Biophysical Journal
Article

Biophysical Society

Contact Statistics Highlight Distinct Organizing Principles of Proteins and RNA

Lei Liu[1] and Changbong Hyeon[1,*]
[1]School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Republic of Korea

ABSTRACT Although both RNA and proteins have densely packed native structures, chain organizations of these two biopolymers are fundamentally different. Motivated by the recent discoveries in chromatin folding that interphase chromosomes have territorial organization with signatures pointing to metastability, we analyzed the biomolecular structures deposited in the Protein Data Bank and found that the intrachain contact probabilities, $P(s)$ as a function of the arc length s, decay in power-law $\sim s^{-\gamma}$ over the intermediate range of s, $10 \lesssim s \lesssim 110$. We found that the contact probability scaling exponent is $\gamma \approx 1.11$ for large RNA ($N > 110$), $\gamma \approx 1.41$ for small-sized RNA ($N < 110$), and $\gamma \approx 1.65$ for proteins. Given that Gaussian statistics is expected for a fully equilibrated chain in polymer melts, the deviation of γ-value from $\gamma = 1.5$ for the subchains of large RNA in the native state suggests that the chain configuration of RNA is not fully equilibrated. It is visually clear that folded structures of large-sized RNA ($N \gtrsim 110$) adopt crumpled structures, partitioned into modular multidomains assembled by proximal sequences along the chain, whereas the polypeptide chain of folded proteins looks better mixed with the rest of the structure. Our finding of $\gamma \approx 1$ for large RNA might be an ineluctable consequence of the hierarchical ordering of the secondary to tertiary elements in the folding process.

INTRODUCTION

RNA and proteins, under appropriate environmental conditions, adopt three-dimensionally (3D) compact native folds that are essential for a variety of biological functions. Despite general similarities of the folding principles that both biopolymers are made of sequences foldable to a functionally competent structure as an outcome of evolutionary selection (1–5), the overall shape of the native RNA differs from that of proteins in several aspects. Proteins are in general more compact, globular, and flexible than RNA (6). Such differences may be originated from the distinct nature of the building block. The energy scale of binary interaction that pairs nucleotides is typically greater than that of amino acids. Furthermore, the requirement of charge neutralization (or screening) along the backbone differentiates the foci of RNA dynamics, especially at the early stage of folding (7), from those of proteins.

Spotlighted in the recent studies of chromatin folding exploiting fluorescence in situ hybridization (8,9) and chromosome conformation capture techniques (10–12), human chromosomes in the interphase have a territorial organization (9) and the individual chromosome is also partitioned into a number of topologically associated domains (TADs), possibly mediated by proteins such as CTCF and cohesin (13). The contact probability $P(s)$ of two loci separated by the genomic distance s can provide glimpses into the arrangement of the chromatin chain. From the polymer perspective, a test chain in a fully equilibrated homogeneous polymer melt is expected to obey the Gaussian statistics because of the screening of excluded volume interaction (14), thus satisfying $P(s) \sim s^{-3/2}$. It was, however, shown that $P(s)$ of human chromatin in cell nucleus displays $P(s) \sim s^{-1.08}$ at the genomic scales of 1 Mb $< s <$ 10 Mb (11). To account for the origins of the human genome organization and its characteristic scaling exponent $\gamma = 1.08$ and patterns of contact map demonstrating TADs, several different models have been put forward, which include the crumpled (fractal) globule (11,15,16), random loop (17), strings and binders switch model (18), and confinement-induced glassy dynamics (19).

Besides the overall shape, chain organizations of the native folds of RNA and proteins are in general visually different from each other. Compared with proteins in which α-helices, β-strands, and loops thread through one another to form a native structure, a folded RNA with large N looks more crumpled; a number of secondary structure elements (helices, bulges, loops) forming independently stable modular contact domains are further assembled into a compact 3D structure. Here, borrowing the several statistical measures that have

Submitted January 28, 2016, and accepted for publication April 1, 2016.
*Correspondence: hyeoncb@kias.re.kr
Editor: Rohit Pappu.
http://dx.doi.org/10.1016/j.bpj.2016.04.020

been used to study the genome/chromosome organization inside cell nucleus, we substantiate the fundamental differences between the chain organizations of RNA and proteins in native states and discuss their significance in connection to their folding mechanisms.

MATERIALS AND METHODS

Calculation of contact probability and extraction of scaling exponent

Using atomic coordinates of RNA and protein from the Protein Data Bank (PDB), we consider that two residues i and j are in contact if the minimum distance between any two heavy atoms of these residues, located at $\vec{r}_i$ and $\vec{r}_j$, is smaller than a cutoff distance d_c (= 4 Å). The contact probability for a biomolecule α with chain length N_α (the number of residues) is thus determined by calculating

$$P_\alpha(s) = \frac{\sum_{i<j}^{N_\alpha}\delta(|i-j|-s)\Theta\left(d_c-\min\left|\vec{r}_i-\vec{r}_j\right|\right)}{\sum_{i<j}^{N_\alpha}\delta(|i-j|-s)}, \quad (1)$$

where $\Theta(x) = 1$ for $x \geq 0$; otherwise, $\Theta(x) = 0$. Two examples of $P(s)$ are given in Fig. 1, *B* and *C*. The power-law relation of $P(s) \sim s^{-\gamma}$ is observed over the intermediate scale. We determined the value of γ by fitting $P(s)$ over the range of $s_{\min} < s < s_{\max}$. The details of fitting procedure are discussed in the Supporting Material.

Mean contact probability

Each structure in the PDB has a different chain size N_α ($\alpha = 1, 2, \ldots, I_{\max}$). Thus, to consider the nonuniform distribution of chain size in computing the mean contact probability, we calculated the following N-dependent probability averaged over the total number of distinct chain sizes:

$$\langle\overline{P}(s)\rangle = \frac{1}{N_{\max}-N_{\min}+1}\sum_{N=N_{\min}}^{N_{\max}}\overline{P}(s\,|\,N), \quad (2)$$

where $\overline{P}(s\,|\,N) \equiv \sum_{\alpha=1}^{I_{\max}}\delta(N_\alpha - N)P_\alpha(s)/\sum_{\alpha=1}^{I_{\max}}\delta(N_\alpha - N)$ is the mean contact probability for the structures with chain size N, and we used the value of $P_\alpha(s)$ only for the range of $4 \leq s \leq N_\alpha^{2/3}$. The value $\langle\overline{P}(s)\rangle$ for RNA and proteins are shown in Fig. 1 *D*. $\langle M(s)\rangle$, $\langle n_s(s)\rangle$, $\langle$DOP$\rangle$, $\langle$DOS$\rangle$, and $\langle R(s)\rangle$, were calculated using similar definitions as Eq. 2. A cautionary note is in place. Unlike the contact probability exponent calculated for each macromolecular structure, these mean properties obtained by averaging over each ensemble of proteins and RNA are meant for understanding the general difference between RNA and proteins as two distinct classes of macromolecules.

