Biophysical Journal, Volume *110*

Supplemental Information

Contact Statistics Highlight Distinct Organizing Principles of Proteins and RNA

Lei Liu and Changbong Hyeon

Supporting Material : Contact statistics highlight distinct organizing principles of proteins and RNA

Lei Liu and Changbong Hyeon[1]

[1]*School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Republic of Korea*

EXTRACTION OF SCALING EXPONENT γ

We obtained the contact probability exponent γ by conducting linear regression on a part of $P(s)$ data that behave as $\sim s^{-\gamma}$ in log-log scale. There are two factors that may affect in determination of γ: (i) d_c, the cut-off distance to define a contact between two residues, affects the overall shape of $P(s)$; (ii) The range of s, $s_{\min} < s < s_{\max}$, to be fitted. Instead of manually tuning the fitting range ($s_{\min} < s < s_{\max}$), we defined a parameter φ ($0 < \varphi < 1$), such that the proportion of fitting range, $(s_{\max} - s_{\min})/N$ where N is the chain length, is at least greater than an allocated threshold value, φ. For instance, if φ is set to 0.3 then the fit is made on more than 30 % of the entire data points. Thus, by fitting $P(s)$ data over all possible pairs of $s_{\min}$ and $s_{\max}$ values which define the range of $(s_{\min}, s_{\max})$ satisfying $(s_{\max} - s_{\min})/N \geq \varphi$, we determine the value of γ from the best fit which gives the smallest standard error relative to the data points.

Fig. S1A shows that the shape of $P(s)$ for 23S-rRNA calculated with different d_c remains effectively identical, giving rise to a similar value of γ: $\gamma = 1.11$ ($d_c = 4$ Å), 1.06 ($d_c = 5$ Å). $p(\gamma)$s for RNA molecules obtained from different d_c are also similar as shown in Fig. S1B.

Next, to study the effect of φ on γ, we set $d_c = 4$ Å and change the value of φ in the fit. We obtain $\gamma = 1.11$ for $\varphi = 0.3$, and $\gamma = 1.01$ with $\varphi = 0.4$ (see Fig. S2A). Fig. S2B also shows that $p(\gamma)$ with different φ are comparable. Analysis applied to protein shows similar results. A series of comparisons in Figs.S1 and S2 indicate that the average value of γ is insensitive to the parameters around the value we have chosen.

In addition, the overall shapes of $p(\gamma)$ and $\langle \overline{P}(s) \rangle$ are insensitive to the two threshold values of sequence similarity (90 and 30 %), which we imposed to select a set of non-homologous proteins (Fig. S3).

We analyzed 186 RNA and 16633 individual proteins whose size satisfies $N \geq 50$, available in PDB as of September 2015. Distributions of γ obtained from the optimal linear fittings on $\log_{10} P(s)$ versus $\log_{10} s$ with a correlation coefficient greater than 0.9 are presented in Fig.1A with $\varphi = 0.3$, $d_c = 4$ Å for both RNAs and proteins. To highlight the robustness of our result presented in Fig.1A (γ vs. N plot), we specified the 95 % confidence interval of γ values using error-bar to each data point in Fig. S4.

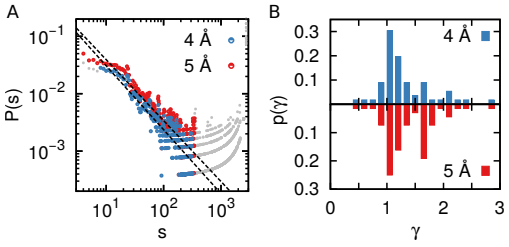

FIG. S1: (A) The contact probability versus sequence distance of 23S rRNA (PDB entry 2O45) with a cut-off distance of contacting d_c of value 4 Å (blue) and 5 Å (red). (B) Distributions of γ in RNA monomers with d_c of 4 Å and 5 Å.

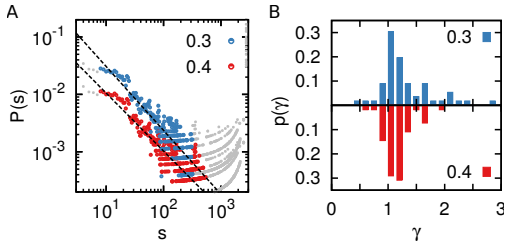

FIG. S2: (A) The contact probability versus sequence distance of 23S rRNA (PDB entry 2O45) with a minimum fraction of all data points used for fitting φ of 0.3 (blue) and 0.4 (red). The data points for $\varphi = 0.4$, as well as the fitted dashed line, are shifted downwards for visual comparison. (B) Distributions of γ in RNA monomers of φ 0.3 and 0.4.

