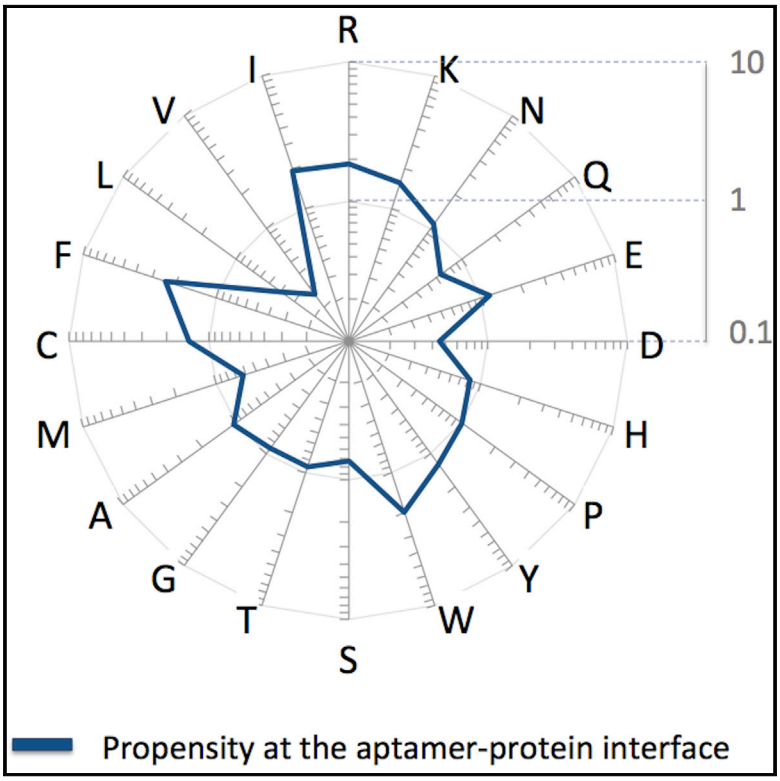
Article

Current Biology

Analysis of Evolutionarily Independent Protein-RNA Complexes Yields a Criterion to Evaluate the Relevance of Prebiotic Scenarios

Graphical Abstract



Highlights

- Protein-aptamer interactions are a useful model for the prebiotic RNA World
- Cationic amino acids (mainly arginine) are critical in protein-aptamer interactions
- The path to the genetic code may have depended on cationic amino acids

Authors

Celia Blanco, Marco Bayas, Fu Yan, Irene A. Chen

Correspondence

chen@chem.ucsb.edu

In Brief

Studying the origin of life requires evaluating proposed prebiotic scenarios. Assuming proteins arose in an RNA World, Blanco et al. analyze protein-aptamer complexes to identify crucial amino acids. Arginine dominates all interaction modes, suggesting that prebiotic syntheses must produce cationic amino acids to be on-pathway to the genetic code.

Blanco et al., 2018, Current Biology *28*, 526–537
February 19, 2018
https://doi.org/10.1016/j.cub.2018.01.014

CellPress

CellPress

Current Biology
Article

Analysis of Evolutionarily Independent Protein-RNA Complexes Yields a Criterion to Evaluate the Relevance of Prebiotic Scenarios

Celia Blanco,[1] Marco Bayas,[2] Fu Yan,[1] and Irene A. Chen[1,3,4,*]
[1]Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, CA 93106-9510, USA
[2]Departamento de Fisica, Escuela Politécnica Nacional, Quito, Ladron de Guevara E11-253, Ecuador
[3]Program in Biomolecular Sciences and Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106-9510, USA
[4]Lead Contact
*Correspondence: chen@chem.ucsb.edu
https://doi.org/10.1016/j.cub.2018.01.014

SUMMARY

A central difficulty facing study of the origin of life on Earth is evaluating the relevance of different proposed prebiotic scenarios. Perhaps the most established feature of the origin of life was the progression through an RNA World, a prebiotic stage dominated by functional RNA. We use the appearance of proteins in the RNA World to understand the prebiotic milieu and develop a criterion to evaluate proposed synthetic scenarios. Current consensus suggests that the earliest amino acids of the genetic code were anionic or small hydrophobic or polar amino acids. However, the ability to interact with the RNA World would have been a crucial feature of early proteins. To determine which amino acids would be important for the RNA World, we analyze non-biological protein-aptamer complexes in which the RNA or DNA is the result of *in vitro* evolution. This approach avoids confounding effects of biological context and evolutionary history. We use bioinformatic analysis and molecular dynamics simulations to characterize these complexes. We find that positively charged and aromatic amino acids are over-represented whereas small hydrophobic amino acids are under-represented. Binding enthalpy is found to be primarily electrostatic, with positively charged amino acids contributing cooperatively to binding enthalpy. Arginine dominates all modes of interaction at the interface. These results suggest that proposed prebiotic syntheses must be compatible with cationic amino acids, particularly arginine or a biophysically similar amino acid, in order to be relevant to the invention of protein by the RNA World.

INTRODUCTION

Understanding the origin of life is an important but thorny problem in biology. A major conceptual difficulty in this field is evaluating the relevance of different proposed prebiotic scenarios. Here, we reason backward from our best knowledge of early life to develop a criterion to evaluate prebiotic synthesis scenarios for amino acids. Perhaps the most robust theory of the origin of life is the presence of an RNA World, in which RNA fulfilled most of the functions of the early cell. The revelation that the catalytic center of the ribosome is a ribozyme is strong evidence that protein coding was invented within the RNA World [1]. Indeed, several authors have suggested that the earliest coded peptides were not catalysts, but instead served to stabilize functional RNA [2–4]. Although the RNA World was likely to contain amino acids and short peptides [5], the invention of coded protein synthesis marked a major evolutionary transition.

Many studies have attempted to outline the order of addition of amino acids to the code, based primarily on estimations of prebiotic availability or chemical or evolutionary hypotheses [6]. However, it is not clear that prebiotic abundance should be correlated with entry into the genetic code, and the connection between various hypotheses and the origin of life is also uncertain. Therefore, we approach the question of early amino acids of the genetic code in terms of biophysical importance to the RNA World. Because proteins presumably joined a protobiology dominated by RNA [1], their earliest functions would have involved close interaction with RNA. We analyze biophysical properties and amino acids that promote interaction with RNA. Because the exact amino acids of early life may differ from those today, we focus on general properties that may characterize classes of amino acids.

Previous analyses of the structures of biological protein-RNA complexes have highlighted the importance of hydrogen bonding in particular, with nonpolar residues aiding packing and aromatic residues stacking in the complex [7–26]. Most studies have focused on the role of hydrogen bonding in base recognition, suggesting that these interactions could be largely sufficient for specificity [24]. Some studies also highlight the role of positively charged amino acids, arginine and lysine, in mediating protein-DNA associations [10, 12, 13, 27, 28]. The positively charged amino acids can interact with nucleic acids via multiple modes [20, 21, 29, 30].

Although many protein-RNA complexes are known from biology, a major confounding factor is that these complexes have been subject to unknown evolutionary and functional constraints that may not be relevant to the binding interaction. With this caveat, algorithms based on a meta-analysis of biological

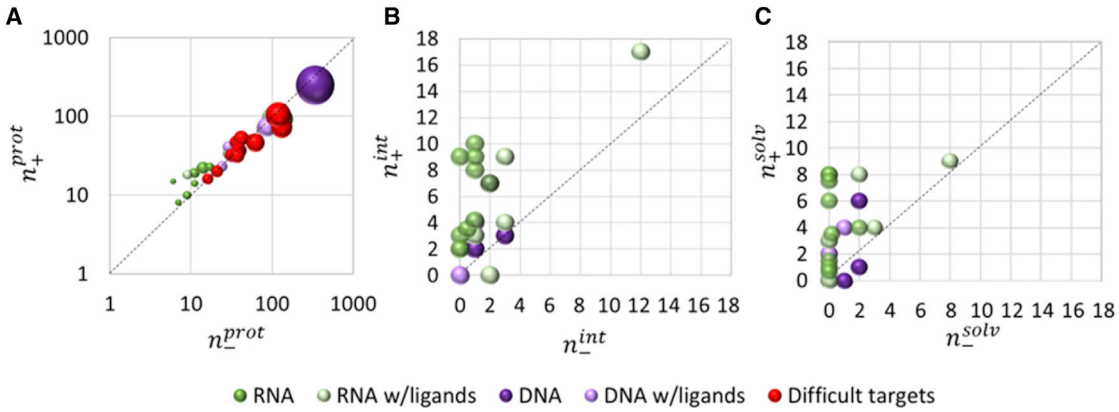


Figure 1. Frequency of Charged Residues per Protein in Aptamer-Binding Proteins

Number of positive and negative residues in each (A) protein, (B) interface, and (C) solvating region of the protein-aptamer complex. The area of the circle is proportional to the protein length. The legend indicates whether the aptamer is RNA (green) or DNA (purple) or whether standard aptamer evolution had failed repeatedly (red). The phrase "w/ligands" indicates that the PDB structure contains ordered ligands (i.e., ions or small molecules) aside from the nucleic acid and protein.

See also Tables S1, S2, and S3.

protein-DNA and protein-RNA complexes have been developed to predict potential DNA- or RNA-binding residues with support vector machines (SVMs). This approach considers calculated biochemical properties of the protein primary sequence (e.g., BindN [31] and Patch Finder Plus [32]) and may include evolutionary information (e.g., BindN+ [33], Pprint [34], and ConSurf [35, 36]). Other considerations include interface residue propensity (KYG [30]), sequence homology (RNABindRPlus [37]), and predicted secondary structure and conservation of physicochemical properties (PRBR [38]).

To reduce these confounding factors, we analyze protein-RNA interactions that evolved *in vitro*, i.e., protein-aptamer complexes. These complexes represent multiple evolutionary experiments that are independent from one another, so themes common among these complexes should reflect biophysical features rather than evolutionary constraints. To understand protein-nucleic acid complexes more generally, we also analyze proteins and DNA aptamers; principles of RNA-protein and DNA-protein interaction appear to be similar [7]. Furthermore, we compare the proteins in these complexes with proteins known to be difficult targets for aptamer binding, which had repeatedly failed SELEX for DNA aptamers [39]. Because aptamers against the difficult targets could be found using SELEX with hydrophobic nucleotides, the prior difficulty was presumably due to biophysical properties of the proteins. Understanding the biophysical characteristics of protein-aptamer complexes is of interest not only for the origin of life but also for improving aptamer engineering [40, 41].

In addition to comparing aptamer-binding proteins with "difficult target" proteins, we compare the aptamer-binding interface of the protein to the non-binding regions of the same protein. We also compare residues that are expected to lower the enthalpy of solvation with those that are not. These comparisons illustrate the amino acid composition and biophysical properties that promote aptamer binding. To characterize chemical interactions between protein and aptamer, we use a structure-based classification of interaction modes [16] and molecular dynamics (MD) simulations to calculate non-covalent electrostatic and van der Waals energies (enthalpies) of each protein-aptamer interface. Our results highlight the importance of electrostatics and particularly arginine. We discuss the implications for the prebiotic chemistry scenarios for the origin of life.

