similar results seen for representative organisms from other domains of life (Figure S8). Finally, it should be also emphasized that the common classification of structured vs. unstructured regions, as given by an IUPRED cutoff of 0.5, is largely retained across the whole exemplary sequence in Figure 4E and, even more importantly, across entire proteomes for average disorder (Figures 4D, S7).

Discussion

Our results suggest that an inherent property of the universal genetic code is that frameshifting yields vastly different protein sequences, which, nevertheless, retain select physicochemical properties of the original sequences in a considerable number of cases. Specifically, this feature is a sign of the robustness of the genetic code with regard to frameshifting, which can be considerably enhanced in realistic biological sequences. A particularly strong example for this is seen in the case of hydrophobicity profiles. Importantly, this finding suggests a plausible novel mechanism for the evolution of protein sequences. In particular, frameshifting insertion and deletion mutations could enable major jumps in protein sequence space, while at the same time ensuring that some of the already optimized physicochemical properties of the original sequences are preserved (Figure 3, 5A). For example, the hydrophobic/hydrophilic sequence patterns are thought to be a key feature of proteins when it comes to determining the nature of their 3-dimensional structures. By keeping the hydrophobicity profile unaffected, the frameshifted sequence increases its chances of being able to adopt a well-defined fold. Recently, Gardner et al. (36) have observed that the predicted secondary and tertiary structure of proteins is relatively robust against point mutations, but, unexpectedly, also against frameshifting insertions and deletions. Our present results provide a potential explanation for this finding: it is possible that frameshift mutations analyzed by Gardner et al. resulted in protein sequences with largely retained hydrophobicity profiles, which, in turn, would have resulted in similar predicted secondary and tertiary structure features.

When it comes to an exact mechanism of how new sequences could be generated via frameshifting, a scenario one could envision involves gene duplication with an accompanying frameshift (Figure 5B). This would likely result in multiple premature stop codons that would need to be mutated, but this burden could be more than compensated for by having a physicochemically optimized starting point for further evolution. Naturally, not the whole sequence would need to be changed: one can also envision local frameshift events resulting in hybrid sequences, which are part wildtype, part new, increasing exponentially the combinatorial richness of the resulting sequences. In this sense, our results capture the most extreme case (i.e. frameshifts of full-length proteins): frameshifts over shorter stretches are expected to result in an even higher correlation

between wildtype and frameshifted profiles. Recently, Tripathi and Deem (37) provided evidence suggesting that the retention of physicochemical properties of amino acids upon point mutations improves the exploration of functional nucleotide sequences at intermediate evolutionary time scales. Our prediction is that similar conclusion applies not only to single-nucleotide substitutions, as explored by these authors, but also to the much more impactful, sequence-altering instances of both local and global frameshifting mutations.

Starting with the pioneering work of Alff-Steinberger and others (17–19), it has already been suggested that compositionally similar codons encode amino acids with similar physicochemical properties, implying that the genetic code may have been optimized for robustness against not only point mutations, but frameshifts as well. While previous studies of frameshift stability have been performed primarily at the genetic code level and only investigated a few amino-acid properties or their indirect correlates, as in the case of substitution matrices, our present results are based on a comprehensive analysis of over 600 different amino-acid properties and, importantly, report on the impact of frameshifting in the case of biologically realistic sequences in multiple organisms. Such generalization is relevant for several reasons. First, even in the best cases, the correlations for different amino-acid property scales at the level of the genetic code are quite weak (Figure 1A) and only in combination with realistic protein sequences can one gauge the full impact of frameshifting. What is more, our results show that for thousands of proteins, frameshift stability is so strong that it indeed might have direct, biologically relevant repercussions. Second, we provide quantitative evidence that frameshift stability applies to a whole category of different hydrophobicity scales and not just select examples. We find it particularly indicative that two consensus hydrophobicity scales, derived previously by considering hundreds of individual scales, rank in the very top when it comes to frameshift stability (Figure 2D, Datasets S1, S2).

Finally, our results suggest that, in addition to hydrophobicity, frameshift stability could apply to several other protein sequence properties including affinity to some nucleobases and structural disorder. It has been suggested that RNA-binding, a process with a strong dependence on hydrophobic forces, was one of the most important functions of ancient proteins (38, 39) and that the genetic code was shaped in response to the physicochemical pressures related to such interactions (29, 40, 41). It is possible that the frameshift invariant properties discussed above all partly reflect protein ability to interact specifically with nucleic acids in an unstructured context. The present results also suggest a generalization of our recent complementarity hypothesis. Namely, we propose that mRNAs bind in a co-aligned manner not only their autologous proteins, if unstructured, but also their frameshifted variants (26–29). Finally, we also observe a bias for disorder propensity to be retained after frameshifts.

Our results open up several directions for future work. An important frontier concerns the investigation of more complex protein properties which are directly dependent on their primary structure. For example, to what extent are secondary and tertiary structures of wildtype proteins related to those of their frameshifted variants? On a more practical note, what are potential implications of our results in a biomedical context? Is it possible that frameshift robustness could actually lead to deleterious gain of function? Finally, can evidence be found that frameshifting has indeed played a relevant role during evolution of real proteins? Future studies should shed light on these exciting questions and possibilities.

Materials and Methods

Complete annotated proteomes of *M. jannaschii*, *E. coli*, *M. musculus* and *H. sapiens* were obtained from the UniProtKB database (42) with the maximal-protein-evidence-level set to 4. The mRNA coding sequences for each protein were downloaded from the European Nucleotide Archive Database (43). Sequences including non-canonical amino acids or nucleobases were not analyzed. The majority of the amino-acid property scales studied were extracted from the AAindex database (44, 45), and were complemented by additional consensus scales derived by Atchley and coworkers (31) and a number of recently derived nucleobase/nucleotide affinity scales (46–49). The full set of scales was categorized following the procedure by Tomii et al. (50). The frameshifted variants of individual protein sequences were generated by removing the first four bases (+1 shift) or the first two bases (resulting in the -1 shift) in their wildtype mRNA coding sequences and translating them using the universal genetic code. Protein sequences were then converted to numerical profiles by exchanging each amino acid with its respective scale value and smoothing using 21-residue/codon windows, as used previously (26). Premature stop codons in frameshifted variants were excluded from the calculation of the average value in a local window, while the size of the window was reduced by their number, except in the case of disorder profiles (see SI). Significance analysis was performed on basis of either randomization of the UGC or scales themselves (see SI for details). The principal component analysis of the scale space was performed by calculating the set of optimized scales as presented previously (28) and using them to calculate the principal components. The experimentally derived scales were subsequently transformed in the already defined space. The analysis of the enrichment of gene ontology (GO) terms was carried out using GOrilla (51), Panther (52) and REVIGO (53) tools. The disorder propensity of a given protein sequence was calculated using IUPred (35) and the ‘long’ setting for the size of disorder detection.

Extended methodological details can be found in Supplementary Information.

Data Availability Statement: All data discussed in the paper will be made available to readers.

Acknowledgements

The authors thank all the members of the Laboratory of Molecular Biophysics at the University of Vienna for useful input. This work was supported by the European Research Council Starting Independent Grant [279408 to B.Z.] and the Austrian Science Fund FWF Standalone Grants [P 30550 and P 30680-B21 to B.Z.]

Figures

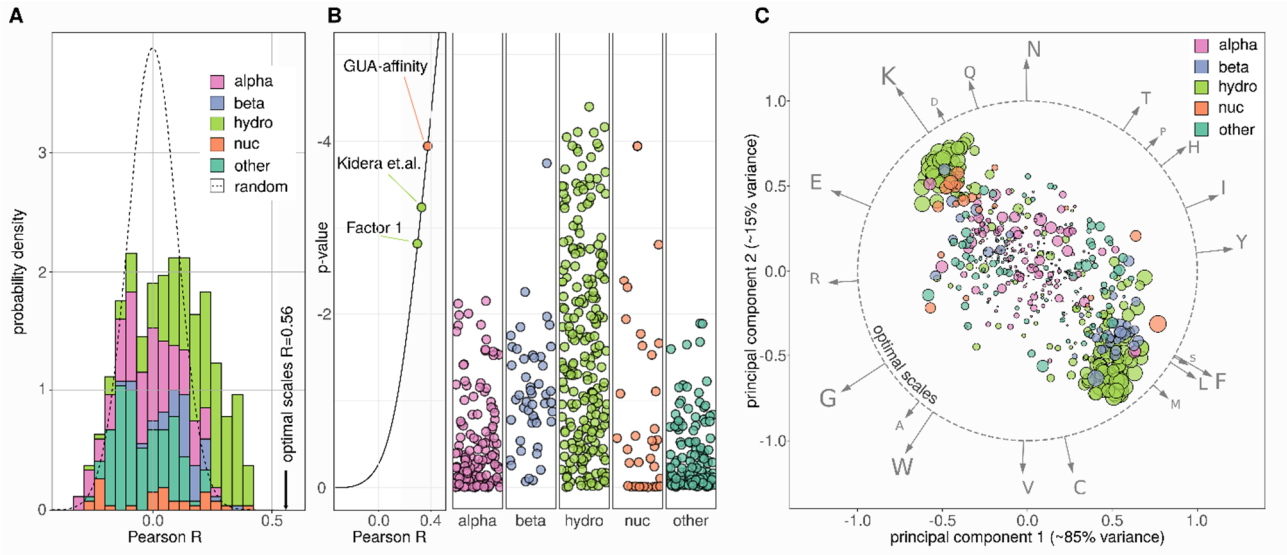


Figure 1. Frameshifting at the level of the genetic code A) Histogram of Pearson correlation coefficients R for UGC vs. its frameshifted version for all 604 scales investigated. Scales are grouped by category and presented as a stacked, normalized histogram. The expected density derived via a random model is shown as a dashed line, while the highest achievable Pearson R obtained for computationally optimized scales is marked by an arrow. B) First panel: Pearson Rs for UGC with the associated p-values, with select scales indicated. Other panels: p-values of 604 studied scales grouped by category (alpha: alpha and turn propensity; beta: beta propensity; hydro: hydrophobicity, nuc: nucleobase/nucleotide affinity; other). C) Clustering of scales according to optimal frameshift stability: the 604 studied scales were transformed into a space defined by the PCA of a set of scales that were computationally optimized for frameshift stability at the UGC level. The first and second principal component of this PCA space, which account for almost 100% of variance of the set of optimal scales, are shown. The 604 physical scales are shown as individual data points whose sizes correspond to the negative logarithm of the associated p-values. The lengths and the directions of arrows correspond to the relative contributions of the respective amino acids to PC1 and PC2.

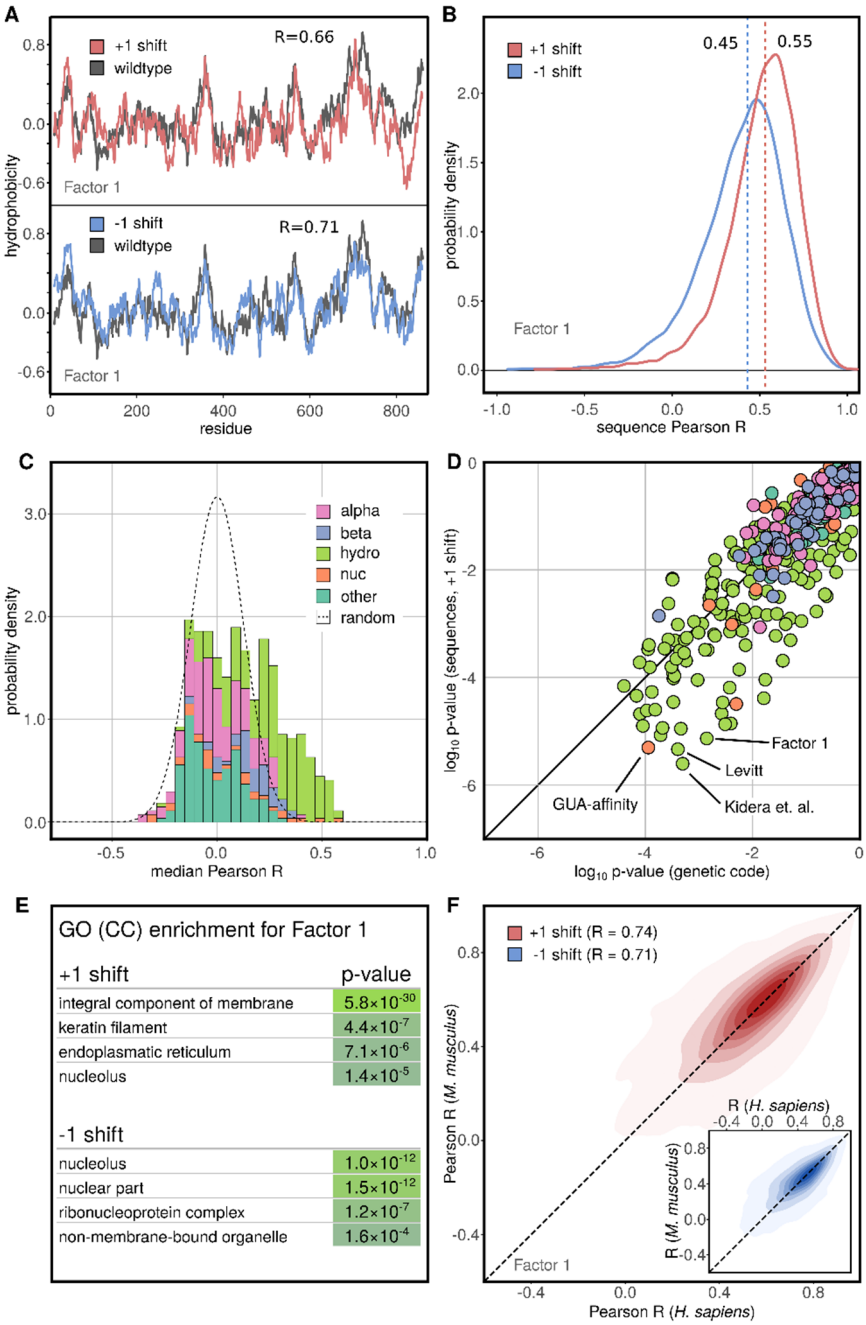


Figure 2. Frameshifting at the level of protein sequences A) Comparison of wildtype and +1 (top) and -1 (bottom) frameshifted Factor 1 hydrophobicity profiles for Ser/Thr phosphatase 4 regulatory subunit 4 protein (UniProtID Q6NUP7) with the associated Pearson Rs. All profiles are treated with a 21-residue smoothing window. B) Distributions of Pearson R for wildtype vs. +1 frameshift (red) and wildtype vs. -1 frameshift (blue) for Factor 1 over all human proteins (N = 17083) with the medians indicated. C) Histogram of median Pearson R for 604 scales when comparing wildtype and +1 frameshifted profiles in human for all investigated scales, grouped by category and presented as a stacked, normalized histogram. The expected density derived via a random model is shown as a dashed line. D) Comparison of p-values for frameshifting at the level of the genetic code and frameshifted human sequences for 604 studied scales. E) Enrichment of

GO cellular compartment terms in the top quartile of human sequences according to Pearson Rs between wildtype and +1 or -1 frameshifted Factor 1 profiles. F) Comparison of Pearson Rs (wildtype vs. +1 frameshifted Factor 1 profiles) between orthologous proteins in *H. sapiens* and *M. musculus* (N = 12174, R = 0.74). The same is shown for -1 frameshifts in the inset (R = 0.71).

