The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA

Kevin Leu, Benedikt Obermayer, Sudha Rajamani, Ulrich Gerland, Irene A. Chen

Supporting Text

Text S1. The relationship between W and the relative error ratio ϕ/ϕ_0

The relative error ratio, i.e., the ratio of the experimental error fraction ϕ to the thermodynamic bound ϕ_0, should increase with W, as the reaction is further from thermodynamic equilibrium with greater W. One way to quantitatively test this relationship is to plot the error ratios ϕ/ϕ_0 against the chemical reaction rates for the 12 mis-incorporation reactions in all four systems. Here, the experimental reaction rates are only crude proxies for the W rates, since all these reactions are not necessarily related to one another, and a small rate compared to other reactions does not necessarily imply closeness to equilibrium. Nevertheless, the plot (Figure S4) reveals a weak positive correlation (r^2=0.11, p=0.019), indicating that, on average, our thermodynamic limits are slightly better approached for the slower processes.

Text S2. Issues in theoretical estimation of experimental error rate

The theoretical estimates of equilibrium thermodynamics do not reflect the kinetics of the experimental situation, where only one base is incorporated at a time and the use of activated nucleotides renders the incorporation essentially irreversible rather than an equilibrium process. However, we expect that evaluating the stability of the extended complexes would give a reliable lower bound for the thermodynamics of a single base. In particular, the equilibrium probability of incorporating a certain nucleotide in a larger complex, which determines the lower bound on the error rate, results from the statistical sampling of all energetically favorable and unfavorable

configurations, and thus should account for interactions with both 5' and 3' neighboring nucleotides. Our theoretical estimates are therefore based on extended complexes.

To accurately predict the actual error rates in an experimental system, rather than the lower bound of the error rate, we would need to account for the microscopic hybridization kinetics, which are unknown. Modeling the microscopic kinetics is beyond the scope of the nearest neighbor model, whose energy parameters are obtained as 'best-fit' values for modeling the measured thermodynamics of an ensemble of polynucleotide complexes. Even the correct treatment of dangling ends, such as those resulting from single mis-incorporations in the experiment, and precise values for the relevant energy contributions are unresolved issues in the recent literature [1,2].

Text S3. Possible optimization of nucleotide supply in mammalian cells

Tuning the nucleotide supply to minimize errors of an exonuclease-deficient human mitochondrial polymerase [3] would entail a similar depletion in G (roughly 33% A, 40% C, 8% G, 20% T) as what we calculate (Figure 4E,F). Although one might expect the nucleotide supply in replicating cells to mirror the genome composition, local dNTP levels are not saturating due to on-the-fly synthesis or accumulation. Enzymatic mutation rates are sometimes smaller than the thermodynamic bounds, possibly due to exclusion of water through conformational switching or other thermodynamically uphill processes [4]. Nevertheless, perhaps base-pairing energetics as we have calculated them also influence the mutation spectrum of polymerase enzymes.

Text S4. Synthesis of activated nucleotides.

In brief, the free acid of the corresponding monophosphate was suspended with imidazole and 2,2'-dithiodipyridine, to which triethylamine and triphenylphosphine were added. The mixture was stirred at room temperature for ~4 h. The resulting solution was precipitated on ice in ether/acetone/TEA with sodium perchlorate. The precipitate was filtered, washed, and dried overnight to give the phosphorimidazolide sodium salt.

Text S5. MALDI-TOF of oligonucleotides.

1 μL containing ~200 pmol of oligonucleotide was mixed with 1 μL of a matrix solution (2:1 mixture of 52.5 mg/mL 3-hydroxypicolinic acid in 50% acetontrile and 0.1 M ammonium citrate in water), and the mixture was spotted onto a stainless steel MALDI-TOF plate and analyzed in positive mode.