RESULTS

Frequency of Charged Residues per Protein in Aptamer-Binding Proteins

A preliminary analysis indicated that aptamer-binding proteins are biophysically distinct from proteins that bind natural biological RNAs (STAR Methods). Because larger proteins appeared to have a disproportionate amount of negative charge and aptamers interact with a specific region within a protein, we attempted to enrich the dataset for aptamer-binding regions by restricting analysis to the smaller aptamer-binding proteins (<500 and <200 amino acids). Indeed, there is a significant trend toward positive charge content for both subsets. For proteins under 500 residues, the average frequency of positively charged amino acids $\rho_+ = 0.13$ is greater than the average $\rho_- = 0.11$ per protein ($p = 0.005$ for a t test comparing two means). The difference is more pronounced for proteins under 200 residues, with average $\rho_+ = 0.14$ and average $\rho_- = 0.11$ per protein ($p = 0.002$ for a t test comparing two means). In contrast, for the known difficult targets (12 proteins comprising 5,435 residues; all examples are <200 amino acids), the average $\rho_+ = 0.11$ was slightly lower than the average $\rho_- = 0.12$, although this difference was not statistically significant from 0 ($p = 0.69$ for a t test comparing two means). These findings are summarized in Figure 1A for the subset of aptamer-binding proteins from the PDB and for difficult targets (Table S1). Overall, the analysis indicates that aptamer-binding regions of proteins are relatively localized within proteins and tend to be positively charged, consistent with the expectation from the negatively charged RNA backbone. Furthermore, a lack of positive charge is associated with difficulty evolving aptamers against the protein.

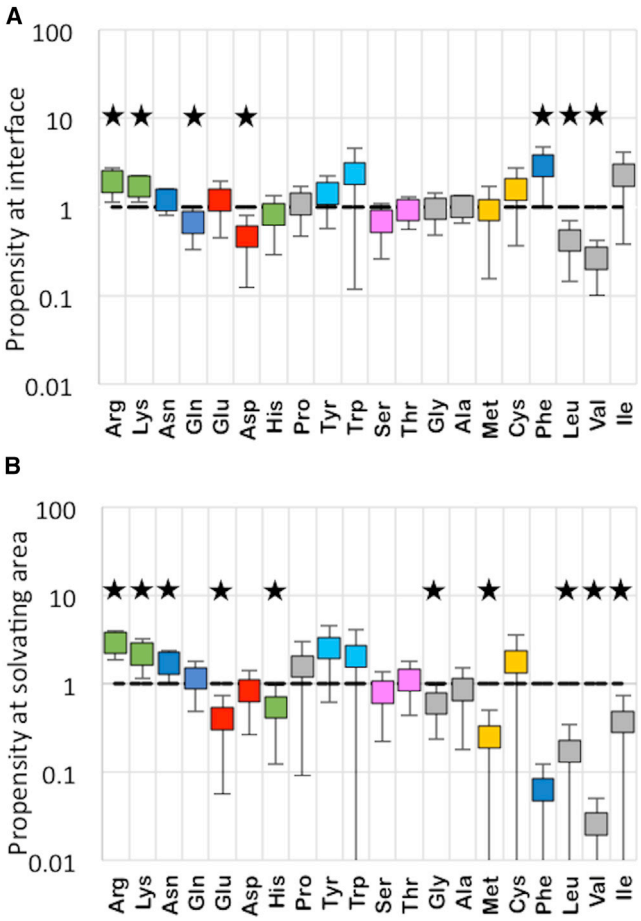


Figure 2. Statistical Significance of Bootstrapped Propensities

Propensities at the interface (A) and the solvating areas (B) with 95% confidence interval determined by bootstrapping method B1, reflecting variation among different complexes. Statistically significant deviations from 1 are noted by the star ($p < 0.05$). Amino acid types are ordered by increasing hydrophobicity. Bars are colored as follows: green, basic; dark blue, amidic; red, acidic; gray, aliphatic; light blue, aromatic; pink, hydroxylic; and yellow, sulfur containing.

See also Tables S1, S2, and S4 and Figures S1, S2, S3, and S7A.

Frequency of Charged Residues at the Protein-Aptamer Interface

Residues composing the protein-nucleic acid interface for protein-aptamer structures in the PDB were identified (STAR Methods; Table S2). The interfaces have a strong tendency to be positively charged, as the average ρ_+^{int} of interfacial residues, 0.41, is much greater than the average ρ_-^{int}, 0.12 ($p = 0.007$ for a t test comparing two means; see Figure 1B and Table S3). Thus, the interfacial residues show a substantially greater tendency toward positive charge compared to the protein as a whole. Of the 21 studied non-redundant interfaces, 18 were overall positively charged (Table S3).

Frequency of Charged Residues in the Solvating Region of the Protein-Aptamer Interface

We identified a subset of interfacial residues that are expected to lower the energy of the bound complex (STAR Methods). As with the protein-aptamer interfaces, the solvating areas have a strong tendency to be positively charged: 16 out of 21 studied solvating areas are positively charged, and the average $\rho_+^{solv} = 0.40$ is significantly greater than the average $\rho_-^{solv} = 0.09$ ($p = 5.5 \times 10^{-4}$ for a t test comparing two means; Figure 1C; Table S3).

Amino Acid Composition of Interfacial and Solvating Regions

To understand the composition of protein regions interacting with aptamers, for each protein-aptamer complex, we estimate the frequency of each amino acid in the whole protein (f_{prot}), the surface (f_{surf}), the protein-aptamer interface (f_{int}), and the solvating region (f_{solv}). Although f_{surf} for the entire protein is usually unknown because the PDB structure does not contain the entire structure, we estimated f_{surf} as described (STAR Methods). Propensity (p_i^{int} or p_i^{solv}) is the ratio of f_{int} (or f_{solv}) to f_{surf} and is a measure of the tendency of amino acid i to participate in the protein-aptamer interface (or in solvation) while controlling for its tendency to be on the surface. The propensities of each amino acid, averaged over all of the protein-aptamer complexes in the PDB, are shown in Table S4.

The statistical significance of the propensities was determined by bootstrap sampling of observed propensity values for the aptamer-protein complexes in the PDB, which reflects the variation observed among different proteins (bootstrap method B1). The propensities at the interface for Arg, Lys, and Phe are significantly greater than 1, whereas the propensities at the interface for Gln, Asp, Leu, and Val are significantly lower than 1 (Figure 2A). The propensities at the solvating area for Arg, Lys, and Asn are found to be significantly greater than 1, whereas the propensities at the solvating area for Glu, His, Gly, Met, Phe, Leu, Val, and Ile are found to be significantly lower than 1 (Figure 2B). Similar trends are seen when considering only the subset of proteins binding RNA and proteins binding RNA without ordered ligands (see Figures S1A and S1B, respectively).

It is possible that propensities could reflect highly unusual protein compositions rather than true preferences of amino acids to be at the interface. Therefore, to account for the composition of a specific protein, we determine whether the propensity of an amino acid in a given protein interface differs from its propensity in a randomly selected subsequence of the same protein (bootstrap method B2). In aptamer-binding proteins (in the absence or presence of ordered ligands), the observed propensities for Arg, Lys, and Trp were found to be significantly greater than those for the random subsequences, whereas the observed propensities for Leu and Val were significantly lower than those for the random subsequences (Figure S2). The combined results from the two different bootstrap methods (B1 and B2) suggest an overall preference for RNA interaction with Arg and Lys and an avoidance of interaction with Val and Leu.

Biophysical Properties of Interfacial and Solvating Regions

To determine whether biophysical properties were correlated with aptamer binding, we estimated charge, average flexibility [42], hydrophobicity using two different scales (Kyte and Doolittle [43] and Black and Mold [44]), hydrophilicity (Hopp and Woods) [45], polarity using two different scales (Zimmerman [46] and Grantham [47]), percentage of buried and accessible residues

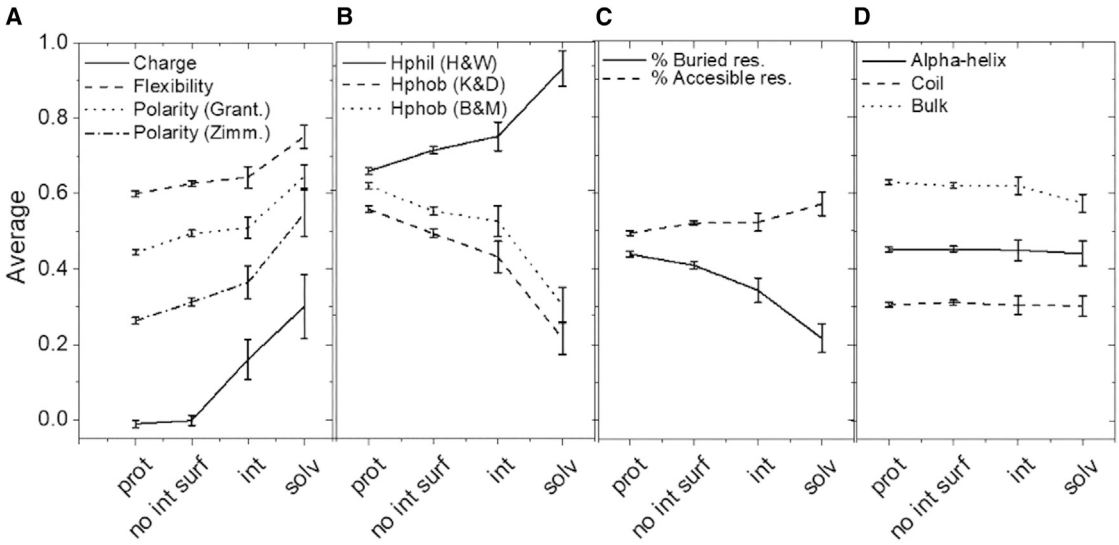


Figure 3. Average Biophysical Properties in Aptamer-Binding Proteins

Normalized biophysical properties for the whole protein (prot), non-interfacial surface (no int surf), interface (int), and solvating areas (solv) of all protein-aptamer complexes. Shown are (A) charge, flexibility, and polarity (following two different scales); (B) hydrophilicity and hydrophobicity (following two different scales); (C) the percentage of buried residues and accessible residues; and (D) the tendency to form alpha helices or coils and bulkiness. Bars correspond to a 95% confidence interval for the average from 19 complexes. See also Figure S4 and Tables S1 and S2.

[48], bulkiness [46], and tendency to form coils [49] or alpha helices [50]. We calculated or estimated these values for each whole-protein sequence, protein-aptamer interface (determined by PDBePISA), and solvating area (determined by PDBePISA). For comparison to the interface, we also attempted to characterize the non-interfacial surface of the entire protein (STAR Methods).

Properties that favor nucleic acid binding are expected to be relatively low in whole sequence and non-interfacial surface compared to protein-aptamer interface or solvating region; those that disfavor binding would show the opposite tendency. Charge, flexibility, polarity, hydrophilicity, and the percent of accessible residues are all increased in the interface and solvating region, whereas hydrophobicity and the percent of buried residues show the opposite (Figures 3A–3C). The tendency to form alpha helices or coils and bulkiness do not show a significant difference (Figure 3D). Similar results were found for protein-RNA complexes in the absence of ligands (Figure S4).