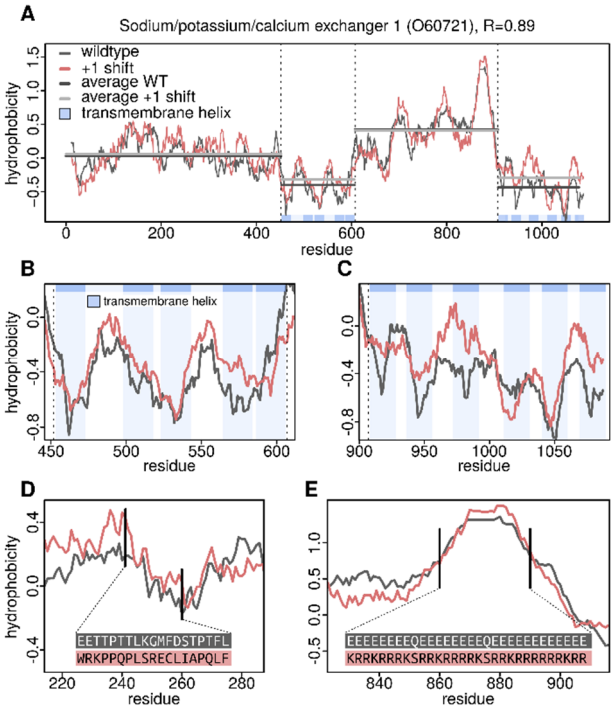


Figure 3. Robustness of a membrane protein's hydrophobicity profile against frameshift A) Factor 1 hydrophobicity profiles of wildtype sodium/potassium/calcium exchanger (UniProtID O60721) and its +1 frameshifted variant with relevant regions indicated with dashed lines. Closeup of the profiles in the first (B) and the second (C) transmembrane domains of the protein. Note that the specific locations of transmembrane helices are matched in all cases but one. D) Comparison of wildtype and +1 frameshifted sequences in a region outside the transmembrane domains together with the associated Factor 1 profiles. The hydrophobic profile shows high similarity despite drastic differences between specific sequences. E) Inversion of the charge pattern upon +1 frameshift with a retained hydrophobicity profile.

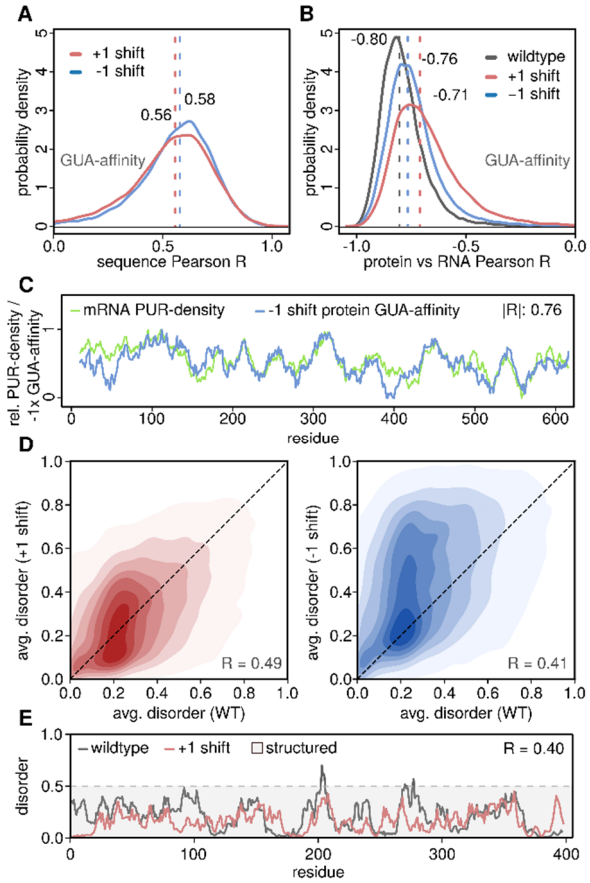


Figure 4. Frameshifting in the context of guanine-affinity and intrinsic disorder A) Distributions of Pearson R between guanine-affinity profiles of wildtype and +1 or -1 frameshifted human protein sequences (N = 17083). B) RNA vs protein: Distributions of Pearson R between mRNA purine-density profiles and their autologous protein's guanine-affinity profiles in human for wildtype, +1 and -1 frameshifted sequences (N = 17083) with the medians indicated. C) RNA vs. protein: comparison of an mRNA purine-density profile and protein guanine-affinity profile for the -1 frameshifted sequence of the nuclear RNA export factor (UniprotID: Q9GZY0) whose Pearson R corresponds to the median given in B). D) Comparison between average disorder in wildtype sequences and +1 (red) or -1 (blue) frameshifted sequences. E) Example IUPRED (35) intrinsic disorder profiles of a wildtype protein and its +1 frameshifted variant. Note that both profiles are predicted to be mostly structured before and after the frameshift event.

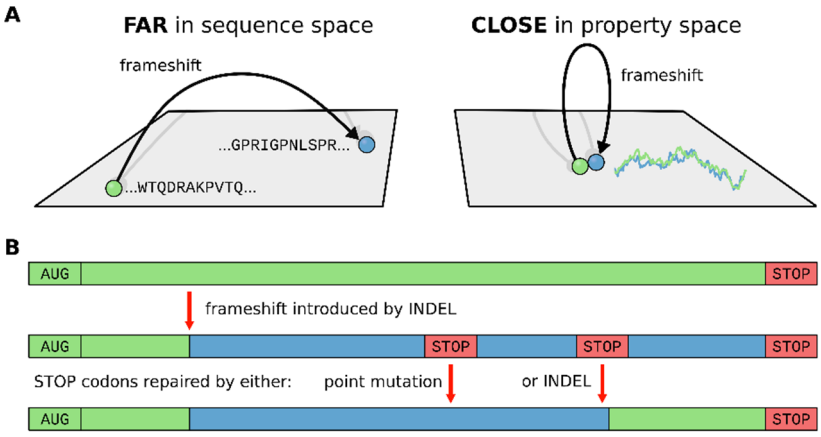


Figure 5. Evolution of protein sequences via frameshifting A) Frameshifts enable major jumps in protein sequence space with little change in the physicochemical properties, like hydrophobicity, select nucleotide affinities or protein disorder of the sequence in question. B) Nucleotide insertion or deletion in an existing gene results in a novel frameshifted gene containing several premature stop codons. These stop codons could then be removed by either single point mutations or another indel-induced frameshift.

References

1. Stenson PD, et al. (2003) Human Gene Mutation Database (HGMD ®): 2003 update. *Hum Mutat* 21(6):577–581.
2. Mertins P, et al. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534(7605):55–62.
3. Garcia-Diaz M, Kunkel TA (2006) Mechanism of a genetic glissando*: structural biology of indel mutations. *Trends Biochem Sci* 31(4):206–214.
4. Maki H (2002) Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. *Annu Rev Genet* 36(1):279–303.
5. Hu J, Ng PC (2012) Predicting the effects of frameshifting indels. *Genome Biol* 13(2):R9.
6. Seligmann H, Pollock DD (2004) The Ambush Hypothesis: Hidden Stop Codons Prevent Off-Frame Gene Reading. *DNA Cell Biol* 23(10):701–705.
7. Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y (2010) Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics* 11(1):491.
8. Lykke-Andersen S, Jensen TH (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* 16(11):665–677.
9. Fernandez M, et al. (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics* 12(Suppl 13):S5.
10. Itzkovitz S, Alon U (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res* 17(4):405–12.
11. Adli M (2018) The CRISPR tool kit for genome editing and beyond. *Nat Commun* 9(1):1911.
12. Claverie J-M (1993) Detecting Frame Shifts by Amino Acid Sequence Comparison. *J Mol Biol* 234(4):1140–1157.
13. Pai H V, et al. (2004) A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine. *J Biol Chem* 279(26):27383–9.
14. Raes J, Van de Peer Y (2005) Functional divergence of proteins through frameshift mutations. *Trends Genet* 21(8):428–431.
15. Dinman JD (2012) Mechanisms and implications of programmed translational frameshifting. *Wiley Interdiscip Rev RNA* 3(5):661–673.

16. Huvet M, Stumpf MP (2014) Overlapping genes: a window on gene evolvability. *BMC Genomics* 15(1):721.
17. Alff-Steinberger C (1969) The genetic code and error transmission. *Proc Natl Acad Sci U S A* 64(2):584–91.
18. Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33(5):412–417.
19. Wagner A (2013) *Robustness and Evolvability in Living Systems* (Princeton University Press, Princeton) doi:10.1515/9781400849383.
20. Wang X, et al. (2015) The shiftability of protein coding genes: the genetic code was optimized for frameshift tolerating. doi:10.7287/peerj.preprints.806v1.
21. Geyer R, Madany Mamlouk A (2018) On the efficiency of the genetic code after frameshift mutations. *PeerJ* 6:e4825.
22. Wnętrzak M, Błażej P, Mackiewicz P (2019) Optimization of the standard genetic code in terms of two mutation types: Point mutations and frameshifts. *Biosystems* 181:44–50.
23. Wang X, Chen G, Zhang J, Liu Y, Cai Y (2019) Frameshifts and wild-type protein sequences are always similar because the genetic code is nearly optimal for frameshift tolerance. *bioRxiv*:67736.
24. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157(1):105–132.
25. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The Protein Folding Problem. *Annu Rev Biophys* 37(1):289–316.
26. Hlevnjak M, Polyansky A, Zagrovic B (2012) Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res* 40(18):8874–8882.
27. Polyansky A, Zagrovic B (2013) Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res* 41(18):8434–8443.
28. Bartonek L, Zagrovic B (2017) mRNA/protein sequence complementarity and its determinants: The impact of affinity scales. *PLoS Comput Biol* 13(7). doi:10.1371/journal.pcbi.1005648.
29. Zagrovic B, Bartonek L, Polyansky AA (2018) RNA-protein interactions in an unstructured context. *FEBS Lett* 592(17). doi:10.1002/1873-3468.13116.
30. Oldfield CJ, Dunker AK (2014) Intrinsically Disordered Proteins and Intrinsically Disordered

Protein Regions. *Annu Rev Biochem* 83(1):553–584.

31. Atchley WR, Zhao J, Fernandes AD, Drüke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A* 102(18):6395–400.
32. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* 4(1):23–55.
33. Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202(4):865–884.
34. Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 104(1):59–107.
35. Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347(4):827–39.
36. Coray DS, Sibaeva N, McGimpsey S, Gardner PP (2018) Evolutionary, structural and functional explorations of non-coding RNA and protein genetic robustness. *bioRxiv*:480087.
37. Tripathi S, Deem MW (2018) The Standard Genetic Code Facilitates Exploration of the Space of Functional Nucleotide Sequences. *J Mol Evol* 86(6):325–339.
38. Alva V, Söding J, Lupas AN (2015) A vocabulary of ancient peptides at the origin of folded proteins. *Elife* 4. doi:10.7554/eLife.09410.
39. Noller HF (2012) Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol* 4(4):a003681.
40. Woese CR (1973) Evolution of the genetic code. *Sci Nat* 60(10):447–459.
41. Koonin E, Novozhilov A (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61(2):99–111.
42. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46(5):2699–2699.
43. Harrison PW, et al. (2019) The European Nucleotide Archive in 2018. *Nucleic Acids Res* 47(D1):D84–D88.
44. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* 28(1):374.
45. Kawashima S, et al. (2007) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36(Suppl 1):D202–D205.
46. Hajnic M, Osorio JI, Zagrovic B (2014) Computational analysis of amino acids and their

sidechain analogs in crowded solutions of RNA nucleobases with implications for the mRNA-protein complementarity hypothesis. *Nucleic Acids Res* 42(21):12984–12994.

47. Hajnic M, Osorio JI, Zagrovic B (2015) Interaction preferences between nucleobase mimetics and amino acids in aqueous solutions. *Phys Chem Chem Phys* 17(33):21414–21422.
48. Hajnic M, Ruiter A de, Polyansky AA, Zagrovic B (2016) Inosine Nucleobase Acts as Guanine in Interactions with Protein Side Chains. *J Am Chem Soc* 138(17):5519–5522.
49. Andrews CT, Campbell BA, Elcock AH (2017) Direct Comparison of Amino Acid and Salt Interactions with Double-Stranded and Single-Stranded DNA from Explicit-Solvent Molecular Dynamics Simulations. *J Chem Theory Comput* 13(4):1794–1811.
50. Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *"Protein Eng Des Sel* 9(1):27–36.
51. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10(1):48.
52. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47(D1):D419–D426.
53. Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* 6(7):e21800.

Supplementary Information for

Frameshifting preserves key physicochemical properties of proteins

Lukas Bartonek, Daniel Braun, and Bojan Zagrovic

Bojan Zagrovic
Email: bojan.zagrovic@univie.ac.at

This PDF file includes:

Extended Materials and Methods
Figures S1 to S8
Table S1
Legends for Datasets S1 to S4
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S4

Extended Materials and Methods

Data sets Complete annotated proteomes of *Methanocaldococcus jannaschii*, *Escherichia coli*, *Mus musculus* and *Homo sapiens* together with their mRNA coding sequences were analyzed. Protein sequences were obtained from the UniProtKB database (1) with the maximal-protein-evidence-level set to 4, including only reviewed Swiss-Prot entries. The mRNA coding sequences for each protein were downloaded from the European Nucleotide Archive Database (2). Sequences including non-canonical amino acids or nucleobases were not analyzed.