Text S6. Oligonucleotide sequences

Primer sequences:

DNA primer: 5′ Cy3 - GG GAT TAA TAC GAC TCA CTG-NH_2

RNA primer: 5' Cy3 - GG GAU UAA UAC GAC UCA CUG-NH_2

Template sequences:

DNA A: 5′ AGT GAT CTA CAG TGA GTC GTA TTA ATC CC

DNA T: 5′ AGT GAT CTT CAG TGA GTC GTA TTA ATC CC

DNA G: 5′ AGT GAT CTG CAG TGA GTC GTA TTA ATC CC

DNA C: 5′ AGT GAT CTC CAG TGA GTC GTA TTA ATC CC

RNA A: 5′ AGU GAU CUA CAG UGA GUC GUA UUA AUC CC

RNA U: 5′ AGU GAU CUU CAG UGA GUC GUA UUA AUC CC

RNA G: 5′ AGU GAU CUG CAG UGA GUC GUA UUA AUC CC

RNA C: 5′ AGU GAU CUC CAG UGA GUC GUA UUA AUC CC

Excess primers:

DNA excess primer: 5′ GG GAT TAA TAC GAC TCA CTG

RNA excess primer: 5’ GG GAU UAA UAC GAC UCA CUG

Text S7. Non-enzymatic polymerization reactions.

The three systems reported here had pairings of RNA template and RNA primer (RNAt/RNAp), RNA template and DNA primer (RNAt/DNAp), or DNA template and RNA primer (DNAt/RNAp). Primer (0.325 μM) and template (1.3 μM) were mixed in water, incubated at 95 ^{o}C for 5 min, and cooled to room temperature for 5 to 7 min. After annealing, Tris (final 100 mM) and NaCl (final 200 mM) were added to the solution. The reaction was begun by adding ImpdN or ImpN for a final volume of 10 μL. For a given time point, 1 μL of the reaction was quenched with 9 μL of loading buffer containing 8 M urea, 100 mM EDTA, and 1.3 μM of a competitor DNA (for DNAp) or RNA (for RNAp) excess primer. The samples were heated to 95 °C for 3 min to disrupt primer–template complexes and run on 20% urea-PAGE.

The gels were scanned using a Typhoon TRIO variable-mode imager (Piscataway, NJ) with excitation wavelength at 532 nm and emission wavelength at 580 nm, and the scans were analyzed with ImageQuant v5.2 software. The fraction of unreacted primer was calculated by dividing the intensity of the unreacted primer band by the sum of intensities of the unreacted and reacted primer. Initial rates were estimated by a linear fit to the first several data points.

Text S8. Thermodynamic estimate of lower limit of error rate.

Our approach for estimating lower bounds on error rates is based on established thermodynamic arguments [5]. Following reaction equation (1) in the main text, the initial rate of formation of *c'* is $r_n = [n_0]/K_n$, where $K_n = (W_n + k_{n,off})/k_{n,on}$ and n_0 is the nucleotide concentration in solution. The analogous equation can be written for r_m. The error rate per site ($\mu_{n,,m}$) for incorporation of *m* instead of *n* is $r_m/(r_n + r_m)$, so in a solution containing equal nucleotide concentrations, we find the following lower limit $\mu^0_{n,m}$:

$$\mu_{n,m} \geq \mu^0{}_{n,m} = \left[1 + \frac{k_{n,on}}{k_{m,on}} \frac{k_{m,off}}{k_{n,off}}\right]^{-1} \approx e^{-\Delta G_{n,m}/RT},$$

where $\Delta G_{n,m}$ is the free energy difference between the correct and incorrect products.

Text S9. Optimization of nucleotide supply

During optimization, relative, not absolute, concentrations determine fidelity, so only three concentrations are allowed to vary. Optimization was performed using Mathematica. The reaction rates are assumed to be first order with respect to nucleotide concentration. For example, to optimize the nucleotide supply using the experimentally determined rates, if the incorporation (and mis-incorporation) reactions have the rates $x_{a,b}$, where *a* is the template base and *b* is the incoming nucleotide, then the rate constants are $x_{a,b}/C$, where *C* is the experimental concentration of the incoming nucleotide, and the reaction rate at monomer concentration *c* is $cx_{a,b}/C$. The mutation rate is the average of the mis-incorporation frequencies, calculated as described in the Methods. The mutation rate is minimized with respect to the monomer concentrations, where [A] is set to 1 and all others ([C], [G], [T] or [U]) are allowed to vary. One concentration may be arbitrarily fixed because multiplying all concentrations by the same factor does not change the

fidelity, i.e., there are only 3 independent concentrations. We also apply the constraint that no two monomers can have concentrations differing by more than a factor of 10, in order to avoid very large discrepancies among the concentrations. The minimum is found by the Mathematica command *FindMinimum*. The concentrations are then converted to percentage composition of the nucleotide supply by dividing by the sum of the concentrations.