Nature of the Chemical Interactions at the Aptamer-Protein Interface

We used ENTANGLE to identify likely hydrogen-bonding, pi-pi stacking, electrostatic, hydrophobic, and van der Waals interactions from PDB structures [16] (STAR Methods). For each mode of chemical interaction, we calculated the fraction of interactions that was attributed to each type of amino acid (Figure 4). Arg was the most frequently found amino acid for all modes of interaction (hydrogen bonding, electrostatic, stacking, hydrophobic, and van der Waals). Both Arg and Lys clearly dominated the electrostatic mode. Lys was also important for hydrophobic and van der Waals interactions. For hydrogen bonding, Ser, Thr, and Gln were prominent (after Arg). For stacking, His and Tyr were prominent (after Arg). For hydrophobic interactions, after Arg and Lys, several amino acids contributed similarly (Asn, Gln, Glu, Pro, Ser, and Thr), although, interestingly, the small hydrophobic amino acids (Leu, Val, and Ile) made only minor contributions. A similar pattern was seen with van der Waals interactions. Overall, Arg and Lys accounted for over half of all interactions.

Electrostatic and van der Waals Contributions to the Internal Energy of Aptamer-Protein Complexes

We used MD simulations of the complexes to calculate the electrostatic and van der Waals energies in the interfaces (Figure 5A). In general, we found that the electrostatic component is much larger than the van der Waals component for all aptamer-protein complexes except one (PDB: 3AHU). Complex 3AHU (RNA-binding protein Hfq) is unusual in this set because it contains three protein chains that contact a single aptamer, with only one positively charged residue found at the interface. In general, the electrostatic energies correspond to more than 80% of the total interfacial energy (absolute numbers are given in Figure S5A).

Energetic Contributions to the Protein-Aptamer Complex from Different Amino Acids

We calculated the contribution of each amino acid and averaged these values for each type (Figure 5B). Two amino acids, Arg and Lys, contributed the majority of electrostatic energy (H_{elec}) to the complex, with the hydroxyl-containing amino acids (Ser and Tyr) also having a minor contribution. The negatively charged residue Glu was destabilizing ($H_{elec} > 0$). In contrast, a variety of amino acids stabilize the complex via van der Waals energies (H_{vdw}). A notable exception is the small hydrophobic residues (Leu, Val, and Ile), which contributed little to the van der Waals energies. For the charged and hydroxylic amino acids, the

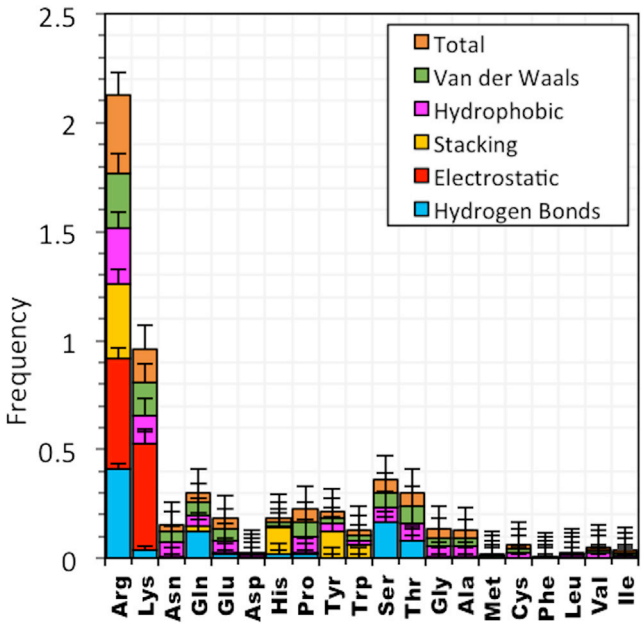


Figure 4. Fraction of Interactions Attributed to Each Type of Amino Acid

Relative frequency of each amino acid in each kind of interaction (hydrogen bonds, electrostatic, stacking, hydrophobic, or Van der Waals and for all together) averaged over 8 protein complexes. The frequency for any amino acid *i* in each complex is calculated as the number of times that amino acid *i* is found in interaction type *j*, divided by the total number of amino acids involved in interaction type *j*. Error bars represent the SE. See also Tables S1 and S2.

magnitude of electrostatic energy was greater than that of van der Waals energy (Figure S5B).

Electrostatic Interactions in Aptamer-Protein Complexes

Based on MD simulations, we further classified the electrostatic interactions into three mutually exclusive groups: (1) hydrogen bonds in the absence of ionic interaction; (2) ionic interactions in the absence of hydrogen bonding; and (3) *mixed* interactions, i.e., hydrogen bonds between donor and acceptor atoms of opposite charge sign.

The proportion of hydrogen bonds (either at the amino acid main chain or side chain), ionic interactions, and mixed interactions present at each complex interface is shown in Figure 5C (absolute numbers are given in Figure S5C). For every case except 3AHU, the number of hydrogen bonds and the number of interactions with ionic character (ionic plus mixed interactions) are roughly equal. The exception to this trend, complex 3AHU, which has a single charged residue at the interface, lacks ionic interactions at the interface and exhibits lower electrostatic enthalpy than the other complexes.

Electrostatic Contributions to the Protein-Aptamer Complex from Different Amino Acids

Given the predominant role of electrostatic interactions in the enthalpy of the complexes, we further classified these interactions between protein and aptamer according to which amino acid was involved in each interaction type (H-bond, ionic, and mixed; Figure 5D). Consistent with their high propensity values at the interface, Arg and Lys are involved in ~5×–10× more electrostatic interactions than any other amino acid type. For amino acids with side chains capable of H bonding, the hydrogen bonds of the interface tend to involve the side chain rather than the main chain (Figure S5D). Overall, H-bonds to the Arg side chain dominate the H-bond landscape. Interestingly, His (pKa ~6) is not involved in hydrogen bonds or ionic interactions in these complexes.

Positively Charged Residues and Electrostatic Energy

We examined the correlation between H_{elec} and the number of positively charged residues in the protein (n_+^{prot}), the interface (n_+^{int}), and the solvating area (n_+^{solv}). Although there is no correlation between H_{elec} and n_+^{prot} ($R^2 = 0.1$; Figure 6A), we find high correlation to n_+^{int} and n_+^{solv} ($R^2 = 0.98$ and $R^2 = 0.92$, respectively; Figure 6B), indicating that the number of positive residues at the interface essentially determines H_{elec}. To determine whether each charged residue contributes a constant energy to the complex, we calculated H_{elec}/n_+^{int} for each protein complex (Figure 6C). We find that, as n_+^{int} increases, H_{elec}/n_+^{int} also increases, i.e., that the energetic contribution of each positive residue is greater in magnitude in complexes with a greater number of positive residues. Similar results are obtained when considering only RNA-binding proteins in the absence of ordered ligands (Figure S6).

Amino Acid Composition of the Last Universal Common Ancestor (LUCA)

In contrast to proteins that interact with aptamers, LUCApedia proteins exhibit a statistically significant excess of negatively charged residues over positively charged residues ($p = 0.022$ and $p < 10^{-6}$ for the high- and medium-confidence sets, respectively). The amino acid composition of LUCApedia proteins is shown in Figure S7A. Interestingly, the positively charged amino acids are found at frequencies similar to the amino acids generally considered prebiotically plausible [51, 52]. A similar profile is found for the protein set composing SwissProt (see Figure S7B).

DISCUSSION

We sought to understand the biophysical features of the earliest coded proteins by reasoning that these proteins must have evolved in the RNA World, and so a major selective pressure would be the ability to interact with RNA. Any protein property, including its sequence, is the product of natural selection for function in a particular environment as well as evolutionary constraints and random factors. We therefore restricted our analysis to proteins that bind aptamers, which represent interactions that evolved *de novo* independently from one another and were selected primarily for binding activity.

For the aptamer-protein complexes analyzed, the interfaces and solvating areas are strongly positively charged. Both positively charged amino acids and aromatic amino acids are over-represented at the interface, whereas small hydrophobic residues are under-represented. Although polarity is similar between the non-interfacial surface and the interface, solvating residues are significantly more polar and hydrophilic than the interface as a whole. This suggests that nonpolar residues at the interface

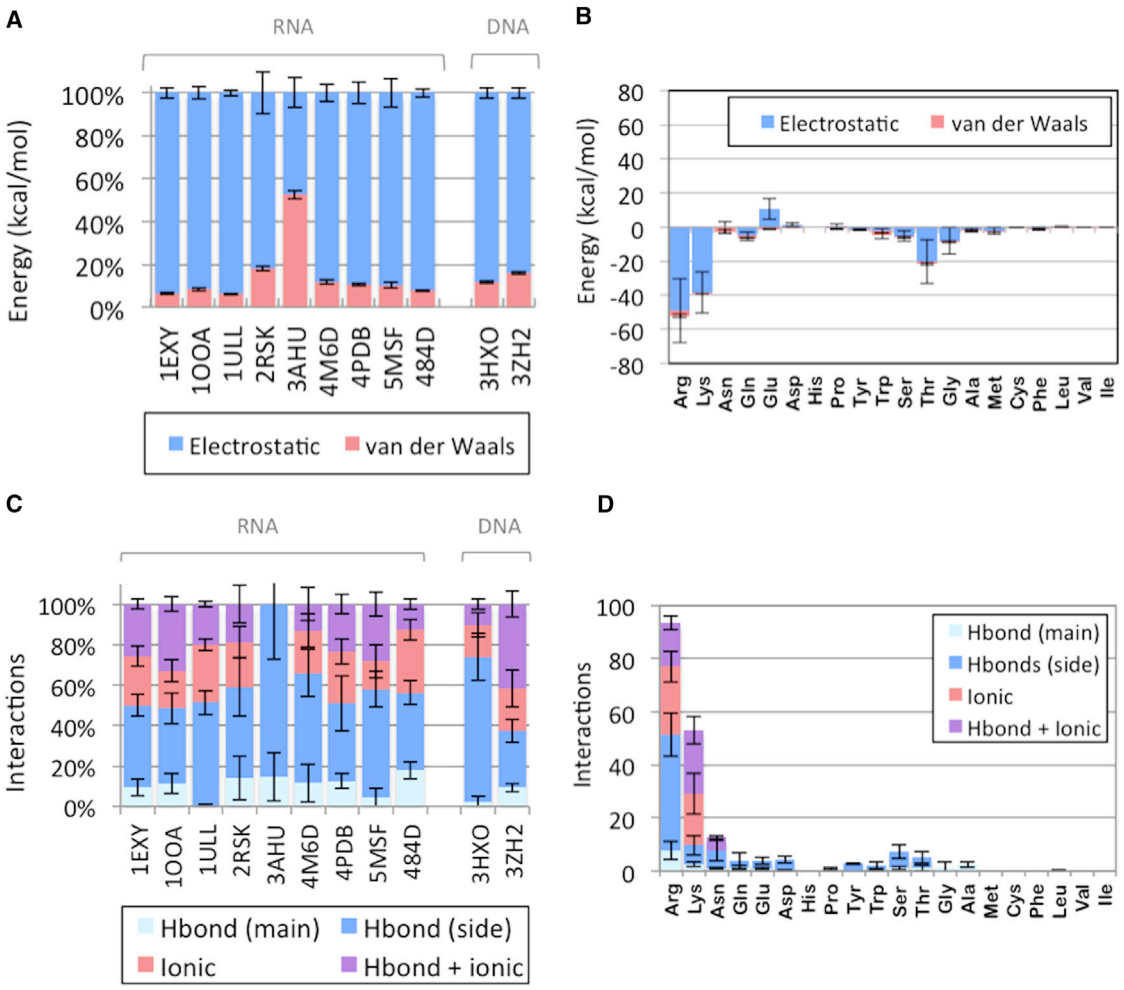


Figure 5. Electrostatic and van der Waals Energies and Contributions from Each Residue Type

(A) Electrostatic and van der Waals energies of the non-covalent interactions for proteins binding aptamers in the absence of ligands, as a proportion of the total non-covalent interaction energy (for absolute contributions instead of percentages, see Figure S5A).