Amino-acid scales The majority of the amino-acid property scales studied were extracted from the AAindex database (3, 4), and were complemented by additional consensus scales derived by Atchley and coworkers (5) and a number of recently derived nucleobase/nucleotide affinity scales (6–9). In the rare cases when individual amino acids were not defined in a given scale, they were treated the same as stop codons (see below). The full set of scales (excluding nucleotide affinity scales) was divided into six categories following the procedure by Tomii et al. (10) who grouped the scales available in the AAindex into six meaningful categories: “Alpha and Turn propensity”, “Beta propensity”, “Composition”, “Hydrophobicity”, “Physical and Chemical Properties” and “Other”. Additional scales not included in the original analysis by Tomii et al. were added to the existing categories using the same method, except for nucleobase-affinity scales which were added as an additional category. In our presentation of the data, categories “Composition”, “Physical and Chemical Properties” and “Other” were collapsed into a single category “Other”, since all showed no significant effect upon frameshifting and the differences between the groups were not evident.

Sequence identity Sequence identity describes the fraction of strictly identical amino acids that are located at the same position in two protein sequences. Specifically, in a wildtype protein sequence and its frameshifted counterpart, this refers to locations in the two sequences defined by the respective wildtype codons and their +1 or -1 frameshifted counterparts in the mRNA coding sequence.

Genetic code analysis For each codon in the genetic code table, four associated frameshift codons were constructed by adding one of the four possible bases to the last two bases of the original codon, thereby creating the +1 frameshift graph, yielding 64×4 pairs of native and respective frameshifted codons. Due to this graph being cyclic, there was no need to separately consider the -1 frameshifts. All pairs containing a stop codon were excluded from the set, resulting in 232 frameshift pairs. Utilizing the universal genetic code (UGC), all codons were translated to their corresponding amino acids. Applying one of the 604 amino acid property scales to both original and frameshifted amino acids produced a set of numerical pairs for which the sample Pearson R correlation coefficient was calculated. To compare the resulting 604 correlation coefficients against an appropriate background, we carried out the same procedure with 10^6 scales each containing 20 values randomly chosen between -1 and 1. Based on the standard deviation of this random background, Z-scores were calculated for each of the 604 scales, which in turn were used to calculate p-values from an analytic Gaussian distribution. Extremely similar results were obtained by randomizing the genetic code while keeping its box-like nature intact (see *genetic code randomization*).

Genetic code randomization Four different methods of generating a random background were used to assess the significance of our observations. First, instead of randomizing the genetic code, the scales themselves were randomized (*“rand. scales”)* as described in the description of genetic code analysis above. Virtually identical significances for individual scales were obtained for a randomization of UGC, while keeping its block structure intact (*“boxes”)*. Specifically, boxes of codons as found in UGC were assigned to new amino acids at random, while keeping the stop codons in their original positions (*“boxes”*). Thirdly, all the assignments of codons to amino acids were shuffled, while retaining the absolute number of codons for each amino acid and keeping the stop codons in their original positions *(“shuffle”)*. Lastly, a complete randomization was performed whereby amino acids were randomly assigned to codons with the only limitations being that each amino acid would appear at least once, and the stop codons would stay in their original places *(“random”)*.

Optimization of scales The optimization of scales for frameshift resistance was performed via a zero-temperature Monte-Carlo scheme in analogy with our previously published analysis in a

different context using Pearson R as an optimization metric (11). Briefly, simulated annealing was performed by scaling the step-size with simulation time. A total of 3000 steps were carried out for each Monte-Carlo run. Scales were initialized with random values. At every step, between 1 and 3 scales values, chosen at random, were perturbed by small time dependent offset. If the resulting scale resulted in a higher Pearson R than the original, the move was accepted, and otherwise it was rejected. This resulted in the highest correlations of up to R=0.558 for the UGC with its frameshifted version.

Principal component analysis The PCA was performed using scikit-learn(12) in python. The PCA space was fitted on 1000 optimal scales for frameshift robustness calculated via Monte-Carlo with an explained variance of 83.2% in PC1 and 16.8% in PC2. The lengths and the directions of arrows correspond to the relative contributions of the respective amino acids to PC1 and PC2. Other local optimization algorithms implemented in scipy (Nelder-Mead, BFGS, trust-constr) yielded theoretical scales resulting in equivalent components in a PCA. Although the ratio of explained variance between PC1 and PC2 varied according to the specific set of optimal scales derived, their total remained ~100% for all cases. In a next step, we transformed the 604 studied 'realistic' scales into this PCA space of optimal scales.

Sequence frameshift generation The frameshifted variants of individual protein sequences were generated by removing the first four bases (+1 shift) or the first two bases (resulting in the -1 shift) in their wildtype mRNA coding sequences and translating them using the UGC. This approach was chosen so that negative shifts could be performed without the need to have additional genomic information beyond the original coding sequence. After frameshifting, the AUG codon at the beginning of each sequence was removed from all original wildtype sequences in order to enable comparison between equally long protein sequences. The effect of this on the obtained results is negligible since the first three bases contain identical information in all sequences (AUG in mRNAs, Met in protein sequences).

Profile Calculation and Comparison Protein sequences were converted to numerical profiles by exchanging each amino acid with its respective scale value for each of the 604 different physicochemical property scales. Protein profiles were subsequently smoothed using a 21-residue window, to reduce noise and highlight global features. Profiles stemming from wildtype and frameshifted protein sequences were compared via the Pearson R correlation coefficient. For each property in question, the full distribution of Pearson Rs between wildtype and frameshifted profiles of all sequences in a given proteome was evaluated and its median used to assess the associated statistical significance in comparison with a randomized background. Specifically, we have derived the distribution of Pearson R between wildtype and frameshifted profiles for a given proteome for 10^5 scales with 20 values randomly chosen between -1 and 1. The standard deviation of the distribution of median values corresponding to this randomized background was subsequently used to calculate the Z-scores and p-values for each of the 604 property scales studied. Finally, in one instance, protein profiles were compared against mRNA nucleobase-density profiles (Fig. 4B, 4C). For this, codons were converted to a number between 0 and 3, representing the number of specific bases they contained and subsequently smoothed using a 21-codon window.

Stop codons In most cases translation of frameshifted mRNA sequences introduces premature stop codons in the resulting protein sequences. For the calculation of sequence profiles corresponding to different physicochemical properties of frameshifted variants, such positions were excluded from the calculation of the average value in a local window, while the size of the window was reduced by their number.

GO term enrichment The analysis of the enrichment of gene ontology (GO) terms in a target set defined as the top quartile of the distribution of Pearson Rs for Factor 1 frameshifts in *H. sapiens* was done using GOrilla (13) and REVIGO (14) tools. In GOrilla, we used the list of 17083 genes in the *H. sapiens* dataset as the background, for which 16281 GO terms were found. In order to reduce the redundancy of the GOrilla output, each ontology enrichment list - cellular compartment (CC), molecular function (MF) and biological process (BP) - was fed into REVIGO, respectively, together with the associated multiple-hypothesis corrected p-values. We used the Lin's measure of similarity with a similarity parameter of 0.5, characterized as a small

allowed similarity. In the main text, we present the top four significant hits in the CC ontology for both +1 and -1 frameshifts, while the total REVIGO output of the GO analysis can be found in Supporting Information (Table S3). The same procedure was applied to *M. musculus*. Since organisms from other domains of life are not supported by GOrilla, we used Panther (15) for their GO analysis, with same parameters whenever possible.

Orthologs Lists of orthologous genes between *H. sapiens* and *M. musculus*, *M. jannaschii* and *E. coli* were downloaded from InParanoid8 (16) and used to compare Pearson Rs of proteins' Factor 1 frameshift stability between different organisms.

Intrinsic disorder The disorder propensity of a given protein sequence was calculated using IUPred (17) and the 'long' setting for the size of disorder detection. Disorder profiles obtained from IUPred were smoothed using the same window of 21 residues as in case of other profiles for consistency. Since IUPred cannot handle stop codons, they were accounted for by creating two sequences *in silico*, one with the previous residue replacing the stop codon and one using the following residue. Disorder was calculated for both versions and averaged.

Data Access All relevant original data is provided as Supplementary Material to this publication. In case of already published data, the original publications are cited, and data should be accessed at the original source. All data necessary to replicate our results is accessible.

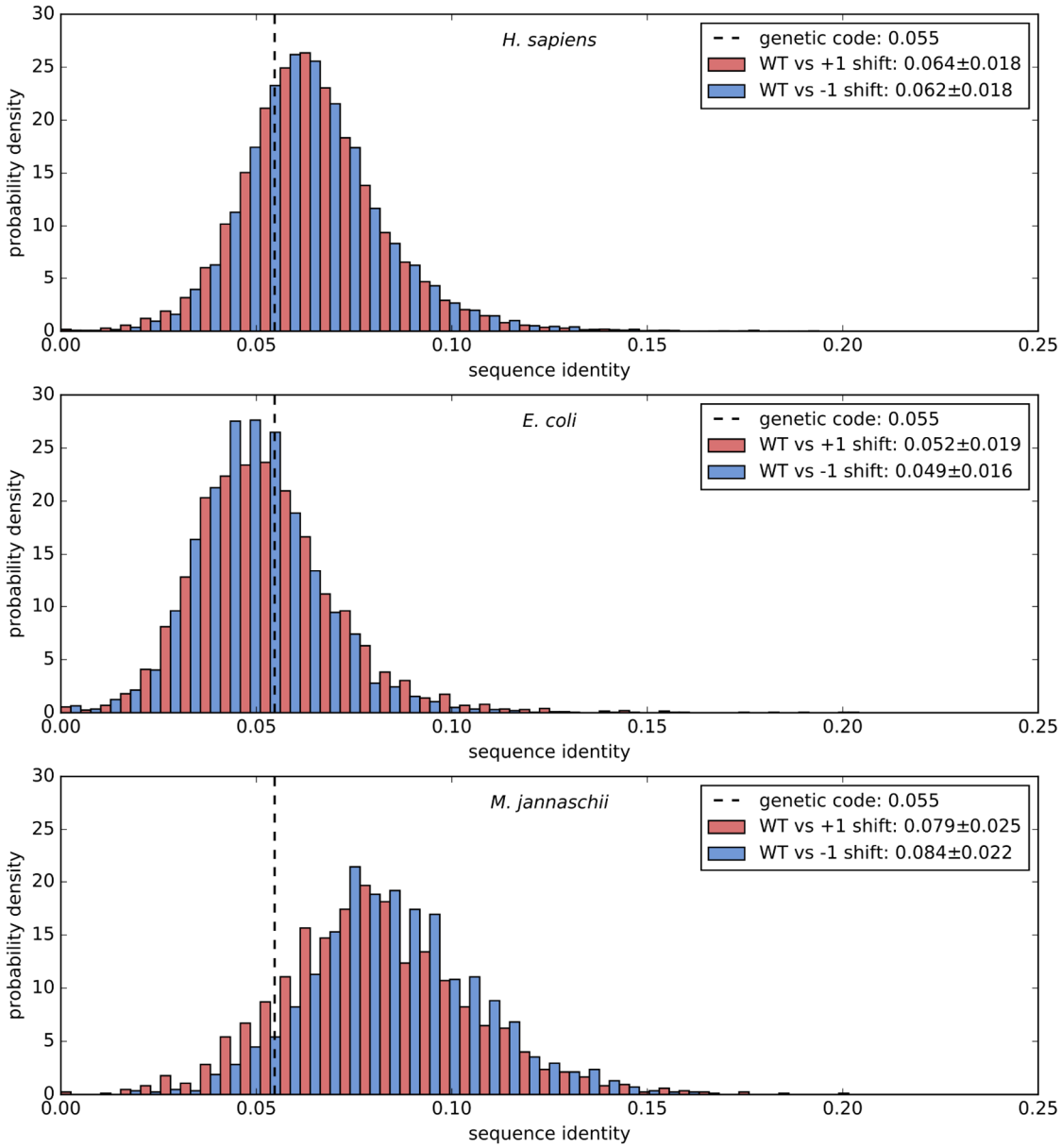


Figure S1. Distributions of sequence identity between wildtype (WT) and frameshifted proteins over complete proteomes of *H. sapiens* (N = 17083), *E. coli* (N = 3944), and *M. jannaschii* (N = 1666). For each distribution, median and standard deviation are shown in the legend. In the universal genetic code, 14 out of 256 frameshift transitions result in the same amino acid, indicated by a dashed line at 5.5%.

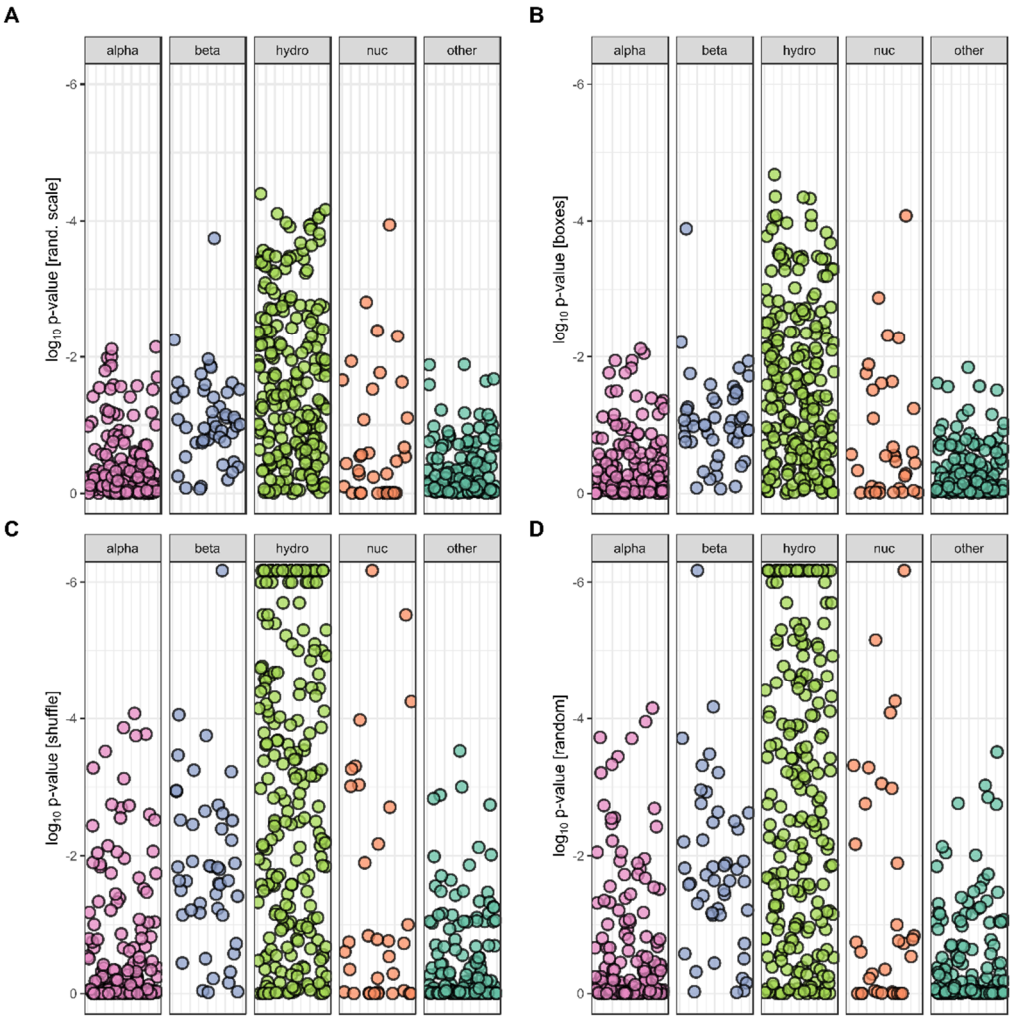


Figure S2. Significance analysis for different randomizations of the universal genetic code (UGC) A) p-values for individual scales as calculated by using a background with randomized scale values. This figure is equivalent to Fig. 1B in the main article. **B)** If the UGC box structure is kept intact while shuffling amino acids in the genetic code, similar results as in **A** are obtained. **C)** By shuffling each amino acid in the genetic code, the box structure of UGC is effectively abolished. Doing so results in a significant decrease in p-values at the cost of unrealistic structures of the genetic code. **D)** instead of shuffling, amino acids are randomly assigned to codons with the only criterion being that each amino acid is assigned at least once. In all cases, STOP codons were not reassigned during the procedure. It is important to notice that the hydrophobicity category is highly significant regardless of the specific randomization procedure. Scales in panels C and D for which not a single randomization performed better than the original code are represented with p-values smaller than 10^{-6}.