CONTACT PROBABILITY BETWEEN TWO SITES OF A POLYMER

In general, the contact probability of two sites in polymer chain is determined by the volume available for the subchain ending with the two sites, $[R(s)]^d$,

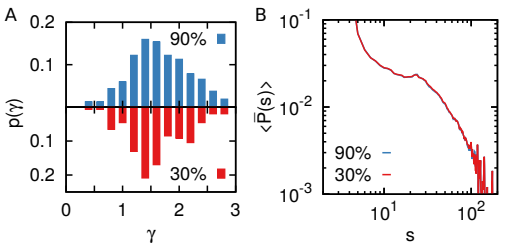

FIG. S3: Effects of imposing different threshold value for the sequence similarity of 90 % and 30 % to the protein structure database to compute $p(\gamma)$ and $\langle \overline{P}(s) \rangle$. No qualitative difference is found in the results.

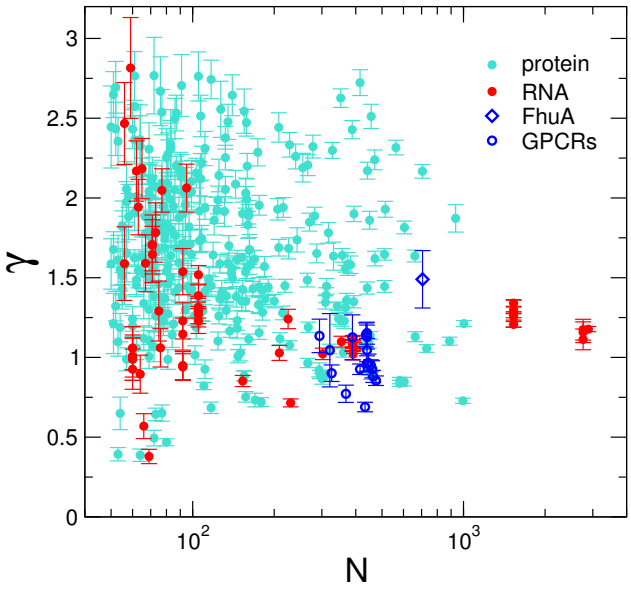

FIG. S4: Scatter plot of γ versus N for RNA (red) and proteins (cyan) with error bars (95 % confidence interval) for γ values.

with normalization condition $\int P_s(r) d^d r = 1$ [14, 29]:

$$P_s(r) = \frac{1}{R(s)^d} \varphi\left(\frac{r}{R(s)}\right) \xrightarrow{r \ll R(s)} P(s) = \frac{1}{R(s)^d} \left(\frac{r}{R(s)}\right)^g, \quad \text{(S1)}$$

where r is the contact distance, $R(s)$ is the size of polymer made of s monomers, d is the dimensionality, and g is the correlation hole exponent. With $R(s) \sim s^\nu$ (see Fig. S5), we obtain the scaling relationship of contact probability, $P(s) \sim s^{-\nu(d+g)}$.

(i) When the excluded volume interaction is fully screened, a test chain (or subchain over a certain length) is ideal. In this case, $\varphi(x) \sim e^{-3x^2/2}$. Thus, the correlation hole exponent $g = 0$ [65] and $R \sim s^\nu$ with $\nu = 1/2$, which leads to $P(s) \sim s^{-\nu d} \sim s^{-3/2}$.

(ii) If the chain adopts an *effectively homogeneous* space-filling configuration, but the interaction between monomers is weak and the excluded volume interaction is still fully screened as in a concentrated melt, then $g = 0$, $d = 3$, and $\nu = 1/3$, which leads to $P(s) \sim s^{-1}$.

(iii) If the chain organization is *inhomogeneous* leading to an anisotropic arrangement because of strong monomer-monomer interactions [26], which for the case of RNA leads to formation of independently stable helices, then $R(s)$ still satisfies $R(s) \sim s^{1/3}$ but the effective dimensionality of the sampling space (d_{eff}) would be less than 3. Thus, $P(s) \sim s^{-d_{\text{eff}}/3}$, and $\gamma = d_{\text{eff}}/3 < 1$, which accounts for the contact probability exponent smaller than 1.