(B) Electrostatic and van der Waals contributions per residue for each residue type, averaged over all protein-aptamer complexes. Error bars represent the SE over measurements from different complexes (for relative frequencies, see Figure S5B).

(C) Proportion of hydrogen bonds, ionic interactions, and mixed interactions between the protein and aptamer, present at the interface of each complex, calculated as the average value over the last 200 ps of the simulation. Error bars correspond to the SD over the simulation frames (for absolute contributions instead of percentages, see Figure S5C).

(D) Number of hydrogen bonds, ionic interactions, and mixed interactions for each amino acid type, calculated as the sum of the average values for the different complexes over the last 200 ps of the simulation. Error bars correspond to the sum of the SDs over the complexes (for relative frequencies, see Figure S5D). See also Tables S1 and S2.

(e.g., aromatic residues) contribute to direct interfacial contacts that may be buried in the complex, rather than having an effect on solvation, an interpretation that is corroborated by the higher calculated tendency of interfacial residues to be buried compared to solvating residues.

Arginine was by far the dominant amino acid at the protein-aptamer interface, being most frequently involved in all modes of interaction (electrostatic, hydrogen bonding, stacking, van der Waals, and hydrophobic). Only lysine came close to the frequency of arginine in any mode (electrostatic). Furthermore, MD simulations indicated that electrostatic energies were generally much greater than van der Waals energies for the complex. For electrostatic interactions, roughly half of these interactions have an ionic component and half have primarily H-bond character. Most H-bonds occur with the amino acid side chains rather than within the main chain, with Arg being again predominant in this mode of interaction. The electrostatic energy was highly correlated with the number of positive residues, as expected. However, a surprising finding was the non-linearity among positively charged residues, as each residue lowered the electrostatic enthalpy by a greater amount if there were more such residues. The mechanism for this effect is presumably not based on the entropic cost of conformational constraint, because these MD calculations are not sensitive to entropic changes. However, one possibility is that each positively charged residue promotes further interaction by increasing the

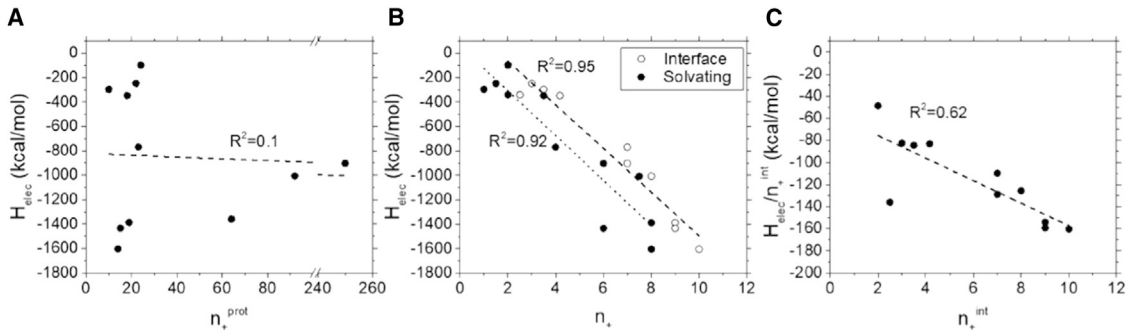


Figure 6. Correlation between Electrostatic Energy and Positively Charged Residues

(A and B) Dependence of electrostatic energy of the interface on the number of positive residues in the (A) primary structure of the entire protein and (B) interface or solvating area.

(C) Cooperativity appears as the electrostatic stabilization per positive residue increases as the number of positive residues in the interface increases. In all cases, the SE of the data across the last 200 ps of the simulation for each data point is under 4% of its value.

See also Figure S6 and Tables S1 and S2.

accessibility of the nucleic acid at other contact sites. The observed non-linearity may thus result from cooperativity among interaction sites. Both cooperativity and anti-cooperativity of salt bridges have been proposed to exist in protein structure; the findings here suggest that cooperativity among salt bridges characterizes protein-aptamer interfaces [53]. These findings, taken together, support the idea that arginine and lysine contribute the majority of binding enthalpy through electrostatic (e.g., hydrogen bonding and polar) interactions.

It has been argued that cations can substitute for positively charged amino acids. It was recently suggested that early peptides could have been overall negatively charged but closely associated with cations (like Mg^{2+} or Fe^{2+}) [4]. We find that complexes containing ordered ligands do not have a notably different charge profile from complexes that lack such ligands, but our dataset of proteins binding aptamers in the presence of ordered ions is too small to draw a conclusion regarding this possibility. It is possible that loosely bound cations may compensate for the positively charged amino acids. However, it is not likely that such cations could provide the same degree of energetic contributions (e.g., in non-electrostatic modes) as positively charged amino acids.

The high propensities of arginine and lysine for binding nucleic acids are supported by previous computational and statistical analysis [20, 21, 30, 54] of biological protein-RNA complexes. These prior studies also found high propensities for tyrosine, phenylalanine, and isoleucine [20]; asparagine and serine [21, 30]; and tyrosine, phenylalanine, methionine, histidine, and glycine [30]. Thus, the over-representation of aromatic amino acids that we observed at the interface is consistent with prior studies on biological complexes. It is important to note that our study did not include biologically evolved protein-RNA interactions and that we used a propensity measure that compares the frequency of an amino acid in the interface to the entire protein, not just to the protein fragment whose structure was determined. Although we are only able to estimate the expected surface based on the probabilities of surface exposure, we believe that the importance of considering the entire protein supersedes the additional precision of a structure-based accessible surface area (ASA) measurement. These differences in design, intended to reduce evolutionary and experimental biases, may contribute to the differences between propensities measured in this study and those in prior studies.

Although our analysis is based on protein-RNA interactions that were evolved *in vitro*, the protein sequences are derived from biology. One may consider whether non-biological proteins or peptides would also exhibit similar trends. Aptamers against single amino acids have been evolved *in vitro* [55]. Interestingly, these include multiple aptamers against arginine [56–60], aptamers against two aromatic amino acids (tyrosine [61] and tryptophan [62, 63]), consistent with our findings. On the other hand, aptamers are also known against two small hydrophobic amino acids (isoleucine [64–66] and valine [67]). Although it is difficult to connect these observations to general tendencies, the apparent ease with which arginine aptamers are found is notable. The converse analysis, i.e., of peptides evolved *in vitro* to bind a biological RNA, may also be illuminating. An analysis of artificial peptides that bind the λ boxB RNA hairpin showed that positively charged amino acids were enriched, with arginine having the highest frequency among amino acids in the selected peptides [68]. Another study of *in vitro* peptide evolution demonstrated that peptides consisting only of arginine, glycine, and serine were capable of binding the Rev response element with affinities similar to that of Rev protein [69]. Finally, a recent analysis of the interaction between ribosomal protein uL23 and its associated rRNA draws the contrast between the more ancient "tail" domain of uL23, which interacts primarily through ionic interactions between cationic amino acids and the rRNA backbone, and the newer globular domain, which interacts through other mechanisms [70]. These studies support the importance of arginine and electrostatic forces in non-biological and ancient protein-RNA interactions.

Several additional caveats should be kept in mind regarding this study. We did not consider histidine to be positively charged, as it has a pKa of ~6 [71] and the pH of the late Archean ocean is generally thought to have been 6.5–8 [72, 73]. The analysis is also limited by the datasets, which include a relatively small number of protein-aptamer complexes that may or may not be representative of protein-aptamer complexes in general. The classification of chemical interactions relies on criteria whose precise

Table 1. Prebiotic Plausibility and Interaction with RNA

Scenario	Amino Acids																			
Suggested chronology [6]	G	A	V	D	E	P	S	L	T	I	R	N	K	Q	C	H	F	M	Y	W
Prebiotic consensus from meteoritic analysis, simulated prebiotic chemistry and simulated hydrothermal vents [52]	G	A	V	D	E	P	S	L	T	I										
Cyanosulfidic protometabolism [74]	G	A	V	D	E	P	S	L	T		R	N		Q						
Interacting with nucleic acids [20, 21, 30]	G						S			I	R[a]	N	K[a]			H	F[a]	M	Y[a]	
Protein-aptamer: interface (this study)			–[b]	–				–[b]			R[b]		K[b]	–			F			W
Protein-aptamer: solvating (this study)	–		–		–			–		–	R	N	K			–	–	–		

Comparison of amino acids suggested to be prebiotically plausible and promoting interaction with RNA. First row: chronology of amino acid entry into the genetic code from a meta-analysis of several studies. Second row: consensus from prebiotic simulation experiments (e.g., Miller-Urey). Third row: amino acids produced by reaction network based on HCN. Fourth row: amino acids identified as interacting with nucleic acids by previous studies. Fifth row: amino acids that are over-represented at the interface with aptamers (dash sign indicated under-represented amino acid). Sixth row: amino acids that lower the solvation energy of the complex with aptamers.

[a]Found in more than one study

[b]Identified by both bootstrapping techniques

choice may affect the classification, given the mixed chemical nature of some interactions. We did not extract information about entropy from MD simulations, so the results reflect only the enthalpic term of free energy. Finally, LUCApedia contains extant proteins, whose composition may or may not reflect that of ancestral proteins.

To understand the emergence of early proteins, several studies have attempted to describe the set of amino acids that were prebiotically available. Different studies (based on analysis of comets or meteorites, Miller-Urey-type spark discharge experiments, and hydrothermal vent synthesis) suggest a consensus set of ten prebiotic amino acids: Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr, and Val (for a summary, see [51, 52]; Table 1). An important caveat to this consensus is that the chemical derivitization used during analytical techniques may not be appropriate to detect Arg and Lys, although their absence could also be due to inefficient synthesis and/or short half-lives [4, 75, 76]. This consensus set may therefore change in light of improvements in analytical techniques [77, 78]. Nevertheless, this set of amino acids shows adequate properties of complexity, secondary structure propensity, hydrophobic-hydrophilic patterning, and core packing potential [52] to be plausible for protein folding. Indeed, proteins or peptides based on a prebiotic 4-amino-acid alphabet (Gly, Ala, Asp, and Val) may have catalytic activity [79].

However, the consensus prebiotic list does not include the amino acids that are of greatest interest for RNA interaction, namely positively charged and aromatic amino acids (Table 1). It is often assumed that these amino acids entered the genetic code at a later stage [6, 80, 81]. Notably, the consensus prebiotic list could not provide ionic interactions, which would be an important interaction mode for macromolecules of high charge density, such as RNA. In addition, a major challenge for this prebiotic set is the inclusion of both negatively charged amino acids, which would disfavor protein folding due to a sharp increase in like-charge density upon collapse [52]. Although peptides or proteins lacking positively charged amino acids can serve structural and biochemical functions, such sequences are notably devoid of interaction with nucleic acid [73]. Our analysis suggests that the interactions of RNA with proteins from this tentative consensus prebiotic set would be seriously hampered.