RESULTS

Power-law exponent γ of contact probability

The contact probability $P(s)$ calculated for individual biopolymers (Eq. 1) exhibit power-law decay over the intermediate range of s, $10 \lesssim s \lesssim \mathcal{O}(10^2)$ (the *left panel* of Fig. 1, *B* and *C*). The scaling exponent γ from the fit using $P(s) \sim s^{-\gamma}$ was obtained for each biopolymer (see text and Figs. S1–S4 in the Supporting Material for details, where we discussed the accuracy of obtaining γ and showed the error bar of γ for each macromolecule) and its distributions, $p(\gamma)$, for RNA, and proteins are contrasted in Fig. 1 *A*. Proteins have $p(\gamma)$ broadly distributed from 0.5 to 2.5 centered around $\gamma \approx 1.5$, whereas $p(\gamma)$ for RNA is sharply peaked at $\gamma \approx 1.1$. No clear correlation is found between γ and the chain length (N) in proteins; however, in RNA while γ-values are broadly distributed at small N, they are sharply centered around $\gamma \approx 1.1$ when $N \gtrsim 100$ (see also Fig. S6).

The distinct scaling exponents, $\gamma \approx 1.11$ for the $P(s)$ of 23S rRNA ($P(s)$ at the *left corner* of Fig. 1 *B*) and $\gamma \approx 1.49$ for FhuA ($P(s)$ at the *left corner* of Fig. 1 *C*), elicit special attention. The value of $\gamma \approx 1.0$, especially for large-sized RNA arises from their characteristic chain organization: Similar to TADs in chromosomes, proximal sequences along the chain are stabilized by basepairing to form independently stable modular contact domains, consisting of hairpin, bulges, and loops. Further assemblies among these contact domains are achieved by a number of tertiary interactions (base triples, kissing loops, coaxial stackings through ribose zipper, A-minor motif, and metal-ion interactions) (20,21). The abundance of distal contacts resulting from the hierarchical chain assembly likely contributes to the greater frequency of the long-range contacts, giving rise to $\gamma \approx 1.11$ for 23S rRNA on the scale of $10 \lesssim s \lesssim 300$ (see the next section). The distinct chain organizations of RNA and proteins become more evident when molecules are visualized using rainbow coloring scheme spanning the chain (Fig. 1, *B* and *C*). The overall chain topology of 23S rRNA resembles a crumpled globule (22,23) that retains clearly demarcated contact domains held by distal interdomain contacts. The territorial organization of contact domains made of proximal sequences is highlighted in large-sized RNA structures (see the large and small subunit of rRNA in Fig. 1 *B*).

In stark contrast to rRNA, typical proteins with $\gamma \approx 1.5$ (indexed with *black labels* from 1 to 5 in Fig. 1, *A* and *C*) retain chain conformations whose subchains look topologically more intermingled with the rest of the structure, lacking visually distinct domains of a similar color. The intermingled chain configurations of native proteins as well as the contact probability scaling exponent $\gamma \approx 1.5$ points to a configuration of equilibrium globule, which is also supported by the same conclusion reached by investigating the loop size distribution of native protein structures (24). Of particular note are the proteins with $\gamma < 1.0$, which are found at the outliers of $p(\gamma)$. For example, $\gamma = 0.73$ is for chrondroitin sulfate ABC lyase I (the protein indexed with 6) (25), the chain configuration of which has clearly demarcated contact domains.

Instead of calculating the s-dependent contact probability for individual molecules ($P_\alpha(s), \alpha = 1, 2, \ldots I_{\max}$), one can also consider ensemble-averaged characteristics of native RNA and protein organizations, $\langle\overline{P}(s)\rangle$ (Eq. 2 and Fig. 1 *D*). The mean contact probability calculated for each ensemble of RNA and proteins exhibits power-law