(iv) Note that when the subchain interactions (repulsion and attraction) are screened ($g = 0$), $P(s)$ and $R(s)$ are related as $P(s) \sim R(s)^{-d}$. This relationship particularly holds good for intermediate range of s: $P(s) \sim s^{-3/2} \leftrightarrow R(s) \sim s^{1/2}$ (ideal chain) and $P(s) \sim s^{-1} \leftrightarrow R(s) \sim s^{1/3}$ (crumpled chain) (see Fig. S5). The scaling exponent of 3/5 at $s < 10$ in Fig. S5 is due to the volume exclusion interaction at short range s.

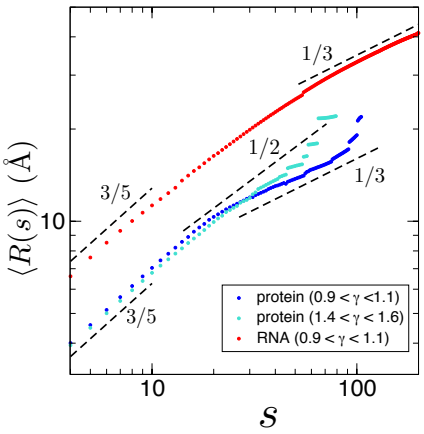

FIG. S5: Mean radius of gyration of subchain as a function of subchain length s for proteins and RNA that display contact probability exponent in a specified range of γ. The structures in the specified range of γ were collected from Fig. 1 and their $R(s)$s were calculated.

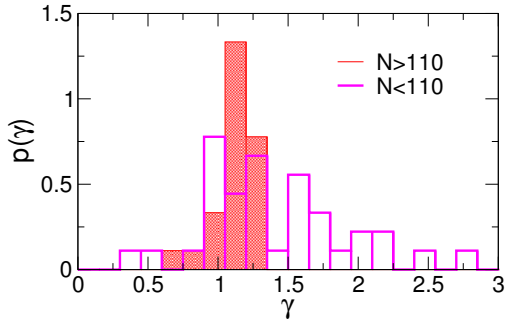

FIG. S6: Distribution of γ value for RNA with $N > 110$ and $N < 110$. $\gamma^{\rm RNA}_{N>110} = 1.12 \pm 0.14$ and $\gamma^{\rm RNA}_{N<110} = 1.41 \pm 0.53$.

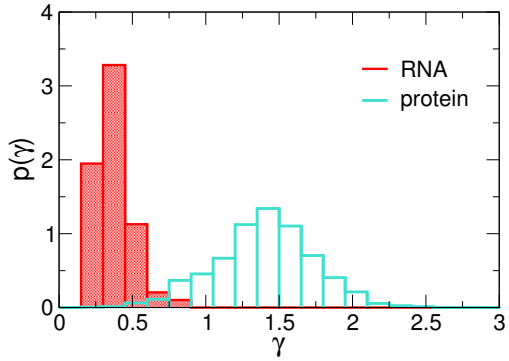

FIG. S7: Distribution of contact probability exponent calculated for the short range of s, $s < 20$. $\gamma^{\rm RNA} = 0.38 \pm 0.13$ and $\gamma^{\rm pro} = 1.40 \pm 0.33$.

2M58	2MIY	1FIR	6TNA	1EHZ	1TRA	4TRA	1TN1
1TN2	3TRA	2TRA	3BBV	1VTQ	4PQV	3A3A	3CW6
2HOP	1I9V	3L0U	2K4C	3D2G	4NYD	2HOM	3GX6
2GIS	3GX2	3IQN	4B5R	2YDH	4RZD	3F2Q	3F2W
3F30	3F2X	3F2T	3F2Y	1U9S	3DHS	1Y0Q	4C4Q
2A2E	3BWP	4FAX	4E8P	4E8R	4E8Q	4E8N	4DS6
4E8M	4FAQ	3J2B	3J2H	3J2D	2YKR	3J28	3J2A
2O45	2O43	2O44	1C2W				

TABLE I: PDB entries of RNA analyzed in Fig. 1.