There have been a few abiotic laboratory syntheses reporting positively charged amino acids [82, 83]. Lysine has been reported under simulated interstellar medium conditions [75], in electric discharge experiments simulating redox-neutral atmospheres [84], and in one carbonaceous chondrite meteorite (trace amounts) [85]. Arginine has been reported in simulated hydrothermal vents using heating as a source of energy [86–88]. Both arginine and lysine have been reported using different initial reagents and heat as a source of energy [89–91]. Other cationic amino acids may also be considered, such as diamino acids (e.g., 2,4-diaminobutanoic acid, found in the Murchison meteorite [92]), and Arg and Lys may have been later additions to biology [4]. Ornithine (arginine's biosynthetic precursor) has been reported in volcanic spark-discharge experiments [93] and in the Murchison meteorite [92] (for a detailed review on the prebiotic plausibility of the different amino acids, see [94]). It has also been suggested that ornithine may have entered and then left the genetic code during early evolution [80, 95].

Significantly, Sutherland and co-workers [74, 96] reported the efficient synthesis of the precursors of ribonucleosides, amino acids, and lipids within a common network of reactions based on hydrogen cyanide. This cyanosulfidic protometabolism does produce arginine. Interestingly, arginine codons are enriched among arginine aptamer sequences [97, 98], and an aptamer that binds two consecutive arginine residues can act as a template for a coupling reaction between them [99], supporting the possibility that arginine was an early entrant into the code. Recently, arginine-lipid conjugates formed from prebiotic reactants were shown to mediate interactions between RNA and lipid [100]. Although we do not know the composition of the early proteome, an analysis of LUCApedia suggests that cationic amino acids may have been fairly abundant (Figure S7A). We suggest that the importance of arginine and other positively charged amino acids may be used as a criterion for evaluating prebiotic synthetic conditions, i.e., that only synthetic conditions that support the formation of arginine or other biophysically similar amino acids are on-pathway toward the genetic code.

Conclusions

The theory that early life was based on RNA is widely supported. If so, a critical property of early proteins would be the ability to interact with RNA. We provide several analyses of aptamer-protein complexes that demonstrate the importance of electrostatic interactions involving positively charged amino acids, particularly arginine. Although many different prebiotic syntheses of amino acids may be proposed, the necessity for early proteins to interact with RNA can be used as a criterion to identify synthetic conditions that are on the pathway toward life.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Preliminary comparison of aptamer-binding proteins versus proteins that bind biological nucleic acids
  - Identification of interfacial residues and solvating residues
  - Amino acid composition of the expected protein surface
  - Propensity of amino acid types at the interface and solvating region
  - Classification of interactions
  - Bioinformatic analysis of biophysical properties of protein sequences in protein-aptamer complexes
  - Molecular dynamics simulations of selected complexes

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and four tables and can be found with this article online at https://doi.org/10.1016/j.cub.2018.01.014.

ACKNOWLEDGMENTS

The authors thank Dr. Margaret Hurley and Dr. Rick Dahlquist for valuable advice. C.B. was supported by an Otis Williams Postdoctoral Fellowship. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. We acknowledge support from the Center for Scientific Computing at the UCSB CNSI and MRL: NSF MRSEC DMR-1121053 and NSF CNS-0960316. This work was supported by the Simons Foundation (grant 290356), NASA (grant NNX16AJ32G), and the Institute for Collaborative Biotechnologies (grant W911NF-09-0001 from the US Army Research Office). The content of the information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

AUTHOR CONTRIBUTIONS

Conceptualization, C.B., M.B., and I.A.C.; Methodology, C.B., M.B., and I.A.C.; Software, M.B. and C.B.; Formal Analysis, C.B.; Investigation, C.B. and F.Y.; Resources, M.B. and I.A.C.; Writing – Original Draft, C.B. and I.A.C.; Writing – Review & Editing, C.B., M.B., and I.A.C.; Visualization, C.B.; Supervision, I.A.C.; Project Administration, I.A.C.; Funding Acquisition, I.A.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 23, 2017
Revised: December 4, 2017
Accepted: January 3, 2018
Published: February 1, 2018

REFERENCES

1. Pressman, A., Blanco, C., and Chen, I.A. (2015). The RNA World as a model system to study the origin of life. Curr. Biol. *25*, R953–R963.
2. Poole, A.M., Jeffares, D.C., and Penny, D. (1998). The path from the RNA world. J. Mol. Evol. *46*, 1–17.
3. Delaye, L., and Lazcano, A. (2000). RNA-binding peptides as early molecular fossils. In Astrobiology: Origins from the Big-Bang to Civilisation, J. Chela-Flores, G.A. Lemarchand, and J. Oró, eds. (Kluwer Academic), pp. 285–288.
4. Raggi, L., Bada, J.L., and Lazcano, A. (2016). On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. Phys. Chem. Chem. Phys. *18*, 20028–20032.
5. Ruiz-Mirazo, K., Briones, C., and de la Escosura, A. (2014). Prebiotic systems chemistry: new perspectives for the origins of life. Chem. Rev. *114*, 285–366.
6. Trifonov, E.N. (2000). Consensus temporal order of amino acids and evolution of the triplet code. Gene *261*, 139–151.
7. Draper, D.E. (1999). Themes in RNA-protein recognition. J. Mol. Biol. *293*, 255–270.
8. Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. (1999). Protein-DNA interactions: a structural analysis. J. Mol. Biol. *287*, 877–896.
9. Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol. *1*, reviews001.1–reviews001.37.
10. Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.
11. Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J. Mol. Biol. *253*, 370–382.
12. Nadassy, K., Wodak, S.J., and Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. Biochemistry *38*, 1999–2017.
13. Pabo, C.O., and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J. Mol. Biol. *301*, 597–624.
14. Jones, S., Shanahan, H.P., Berman, H.M., and Thornton, J.M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res. *31*, 7189–7198.
15. Tsuchiya, Y., Kinoshita, K., and Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins *55*, 885–894.
16. Allers, J., and Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J. Mol. Biol. *311*, 75–86.
17. Cheng, A.C., Chen, W.W., Fuhrmann, C.N., and Frankel, A.D. (2003). Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. J. Mol. Biol. *327*, 781–796.
18. Cheng, A.C., and Frankel, A.D. (2004). Ab initio interaction energies of hydrogen-bonded amino acid side chain[bond]nucleic acid base interactions. J. Am. Chem. Soc. *126*, 434–435.
19. Hermann, T., and Westhof, E. (1999). Non-Watson-Crick base pairs in RNA-protein recognition. Chem. Biol. *6*, R335–R343.

20. Jones, S., Daley, D.T., Luscombe, N.M., Berman, H.M., and Thornton, J.M. (2001). Protein-RNA interactions: a structural analysis. Nucleic Acids Res. *29*, 943–954.
21. Treger, M., and Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. J. Mol. Recognit. *14*, 199–214.
22. Walberer, B.J., Cheng, A.C., and Frankel, A.D. (2003). Structural diversity and isomorphism of hydrogen-bonded base interactions in nucleic acids. J. Mol. Biol. *327*, 767–780.
23. Houser-Scott, F., and Engelke, D.R. (2001). Protein–RNA Interactions. In eLS (John Wiley & Sons).
24. Morozova, N., Allers, J., Myers, J., and Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. Bioinformatics *22*, 2746–2752.
25. Nagai, K. (1996). RNA-protein complexes. Curr. Opin. Struct. Biol. *6*, 53–61.
26. Patel, D.J., Suri, A.K., Jiang, F., Jiang, L., Fan, P., Kumar, R.A., and Nonin, S. (1997). Structure, recognition and adaptive binding in RNA aptamer complexes. J. Mol. Biol. *272*, 645–664.
27. Anderson, K.M., Esadze, A., Manoharan, M., Brüschweiler, R., Gorenstein, D.G., and Iwahara, J. (2013). Direct observation of the ion-pair dynamics at a protein-DNA interface by NMR spectroscopy. J. Am. Chem. Soc. *135*, 3613–3619.
28. Esadze, A., Chen, C., Zandarashvili, L., Roy, S., Pettitt, B.M., and Iwahara, J. (2016). Changes in conformational dynamics of basic side chains upon protein-DNA association. Nucleic Acids Res. *44*, 6961–6970.
29. Kim, H., Jeong, E., Lee, S.W., and Han, K. (2003). Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. FEBS Lett. *552*, 231–239.
30. Kim, O.T., Yura, K., and Go, N. (2006). Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nucleic Acids Res. *34*, 6450–6460.
31. Wang, L., and Brown, S.J. (2006). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. *34*, W243–W248.
32. Shazman, S., Celniker, G., Haber, O., Glaser, F., and Mandel-Gutfreund, Y. (2007). Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res. *35*, W526–W530.
33. Wang, L., Huang, C., Yang, M.Q., and Yang, J.Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. *4*, S3.
34. Kumar, M., Gromiha, M.M., and Raghava, G.P. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins *71*, 189–194.
35. Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013). ConSurf: using evolutionary data to raise testable hypotheses about protein function. Isr. J. Chem. *53*, 199–206.
36. Miao, Z., and Westhof, E. (2015). Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. Nucleic Acids Res. *43*, 5340–5351.
37. Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2014). RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS ONE *9*, e97725.
38. Ma, X., Guo, J., Wu, J., Liu, H., Yu, J., Xie, J., and Sun, X. (2011). Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins *79*, 1230–1239.
39. Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E.N., Carter, J., Dalby, A.B., Eaton, B.E., Fitzwater, T., et al. (2010). Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS ONE *5*, e15004.
40. Famulok, M., Mayer, G., and Blind, M. (2000). Nucleic acid aptamers-from selection in vitro to applications in vivo. Acc. Chem. Res. *33*, 591–599.
41. Keefe, A.D., Pai, S., and Ellington, A. (2010). Aptamers as therapeutics. Nat. Rev. Drug Discov. *9*, 537–550.
42. Bhaskaran, R., and Ponnuswamy, P.K. (1988). Positional flexibilities of amino acid residues in globular proteins. Int. J. Pept. Protein Res. *32*, 241–255.
43. Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.
44. Black, S.D., and Mould, D.R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Anal. Biochem. *193*, 72–82.
45. Hopp, T.P., and Woods, K.R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA *78*, 3824–3828.
46. Zimmerman, J.M., Eliezer, N., and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. J. Theor. Biol. *21*, 170–201.
47. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862–864.
48. Janin, J. (1979). Surface and inside volumes in globular proteins. Nature *277*, 491–492.
49. Deléage, G., and Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. Protein Eng. *1*, 289–294.
50. Chou, P.Y., and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. *47*, 45–148.
51. Longo, L.M., and Blaber, M. (2014). Prebiotic protein design supports a halophile origin of foldable proteins. Front. Microbiol. *4*, 418.
52. Longo, L.M., and Blaber, M. (2012). Protein design at the interface of the pre-biotic and biotic worlds. Arch. Biochem. Biophys. *526*, 16–21.
53. Gvritishvili, A.G., Gribenko, A.V., and Makhatadze, G.I. (2008). Cooperativity of complex salt bridges. Protein Sci. *17*, 1285–1290.
54. Barik, A., C, N., Pilla, S.P., and Bahadur, R.P. (2015). Molecular architecture of protein-RNA recognition sites. J. Biomol. Struct. Dyn. *33*, 2738–2751.
55. Yarus, M. (1998). Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. J. Mol. Evol. *47*, 109–117.
56. Connell, G.J., Illangesekare, M., and Yarus, M. (1993). Three small ribooligonucleotides with specific arginine sites. Biochemistry *32*, 5497–5502.
57. Connell, G.J., and Yarus, M. (1994). RNAs with dual specificity and dual RNAs with similar specificity. Science *264*, 1137–1141.
58. Famulok, M. (1994). Molecular recognition of amino acids by RNA-aptamers: an L-citrulline binding RNA motif and its evolution into an L-arginine binder. J. Am. Chem. Soc. *116*, 1698–1706.
59. Geiger, A., Burgstaller, P., von der Eltz, H., Roeder, A., and Famulok, M. (1996). RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. Nucleic Acids Res. *24*, 1029–1036.
60. Tao, J., and Frankel, A.D. (1996). Arginine-binding RNAs resembling TAR identified by in vitro selection. Biochemistry *35*, 2229–2238.
61. Manironi, C., Scerch, C., Fruscoloni, P., and Tocchini-Valentini, G.P. (2000). Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif. RNA *6*, 520–527.
62. Majerfeld, I., Chocholousova, J., Malaiya, V., Widmann, J., McDonald, D., Reeder, J., Iyer, M., Illangasekare, M., Yarus, M., and Knight, R.