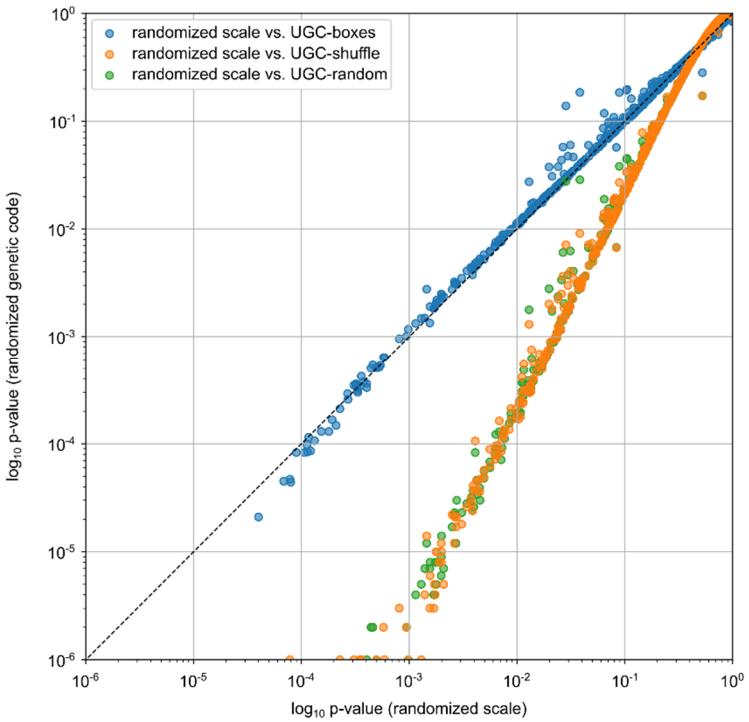


Figure S3. Comparison of different randomization procedures of generating a background. Randomizing the universal genetic code while keeping the box structure intact results in quantitatively highly similar results to randomizing the specific scale values (blue). Similarly, the other two methods, shuffling the codon assignments and randomizing the assignments, results in highly similar results (orange and green). Although these two groups differ in values, they still show the same behavior qualitatively.

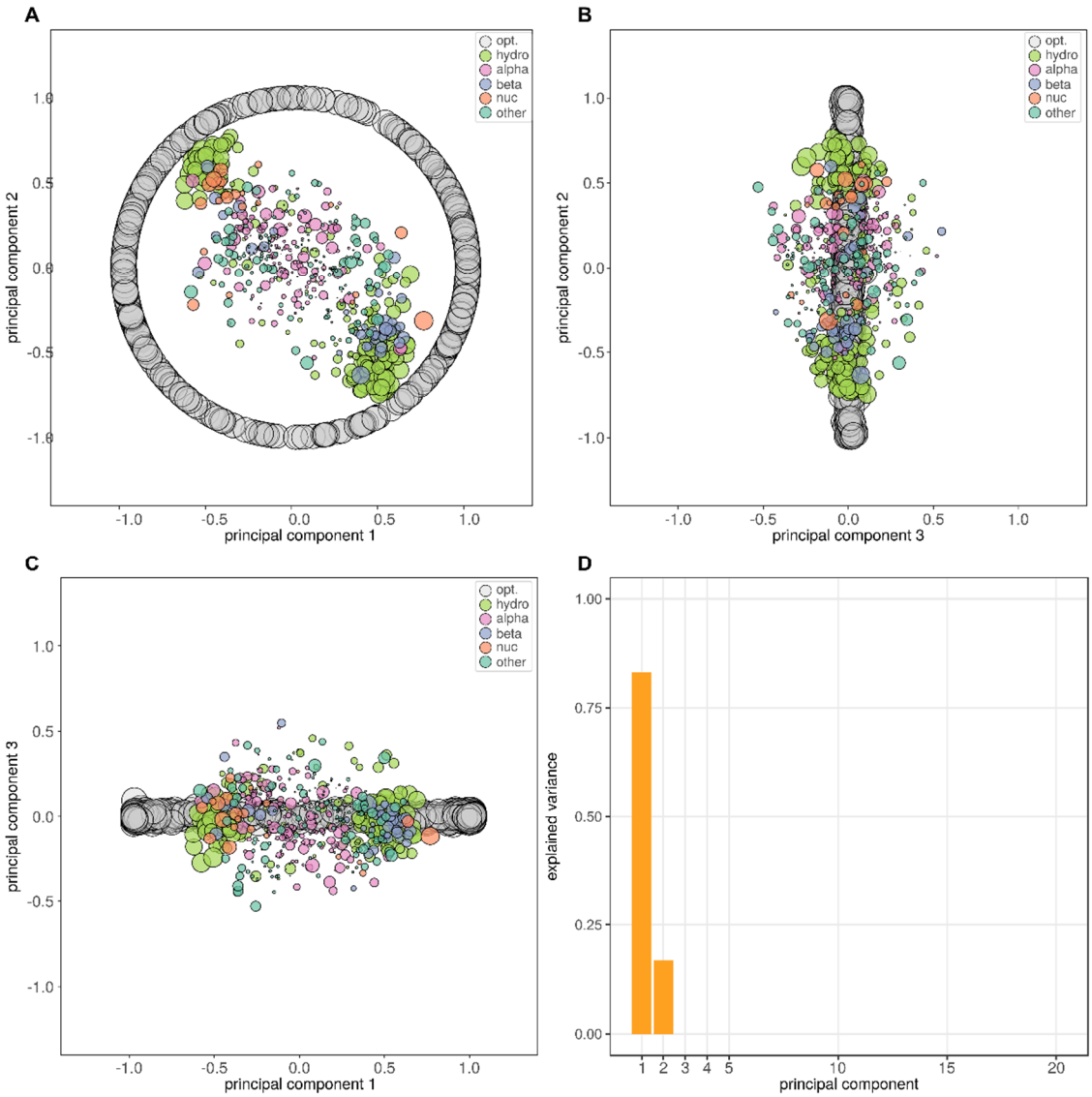


Figure S4. Principal component analysis (PCA) performed on 1000 theoretical scales derived for optimal frameshift stability (grey), together with 604 studied scales (colored according to their category) transformed into this PCA space. Panel **A** is similar to Fig 1C, except that 1000 scales from the Monte-Carlo optimization are explicitly shown. Panels **B** and **C** show PC 3 (0% explained variance of optimal scales) together with either PC 2 (16.8 % explained variance of optimal scales) or PC 1 (83.2 % explained variance of optimal scales). Panel **D** shows the explained variance of the Monte-Carlo optimized scales for each principal component in the PCA.

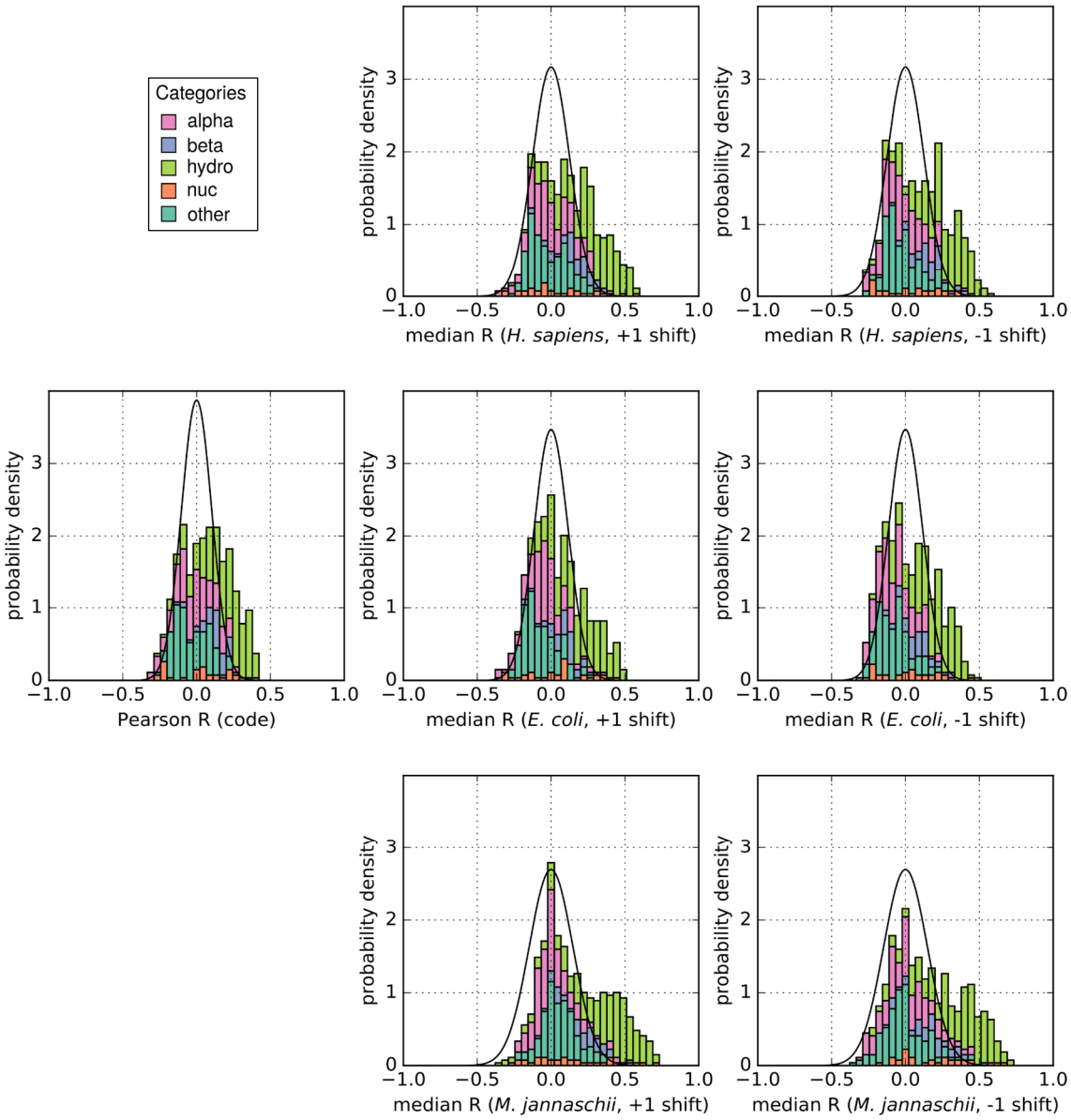


Figure S5. First column: Distribution (stacked histogram) of Pearson R correlation coefficients between UGC and its frameshifted variant for 604 different amino acid property scales grouped by category. The Gaussian curve with a standard deviation of 0.103 corresponds to the random background derived from 10^6 scales with random values. **Second and third column:** Distributions of the median Pearson R correlation coefficients over complete proteomes of the indicated organisms for either +1 frameshifted or -1 frameshifted sequences. The random backgrounds were derived from 10^5 scales with random values resulting in standard deviations of 0.126 (*H. sapiens*), 0.115 (*E. coli*), and 0.148 (*M. jannaschii)*. All distributions integrate to 1.