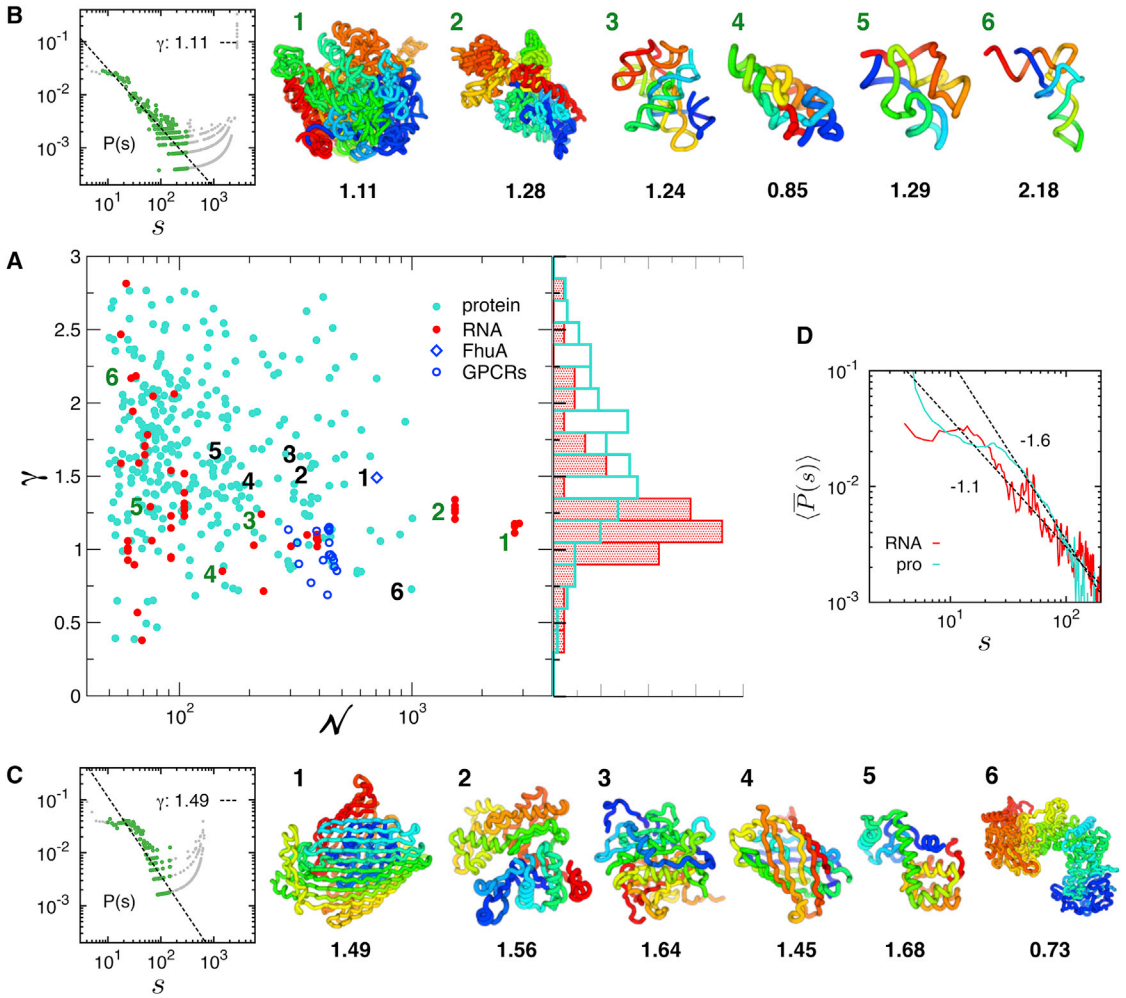

FIGURE 1 Contact probability scaling exponent, γ, and chain configurations of RNA and proteins. (*A*) The value γ versus N obtained for 60 individual RNA (data in *red*) and 324 proteins (data in *cyan*), whose γ-value is obtained from the power-law fit to $P(s)$ with the correlation coefficient (*c.c.*) > 0.9 (see Fig. S4 for γ versus N plot with error bars, and Tables S1 and S2 in the Supporting Material for PDB entries used here). The data points for FhuA and GPCRs are included for further discussion, although *c.c.* <0.9. Histograms of γ, $p(\gamma)$ for RNA ($\gamma \approx 1.30 \pm 0.44$) and proteins ($\gamma = 1.65 \pm 0.50$) are shown on the right. (*B*) Representative structures of RNA in the rainbow coloring scheme from 5′ (*blue*) to 3′ (*red*), indexed with the number in γ versus N plot. Depicted are the structures of 1) a large subunit of rRNA (PDB: 2O45 (66), $\gamma = 1.11$); 2) a small subunit of rRNA (PDB: 2YKR, $\gamma = 1.28$) (67); 3) Twort group I ribozyme (PDB: 1Y0Q, $\gamma = 1.24$) (68); 4) A-type ribonuclease P (PDB: 1U9S, $\gamma = 0.85$) (69); 5) TPP-riboswitch (PDB: 3D2G, $\gamma = 1.29$) (70); and 6) tRNA (PDB: 1VTQ, $\gamma = 2.18$). $P(s)$, which provides γ-value, is shown for a large subunit of rRNA on the left corner. The scaling exponent (γ) of $P(s) \sim s^{-\gamma}$ is obtained from the fit (*dashed line*) to the data points in green; the data in gray are excluded from the fit (see the Supporting Material for details of fitting procedure). (*C*) Protein structures in the rainbow coloring scheme from the N- (*blue*) to the C-terminus (*red*). Depicted in (*C*) are the structures of 1) FhuA (PDB: 1QJQ, $\gamma = 1.49$) (71); 2) an actin monomer (PDB: 1J6Z, $\gamma = 1.56$) (72); 3) metacaspase (PDB: 4AF8, $\gamma = 1.64$) (73); 4) green fluorescent protein (PDB: 1EMA, $\gamma = 1.45$) (74); 5) T4 lysozyme (PDB: 2LZM, $\gamma = 1.68$) (32,75); and 6) Chondroitin Sulfate ABC lyase I (PDB: 1HN0, $\gamma = 0.73$) (25). $P(s)$ for FhuA is shown on the left corner. (*D*) The mean contact probabilities, $\langle \overline{P}(s) \rangle$, calculated over the RNA and protein structures in the PDB. To see this figure in color, go online.

decay $\langle \overline{P}(s) \rangle$ with $\gamma \approx 1.1$ for RNA and $\gamma \approx 1.6$ for proteins on the scale of $(20 - 30) \lesssim s \lesssim 100$, which helps us in understanding the general difference of structural ensemble between RNA and proteins as two distinct classes of macromolecules.

Cautionary remarks are in place in regard to the power-law scaling of $P(s)$. The characteristic power-law decay behavior of 23S rRNA with $\gamma \approx 1.1$ is only valid for the intermediate range of s. For small s, $P(s)$ decays with a different power-law exponent (see the two panels of $P(s)$ in Fig. 1, *B* and *C*). As reported by Lua and Grosberg (22), on local scales both RNA and proteins have a chain organization different from the one on a larger scale, which is also confirmed in our study by the distinct scaling exponent $\gamma \approx 0.4$ for RNA and $\gamma \approx 1.4$ for proteins with $s < 20$ (Fig. S7). Hence, in the strict sense the chain organizations of both RNA and proteins are not scale-invariant, which is not the case for any real polymer either. Depending on the length scale of interest, a different picture is revealed from real polymer chains. Of note, the new scaling exponent $\gamma = 0.75$ recently discovered for chromatin organization at a resolution (10 kb $\lesssim s \lesssim$ 1 Mb) (26) higher than the previous study ($s \gtrsim 700$ kb) (11) implies that

the self-similarity found at the intermediate resolution ($P(s) \sim s^{-1.08}$) cannot be extended to the internal structure of contact domain.