2MGW	2JY5	2CR8	2RRU	2KAK	2DAH	2EPS	1JJR
1KMX	2KQB	1YSM	2ECM	2KMU	1KFT	2KPI	2M8E
2K2T	2REL	2YSD	2L4E	2MWR	2YRG	3GOH	1Z60
2KKJ	1A7I	1VYX	2M2F	2JXD	2DAL	3WIT	2M9W
2YSJ	1UEO	1AA3	4A3N	1WG2	2D8U	1WFH	1HYI
1BW5	2DZL	1X4P	1VFY	1X4W	1HTA	1SF0	1H0Z
2EA6	2MFK	2DI0	2EWT	2RMR	3H33	1RIY	4TXA
2DA7	2LGW	2JVG	1X61	1WEE	1X4K	2DJB	4P3V
2CT5	2LEK	2HI3	1G33	2EP4	1NEQ	1APJ	1WFP
2JXW	2KW9	1SIG	2M4G	2LT1	1WYS	1X68	2ENN
2E6S	2D9H	2ECT	1E4U	1JQ0	1J3C	1MJ4	4U12
2MLB	1UHC	2CR7	1KDU	1QRY	1X3H	2CSY	2ECL
1RWJ	2LDR	4CIK	3J0R	1UHA	4EIF	1X63	2DOE
2LQL	1CC5	1XFE	2L0S	3CP1	3ZJ1	3BT4	2LRQ
1IPG	2Q18	4IYL	2ECW	2LV2	1LMJ	1ABA	1C9F
1F1F	2CT2	1C6R	1FP0	2KW1	4GPS	1CTJ	2M5W
1Y02	2D8Y	2E6R	1WEO	2CS3	1FBR	2LGX	2LGP
2MIQ	1SJ6	1WIA	2JSN	2DMD	2VTK	3PO8	1OPC
2YRE	2LGV	1T1D	3H6N	1JHG	4BGC	2OA4	2CQK
2CTK	3GCE	2K4J	3DQY	1X0T	2JVL	1HKF	2CS8
3O8V	3DVI	2CTW	4EEU	2MLK	1ZOX	2XXC	2EO3
4TVM	2IVW	2LW4	4HWM	2KQR	2JXN	2HC5	1T6A
4ZBH	1UJX	2MMZ	2LHT	1JUG	2RA9	2XWS	1G3P
2QYZ	2FYG	3O5E	2ES0	4NAZ	3E2I	1DQG	1VSR
1KQW	1E29	2FVV	3W9K	1NL1	1WK0	1XN5	2IN0
2NWF	2L5Q	2P0B	2MO5	3ZUI	2HNA	2JY9	4MYM
3N9D	2N48	4M4Z	3FME	1ENV	2D37	2XB3	1ZND
4GNY	4LD1	3UF4	1D7P	1EW3	3OUQ	1E88	2LFU
2KIG	2KFU	1KLO	3NZM	2M47	4JHG	1RL6	3TXO
2LZM	2NN5	3W9R	2CP6	4F47	1EH6	1CDY	2R6V
3K21	3WJT	1WV3	4M6T	2D5M	3KBG	1J3G	1EJE
1JM1	3TFM	4QA8	1HXN	4E1B	4IT3	4JZC	1EMA
2K18	3HBK	3NO3	4PQ0	2PNN	1LVA	3LTI	4JS8
4DWO	2A1L	4NW4	3V75	5BN7	1DUW	3JRP	2QLU
2LQW	4X36	2HES	4GGC	4GGA	4V16	4AA8	2FGQ
4AF8	1VPR	2PMN	2XE1	2ASI	2ZYL	1T6E	3BA0
1J6Z	4QDC	4GQ1	1FEP	3GRE	4UQE	4MSX	3R1K
3ACP	2DH2	4COT	3DWO	1QCF	1FMK	1W52	1DQ3
1G0D	3K5W	2OBD	4NOX	4FWW	2E84	1Z1N	4AW7
1XEZ	4TLW	1PI6	4UMW	4BBJ	3OKT	1QFG	4MHC
2OAJ	4UP5	1HN0	3KLK				

TABLE II: PDB entries of proteins analyzed in Fig. 1.

2JX9	1ISR	2LNL	2RH1	2YDV	2ZIY	3C9L	3EHS
3EML	3N94	3OE6	3RZE	3UON	3V2Y	3VW7	4DKL
4EJ4	4F11	4IB4	1QJQ				

TABLE III: PDB entries of GPCRs analyzed in Fig. 1.