(2010). Nucleotides that are essential but not conserved; a sufficient L-tryptophan site in RNA. RNA *16*, 1915–1924.

63. Famulok, M., and Szostak, J.W. (1992). Stereospecific recognition of tryptophan agarose by in vitro selected RNA. J. Am. Chem. Soc. *114*, 3990–3991.
64. Majerfeld, I., and Yarus, M. (1998). Isoleucine:RNA sites with associated coding sequences. RNA *4*, 471–478.
65. Lozupone, C., Changayil, S., Majerfeld, I., and Yarus, M. (2003). Selection of the simplest RNA that binds isoleucine. RNA *9*, 1315–1322.
66. Legiewicz, M., and Yarus, M. (2005). A more complex isoleucine aptamer with a cognate triplet. J. Biol. Chem. *280*, 19815–19822.
67. Majerfeld, I., and Yarus, M. (1994). An RNA pocket for an aliphatic hydrophobe. Nat. Struct. Biol. *1*, 287–292.
68. Barrick, J.E., and Roberts, R.W. (2002). Sequence analysis of an artificial family of RNA-binding peptides. Protein Sci. *11*, 2688–2696.
69. Harada, K., Martin, S.S., and Frankel, A.D. (1996). Selection of RNA-binding peptides in vivo. Nature *380*, 175–179.
70. Lanier, K.A., Roy, P., Schneider, D.M., and Williams, L.D. (2017). Ancestral interactions of ribosomal RNA and ribosomal proteins. Biophys. J. *113*, 268–276.
71. Ballin, J.D., Prevas, J.P., Ross, C.R., Toth, E.A., Wilson, G.M., and Record, M.T., Jr. (2010). Contributions of the histidine side chain and the N-terminal alpha-amino group to the binding thermodynamics of oligopeptides to nucleic acids as a function of pH. Biochemistry *49*, 2018–2030.
72. Grotzinger, J.P., and Kasting, J.F. (1993). New constraints on Precambrian ocean composition. J. Geol. *101*, 235–243.
73. McDonald, G.D., and Storrie-Lombardi, M.C. (2010). Biochemical constraints in a protobiotic earth devoid of basic amino acids: the "BAA(-) world". Astrobiology *10*, 989–1000.
74. Patel, B.H., Percivalle, C., Ritson, D.J., Duffy, C.D., and Sutherland, J.D. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. Nat. Chem. *7*, 301–307.
75. Nuevo, M., Auger, G., Blanot, D., and d'Hendecourt, L. (2008). A detailed study of the amino acids produced from the vacuum UV irradiation of interstellar ice analogs. Orig. Life Evol. Biosph. *38*, 37–56.
76. Cleaves, H.J., 2nd. (2010). The origin of the biologically coded amino acids. J. Theor. Biol. *263*, 490–498.
77. Callahan, M.P., Martin, M.G., Burton, A.S., Glavin, D.P., and Dworkin, J.P. (2014). Amino acid analysis in micrograms of meteorite sample by nanoliquid chromatography-high-resolution mass spectrometry. J. Chromatogr. A *1332*, 30–34.
78. Burton, A.S., Stern, J.C., Elsila, J.E., Glavin, D.P., and Dworkin, J.P. (2012). Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. Chem. Soc. Rev. *41*, 5459–5472.
79. Oba, T., Fukushima, J., Maruyama, M., Iwamoto, R., and Ikehara, K. (2005). Catalytic activities of [GADV]-peptides. Formation and establishment of [GADV]-protein world for the emergence of life. Orig. Life Evol. Biosph. *35*, 447–460.
80. Jukes, T.H. (1973). Arginine as an evolutionary intruder into protein synthesis. Biochem. Biophys. Res. Commun. *53*, 709–714.
81. Jukes, T.H. (1974). On the possible origin and evolution of the genetic code. Orig. Life *5*, 331–350.
82. Yoshino, D., Hayatsu, K., and Anders, E. (1971). Origin of organic matter in early solar system—III. Amino acids: catalytic synthesis. Geochim. Cosmochim. Acta *35*, 927–938.
83. Hayatsu, R., Studier, M.H., and Anders, E. (1971). Origin of organic matter in early solar system—IV. Amino acids: confirmation of catalytic synthesis by mass spectrometry. Geochim. Cosmochim. Acta *35*, 939–951.
84. Plankensteiner, K., Reiner, H., and Rode, B.M. (2006). Amino acids on the rampant primordial Earth: electric discharges and the hot salty ocean. Mol. Divers. *10*, 3–7.
85. Kotra, R.K., Shimoyama, A., Ponnamperuma, C., and Hare, P.E. (1979). Amino acids in a carbonaceous chondrite from Antarctica. J. Mol. Evol. *13*, 179–184.
86. Sakurai, M., and Yanagawa, H. (1984). Prebiotic synthesis of amino acids from formaldehyde and hydroxylamine in a modified sea medium. Orig. Life *14*, 171–176.
87. Kamaluddin, Yanagawa, H., and Egami, F. (1979). Formation of molecules of biological interest from formaldehyde and hydroxylamine in a modified sea medium. J. Biochem. *85*, 1503–1507.
88. Hatanaka, H., and Egami, F. (1977). Formation of amino acids and related oligomers from formaldehyde and hydroxylamine in modified sea mediums related to prebiotic conditions. Bull. Chem. Soc. Jpn. *50*, 1147–1156.
89. Ferris, J.P., and Hagan, W.J., Jr. (1984). HCN and chemical evolution: the possible role of cyano compounds in prebiotic synthesis. Tetrahedron *40*, 1093–1120.
90. Ferris, J.P., Joshi, P.C., Edelson, E.H., and Lawless, J.G. (1978). HCN: a plausible source of purines, pyrimidines and amino acids on the primitive earth. J. Mol. Evol. *11*, 293–311.
91. Lowe, C.U., Rees, M.W., and Markham, R. (1963). Synthesis of complex organic compounds from simple precursors: formation of amino-acids, amino-acid polymers, fatty acids and purines from ammonium cyanide. Nature *199*, 219–222.
92. Meierhenrich, U.J., Muñoz Caro, G.M., Bredehöft, J.H., Jessberger, E.K., and Thiemann, W.H.P. (2004). Identification of diamino acids in the Murchison meteorite. Proc. Natl. Acad. Sci. USA *101*, 9182–9186.
93. Johnson, A.P., Cleaves, H.J., Dworkin, J.P., Glavin, D.P., Lazcano, A., and Bada, J.L. (2008). The Miller volcanic spark discharge experiment. Science *322*, 404.
94. Zaia, D.A.M., Zaia, C.T.B.V., and De Santana, H. (2008). Which amino acids should be used in prebiotic chemistry studies? Orig. Life Evol. Biosph. *38*, 469–488.
95. Hartman, H., and Smith, T.F. (2014). The evolution of the ribosome and the genetic code. Life (Basel) *4*, 227–249.
96. Sutherland, J.D. (2016). The origin of life–out of the blue. Angew. Chem. Int. Ed. Engl. *55*, 104–121.
97. Janas, T., Widmann, J.J., Knight, R., and Yarus, M. (2010). Simple, recurring RNA binding sites for L-arginine. RNA *16*, 805–816.
98. Yarus, M. (2017). The genetic code and RNA-amino acid affinities. Life (Basel) *7*, E13.
99. Harada, K., Aoyama, S., Matsugami, A., Kumar, P.K.R., Katahira, M., Kato, N., and Ohkanda, J. (2014). RNA-directed amino acid coupling as a model reaction for primitive coded translation. ChemBioChem *15*, 794–798.
100. Izgu, E.C., Björkbom, A., Kamat, N.P., Lelyveld, V.S., Zhang, W., Jia, T.Z., and Szostak, J.W. (2016). N-carboxyanhydride-mediated fatty acylation of amino acids and peptides for functionalization of protocell membranes. J. Am. Chem. Soc. *138*, 16669–16676.
101. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
102. Aptagen (2016). Apta-Index. https://aptagen.com/aptamer-index/aptamer-list.aspx.
103. Kirsanov, D.D., Zanegina, O.N., Aksianov, E.A., Spirin, S.A., Karyagina, A.S., and Alexeevski, A.V. (2013). NPIDB: nucleic acid-protein interaction database. Nucleic Acids Res. *41*, D517–D523.
104. Goldman, A.D., Bernhard, T.M., Dolzhenko, E., and Landweber, L.F. (2013). LUCApedia: a database for the study of ancient life. Nucleic Acids Res. *41*, D1079–D1082.

105. Walker, J.M. (2005). The Proteomics Protocols Handbook (Humana Press).

106. Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774–797.

107. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

108. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

109. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

110. Schein, C.H. (1990). Solubility as a function of protein structure and solvent components. Biotechnology (N. Y.) *8*, 308–317.

111. Ellis, J.J., Broom, M., and Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. Proteins *66*, 903–911.

112. UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Protein Data Bank (PDB)	[101]	https://www.rcsb.org/
Apta-Index aptamer database	[102]	https://www.aptagen.com/
Nucleic acid – Protein Interaction DataBase (NPIDB)	[103]	http://npidb.belozersky.msu.ru/
LUCApedia	[104]	http://eeb.princeton.edu/lucapedia/
Universal Protein Resource (UniProt)	[105]	https://www.uniprot.org/
Software and Algorithms		
ENTANGLE	[16]	http://www.bioc.rice.edu/~shamoo/entangle.html/
PDBePISA	[106]	https://www.ebi.ac.uk/pdbe/pisa/
VMD	[107]	https://www.ks.uiuc.edu/Research/vmd/
NAMD	[108]	http://www.ks.uiuc.edu/Research/namd/
CHARMM 27 force field	[109]	http://www.gromacs.org/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Irene A. Chen (chen@chem.ucsb.edu).