Long-range contacts from contact map

Contact maps along with the 3D structure offer a more concrete insight into the distinct chain organization of biopolymers with different γ. For instance, the contact maps of 23S-rRNA ($\gamma = 1.11$; Fig. 2 *A*) and FhuA ($\gamma = 1.49$; Fig. 2 *B*) reveal that 23S rRNA has a greater density of long-range contacts than FhuA. Interestingly, in 23S rRNA the modular contact domains made of sequences, spanning $i = 500$–1000 (*magenta*) and 1500–1750 (*orange*) or between $i = 500$–1000 (*magenta*) and 2000–2500 (*cyan*), form extensive interfaces (Fig. 2 *A*). In comparison, FhuA has β-barrel structure with the long-range tertiary contacts formed between the subdomain (*blue*) made of N-terminal sequences ($i = 1$–150) and β-strands ($i = 200$–700) surrounding it (Fig. 2 *B*).

To generalize this finding for RNA and proteins, for each structure we calculated the proportion of long-range contacts (ϕ), between any sites i and j, satisfying $j - i \geq s_{\min}$, as the ratio between the observed number of long-range contacts and the maximum possible number of long-range contacts, i.e., $\phi = \mathcal{N}\sum_{j-i\geq s_{\min}}^{N} \Theta(d_c - |\vec{r}_i - \vec{r}_j|)$, where $\Theta(\bullet\bullet\bullet)$ is the Heaviside step function and the normalization constant $\mathcal{N} = (N - s_{\min} + 1)(N - s_{\min})/2$. The corresponding histograms $p(\phi)$ for RNA and proteins are shown in Fig. 2 *C* with $s_{\min} = 30$. The finding that RNA has $p(\phi)$ distributed to larger ϕ-values than proteins indicates that a significant number of tertiary contacts are used for assembling the secondary structure elements abundant in RNA. This result is robust to the variation of $s_{\min}$ value.

Inter-subchain interactions and surface roughness

To quantify further the distinct chain organization of RNA and proteins, we borrow analytic tools developed in the studies of chromosome organization (22,23). The number of contacts, $M(s)$, that a subchain has with the rest of the structure (see Fig. 3 *A*) (16) scales as $\langle M(s)\rangle \sim s^{\beta_1}$ for both RNA and proteins, where $\langle\bullet\bullet\bullet\rangle$ denotes an average over the chain size frequency (see Materials and Methods). The exponent β_1 is different for RNA ($\beta_1^{\mathrm{RNA}} = 0.9$) and protein ($\beta_1^{\mathrm{prot}} = 0.6$), and $\langle M\rangle$ is greater for RNA when $s \gtrsim 40$, indicating that RNA has more number of inter-subchain contacts for $s \gtrsim 40$. The same conclusion was drawn by computing the roughness of the sub-chain surface (23), which is quantified using $n_s(s)$, the number of monomers in a subchain that are in contact with at least one monomer belonging to other subchains (see Fig. 3 *B*). $\langle n_s(s)\rangle_L \sim s^{\beta_2}$ with $\beta_2^{\mathrm{RNA}} = 0.9 > \beta_2^{\mathrm{prot}} = 0.7$, suggesting that RNAs have rougher sub-chain surfaces. The scaling relationships of the inter-subchain interactions ($M \sim s^{0.9}$) and the surface

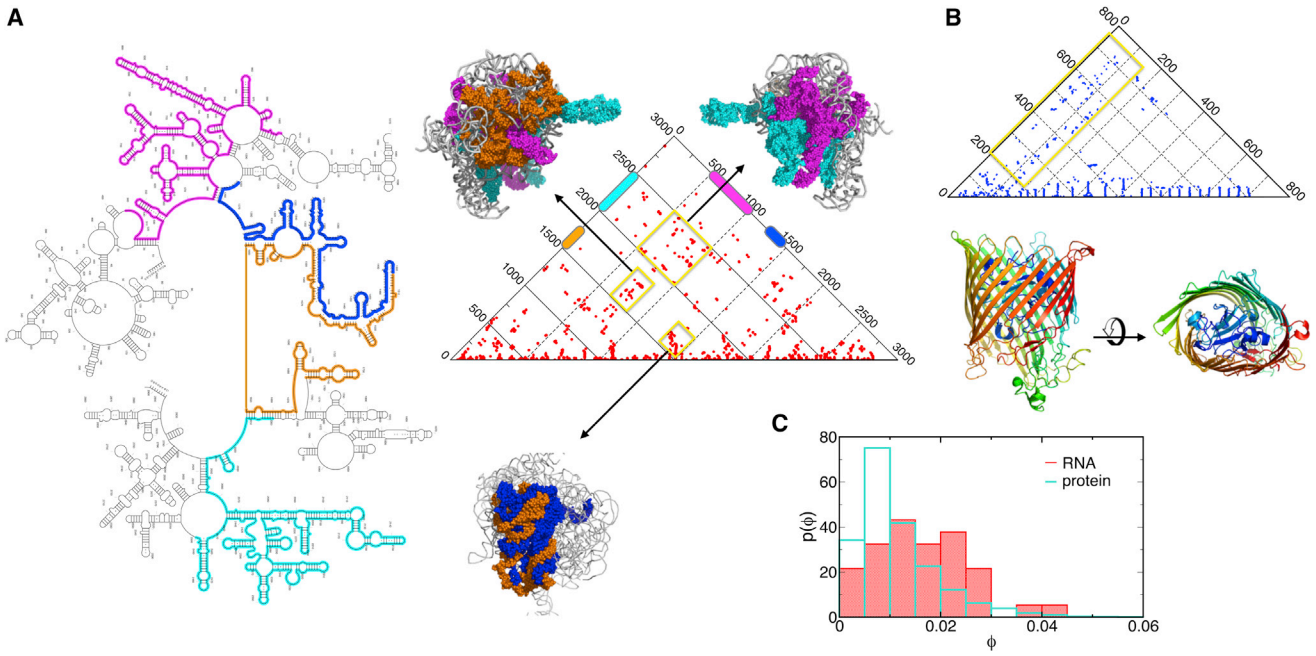

FIGURE 2 Analysis of long-range contacts. (*A* and *B*) Contact maps of 23S rRNA and FhuA, whose $P(s)$ values are provided in Fig. 1. In FhuA, the long-range contacts ($s \gtrsim 100$), enclosed by a yellow box, are formed between the structure made of N-terminal sequences ($i = 1$–150) and surrounding β-strands forming the barrel. In 23S-rRNA, the locations of the clusters of long-range contacts formed at the interfaces between contact domains are highlighted using different colors on each range of sequences along with the 3D structures. (*C*) Histogram of the density of long-range contacts calculated for RNA and protein structures in the PDB. To see this figure in color, go online.