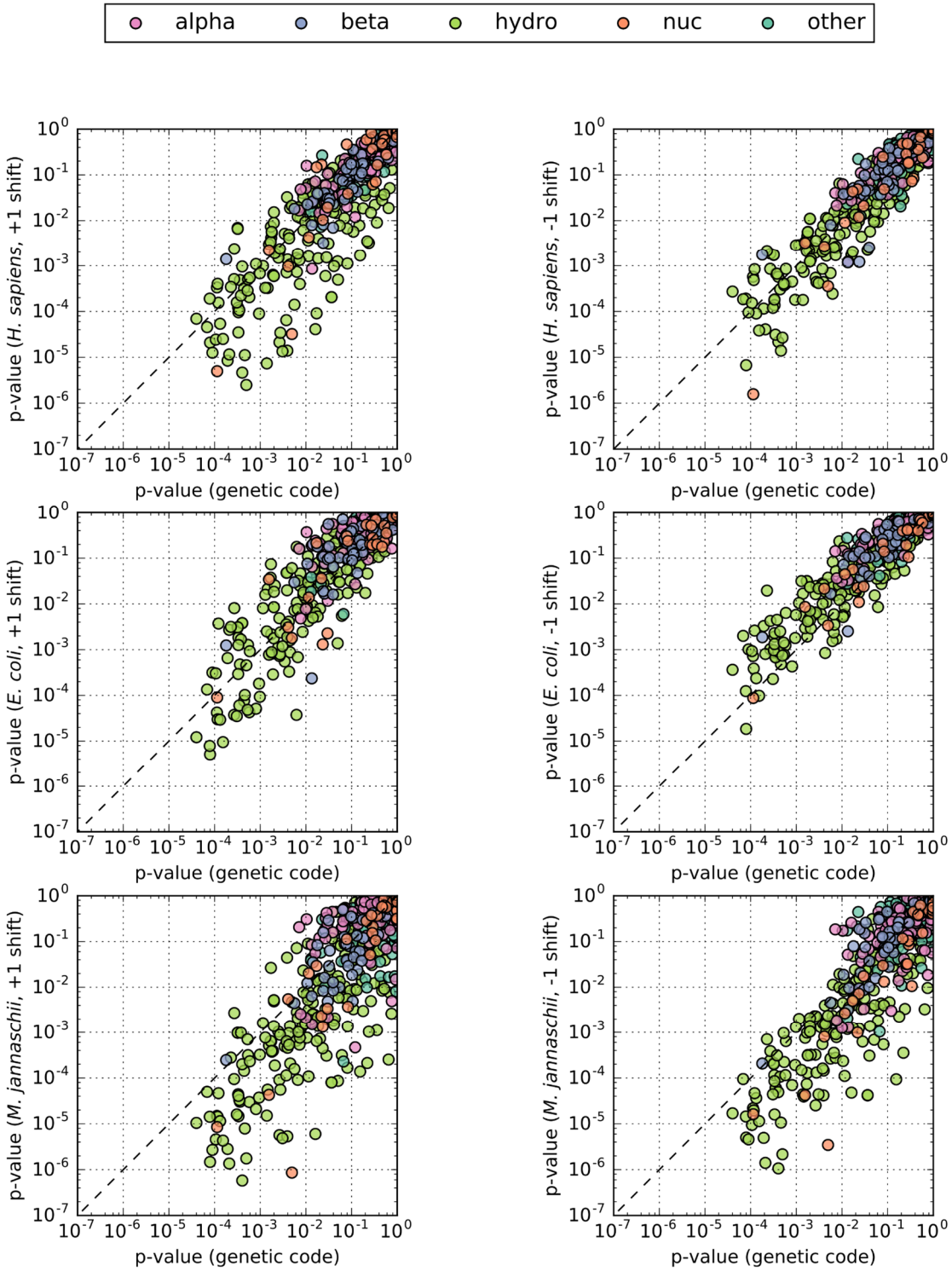


Figure S6. Scatter plots for p-values of 604 different amino acid property scales in the context of UGC and real biological sequences of *H. sapiens*, *M. jannaschii*, and *E. coli*. The first and second column compare the genetic code with +1 shifts and -1 shifts, respectively.

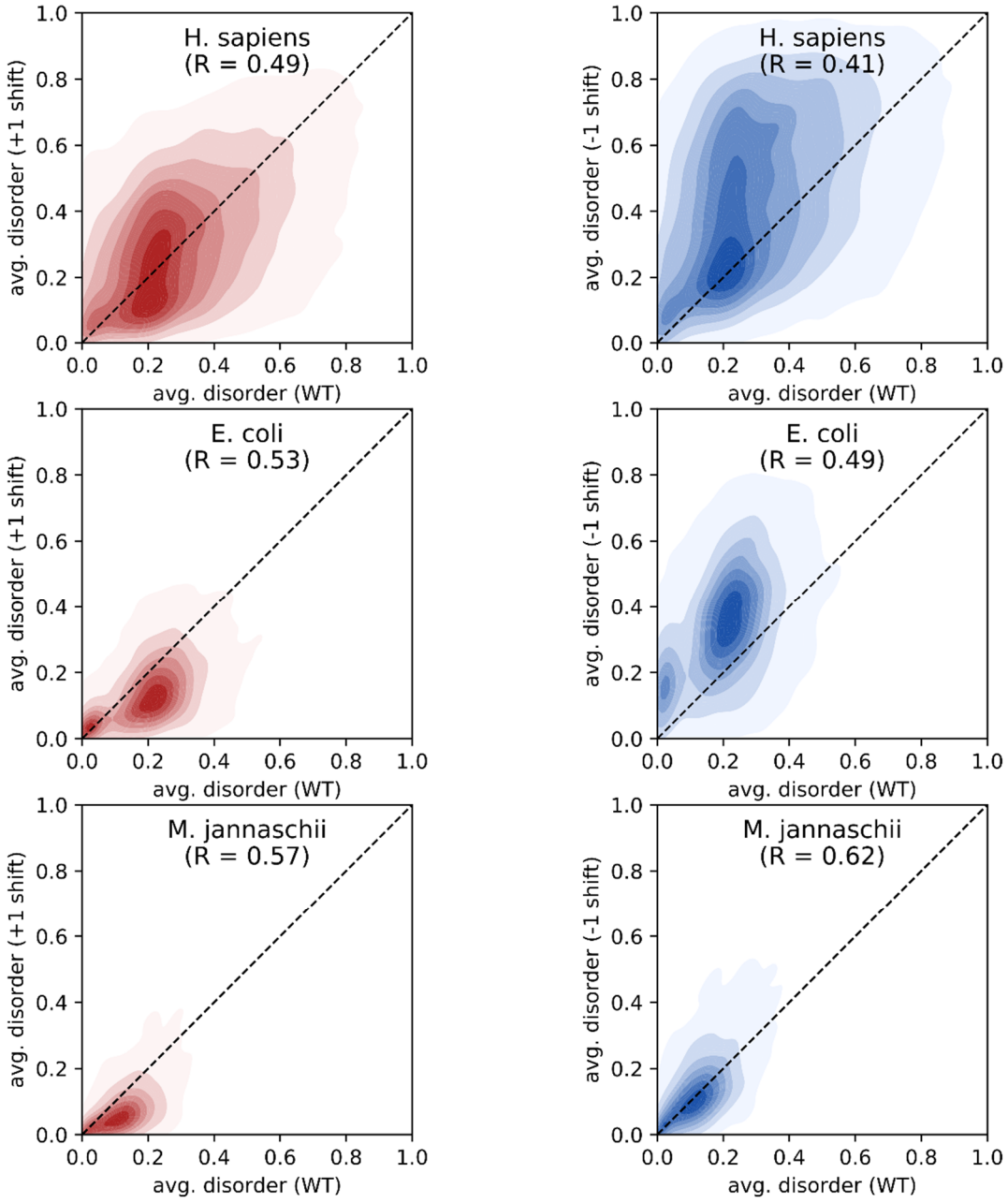


Figure S7. Correlation between the average IUPred(17) disorder of wildtype (WT) proteins and +1/ -1 frameshifted proteins over complete proteomes of *H. sapiens* (N = 17083), *E. coli* (N = 3944), and *M. jannaschii* (N = 1666).

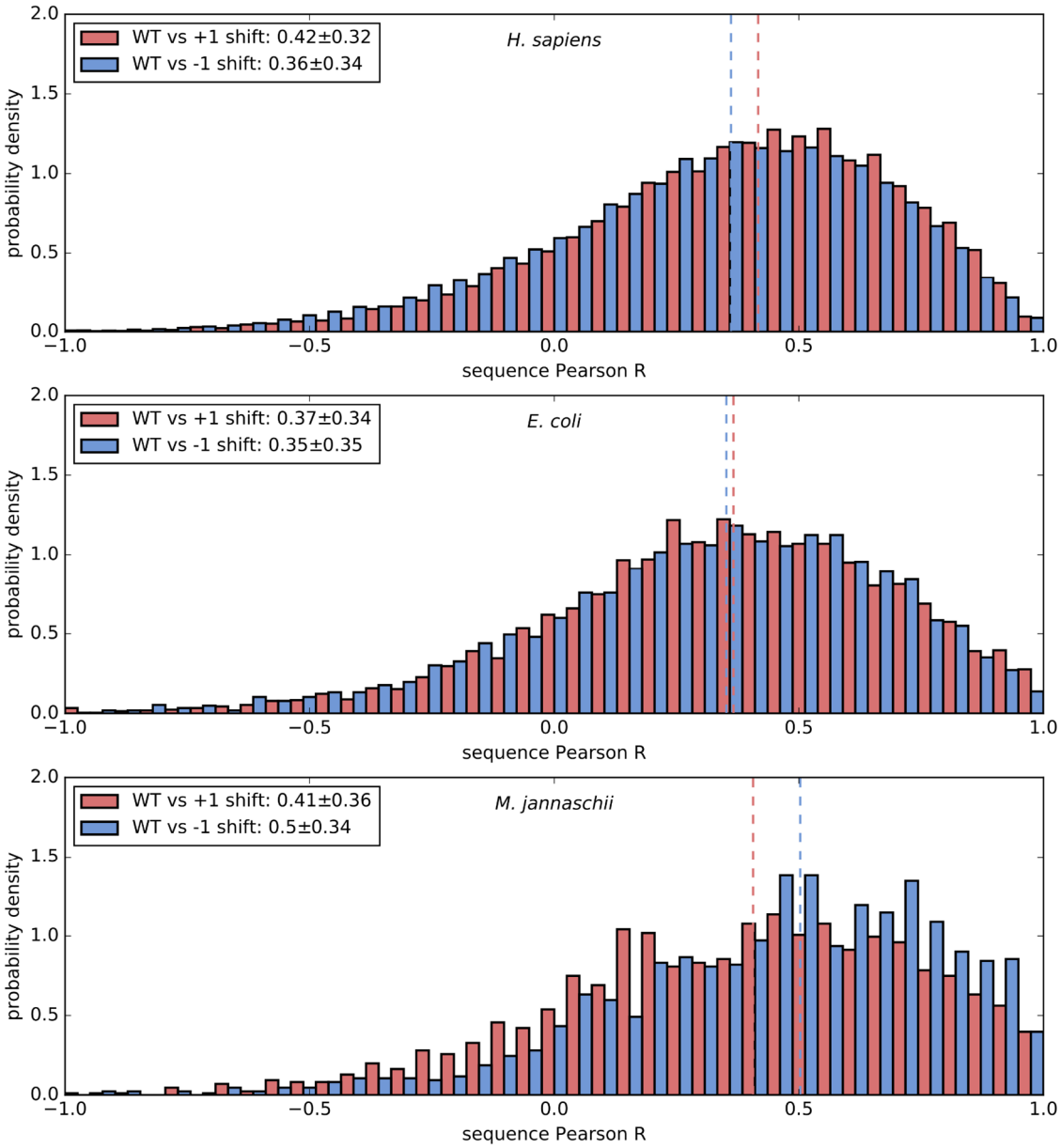


Figure S8. Distributions of Pearson R correlation coefficients between intrinsic disorder profiles of wildtype (WT) and frameshifted proteins over complete proteomes of *H. sapiens* (N = 17083), *E. coli* (N = 3944), and *M. jannaschii* (N = 1666). For each distribution, median and standard deviation are shown in the legend, and the median is indicated as dashed line.

Table S1. Comparison of Pearson Rs (for Factor 1 wildtype vs. frameshifted protein profiles) between orthologous proteins in *H. sapiens* and multiple species. The Pearson R correlation coefficient between two organisms is shown for the +1 shifts and -1 shifts respectively. Poor correlations are observed with the distantly related species *M. jannaschii* (N = 260) and *E. coli* (N = 485) while the closer relative *M. musculus* shows strong correlations with *H. sapiens* (N = 12174).

		M. jannaschii	***E. coli***	***M. musculus***
H. sapiens	***+1 shift***	-0.01	0.10	0.74
	-1 shift	-0.06	0.06	0.71

Dataset S1 (separate file). Summary table for 604 different amino acid property scales, including scale categories as defined in this study, Pearson R correlation coefficients and p-values for a frameshift of UGC, as well as medians of Pearson R correlation coefficients over complete proteomes of *H. sapiens* (N = 17083), *E. coli* (N = 3944), and *M. jannaschii* (N = 1666) for either +1 frameshifted or -1 frameshifted sequences.

Dataset S2 (separate file). Summary table with descriptions and literature references for all amino acid property scales listed in Table S1.

Dataset S3 (separate file). Summary table of the output of the gene ontology analysis for enrichment in the top quartile of the distribution of Pearson R correlation coefficients for Factor 1 WT/+1 frameshifts and WT/-1 frameshifts in *H. sapiens*.

Dataset S4 (separate file). Observed correlations after +1 and -1 frameshifts, for Factor 1, Kidera et.al., Levitt hydrophobic parameter and GUA-affinity scale, as well as disorder and observed sequence identity for all proteins in the *H. sapiens* proteome.

SI References

1. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46(5):2699–2699.

2. Harrison PW, et al. (2019) The European Nucleotide Archive in 2018. *Nucleic Acids Res* 47(D1):D84–D88.

3. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* 28(1):374.

4. Kawashima S, et al. (2007) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36(Suppl 1):D202–D205.

5. Atchley WR, Zhao J, Fernandes AD, Drüke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A* 102(18):6395–400.

6. Hajnic M, Osorio JI, Zagrovic B (2014) Computational analysis of amino acids and their sidechain analogs in crowded solutions of RNA nucleobases with implications for the mRNA-protein complementarity hypothesis. *Nucleic Acids Res* 42(21):12984–12994.

7. Hajnic M, Osorio JI, Zagrovic B (2015) Interaction preferences between nucleobase mimetics and amino acids in aqueous solutions. *Phys Chem Chem Phys* 17(33):21414–21422.

8. Hajnic M, Ruiter A de, Polyansky AA, Zagrovic B (2016) Inosine Nucleobase Acts as Guanine in Interactions with Protein Side Chains. *J Am Chem Soc* 138(17):5519–5522.

9. Andrews CT, Campbell BA, Elcock AH (2017) Direct Comparison of Amino Acid and Salt Interactions with Double-Stranded and Single-Stranded DNA from Explicit-Solvent Molecular Dynamics Simulations. *J Chem Theory Comput* 13(4):1794–1811.

10. Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *"Protein Eng Des Sel* 9(1):27–36.

11. Bartonek L, Zagrovic B (2017) mRNA/protein sequence complementarity and its determinants: The impact of affinity scales. *PLOS Comput Biol* 13(7):e1005648.

12. Pedregosa F, et al. (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12:2825–2830.

13. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10(1):48.

14. Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* 6(7):e21800.

15. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47(D1):D419–D426.

16. Sonnhammer ELL, Östlund G (2015) InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43(D1):D234–D239.

17. Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content

estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347(4):827–39.

Part III

DISCUSSION

6 DISCUSSION

A common thread connecting all the work in this thesis concerns a way of approaching biological sequences as physicochemical entities in a highly simplified manner. Importantly, none of the main results in the thesis could have been achieved by using more complicated all-atom approaches such as molecular dynamics or quantum mechanical simulations. What is more, the questions asked would also have been impossible to investigate using standard sequence-based bioinformatics methods. Here, we provide a panoptic discussion of the principal results of the thesis in order to allow the reader to get a deeper understanding of its main implications. Some new ideas are presented which, although of a potentially far-reaching importance, remain to be explored.