METHOD DETAILS

Datasets

RNA aptamer-protein complexes

All the protein-RNA aptamer complexes whose structures have been determined were extracted from the Protein Data Bank, RCSB PDB [101] (July 2016), by using the term 'aptamer' as a keyword for Text Search and including exclusively protein and RNA as macromolecules types. Structures were manually verified to include an aptamer and the protein target against which it was selected through *in vitro* evolution (e.g., not bound to a crystallization-promoting protein). This search resulted in a list of 19 protein-RNA aptamer complexes (Table S1). Of the 19 complex structures, 14 had been solved by X-ray crystallography and 5 by NMR spectroscopy. Some structures included ligands (ions and small molecules), which are detailed in Table S1.

In cases in which multiple structures had been solved, we chose the highest resolution structure for analysis, resulting in 14 protein-RNA complex structures (bold face PDB ID in Table S1). Two additional complexes (PDB: 3UZS and 2B63) involved 4 and 13 different types of polymeric chains respectively (i.e., multiple proteins). Since the presence of protein-protein interactions may alter the profile of protein-nucleic acid interactions, these two complexes were not considered in our study.

DNA aptamer-protein complexes

There are 16 protein-DNA aptamer structures in the PDB (July 2016). These structures were extracted as described above, with the exception that DNA was used instead of RNA as one macromolecule type. When multiple structures were available, the structure with the fewest ligands and highest resolution was selected for analysis. This resulted in a dataset of 5 protein-ssDNA complexes (bold face in Table S1). As with the protein-RNA aptamer complexes, protein-DNA aptamer structures containing more than one protein chain were excluded from our analysis. Two structures (PDB: 4HQU and 4NI7) contained SOMAmers, and the remainder (12) contained aptamers derived from *in vitro* selection.

Difficult aptamer targets

While many proteins have been successful targets of *in vitro* selection for aptamers, anecdotes of failed selections suggest that some proteins may be difficult targets for conventional aptamers. A list of such difficult targets was given by Gold et al. [39] (Table S1), and was used as a comparison group in our analysis. Aptamers to some of these targets have been reported, illustrating that selection was difficult but not impossible, in some cases.

Aptamer-binding proteins

We obtained a list of proteins that are reported to bind aptamers from the Apta-Index aptamer database [102], and the primary sequence for each protein was identified. The list has 81 entries, with 36 proteins binding to RNA aptamers, and 45 binding to DNA aptamers. Aptamer-binding proteins are similar in overall composition to proteins from the SwissProt database (Release 2017_11 of 22-Nov-17 of UniProtKB/Swiss-Prot) (Figure S7B).

Proteins that bind biological nucleic acids

To obtain a list of proteins that bind biological nucleic acids (either DNA or RNA), we used the Nucleic acid – Protein Interaction DataBase (NPIDB). To obtain a list of protein sequences that bind biological RNAs, we extracted the subset of RNA-binding proteins from the NPIDB [103].

Proteins likely to have been in the last universal common ancestor (LUCA)

We used the LUCApedia [104] to identify the set of proteins that is predicted to have been in LUCA. At high confidence (6 out of 6 studies), 180 protein sequences (170,524 amino acids) are included. At moderate confidence (3 out of 6 studies), 14,085 protein sequences are included (8,898,430 amino acids).

QUANTIFICATION AND STATISTICAL ANALYSIS

Preliminary comparison of aptamer-binding proteins versus proteins that bind biological nucleic acids

Validation of treatment of aptamer-binding proteins as a distinct set

To determine whether the biophysical profile of aptamer-binding proteins differed significantly from that of proteins that bind biologically derived nucleic acids, we calculated the frequency of positively and negatively charged amino acids among proteins that bind biological RNAs (NPIDB_rna). From 1132 proteins analyzed, with a total of 365,690 residues, we found that the frequency of positively charged residues was $P_{pos} = 0.138$, and the frequency of negatively charged residues was $P_{neg} = 0.120$. However, we find substantially lower frequencies for the 47 proteins that are known to bind RNA aptamers (from a combination of the Apta-index and the PDB subsets, discarding homologous proteins between subsets), for which $P_{pos} = 0.112$ and $P_{neg} = 0.114$. Interestingly, positively charged residues appear to be more depleted in aptamer-binding proteins compared to negatively charged residues. A chi-square test (comparing the number of positive, negative and non-charged residues in RNA aptamer-binding proteins to the numbers expected based on proteins binding biological RNAs) shows a highly significant difference ($p < 10^{-6}$). This effect was seen regardless of whether ordered ligands (e.g., ions) were present, as a similar analysis of RNA aptamer-binding proteins in the PDB whose structures lack ordered ligands also showed a significant difference from expectation based on proteins binding biological RNAs ($p = 5x10^{-6}$). Thus, proteins that bind RNA aptamers differ in overall charged residue content from proteins that bind biological RNAs, highlighting the importance of analyzing aptamer-binding proteins as a distinct group.

In an analogous analysis with proteins that bind either DNA or RNA, 2113 proteins were analyzed from the NPIDB (838,337 residues), for which $P_{pos} = 0.129$, $P_{neg} = 0.123$. For the 95 different aptamer-binding proteins (37,504 residues from a combination of the Apta-index and the PDB subsets, discarding homologous proteins between subsets), we again found substantially fewer charged residues than expected, with $P_{pos} = 0.112$ and $P_{neg} = 0.117$. These frequencies are similar to those found when only RNA aptamer-binding proteins are analyzed, with charged residues being lower in aptamer-binding proteins compared to proteins that bind biological nucleic acids. A chi-square test (comparing the number of positive, negative and non-charged residues in proteins binding to aptamers to the numbers expected based on proteins binding biological DNA or RNA) shows a highly significant difference ($p < 10^{-6}$). As before, similar results are found when considering the subset of proteins that bind aptamers without ordered ligands in their PDB structures ($p < 10^{-6}$). Again, the differences between proteins that bind aptamers and proteins that bind biological nucleic acids indicate that aptamer-binding proteins are a biophysically distinct group and motivated further analysis.

Thus, a simple summation of the number of positive and negative residues indicated that the total number of positive residues was lower in aptamer-binding proteins compared to proteins that bind biological DNA and RNA. A similar effect was seen for negative residues, to a lesser extent. These decreases could occur if the charged residue content was uniformly lower among aptamer-binding proteins, or if charged residues were disproportionately depleted in larger aptamer-binding proteins, which contribute disproportionately to the total. To distinguish these possibilities, we calculated the positive and negative charge content per protein (defined as $\rho_+ = n_+^{prot} / N_{prot}$ and $\rho_- = n_-^{prot} / N_{prot}$, where n_+^{prot} and n_-^{prot} are the number of positively and negatively charged residues in a given protein, respectively, and N_{prot} is the total number of residues in that protein).

Validation of importance of protein size when analyzing charge content

For the 95 aptamer-binding proteins (PDB and Apta-Index), the average $\rho_+ = 0.12$ was greater than $\rho_- = 0.11$, with borderline statistical significance ($p = 0.052$ for a t test comparison of two means). For the 47 RNA aptamer-binding proteins, the average $\rho_+ = 0.13$ was greater than the average $\rho_- = 0.11$ ($p = 0.026$ for a t test comparison of two means), with 29 proteins being overall positively charged and 18 overall negatively charged. Although the total number of negatively charged residues in this subset (2050) is greater than the total number of positive residues (2021), the average charge per protein tends to be positive, indicating that smaller proteins contain disproportionately more positive residues.

Comparison of structures with ligands vs. without ligands

We considered whether highly ordered charged ligands, such as ions, may counter simple trends in electrostatics. We therefore compared the aptamer-protein structures in the PDB that contain ordered ligands with those lacking ordered ligands. With one exception (PDB: 1OOA), RNA-binding proteins lacking ordered ligands were positively charged (8 out of 9 complexes), as expected. However, for five RNA-binding proteins with ordered ligands, two were positively charged (Figure 1A and Table S3). There were too few complexes involving DNA aptamers to draw a conclusion about the role of ligands. Although the number of complexes is small, these findings suggest that tightly bound cations might sometimes fulfill the electrostatic role of positively charged residues for RNA-binding.

Identification of interfacial residues and solvating residues

Residues composing the protein-nucleic acid interface for 19 protein-aptamer structures (14 RNA aptamer-protein complexes and 5 DNA aptamer-protein complexes from the PDB; Table S1) were identified as follows. Among these complexes, two of them (PDB ID: 3AGV and 3ZH2) contribute two non-redundant interfaces, yielding a total of 21 different interfaces. We define the positive (or negative) charge content at each protein-aptamer interface as the number of positive (or negative) residues in the interface divided by the total number of residues at the interface, i.e., $\rho_+^{int} = n_+^{int} / N_{int}$ and $\rho_-^{int} = n_-^{int} / N_{int}$. Similarly, we define the positive and negative charge contents in the solvating area of each protein as: $\rho_+^{solv} = n_+^{solv} / N_{solv}$ and $\rho_-^{solv} = n_-^{solv} / N_{solv}$, where n_+^{solv} and n_-^{solv} are the number of positively or negatively charged solvating residues, and N_{solv} is the total number of solvating residues.

To analyze the interfaces and solvating areas in the different complexes we used PDBePISA [106] (Protein Interfaces, Surfaces and Assemblies) from the European Bioinformatics Institute, (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html). PDBePISA is an interactive tool to explore macromolecular interfaces and perform calculations of structural and chemical properties of macromolecular surfaces and interfaces [106]. For a selected interface, we identified the involved residues and the solvation energy (ΔG) contributed by each residue to the complex. In PDBePISA, interfacial residues are defined as those with solvation energy different from zero ($\Delta G \neq 0$), and we further define 'solvating residues' as those that lower the solvation energy of the complex ($\Delta G < 0$).

Most of the complexes were composed of only two molecules, a protein and an aptamer, generating only one interface between them (PDB: 1EXY, 1ULL, 4PDB, 484D, 1EXD, 3DD2, 4M4O, 3UZT, 1AHO, 4HQU, 4NI7). In some cases, the complexes were composed of more than one protein and one aptamer, generating two or more interfaces between them (PDB: 2RSK, 3AHU, 5MSF, 3AGV, 3ZH2). In some cases, the complex was composed of either several biological assemblies that form an asymmetric unit (PDB: 1OOA, 4M6D) or by several asymmetric units that form a biological assembly (PDB: 5MSF, 3EGZ), forming several interfaces that are similar but not identical. Table S2 summarizes the total number of interfaces per complex. In these cases, we compute average quantities for all similar interfaces from the same complex, such that all complexes are weighted equally in overall metrics. In our analysis of charged residues, if multiple equivalent interfaces were present in the same PDB structure, we averaged the frequency of charged residues among the interfaces, such that only non-redundant interfaces are included (i.e., each distinct interface is equally weighted in the average).

Relationship between interfacial and solvating residues and charge

We calculated that 46% of the interfacial residues are solvating and 34% of the interfacial residues are charged (26% are positive and 8% are negative), so one would expect around 15% of the charged interfacial residues to be solvating if charge is independent of solvation effect. However, in fact, 68% of the charged interfacial residues are solvating, indicating that charged residues have a disproportionate solvating effect. Positively charged interfacial residues are more likely to be solvating compared to negatively charged interfacial residues (72% versus 54%). Conversely, 50% of the solvating residues are charged (compared with 15% as the random expectation), and the majority (81%) of these are positively charged, again indicating that charged residues (particularly positively charged residues) lower the solvation energy of aptamer-nucleic acid complexes.