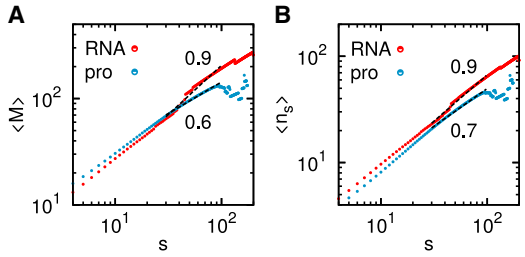

FIGURE 3 Chain-size frequency weighted number of inter-subchain interactions M (A) and the number of surface monomers n_s (B) as a function of subchain size s in RNA (*red*) and proteins (*blue*). To see this figure in color, go online.

monomers ($n_s \sim s^{0.9}$) for RNA compare well with those of crumpled globules (M, $n_s \sim s^1$) (16,23).

The values $\langle M(s)\rangle$ and $\langle n_s(s)\rangle$ are related to each other with $\langle M(s)\rangle \approx Q\langle n_s(s)\rangle$, where $Q \sim s^{\nu d}/s$ is the proportionality constant, the total number of possible monomers ($\sim s^{\nu d}$) that can fill the volume defined by a blob consisting of s monomers, thus giving a scaling relation $\beta_1 = \nu d - 1 + \beta_2$ (23). From this relation and $\beta_{1,2}$, we obtain the Flory exponent $\nu = 1/3$ for native RNA and $\nu = 0.3$ for proteins, which is in perfect agreement with the values of ν obtained from an independent analysis of macromolecular structures in the PDB, $\nu \approx 0.33$ for RNA and $\nu = 0.31$ for proteins in $R_G \sim N^{\nu}$ (6).

Degree of interpenetration and segregation

Next, we calculate the fraction of residues from other subchains found in the ellipsoidal volume enclosing a subchain averaged over all subchains of length s, which corresponds to the degree of interpenetration (DOP) (22). The degree of segregation (DOS), $\mathrm{DOS} = \langle d_{A,B}/(2R_G^{A\cup B})\rangle$, is defined by the ratio between $d_{A,B}$ and $(2R_G^{A\cup B})$, where $d_{A,B}$ is the distance between the center positions of two nonoverlapping subchains A and B, and $R_G^{A\cup B}$ is the gyration radius of the union of these two subchains. DOS is defined by the ratio of these two values ($d_{A,B}$ and $R_G^{A\cup B}$) averaged over all the pairs of subchains A and B with the same length s. DOP and DOS as a function of s for both RNA and proteins (Fig. 4) indicate that while subchains separated by a large arc length s are well separated from each other in RNA, the subchains in RNA penetrate the volume of other subchains more deeply than proteins can. This explains why the decline of $P(s)$ for RNA is slower than for proteins (Fig. 2), which leads to a smaller exponent, γ.

The number of long-range contacts

The total number of contacts over a given range of s, $s_{\min} < s < s_{\max}$ is considered with $P(s) \approx qs^{-\gamma}$: $n_c(N) = \int_{s_{\min}}^{s_{\max}} (N-s)P(s)ds \approx q\int_{s_{\min}}^{s_{\max}} (N-s)s^{-\gamma}ds$, and hence

$$n_c(N)\Big/q \approx N\left(\frac{s_{\max}^{1-\gamma} - s_{\min}^{1-\gamma}}{1-\gamma}\right) + \left(\frac{s_{\max}^{2-\gamma} - s_{\min}^{2-\gamma}}{2-\gamma}\right). \quad (3)$$

Notably, (1) n_c/q scales linearly with N for both RNA and proteins, regardless of γ-value; and (2) the prefactor of n_c/q depends only on γ. For $s_{\min} = 30$ and $s_{\max} = 100$, Eq. 3 leads to $n_c^{\gamma=1.1}(N)/q \approx 0.81N + 46$, and $n_c^{\gamma=1.6}(N)/q \approx 0.11N + 6.0$.

Meanwhile, from the plots of $n_c(N)$ using structures in PDB (see Fig. 5), we obtain

$$\begin{aligned} n_c^{\mathrm{RNA}}(N)\big/q^{\mathrm{RNA}} &\approx 0.77N + 65, \\ n_c^{\mathrm{pro}}(N)\big/q^{\mathrm{pro}} &\approx 0.11N + 4.7, \end{aligned} \quad (4)$$

where the prefactors $q^{\mathrm{RNA}} \approx 0.48$ and $q^{\mathrm{pro}} \approx 4.71$ from the fits to $\langle\overline{P}(s)\rangle$ in Fig. 1 are used. Note that for a given N, $n_c^{\gamma=1.1}(N) > n_c^{\gamma=1.6}(N)$ and $n_c^{\mathrm{RNA}}(N) > n_c^{\mathrm{pro}}(N)$. Together with other quantities, i.e., $\langle\mathrm{DOP}\rangle$, $\langle\mathrm{DOS}\rangle$, $\langle M(s)\rangle$, and $\langle n_s(s)\rangle$, the number of contacts, $n_c(N)$, calculated here persistently assert that RNA has a greater number of long-range contacts than proteins of the same size.

It is of note that the analyses presented in Figs. 3, 4, and 5 are different from investigating each macromolecule one by one (Fig. 1) and finding the structure-function relationship. Given that the ensemble in question is the product of evolution, clarifying the difference between two classes of macromolecules (RNA and proteins) is promising as soon as the evolutionary questions are concerned.

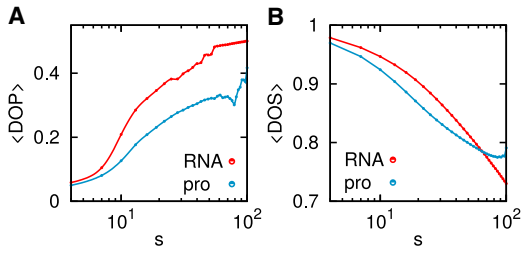

FIGURE 4 The mean (A) DOP and (B) DOS as a function of sub-chain size s in RNA (*red*) and in proteins (*blue*). To see this figure in color, go online.