6.1 EXPLORATIONS IN THE SCALE SPACE

Method to explore the scale space efficiently while gaining insight about the properties of the space itself The chief aspect that separates physicochemical sequence profile treatment from classical bioinformatics is the incorporation of property scales in the analysis. A prerequisite for this, however, is a detailed understanding of the space in which such scales reside. In "mRNA/protein sequence complementarity and its determinants: The impact of affinity scales"[34], a generalizable approach was introduced to sample such a space by using a specific fitness function. In particular, we focused on understanding the impact of nucleobase/amino-acid affinity scales in the context of cognate mRNA/protein complementarity hypothesis.

Before this work, several publications have been published in the host group in which nucleobase/amino-acid affinity scales were derived on the basis of different assumptions and methods [33, 40–42]. While the results obtained in these studies qualitatively agreed with the original complementarity hypothesis as stated in [32], the similarity between individual scales was not overly convincing. The question arose whether the observed variance is simply a result of the space these scales reside in or rather suggests that high levels of profile matching do not necessarily provide definitive support for the hypothesis and can be obtained by using very different scales.

A key result of the study was to show that the largest part of the explainable variance in the matching between mRNA nucleobase-density profiles and their cognate proteins' nucleobase-affinity profiles is due to placing several key amino acids at the extreme posi-

tions in the nucleobase/amino-acid affinity scales employed. In other words, not all relative positions in a given affinity scale contribute equally to the observed matching. Rather, some amino acids contribute significantly, and other little or not at all. Importantly, the methodology used in the study can be applied to any situation where the impact of a given scale on certain quantitative descriptors of sequence profiles, such as the degree mRNA/protein profile matching as studied here, needs to be analyzed. It is important, on the other hand, to remember that while such an analysis gives quantifiable answers as to the contributions of individual scale elements to a given value of the fitness function, assigning significance to this value is a different matter, which should be considered by the researcher case by case. In the original publications concerning cognate mRNA/protein matching, results close to a Pearson R value of 0.7 were reported, suggesting 50% of the variance in the mRNA nucleobase-density profiles could be explained by their cognate proteins' nucleobase-affinity profiles. Whether this result constitutes relevant support of the stated hypothesis or not has to be assessed on a case-by-case basis by the researcher.

Finally, the applied Monte Carlo method of sampling the space of scales is general and the associated computational effort independent of any specific scale. Moreover, deriving scales in this way could be computationally cheaper as compared with existing physical scales. The method could, therefore, also be used to define scales, which affect a given fitness function in a desired manner. For example, one could in this way derive an amino-acid scale, which would result in protein profiles that optimally match their cognate mRNA secondary structure profiles. Kevin Weeks and co-workers have proposed that, indeed, mRNA secondary structure might encode certain structural features of their cognate proteins such as, for example, their degree of intrinsic disorder ("speed-bump hypothesis")[77]. By analyzing amino-acid scales, which are computationally optimized as discussed above, one might directly explore the limits of what can or cannot be encoded in the mRNA secondary structure profiles.

6.2 MAKING PHYSICOCHEMICAL BIOINFORMATICS ACCESSIBLE TO EVERYONE

While our exploration of the space of amino-acid property scales did delineate one way of testing relevant hypotheses within a simple framework, at that point there existed no suitable, easy-to-use tool to evaluate and compare physicochemical property sequence profiles in a more general way. This challenge was addressed by our release of VOLPES, a visualization and analysis server, which allows biomolecular sequence information to be represented in the form of physicochemical property profiles to be analyzed and compared in an in-

teractive way. In the original VOLPES publication, however, it was not possible to fully compare and contrast the tool with other available solutions due to space constraints. Here, we first provide such a comparison, followed by a discussion of the wider set of potential applications of the JavaScript library Volpini, associated with VOLPES.

6.2.1 *Comparison to other tools*

The interactive analysis and visualization server VOLPES is neither the first nor the only tool allowing users to visualize physicochemical profiles of biomolecular sequences, but to our knowledge it is the only tool that allows the exploration of parameters, required for visualization and analysis, in an interactive way without installing any software besides a reasonably modern web browser. While there are many tools available that overlap with VOLPES in some respects [29, 48, 78], here we discuss its most popular competitor, ExPASy ProtScale[28], in order to demonstrate in detail how VOLPES elevates the idea of physicochemical bioinformatics and goes from a simple visualization aid to a tool that can be used for visual analytics. The discussion here in no way intends to criticize or marginalize the work done by competing developers. Rather, it is our aim to highlight the fact that online scientific tools have a general tendency to be based on older technology, while an adoption of more recent technologies potentially allows for a much more efficient and engaging experience for the user.

In Figure 1, we show the results of visualizing the hydrophobicity profile based on the Kyte&Doolittle hydrophobicity scales in the case of a chloroplastic protein plastocyanin from *Pisum sativum* (UniProtKB accession number P16002) using ExPASy ProtScale. After inputting the sequence and choosing the scale from one of 57 alternatives provided on the starting page, the visualization is provided as a static image with some export options including a rasterized GIF image, a PostScript file and raw data in form of a text file. Most notably, no option for modifying the visualization parameters (e.g. size of averaging window, sequence shift, axis scaling etc.) is available, besides restarting the whole procedure. While the selected parameters are clearly listed in a very prominent location, the user does not gain any feeling for the parameters that were just selected due to the lack of options to modify them. Furthermore, only a single sequence can be visualized at once, prohibiting the user to look for similarities in two different sequences using the same property or comparing multiple properties of one and the same sequence.

In comparison, the VOLPES server is not restricted by the slow communication and calculation time resulting from a classical back-end/front-end model for generating the results by outsourcing most of the calculation to the client's computer by design. Having most

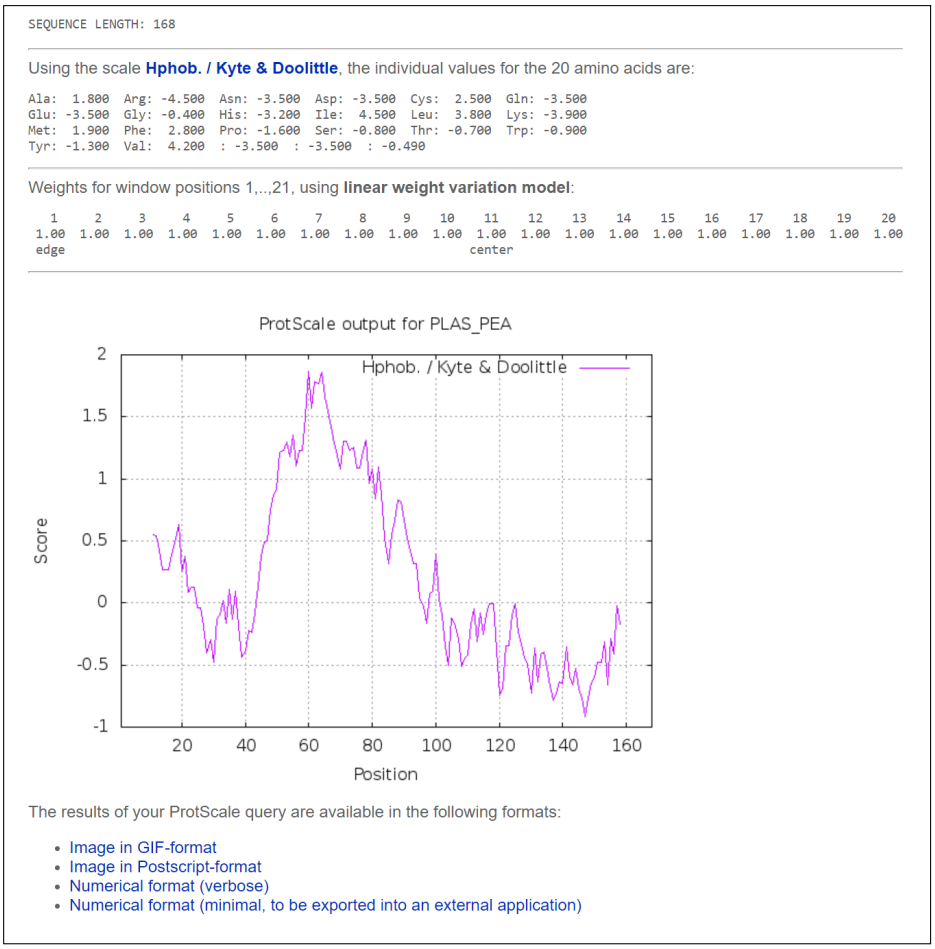


Figure 1: *Result of the visualization process in ExPASy ProtScale* After providing all the necessary details in multiple steps, a simple visualization is created as a GIF formatted rasterized image. Further export options include PostScript, and raw data in a text file. Note, however, that none of the visualization parameters can be changed in this state. Any change to input parameters requires a complete restart of the visualization process. Finally, using the ProtScale server, one can analyze only a single sequence at a time.

of the performance-critical part written in JavaScript to be executed by the client's browser, all the parameters can be changed in real-time and the user is able to explore directly the influence of different input parameters on the final result.

	1	2	3
1	1	-0.98	0.91
2	-0.98	1	-0.86
3	0.87	-0.82	1

Figure 2: *Example of the VOLPES interface* The same protein sequence as in the previous examples was used (Plastocyanin from *Pisum sativum*), but since VOLPES supports parallel analysis of multiple profiles, three different protein hydropathy measures were used at the same time -the Kyte & Doolittle scale, as used in the previous examples, the Eisenberg consensus hydrophobicity scale and Factor 1 from Atchley and coworkers[44]. To make comparisons between different properties easier, VOLPES allows one to rescale profiles by their standard deviation as used here. Details about a profile can be seen by simply hovering the mouse over a profile, while correlation coefficients, capturing the similarity between different profiles, can be accessed in the table on the bottom left. A new box in the interface has been created for each profile. Importantly, all parameters, including the sequence, the scale used, the smoothing window and the shift relative to other sequences, can be changed interactively without having to exit the current state of analysis.

Another design limitation of most competing tools that has been removed in VOLPES is the ability to compare multiple profiles at the same time. By including several broadly used similarity measures, an arbitrary number of profiles can be compared (we have tested up to 500, although a reasonable maximum is about 10). This could entail a comparison of different sequences, but also a comparison of different properties for one and the same sequence. Furthermore, VOLPES

supports gaps in sequences, thereby allowing for multiple sequence alignments to be submitted, a feature rarely seen in comparable tools.

The interface of VOLPES is shown in Figure 2 using again plastocyanin as example. While most tools, such as ExPASy ProtScal, are limited to a single profile at a time, VOLPES is designed around the idea of comparing multiple profiles with each other. While before only a single property could be shown, here three different definitions of hydrophobicity are employed and compared. Specifically, we again use the Kyte & Doolittle scale, but this time it is accompanied by the Eisenberg consensus hydrophobicity scale and the Factor 1 scale as derived by Atchley and coworkers. Numerical scores of similarity (Pearson and Spearman Rs) are shown in the bottom right corner, with an option to change the scoring functions anytime to any other supported one. Each profile has its own interface element in the shape of a minimizable box, containing the most popular parameters to be changed interactively and a menu in which every single visualization parameter can be changed. The size of the smoothing window and the shift along the x-axis can be adjusted interactively without having to open a single window. The name and color of the profile can be changed by interacting with the respective elements directly. In the individual menus, the sequence, the scale, the polymer type as well as multiple visualization parameters such as linewidth and linetype can be changed. Unlike in ExPASy ProtScale, all these actions can be performed without ever having to reload the website or leaving the current state of analysis. To support interactions with the visualization window, panning and zooming are supported. Finally, detailed information on the profiles is given when hovering a mouse over the profile. Importantly, we are strongly convinced that the technical advantages of VOLPES, as described above, not only improve the users' experience, but rather result in a qualitatively different approach to sequence information, which in turn could result in a detection of previously inaccessible, yet biologically meaningful patterns. For example, simply changing the size of the averaging window in an interactive way, one immediately develops an intuitive feeling for sequence organization on multiple length-scales. Moreover, by directly comparing different sequence profiles in a quantitative manner, one rapidly and in a parallel fashion gleans key information not only about sequence properties in question, but also about the impact of the individual scales used. It is our hope that future users of VOLPES will share this conviction and that our server will help make new discoveries concerning the basic organization and properties of biological sequences.

6.2.2 *Interactive publications*

There have been many ideas discussed on how to make publications - in the classical sense, in printed papers, but also online-only media publications - interactive and fit for the 21st century[79–81]. Most printed journals in the area of natural sciences offer a highly styled and formatted online versions of publications, while at the same time only allowing for high resolution, rasterized images fit for print. Some journals allow for movies to be submitted as part of the supplementary material in their online versions, but that is in most cases as far as attempts towards interactivity go.

With the release of VOLPES, the associated JavaScript library Volpini has been released as an open-source library to GitHub for everyone to use in their individual projects, which could go beyond classical research publications (see github.com/BarLuk/volpini). Some examples in this direction, ranging from proof-of-concept to actual finished implementations, are available as part of the repository and/or are highlighted on the VOLPES server. Most importantly, the library has so far been used to produce interactive presentation slides as well as interactive online versions of conference posters, with the latter - in terms of technical implementation - being identical to an interactive version of an online research publication. Once JavaScript code is acceptable by journals as part of research articles, the library can be immediately utilized as is. Until this is realized, a link to a different host needs to be provided as part of the submission. In general, we are confident that the Volpini library will be recognized and used in a wide range of applications, going significantly beyond its original use for understanding the properties of biological sequences in a basic research context.

6.3 IMPACT OF FRAMESHIFT INVARIANCE

Finally, the physicochemical view of biomolecular sequences resulted in a novel finding regarding the invariance of certain types of protein sequence profiles with respect to frameshifts in the respective mRNA coding sequences. Most importantly, we could show that hydrophobicity profiles of frameshifted variants are highly similar to the original profiles for a significant number of sequences. While details of the analysis are presented in the accompanying preprint (please, see above), here we discuss in more detail a particular implication of this finding. Namely, if the idea that frameshifts represent a viable evolutionary strategy to generate novel protein sequences with invariant physicochemical properties indeed holds, this could carry significant implications when it comes protein sequence alignment. In particular, sequences that on a first sight may be extremely different, could be related to each other by a simple frameshift in the

underlying mRNA and, therefore, be much closer in terms of evolutionary distance than implied by classical alignment. Importantly, when it comes to alignments of protein sequences with medium to high similarity, frameshift-based analysis is not going to contradict classical results, but may result in more sequences being alignable in general, especially in situations of high sequence divergence. As a case in point, sequence identity of the original wildtype and the +1 frameshifted protein sequences in human is only 6% on average [76]. Of course, in an idealized case where the complete mRNA is frameshifted, the relationship between two such sequences would be immediately apparent if one aligned the associated genes. However, in the case of multiple indel mutations, resulting in local frameshifts, together with classical substitutions, such an alignment may not always be straightforward. Moreover, protein alignment is typically preferred over DNA/RNA alignment due to higher signal-to-noise ratio. There already exist services [82–84] that enable alignment of back-translated sequences, but there are currently no programs or even algorithms that would allow fast, high-throughput analysis of distant proteins related by frameshifts. Moreover, there have been publications analyzing the evolutionary relationship between frameshifted protein variants in the case of snake venoms [85, 86], but it has to be noted that this class of proteins is known to be highly mutable. Development of an efficient method to search for pairs of protein sequences related by frameshifts in multiple genomes would, therefore, be key for understanding the relevance of frameshifts in evolutionary processes in general, but could also help identify specific, biologically relevant cases in particular.