Amino acid composition of the expected protein surface

Although the entire protein is usually not present in the PDB structure, we estimate the frequency of each amino acid at the surface (f_{surf}) for the entire protein based on the empirical Schein scale [110], s_i^{bur} (%). We calculated the fraction of amino acids of type i expected to be exposed on the surface as $s_i^{exp} = 1 - s_i^{bur}$ (%) / 100. For each protein, the number of times that amino acid i was expected to be present at the surface (N_i^{surf}) was calculated as $N_i^{surf} = N_i^{prot} \cdot s_i^{exp}$, where N_i^{prot} is the number of times i is present in the whole protein. Thus f_{surf} was approximated as $N_i^{surf} / \sum_i N_i^{surf}$.

To estimate the composition of the non-interfacial surface, we first estimate the expected composition of the surface based on the amino acid composition of the whole protein and the Schein scale. Then, we subtract the composition of the known interface, which leaves the estimated non-interfacial surface, for which the biophysical properties can be calculated.

Propensity of amino acid types at the interface and solvating region

Preliminary consideration of general trends

We expect that amino acids that contribute to RNA or DNA binding would follow the general increasing trend of $f_{prot} < f_{surf} < f_{int} < f_{solv}$, and amino acids that are generally unfavorable for interaction should show the opposite trend. We find an increasing trend for the positively charged residues (Arg and Lys), and, to a lesser extent, for two polar residues (Asn and Gln) (Table S4). We find a decreasing trend for a negatively charged residue (Glu) and two nonpolar residues (Leu and Val) (Table S4). While these qualitative trends are suggestive, it is also possible that an amino acid important for interaction with RNA would be over-represented at either the interface or among solvating residues without following this trend. Therefore we focus on propensity as follows.

Propensity

To determine whether the amino acid is over-represented at the interface or solvating region, one might calculate the ratio of f_{int} (or f_{solv}) to f_{prot}. However, these ratios could merely reflect the tendency of an amino acid to be present at the surface in general, rather than at the protein-aptamer interface. To control for this effect, we calculate the ratio of f_{int} (or f_{solv}) to f_{surf}, which we define as the propensity of amino acid *i* to be present at the interface or solvating area (p_i^{int} and p_i^{solv}). Thus we define 'propensity' to represent the tendency of each amino acid to be present in the interface or the solvating region. The propensity was calculated as the ratio

of the observed frequency at the interface or solvating area to the frequency at the expected surface of the protein. The propensity of amino acid *i* to be present at the interface or solvating area (p_i^{int} and p_i^{solv}) is:

$$p_i^{int} = \frac{N_i^{int} / \sum_i N_i^{int}}{N_i^{surf} / \sum_i N_i^{surf}} \qquad p_i^{solv} = \frac{N_i^{solv} / \sum_i N_i^{solv}}{N_i^{surf} / \sum_i N_i^{surf}}$$

where N_i^{int} and N_i^{sol} are the number of amino acids of type i at the interface or solvating region, respectively, and N_i^{surf} (the number of amino acids of type i at the surface) is estimated using the expected surface, as described above.

To determine whether a particular propensity value is significantly different from 1 given the spread of values among different complexes, we implemented a statistical bootstrapping method. This procedure (method B1) consists of simulating propensity values for each amino acid through random sampling of the calculated propensities for each complex. We generated 1000 random samples and then computed the 95% confidence interval of the mean bootstrap.

To determine whether the propensity of an amino acid in a given protein is significantly different from its propensity in a randomly selected subsequence of the same protein, we implemented a second bootstrapping procedure (method B2). For each protein, knowing the number of residues at the interface, we selected 1000 random subsequences (having the same number of contiguous amino acids as the true interface) and computed the propensity of each amino acid to be in these subsequences. The distribution of p_i for the random samples created a null expectation. We calculated statistical significance for whether the true value differed from the random subsamples.

Propensity using abundance vs. ASA

We defined the propensity of an amino acid as the ratio of its abundance in interfacial (or solvating) residues to its expected abundance at the surface. Propensity has been previously calculated as the ratio of its frequency in the ASA (accessible surface area) of the interface to its frequency in the ASA of the protein [8, 111]. Although the ASA-based method could give a more accurate quantification of the propensity in theory, a major caveat is that the ASA is only calculable for those residues whose crystallographic structure has been resolved (not for truncated or disordered regions of the protein). All of the proteins of the aptamer-protein complexes studied here possessed regions that were not represented in the crystal structure. Quantifying propensity based on ASA values for the incomplete protein could lead to artificial biases of unknown magnitude and direction because the *in vitro* selection of aptamers was conducted against the whole protein. Indeed, a comparison of propensity at the interface, calculated by ASA versus by expected frequency at the surface, demonstrates significant differences (Figure S3). Nevertheless, both types of calculation support the over-representation of Arg and Lys and the under-representation of Leu and Val among interfacial residues.

Propensity of dipeptide sequences

An analogous analysis was performed to calculate the propensity of dipeptide sequences, but no propensities were found to be statistically different from 1; this is likely because the dataset is small compared to the large number of possible dipeptide sequences.

Classification of interactions

To analyze and classify the different types of chemical interactions, we used ENTANGLE [16]. ENTANGLE identifies hydrogen bonds if the donor and acceptor atoms are within a distance of 3.9 Å, the hydrogen (inferred position) and acceptor are within a distance of 2.5 Å, and the donor-H-acceptor angle is > 90°. Ionic interactions are identified if two atoms of opposite charge are within a distance of 7 Å (i.e., Arg and Lys with phosphate backbone) and do not meet the criteria for hydrogen bonding. ENTANGLE identifies pi-pi stacking interactions if the center-to-center distance between aromatic amino acid side chain and the base of the RNA is < 3.8 Å with a dihedral angle of < 30°. Hydrophobic interactions are identified between non-polar atoms that are < 5 Å apart. Van der Waals interactions are identified if the distance between the atoms is less than the sum of the two atoms van der Waals radii plus 0.8 Å. ENTANGLE cannot analyze alternate conformers for amino acids or many modified nucleotides, so this analysis was limited to 8 protein-RNA aptamer complexes (PDB: 1EXY, 100A, 1ULL, 5MSF, 2RSK, 3AHU, 484D, 4PDB and 4M6D). All the error bars correspond to the standard error of the data across multiple protein complexes.

Bioinformatic analysis of biophysical properties of protein sequences in protein-aptamer complexes

We characterized biophysical properties for each protein using several scales. Each scale assigns a numerical value to each type of amino acid in order to estimate various biophysical properties. To calculate properties along a sequence, values are calculated for each residue within a sliding window and then averaged. A window size of 5 to 7 is appropriate for finding hydrophilic regions that are likely to be exposed on the surface and may potentially be antigenic, and window sizes of 19 or 21 will make hydrophobic, membrane-spanning domains stand out clearly (typically > 1.6 on the Kyte & Doolittle scale) [105]. Initially, we computed our results on a window size of 11 residues. Then, we decreased the window size (to 5 and 1) and computed the correlation coefficient (R^2) between the values calculated for different window sizes for the same property. For RNA-binding proteins, R^2 for any biophysical property between a window size 11 and a window size 1 was found to be greater than 0.9. Thus, we performed our subsequent calculations using a window size of 1, which allows us to also calculate the same biophysical properties for the expected surface of the protein (i.e., with an estimate of composition but without precise knowledge of the order of residues on the surface). We normalize values from different scales so that they range from 0 to 1, to facilitate comparison of the results obtained with different scales.

To estimate biophysical properties of an expected surface of a protein, we first calculate the number of each amino acid i on the expected surface for each protein, N_i^{surf}, and then compute the biophysical properties of the expected surface using its

composition and a window size of 1 residue (i.e., $\sum_i N_i^{surf} b_i$, where b_i corresponds to the value of a certain biophysical property for residue type i).

To estimate biophysical properties along the non-interfacial surface, we calculate the number of amino acid i in the non-interfacial surface as $N_i^{nonint} = N_i^{surf} - N_i^{int}$ for each protein, and then compute the biophysical properties of the non-interfacial surface using its composition and a window size of 1 residue (i.e., $\sum_i N_i^{nonint} b_i$, where b_i corresponds to the value of a certain biophysical property for residue type i).

The PDB files (including FASTA files available at the PDB) usually contain information for only a fragment of the protein (for which the structure was solved), so the sequences of entire proteins were extracted from the Universal Protein Resource (UniProt) [112]. For every analysis we perform on protein sequences, we have used the entire sequence. As before, error bars correspond to the standard error of the data across multiple protein complexes.

Molecular dynamics simulations of selected complexes

To understand the energetic contribution of interfacial interactions, we ran molecular dynamics (MD) simulations of the complex with explicit water and spherical boundary conditions. For each complex, the interaction energies between the protein and nucleic components were calculated during the stable phase of the simulations. The complexes that were simulated are PDB: 1EXY, 1OOA, 1ULL, 2BU1, 2RSK, 2V2T, 3AHU, 3HXO, 3ZH2, 4M6D, 4PDB, 5MSF and 484D. These complexes were chosen because they do not contain additional ligands (e.g., small molecules).

The molecular systems were built from the original complex PDB files, using the program VMD [107]. Every complex was solvated by placing it in a water sphere large enough to ensure that the smallest distance between its surface and any atom from the complex was at least 10 Å. Na^+ and Cl^- ions were added to the system to ensure electric neutrality, and the resulting molecular systems had radii ranging from 22 Å to 55 Å. Once the solvated complexes were prepared, MD simulations were performed using NAMD [108]. The energy of each system was minimized before heating it up to 300 K and the systems were equilibrated for 1 ns at 300 K. All the simulations were performed using the CHARMM 27 force field [109].

For each simulation, the RMSD (Root Mean Square Deviation) of the structure along consecutive trajectory frames was used to monitor the stability of the system. Since the complexes were relatively stable during the last 200 ps of the equilibration phase, the 20 structures corresponding to this period were used for the energy calculations (1 frame per 10 ps). This analysis was done using the plugin NAMDEnergy of the program VMD [107].

These structures were also used to categorize the different interactions between amino acids and nucleic acids at the interface, as well as the contribution from each amino acid residue to each type of interaction. Using VMD [107] we identified hydrogen bonds and ion-pair interactions between protein and aptamer at the interface for each complex. Hydrogen bonds were identified if the donor and acceptor atoms were within a distance of 3 Å and the donor-H-acceptor angle was sufficiently linear (> 130°). Ionic interactions were identified if two atoms of opposite charge were within a distance of 3 Å. We classified the interactions into three mutually exclusive groups: 1) hydrogen bonds between pairs of atoms with the same charge sign or in which at least one atom of the pair is non-charged (i.e., primarily H-bond character), 2) ionic interactions between pairs of atoms with opposite charge signs and a nonlinear donor-H-acceptor angle (i.e., primarily ionic character), and 3) hydrogen bonds between pairs of atoms with opposite charge signs (i.e., mixed H-bond and ionic character).