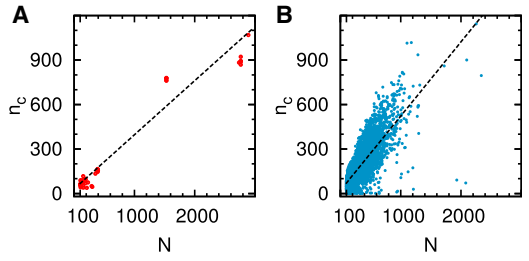

FIGURE 5 Scatter plots of the number of contacts for (A) RNA and (B) proteins over the intermediate range with $s_{\min} = 30$ and $s_{\max} = 100$. To see this figure in color, go online.

DISCUSSION

Due to intramolecular forces stabilizing the chain molecule, both native RNA and protein molecules retain compact and space-filling structures, satisfying $R_G \sim N^{1/3}$ (6,27), which, from the polymer physics perspective, is regarded as the property of polymers in poor solvent conditions. It is, however, critical to note that the size of a subchain surrounded by other subchains should scale as $R(s) \sim s^{1/2}$, which is indeed confirmed for the proteins with $\gamma = 1.5$ (Fig. S5). According to the "Flory theorem" (14,28), a test chain in a fully equilibrated homogeneous semidilute or concentrated polymer melt (29), in spherical confinement (30), or even in globule, is expected to obey the Gaussian statistics because of the screening of excluded volume interaction or counterbalance between attraction and repulsion (14), thus satisfying $R(s) \sim s^{1/2}$ or $P(s) \sim s^{-3/2}$ (see the Supporting Material). The distinct contact probability exponent is highlighted by our analysis that $\gamma \sim 1.0$ for large RNA and $\gamma \sim 1.5$ for small RNA or globular proteins over the intermediate range of $20 \lesssim s \lesssim 100$. Evident from rRNA structure (Fig. 1), subchains of RNA at scales $s > 20$ are assembled into modular contact domains, which are better demarcated in the form of stem-loop helices than proteins, and stitched together through long-range tertiary contacts (Fig. 2 *A*). The evidence of this characteristic architecture of RNA with multimodular domains is visualized vividly in the form of multiple rupture events in single-molecule pulling experiments of Tetrahymena ribozymes (31), while many proteins display a cooperative and effectively all-or-none unfolding under force (32,33).

What causes the crumpled structures of large RNA at the scale of $20 \lesssim s \lesssim 100$? Here, the statistical rarity of knots in native RNA (34,35), which is unparalleled by proteins or DNA (22,36), is worth noting. In general, knots are unavoidable when a long polymer chain ($N \gg N_e \sim 200$, where N_e is the entanglement length (29)) is folded to an equilibrium globule (16,37). Topological knot-free constraints inherent to the ring polymers, however, have been shown to organize melts of unconcatenated polymer rings or a single long polymer ring into crumpled globules, preventing entanglements (23,38). Because large RNA molecules, assembled by a number of secondary structural elements (hairpin loops, stems), resemble a collection of small and large rings, it can be surmised during the folding process, the knot-free constraints are effectively imposed. The knot-free constraints are more likely applied for RNA because the energy scale associated with secondary structure elements ($\varepsilon^{\mathrm{sec}}$), is in general well separated from that of tertiary interactions ($\varepsilon^{\mathrm{ter}}$), such that $\sum_i \varepsilon_i^{\mathrm{sec}} \gg \sum_k \varepsilon_k^{\mathrm{ter}} \gg k_{\mathrm{B}}T$ (39), which makes secondary structure elements independently stable. By contrast, to fold, proteins undergo a reptation-like process, after the initial collapse (40), which may take place with ease because secondary structure elements of proteins (α-helix, β-sheet) are only marginally stable relative to the thermal energy. If necessary, these motifs can be reassembled into thermodynamically more stable structures.

While local and remote contacts are mixed in the folding nuclei of proteins, the formation of secondary structures in RNA folding usually precedes the formation of tertiary contacts, so that the folding of RNA is hierarchical (2,41). Folding under kinetic control produces thermodynamically metastable and kinetically trapped intermediates, which occurs ubiquitously in RNA folding (42), especially in cotranscriptional folding of RNA (43,44). A decision, made at an early stage of folding, involved with the formation of independently stable secondary structure elements is difficult to reverse, although in a worst-case scenario, cofactors such as metal-ions (45,46), metabolites (47), and RNA chaperones (48) still can induce a secondary structure rearrangement. Hence, a more proper way to understand conformational dynamics of a large RNA molecule with $N \gtrsim 100$ is to consider an ensemble of multiple functional states (49–52) instead of a thermodynamically driven, unique native state. It is noteworthy that RNA secondary structure prediction algorithms, which use the strategy of searching the minimum free energy structure (53–55), fail to predict the correct secondary structure when $N \gtrsim 100$, and require the comparative sequence analysis or experimental constraints (56,57). This could be ascribed to the consequence of error accumulated in predicting RNA structures with large N, but it is also suspected that the (free) energy minimization principle cannot be extended to account for the folding process of large RNA. The contact statistics of large RNA, $P(s) \sim s^{-1}$, can be used as an additional constraint or guideline for structure prediction.

A situation analogous to the hierarchical folding of large RNA is prevalent in the two-stage membrane protein folding where the insertion of transmembrane α-helices, guided by translocons, is followed by the postinsertion folding (58,59). We indeed find that the contact probabilities of class A G-protein coupled receptors (GPCRs) give $\gamma \approx 1$ (*blue circles* in the *middle panel* of Fig. 1). Because $\gamma \approx 1$ means the chain organization of native GPCRs is not in entropy-maximum state, a thermodynamically guided, spontaneous in vitro refolding of GPCRs into the native form is expected to be nonpermissible. An atomic force microscopy experiment on an α-helical membrane protein, antiporter ($N \approx 380$), whose γ-value we find is ≈ 1.1, could not be refolded to the original form after mechanically unfolded (60). However, a recent single-molecule force experiment (61) has shown that GlpG, an α-helical transmembrane protein with $N \approx 270$, can reversibly fold in bicelles even after the entire structure including transmembrane helices is disrupted by mechanical forces. Remarkably, we find $\gamma \approx 1.5$ for GlpG. For membrane proteins of known native structures, their γ-values can be used to judge whether or not spontaneous in vitro refolding is possible.