Two different approaches seem to be viable and should be explored further in this regard. Firstly, the two protein sequences to be compared could each be back-translated to a set of potential mRNA sequences coding for them. These sets could be compared and the closest common coding sequence found, as already implemented in the webserver PATH by Gîrdea et.al. [82]. The closest common coding sequence could then be analyzed to look for traces of frameshifts that could have occurred in the past. The problem, however, is that the implementation of PATH is far too slow to be used on a genomic scale. Specifically, a single analysis on current server infrastructure and using the current implementation can take on the order of 10 minutes to complete, which is much too slow to explore all sequence pairs in a typical proteome. Parallelization and more advanced hardware could help, but ultimately an improved algorithm needs to be designed. Secondly, one could follow up on the idea of using physicochemical profiles even further and design a profile alignment algorithm that allows for an adequate treatment of gaps. In this case, one would align physicochemical profiles and not sequences, enabling exploration of evolutionary relationships in a completely novel way. This

seems to be far less complicated than the first option since a significant amount of research on multiple sequence alignments and alignment algorithms has been done in the past, making very fast analyses possible[58, 59, 87, 88]. The key change that needs to be considered involves replacing the scoring matrices used in standard alignment with equivalent matrices that describe the relationship of amino acids with each other in terms of physicochemical properties of interest. Furthermore, the importance of smoothing windows in profiles needs to be investigated further. If the smoothing window should be considered as a part of the model, a more complicated scoring matrix needs to be devised that would include these effects. Once that has been implemented, all the classical alignment algorithms could be used and analysis of full proteomes could be performed very quickly.

Having access to such a method would allow for a high-throughput analysis of proteomes to find proteins, which are related by a frameshift, in a single proteome or in different proteomes. This could answer, on the one hand, how frequently the frameshift-based *de novo* gene generation occurs in real biological systems, but also help determine the evolutionary distance between organisms more precisely. Finally, such an alignment strategy could yield relevant results going beyond the question of frameshifts i.e. it could produce sequence alignments based on the similarity in physicochemical property profiles rather than just plain sequence. In this way, one would not necessarily capture evolutionary distance, but rather the similarity in the general physicochemical properties between two sequences, which could, but does not have to be a result of close evolutionary distance. Future research should shed light on these exciting possibilities.

6.3.1 *Frameshifts in a primordial context*

As detailed above, our analysis of frameshifts gives support to a novel evolutionary mechanism for increasing protein sequence diversity. However, another conclusion could also be drawn from our results, especially if one considers a primordial, origin-of-life setting in which the impact of frameshift errors may have been very different. In such a setting, stable translation and transcription were likely not possible. Note that here we remain agnostic regarding the RNA world debate [89–91] and refer to transcription as any kind of primordial replication of genetic information, regardless of whether it was encoded in DNA, RNA or some other biopolymer. In any case, it can be assumed that error rates were much higher in primordial than in modern systems, which evolved a wide range of compensatory control mechanisms. Importantly, in such primordial systems - regardless of the specifics of how they came to be - a wide range of protein variants, corresponding to one and the same piece of genetic information, were likely to be present. Such a state, described by Woese

and coworkers as a *statistical protein*[92–94], analogous to the quasispecies model in RNA/DNA sequences [95–98], has the potential to exhibit a high diversity of sequences at any given time. This diversity may be the result of translation errors, transcription errors, frameshift errors or most likely a combination of all three. The question arises how such an ensemble of related, but different protein products, all encoded by the same gene, could properly function in such an environment. Therefore, in addition to the properties of frameshifted variants of protein products, we also modelled the impact of transcription and translation errors on the nature of the final proteins produced. In particular, we specifically investigated the correlation between the nucleobase affinity patterns on the side of proteins and the nucleobase density patterns of the respective mRNAs, as performed previously [32–34, 75, 76], for different levels of translational and transcriptional error rates. The underlying assumption is that a direct binding between cognate mRNA/protein pairs was the most important function of proteins in such a scenario, an idea discussed in different contexts by many authors [99–104]. If such binding were retained to a similar degree for most members of the statistical protein ensemble, this could explain why such a diverse system could still perform well enough to survive and evolve, despite high error rates involved.

In our analysis, error rates for both translation and transcription errors were defined as $E = \{r \in \mathbb{R} \mid 0 \leqslant r \leqslant 0.75\}$ in order to correspond to experimental error rates. Namely, rates greater than 0.75 would result in a reduction of randomness, leading to a bias towards inverted sequences. The space of possible combinations of error rates was sampled on a semi-regular grid, with every grid point representing the average result obtained for 100 copies of the human proteome each with stochastic errors introduced.

To our surprise, even excessively high error rates of up to 0.25 result, on average, in very high, significant correlations between the nucleobase affinity profiles of proteins and the nucleobase-density profiles of the their cognate mRNAs (3). It also becomes apparent that at lower, and arguably more likely, error rates virtually no change is observed in comparison to the native sequences. When exploring the potential space of error rates in the very high range of up to 0.75, a stronger effect on the correlation can be observed in case of translation errors as opposed to transcription errors (3).

These observations open up a range of possible directions to be researched further. In particular, an immediate, important question is how this result should be interpreted in the framework of the *complementartity hypothesis*. Can the previous results regarding modern day sequences be interpreted as a reflection of a once essential property, which allowed more advanced biological organisms to evolve out of the primordial context? Should our findings regarding frameshift invariance be interpreted as an essential property that was kept intact

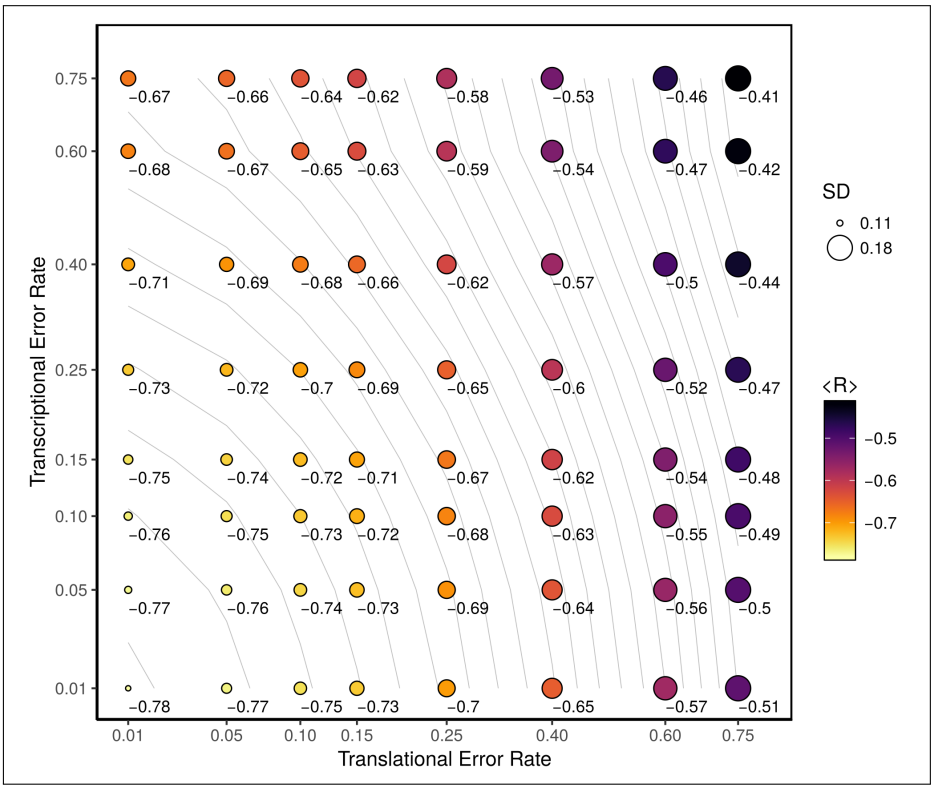


Figure 3: Even excessively high error rates of up to 0.25 result in very high correlations between nucleobase affinity profiles and nucleobase-density profiles of their cognate mRNAs. In this example the affinity for the purine base Guanine [33] is compared with the mRNA purine-density.

throughout evolution? Most likely the true mechanisms at play were a combination of the aforementioned possibilities, which can never be properly decoupled and analyzed. Nevertheless, it is our firm belief that such obstacles should not keep future research from attempting to do exactly that.

BIBLIOGRAPHY

[1] Douglas R Hartree. "The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods." In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 24. 1. Cambridge University Press. 1928, pp. 89–110.

[2] Vladimir Fock. "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems." In: *Zeitschrift für Physik* 61.1-2 (1930), pp. 126–148.

[3] Pierre Hohenberg and Walter Kohn. "Inhomogeneous electron gas." In: *Physical review* 136.3B (1964), B864.

[4] Walter Kohn and Lu Jeu Sham. "Self-consistent equations including exchange and correlation effects." In: *Physical review* 140.4A (1965), A1133.

[5] Miguel AL Marques, Xabier López, Daniele Varsano, Alberto Castro, and Angel Rubio. "Time-dependent density-functional approach for biological chromophores: the case of the green fluorescent protein." In: *Physical review letters* 90.25 (2003), p. 258101.

[6] Shina CL Kamerlin, Maciej Haranczyk, and Arieh Warshel. "Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of p K a, redox reactions and solvation free energies." In: *The Journal of Physical Chemistry B* 113.5 (2008), pp. 1253–1272.

[7] Hans Martin Senn and Walter Thiel. "QM/MM methods for biomolecular systems." In: *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229.

[8] Walter Thiel. "QM/MM methodology: fundamentals, scope, and limitations." In: *Multiscale simulation methods in molecular sciences* 42 (2009), pp. 203–214.

[9] Heather J Kulik, Jianyu Zhang, Judith P Klinman, and Todd J Martinez. "How large should the QM region be in QM/MM calculations? The case of catechol O-methyltransferase." In: *The Journal of Physical Chemistry B* 120.44 (2016), pp. 11381–11394.

[10] Michael Levitt and Arieh Warshel. "Computer simulation of protein folding." In: *Nature* 253.5494 (1975), p. 694.

[11] JA McCammon and M Karplus. "Simulation of protein dynamics." In: *Annual Review of Physical Chemistry* 31.1 (1980), pp. 29–45.

[12] Raviprasad Aduri, Brian T Psciuk, Pirro Saro, Hariprakash Taniga, H Bernhard Schlegel, and John SantaLucia. "AMBER force field parameters for the naturally occurring modified nucleosides in RNA." In: *Journal of Chemical Theory and Computation* 3.4 (2007), pp. 1464–1475.

[13] Ron O Dror, Morten Ø Jensen, David W Borhani, and David E Shaw. "Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations." In: *The Journal of general physiology* 135.6 (2010), pp. 555–562.

[14] Isseki Yu, Takaharu Mori, Tadashi Ando, Ryuhei Harada, Jaewoon Jung, Yuji Sugita, and Michael Feig. "Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm." In: *Elife* 5 (2016), e19274.

[15] Richard A Friesner and Victor Guallar. "Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis." In: *Annu. Rev. Phys. Chem.* 56 (2005), pp. 389–427.

[16] W Patrick Walters, Matthew T Stahl, and Mark A Murcko. "Virtual screening—an overview." In: *Drug discovery today* 3.4 (1998), pp. 160–178.

[17] Peter Willett. "Similarity-based virtual screening using 2D fingerprints." In: *Drug discovery today* 11.23-24 (2006), pp. 1046–1053.

[18] Sean Ekins, Jordi Mestres, and Bernard Testa. "In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling." In: *British journal of pharmacology* 152.1 (2007), pp. 9–20.

[19] Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. "Software for molecular docking: a review." In: *Biophysical reviews* 9.2 (2017), pp. 91–102.

[20] Paweł Śledź and Amedeo Caflisch. "Protein structure-based drug design: from docking to molecular dynamics." In: *Current opinion in structural biology* 48 (2018), pp. 93–102.

[21] Wouter Steyaert, Steven Callens, Paul Coucke, Bart Dermaut, Dimitri Hemelsoet, Wim Terryn, and Bruce Poppe. "Future perspectives of genome-scale sequencing." In: *Acta Clinica Belgica* 73.1 (2018), pp. 7–10.

[22] Hidewaki Nakagawa and Masashi Fujita. "Whole genome sequencing analysis for cancer genomics and precision medicine." In: *Cancer science* 109.3 (2018), pp. 513–522.

[23] Katharina Schwarze, James Buchanan, Jilles M Fermont, Helene Dreau, Mark W Tilley, John M Taylor, Pavlos Antoniou, Samantha JL Knight, Carme Camps, Melissa M Pentony, et al. "The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom." In: *Genetics in Medicine* (2019), pp. 1–10.

[24] Jack Kyte and Russell F Doolittle. "A simple method for displaying the hydropathic character of a protein." In: *Journal of molecular biology* 157.1 (1982), pp. 105–132.

[25] Mauro Degli Esposti, Massimo Crimi, and Giovanni Venturoli. "A critical evaluation of the hydropathy profile of membrane proteins." In: *European journal of biochemistry* 190.1 (1990), pp. 207–219.

[26] Peter Rice, Ian Longden, and Alan Bleasby. *EMBOSS: the European molecular biology open software suite*. 2000.

[27] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." In: *Bioinformatics* 21.16 (2005), pp. 3433–3434.

[28] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Marc R Wilkins, Ron D Appel, Amos Bairoch, et al. "Protein identification and analysis tools on the ExPASy server." In: *The proteomics protocols handbook*. Springer, 2005, pp. 571–607.

[29] Craig Snider, Sajith Jayasinghe, Kalina Hristova, and Stephen H White. "MPEx: a tool for exploring membrane proteins." In: *Protein Science* 18.12 (2009), pp. 2624–2628.