Because the time required for equilibrium sampling of conformations (τ_{eq}) increases exponentially with the system size

(N) as $\tau_{eq} \sim e^N$ (62), signatures of metastability or nonequilibration in chain conformation could be ubiquitous in a macromolecular structure with large N. Through the statistical analysis of structures in PDB, our study puts forward that these forms of crumpled chain organization with $\gamma \approx 1$ of large native RNA and some classes of proteins are an ineluctable outcome of the folding mechanism under kinetic control.

Our results, based on the structures available in PDB, might be fraught with a possible sample bias because the current structural information available in the PDB is limited, underrepresenting intrinsically disordered proteins or membrane proteins for proteins, and long noncoding intron RNA abundant in the cell for RNA (63,64). Nevertheless, our general conclusions on the difference in the organization principle between proteins and RNA will still hold even when the database of PDB is further expanded. Especially, we expect that an inclusion of long noncoding RNA structures ($N > 200$), which should be possible in the near future, will make our conclusions more robust because the hierarchical nature of RNA folding process would become more evident for RNA with larger N and reinforce the territorial (crumpled-like) organization in RNA.

SUPPORTING MATERIAL

Supporting Materials and Methods, seven figures, and three tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(16)30215-6.

AUTHOR CONTRIBUTIONS

L.L. and C.H. designed and performed the research, analyzed the data, and wrote the article.

ACKNOWLEDGMENTS

We thank the Korea Institute for Advanced Study for providing computing resources (KIAS Center for Advanced Computation, Linux Cluster System) for this work.

SUPPORTING CITATIONS

Reference (65) appears in the Supporting Material.

REFERENCES

1. Schuster, P., W. Fontana, …, I. L. Hofacker. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.* 255:279–284.
2. Tinoco, I., Jr., and C. Bustamante. 1999. How RNA folds. *J. Mol. Biol.* 293:271–281.
3. Thirumalai, D., and C. Hyeon. 2005. RNA and protein folding: common themes and variations. *Biochemistry.* 44:4957–4970.
4. Chen, S. J., and K. A. Dill. 2000. RNA folding energy landscapes. *Proc. Natl. Acad. Sci. USA.* 97:646–651.
5. Morcos, F., N. P. Schafer, …, P. G. Wolynes. 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. USA.* 111:12408–12413.
6. Hyeon, C., R. I. Dima, and D. Thirumalai. 2006. Size, shape, and flexibility of RNA structures. *J. Chem. Phys.* 125:194905.
7. Thirumalai, D., N. Lee, …, D. Klimov. 2001. Early events in RNA folding. *Annu. Rev. Phys. Chem.* 52:751–762.
8. Langer-Safer, P. R., M. Levine, and D. C. Ward. 1982. Immunological method for mapping genes on *Drosophila* polytene chromosomes. *Proc. Natl. Acad. Sci. USA.* 79:4381–4385.
9. Cremer, T., and C. Cremer. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* 2:292–301.
10. Dekker, J., K. Rippe, …, N. Kleckner. 2002. Capturing chromosome conformation. *Science.* 295:1306–1311.
11. Lieberman-Aiden, E., N. L. van Berkum, …, J. Dekker. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293.
12. Dekker, J., M. A. Marti-Renom, and L. A. Mirny. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14:390–403.
13. Zuin, J., J. R. Dixon, …, K. S. Wendt. 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA.* 111:996–1001.
14. Grosberg, A. Y., and A. R. Khokhlov. 1994. Statistical Physics of Macromolecules. AIP Press, New York.
15. Grosberg, A., S. Nechaev, and E. Shakhnovich. 1988. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.* 49:2095–2100.
16. Mirny, L. A. 2011. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* 19:37–51.
17. Bohn, M., D. W. Heermann, and R. van Driel. 2007. Random loop model for long polymers. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76:051805.
18. Barbieri, M., M. Chotalia, …, M. Nicodemi. 2012. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA.* 109:16173–16178.
19. Kang, H., Y.-G. Yoon, …, C. Hyeon. 2015. Confinement-induced glassy dynamics in a model for chromosome organization. *Phys. Rev. Lett.* 115:198102.
20. Batey, R. T., R. P. Rambo, and J. A. Doudna. 1999. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed. Engl.* 38:2326–2343.
21. Nissen, P., J. A. Ippolito, …, T. A. Steitz. 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci. USA.* 98:4899–4903.
22. Lua, R. C., and A. Y. Grosberg. 2006. Statistics of knots, geometry of conformations, and evolution of proteins. *PLOS Comput. Biol.* 2:e45.
23. Halverson, J. D., J. Smrek, …, A. Y. Grosberg. 2014. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.* 77:022601.
24. Berezovsky, I. N., A. Y. Grosberg, and E. N. Trifonov. 2000. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.* 466:283–286.
25. Huang, W., V. V. Lunin, …, M. Cygler. 2003. Crystal structure of *Proteus vulgaris* chondroitin sulfate ABC lyase I at 1.9Å resolution. *J. Mol. Biol.* 328:623–634.
26. Sanborn, A. L., S. S. Rao, …, E. L. Aiden. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA.* 112:E6456–E6465.
27. Flory, P. J. 1969. Statistical Mechanics of Chain Molecules. Interscience Publishers, New York.
28. Flory, P. J. 1949. The configuration of real polymer chains. *J. Chem. Phys.* 17:303–310.
29. de Gennes, P. G. 1979. Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca, NJ.
30. Cacciuto, A., and E. Luijten. 2006. Self-avoiding flexible polymers under spherical confinement. *Nano Lett.* 6:901–905.

31. Onoa, B., S. Dumont, …, C. Bustamante. 2003. Identifying kinetic barriers to mechanical unfolding of the *T. thermophila* ribozyme. *Science.* 299:1892–1895.

32. Shank, E. A., C. Cecconi, …, C. Bustamante. 2010. The folding cooperativity of a protein is controlled by its chain topology. *Nature.* 465:637–640.

33. Mickler, M., R. I. Dima, …, M. Rief. 2007. Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations. *Proc. Natl. Acad. Sci. USA.* 104:20268–20273.