[30] Matteo Bellucci, Federico Agostini, Marianela Masin, and Gian Gaetano Tartaglia. "Predicting protein associations with long noncoding RNAs." In: *Nature methods* 8.6 (2011), p. 444.

[31] Lukas Bartonek and Bojan Zagrovic. "VOLPES: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences." In: *Nucleic acids research* 47.W1 (2019), W632–W635.

[32] Mario Hlevnjak, Anton A Polyansky, and Bojan Zagrovic. "Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels." In: *Nucleic acids research* 40.18 (2012), pp. 8874–8882.

[33] Anton A Polyansky and Bojan Zagrovic. "Evidence of direct complementary interactions between messenger RNAs and their cognate proteins." In: *Nucleic acids research* 41.18 (2013), pp. 8434–8443.

[34] Lukas Bartonek and Bojan Zagrovic. "mRNA/protein sequence complementarity and its determinants: The impact of affinity scales." In: *PLoS computational biology* 13.7 (2017), e1005648.

[35] Oliviero Carugo and Sándor Pongor. "A normalized root-mean-spuare distance for comparing protein three-dimensional structures." In: *Protein science* 10.7 (2001), pp. 1470–1473.

[36] Xueyi Wang and JA Dong. "A Normalized Weighted RMSD for Measuring Protein Structure Superposition." In: *IPCBEE* 34 (2012), pp. 68–72.

[37] Carl R Woese, DH Dugre, SA Dugre, M Kondo, and WC Saxinger. "On the fundamental nature and evolution of the genetic code." In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 31. Cold Spring Harbor Laboratory Press. 1966, pp. 723–736.

[38] CR Woese, DH Dugre, WC Saxinger, and SA Dugre. "The molecular basis for the genetic code." In: *Proceedings of the National Academy of Sciences of the United States of America* 55.4 (1966), p. 966.

[39] Carl R Woese. "Evolution of the genetic code." In: *Naturwissenschaften* 60.10 (1973), pp. 447–459.

[40] Matea Hajnic, Juan Iregui Osorio, and Bojan Zagrovic. "Computational analysis of amino acids and their sidechain analogs in crowded solutions of RNA nucleobases with implications for the mRNA–protein complementarity hypothesis." In: *Nucleic acids research* 42.21 (2014), pp. 12984–12994.

[41] Matea Hajnic, Juan I Osorio, and Bojan Zagrovic. "Interaction preferences between nucleobase mimetics and amino acids in aqueous solutions." In: *Physical Chemistry Chemical Physics* 17.33 (2015), pp. 21414–21422.

[42] Anita de Ruiter and Bojan Zagrovic. "Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments." In: *Nucleic acids research* 43.2 (2014), pp. 708–718.

[43] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A Scheraga. "Statistical analysis of the physical properties of the 20 naturally occurring amino acids." In: *Journal of Protein Chemistry* 4.1 (1985), pp. 23–55.

[44] William R Atchley, Jieping Zhao, Andrew D Fernandes, and Tanja Drüke. "Solving the protein sequence metric problem." In: *Proceedings of the National Academy of Sciences* 102.18 (2005), pp. 6395–6400.

[45] Michael Levitt. "A simplified representation of protein conformations for rapid simulation of protein folding." In: *Journal of molecular biology* 104.1 (1976), pp. 59–107.

[46] David Eisenberg. "Three-dimensional structure of membrane and surface proteins." In: *Annual review of biochemistry* 53.1 (1984), pp. 595–623.

[47] DM Engelman, TA Steitz, and A Goldman. "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins." In: *Annual review of biophysics and biophysical chemistry* 15.1 (1986), pp. 321–353.

[48] Charles M Deber, Chen Wang, Li-Ping Liu, Andrew S Prior, Shuchi Agrawal, Brenda L Muskat, and A Jamie Cuticchia. "TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales." In: *Protein Science* 10.1 (2001), pp. 212–219.

[49] Jonathan M Cuthbertson, Declan A Doyle, and Mark SP Sansom. "Transmembrane helix prediction: a comparative evaluation and analysis." In: *Protein Engineering Design and Selection* 18.6 (2005), pp. 295–308.

[50] Gang Zhao and Erwin London. "An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity." In: *Protein science* 15.8 (2006), pp. 1987–2001.

[51] Rahul K Das and Rohit V Pappu. "Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues." In: *Proceedings of the National Academy of Sciences* 110.33 (2013), pp. 13392–13397.

[52] Rahul K Das, Yongqi Huang, Aaron H Phillips, Richard W Kriwacki, and Rohit V Pappu. "Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling." In: *Proceedings of the National Academy of Sciences* 113.20 (2016), pp. 5616–5621.

[53] T Lesnik and C Reiss. "Detection of transmembrane helical segments at the nucleotide level in ukarayotic membrane protein genes." In: *IUBMB Life* 44.3 (1998), pp. 471–479.

[54] Jaime Prilusky and Eitan Bibi. "Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs." In: *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 6662–6666.

[55] Timothy J Nott, Evangelia Petsalaki, Patrick Farber, Dylan Jervis, Eden Fussner, Anne Plochowietz, Timothy D Craggs, David P Bazett-Jones, Tony Pawson, Julie D Forman-Kay, et al. "Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles." In: *Molecular cell* 57.5 (2015), pp. 936–947.

[56] Jie Wang, Jeong-Mo Choi, Alex S Holehouse, Hyun O Lee, Xiaojie Zhang, Marcus Jahnel, Shovamayee Maharana, Régis Lemaitre, Andrei Pozniakovsky, David Drechsel, et al. "A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins." In: *Cell* 174.3 (2018), pp. 688–699.

[57] Federico Agostini, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. "cat RAPID omics: a web server for large-scale prediction of protein–RNA interactions." In: *Bioinformatics* 29.22 (2013), pp. 2928–2930.

[58] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[59] Temple F Smith, Michael S Waterman, et al. "Identification of common molecular subsequences." In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[60] Juke S Lolkema and Dirk-Jan Slotboom. "Estimation of structural similarity of membrane proteins by hydropathy profile alignment." In: *Molecular membrane biology* 15.1 (1998), pp. 33–42.

[61] Igor B Kuznetsov. "Protein sequence alignment with family-specific amino acid similarity matrices." In: *BMC Research Notes* 4.1 (2011), p. 296.

[62] Marcus Stamm, René Staritzbichler, Kamil Khafizov, and Lucy R Forrest. "AlignMe—a membrane protein sequence alignment web server." In: *Nucleic acids research* 42.W1 (2014), W246–W251.

[63] Igor B Kuznetsov and Michael McDuffie. "PR2ALIGN: a standalone software program and a web-server for protein sequence alignment using weighted biochemical properties of amino acids." In: *BMC research notes* 8.1 (2015), p. 187.

[64] Kamil Khafizov, René Staritzbichler, Marcus Stamm, and Lucy R Forrest. "A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe." In: *Biochemistry* 49.50 (2010), pp. 10702–10713.

[65] Marcus Stamm, René Staritzbichler, Kamil Khafizov, and Lucy R Forrest. "Alignment of helical membrane protein sequences using AlignMe." In: *PloS one* 8.3 (2013), e57731.

[66] Abraham Savitzky and Marcel JE Golay. "Smoothing and differentiation of data by simplified least squares procedures." In: *Analytical chemistry* 36.8 (1964), pp. 1627–1639.

[67] Eörs Szathmáry. "The origin of the genetic code: amino acids as cofactors in an RNA world." In: *Trends in genetics* 15.6 (1999), pp. 223–229.

[68] Michael Yarus, Jeremy Joseph Widmann, and Rob Knight. "RNA–amino acid binding: a stereochemical era for the genetic code." In: *Journal of molecular evolution* 69.5 (2009), p. 406.

[69] Stanley L Miller and Harold C Urey. "Organic compound synthesis on the primitive earth." In: *Science* 130.3370 (1959), pp. 245–251.

[70] Francis HC Crick. "The origin of the genetic code." In: *Journal of molecular biology* 38.3 (1968), pp. 367–379.

[71] Yuri I Wolf and Eugene V Koonin. "On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization." In: *Biology Direct* 2.1 (2007), p. 14.

[72] Eors Szathmary. "Coding coenzyme handles: a hypothesis for the origin of the genetic code." In: *Proceedings of the National Academy of Sciences* 90.21 (1993), pp. 9916–9920.

[73] Michael Yarus. "The meaning of a minuscule ribozyme." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1580 (2011), pp. 2902–2909.

[74] Linus Pauling and Max Delbrück. "The nature of the intermolecular forces operative in biological processes." In: *Science* 92.2378 (1940), pp. 77–79.

[75] Bojan Zagrovic, Lukas Bartonek, and Anton A Polyansky. "RNA-protein interactions in an unstructured context." In: *FEBS letters* 592.17 (2018), pp. 2901–2916.

[76] Lukas Bartonek, Daniel Braun, and Bojan Zagrovic. "Invariants of Frameshifted Variants." In: *bioRxiv* (2019), p. 684076.

[77] Anthony M Mustoe, Steven Busan, Greggory M Rice, Christine E Hajdin, Brant K Peterson, Vera M Ruda, Neil Kubica, Razvan Nutiu, Jeremy L Baryza, and Kevin M Weeks. "Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing." In: *Cell* 173.1 (2018), pp. 181–195.

[78] Andrew M Waterhouse, James B Procter, David MA Martin, Michèle Clamp, and Geoffrey J Barton. "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." In: *Bioinformatics* 25.9 (2009), pp. 1189–1191.

[79] Tracey L Weissgerber, Natasa M Milic, Stacey J Winham, and Vesna D Garovic. "Beyond bar and line graphs: time for a new data presentation paradigm." In: *PLoS biology* 13.4 (2015), e1002128.

[80] Amanda J Fosang and Roger J Colbran. "Transparency is the key to quality." In: *Journal of Biological Chemistry* 290.50 (2015), pp. 29692–29694.

[81] Tracey L Weissgerber, Vesna D Garovic, Marko Savic, Stacey J Winham, and Natasa M Milic. "From static to interactive: transforming data visualization to improve transparency." In: *PLoS biology* 14.6 (2016), e1002484.

[82] Marta Gîrdea, Laurent Noé, and Gregory Kucherov. "Back-translation for discovering distant protein homologies in the presence of frameshift mutations." In: *Algorithms for Molecular Biology* 5.1 (2010), p. 6.

[83] Sergey L Sheetlin, Yonil Park, Martin C Frith, and John L Spouge. "Frameshift alignment: statistics and post-genomic applications." In: *Bioinformatics* 30.24 (2014), pp. 3575–3582.

[84] Safa Jammali, Esaie Kuitche, Ayoub Rachati, François Bélanger, Michelle Scott, and Aïda Ouangraoua. "Aligning coding sequences with frameshift extension penalties." In: *Algorithms for Molecular Biology* 12.1 (2017), p. 10.

[85] Bryan G Fry, Holger Scheib, Louise van der Weerd, Bruce Young, Judith McNaughtan, SF Ryan Ramjan, Nicolas Vidal, Robert E Poelmann, and Janette A Norman. "Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (Caenophidia)." In: *Molecular & Cellular Proteomics* 7.2 (2008), pp. 215–246.

[86] Yasuo Yamazaki, Yukiko Matsunaga, Yuko Tokunaga, Shinya Obayashi, Mai Saito, and Takashi Morita. "Snake venom Vascular Endothelial Growth Factors (VEGF-Fs) exclusively vary their structures and functions among species." In: *Journal of Biological Chemistry* 284.15 (2009), pp. 9885–9891.

[87] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. "Basic local alignment search tool." In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[88] W James Kent. "BLAT—the BLAST-like alignment tool." In: *Genome research* 12.4 (2002), pp. 656–664.

[89] Harold S Bernhardt. "The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others) a." In: *Biology direct* 7.1 (2012), p. 23.

[90] Marc Neveu, Hyo-Joong Kim, and Steven A Benner. "The "strong" RNA world hypothesis: Fifty years old." In: *Astrobiology* 13.4 (2013), pp. 391–403.

[91] Li Li, Christopher Francklyn, and Charles W Carter. "Aminoacylating urzymes challenge the RNA world hypothesis." In: *Journal of Biological Chemistry* 288.37 (2013), pp. 26856–26863.

[92] Carl Woese. "The universal ancestor." In: *Proceedings of the national academy of Sciences* 95.12 (1998), pp. 6854–6859.

[93] Carl R Woese. "On the evolution of the genetic code." In: *Proceedings of the National Academy of Sciences of the United States of America* 54.6 (1965), p. 1546.

[94] Carl R Woese and George E Fox. "The concept of cellular evolution." In: *Journal of molecular evolution* 10.1 (1977), pp. 1–6.

[95] Manfred Eigen and Peter Schuster. "A principle of natural self-organization." In: *Naturwissenschaften* 64.11 (1977), pp. 541–565.

[96] Manfred Eigen and Peter Schuster. "The hypercycle." In: *Naturwissenschaften* 65.1 (1978), pp. 7–41.

[97] Manfred Eigen, John McCaskill, and Peter Schuster. "Molecular quasi-species." In: *The Journal of Physical Chemistry* 92.24 (1988), pp. 6881–6891.

[98] Manfred Eigen. "On the nature of virus quasispecies." In: *Trends in microbiology* 4.6 (1996), pp. 216–218.

[99] John D Sutherland and Jonathan M Blackburn. "Killing two birds with one stone: a chemically plausible scheme for linked nucleic acid replication and coded peptide synthesis." In: *Chemistry & biology* 4.7 (1997), pp. 481–488.

[100] Vinciane Borsenberger, Michael A Crowe, Jörg Lehbauer, Jim Raftery, Madeleine Helliwell, Kajal Bhutia, Tony Cox, and John D Sutherland. "Exploratory studies to investigate a linked prebiotic origin of RNA and coded peptides." In: *Chemistry & biodiversity* 1.2 (2004), pp. 203–246.

[101] Lee B Mullen and John D Sutherland. "Simultaneous Nucleotide Activation and Synthesis of Amino Acid Amides by a Potentially Prebiotic Multi-Component Reaction." In: *Angewandte Chemie International Edition* 46.42 (2007), pp. 8063–8066.

[102] Bhavesh H Patel, Claudia Percivalle, Dougal J Ritson, Colm D Duffy, and John D Sutherland. "Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism." In: *Nature chemistry* 7.4 (2015), p. 301.

[103] Michael Yarus. "Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin." In: *Journal of molecular evolution* 47.1 (1998), pp. 109–117.

[104] Michael Yarus. "Getting past the RNA world: the initial Darwinian ancestor." In: *Cold Spring Harbor perspectives in biology* 3.4 (2011), a003590.