34. Micheletti, C., M. Di Stefano, and H. Orland. 2015. Absence of knots in known RNA structures. *Proc. Natl. Acad. Sci. USA.* 112:2052–2057.

35. Burton, A. S., M. Di Stefano, …, C. Micheletti. 2016. The elusive quest for RNA knots. *RNA Biol.* 13:134–139.

36. Noel, J. K., J. I. Sułkowska, and J. N. Onuchic. 2010. Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Natl. Acad. Sci. USA.* 107:15403–15408.

37. Grosberg, A. Y. 2000. Critical exponents for random knots. *Phys. Rev. Lett.* 85:3858–3861.

38. Imakaev, M. V., K. M. Tchourine, …, L. A. Mirny. 2015. Effects of topological constraints on globular polymers. *Soft Matter.* 11:665–671.

39. Thirumalai, D., and C. Hyeon. 2008. Theory of RNA folding: from hairpins to ribozymes. *In* Non-Protein Coding RNAs. Springer, New York.

40. Thirumalai, D. 1995. From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys. I (Fr.).* 5:1457–1467.

41. Greenleaf, W. J., K. L. Frieda, …, S. M. Block. 2008. Direct observation of hierarchical folding in single riboswitch aptamers. *Science.* 319:630–633.

42. Treiber, D. K., and J. R. Williamson. 2001. Beyond kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* 11:309–314.

43. Repsilber, D., S. Wiese, …, G. Steger. 1999. Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (−)-stranded RNA by temperature-gradient gel electrophoresis. *RNA.* 5:574–584.

44. Lutz, B., M. Faber, …, A. Schug. 2014. Differences between cotranscriptional and free riboswitch folding. *Nucleic Acids Res.* 42:2687–2696.

45. Wu, M., and I. Tinoco, Jr. 1998. RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci. USA.* 95:11555–11560.

46. Koculi, E., S. S. Cho, …, S. A. Woodson. 2012. Folding path of P5abc RNA involves direct coupling of secondary and tertiary structures. *Nucleic Acids Res.* 40:8011–8020.

47. Montange, R. K., and R. T. Batey. 2008. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* 37:117–133.

48. Russell, R., I. Jarmoskaite, and A. M. Lambowitz. 2013. Toward a molecular understanding of RNA remodeling by DEAD-box proteins. *RNA Biol.* 10:44–55.

49. Al-Hashimi, H. M., and N. G. Walter. 2008. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* 18:321–329.

50. Solomatin, S. V., M. Greenfeld, …, D. Herschlag. 2010. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature.* 463:681–684.

51. Hyeon, C., J. Lee, …, D. Thirumalai. 2012. Hidden complexity in the isomerization dynamics of Holliday junctions. *Nat. Chem.* 4:907–914.

52. Hyeon, C., M. Hinczewski, and D. Thirumalai. 2014. Evidence of disorder in biological molecules from single molecule pulling experiments. *Phys. Rev. Lett.* 112:138101.

53. Rivas, E., and S. R. Eddy. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053–2068.

54. Hofacker, I. L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* 31:3429–3431.

55. Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

56. Gutell, R. R., J. C. Lee, and J. J. Cannone. 2002. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12:301–310.

57. Mathews, D. H., J. Sabina, …, D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

58. Popot, J.-L., S.-E. Gerchman, and D. M. Engelman. 1987. Refolding of bacteriorhodopsin in lipid bilayers. A thermodynamically controlled two-stage process. *J. Mol. Biol.* 198:655–676.

59. Bowie, J. U. 2005. Solving the membrane protein folding problem. *Nature.* 438:581–589.

60. Kedrov, A., C. Ziegler, …, D. J. Müller. 2004. Controlled unfolding and refolding of a single sodium-proton antiporter using atomic force microscopy. *J. Mol. Biol.* 340:1143–1152.

61. Min, D., R. E. Jefferson, …, T.-Y. Yoon. 2015. Mapping the energy landscape for second-stage folding of a single membrane protein. *Nat. Chem. Biol.* 11:981–987.

62. Palmer, R. 1982. Broken ergodicity. *Adv. Phys.* 31:669–735.

63. Rinn, J. L., and H. Y. Chang. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81:145–166.

64. Carninci, P., T. Kasukawa, …, Y. Hayashizaki. 2005. The transcriptional landscape of the mammalian genome. *Science.* 309:1559–1563.

65. Friedman, B., and B. O'Shaughnessy. 1991. Short time behavior and universal relations in polymer cyclization. *J. Phys. II.* 1:471–486.

66. Pyetan, E., D. Baram, …, A. Yonath. 2007. Chemical parameters influencing fine-tuning in the binding of macrolide antibiotics to the ribosomal tunnel. *Pure Appl. Chem.* 79:955–968.

67. Guo, Q., Y. Yuan, …, N. Gao. 2011. Structural basis for the function of a small GTPase RsgA on the 30S ribosomal subunit maturation revealed by cryoelectron microscopy. *Proc. Natl. Acad. Sci. USA.* 108:13100–13105.

68. Golden, B. L., H. Kim, and E. Chase. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat. Struct. Mol. Biol.* 12:82–89.

69. Krasilnikov, A. S., Y. Xiao, …, A. Mondragón. 2004. Basis for structural diversity in homologous RNAs. *Science.* 306:104–107.

70. Thore, S., C. Frick, and N. Ban. 2008. Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. *J. Am. Chem. Soc.* 130:8116–8117.

71. Ferguson, A. D., V. Braun, …, W. Welte. 2000. Crystal structure of the antibiotic albomycin in complex with the outer membrane transporter FhuA. *Protein Sci.* 9:956–963.

72. Otterbein, L. R., P. Graceffa, and R. Dominguez. 2001. The crystal structure of uncomplexed actin in the ADP state. *Science.* 293:708–711.

73. McLuskey, K., J. Rudolf, …, J. C. Mottram. 2012. Crystal structure of a *Trypanosoma brucei* metacaspase. *Proc. Natl. Acad. Sci. USA.* 109:7469–7474.

74. Ormö, M., A. B. Cubitt, …, S. J. Remington. 1996. Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science.* 273:1392–1395.

75. Weaver, L. H., and B. W. Matthews. 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* 193:189–199.