Text S10. Estimation of unknown energy terms for RNA/DNA hybrids

Some energy parameters for stacks involving wobble pairs and mismatch energies are not known, so we estimated these values as the average of the pure RNA and DNA values. This appeared to be a reasonable approximation, as verified by known stacking energies and for the mismatches with different flanking base pairs that have been measured [6,7].

Supporting Figures

Figure S1. Examples of PAGE analysis of polymerization reactions. Particularly slow reactions are shown in panels B-E. Time points in hours are labeled above each lane. The original primer (n) and single extension product (n+1) are labeled; impurities in the primer sometimes cause minor faster migrating bands. (A) Incorporation of ImpG across from template base C in the RNAt/RNAp system. (B) Mis-incorporation of ImpC across template base C in the DNAt/RNAp system. (C) Mis-incorporation of ImpA across template base C in the RNAt/RNAp system. (D) Mis-incorporation of ImpdA across from template base C in the RNAt/DNAp system. (E) Mis-incorporation of ImpdT across from template base C in the RNAt/DNAp system.

Figure S1 continued

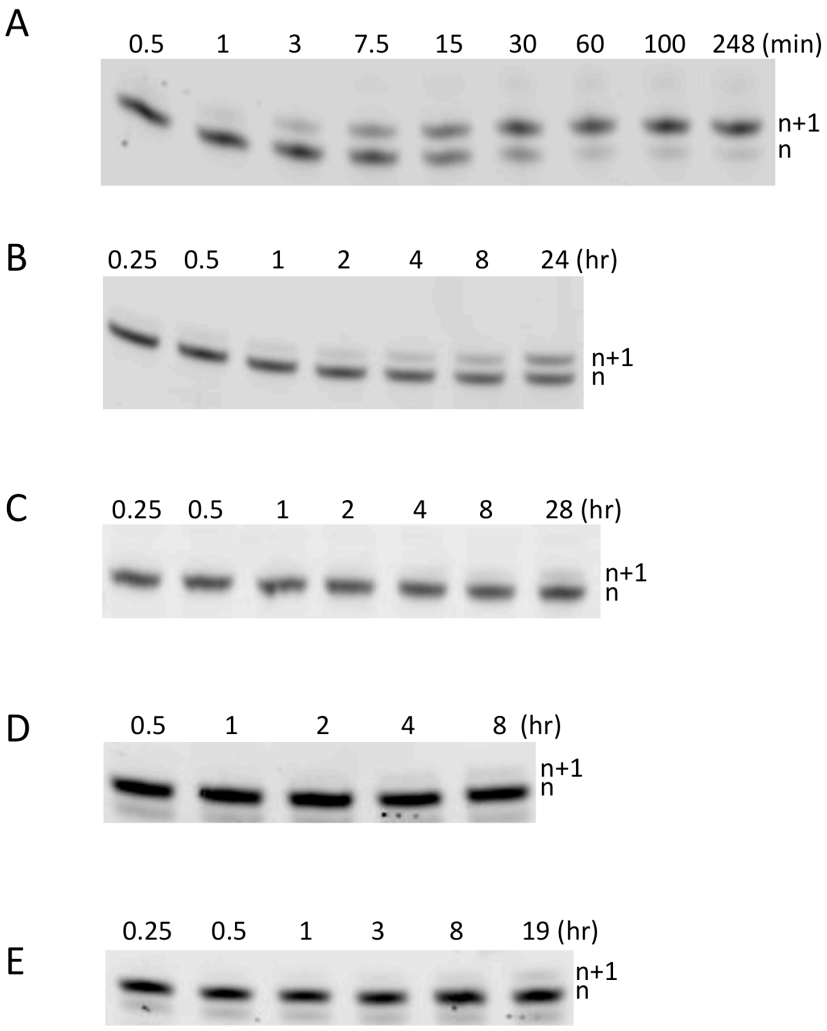


Figure S2. Correlation between RNA replication incorporation and mis-incorporation rates in non-enzymatic vs. previously described ribozyme-catalyzed replication [8]. Frequencies shown for the non-enzymatic system were calculated for equimolar nucleotide concentrations for comparison to the ribozyme system (measured under equimolar conditions); $r^2 = 0.75$ (line: linear regression on log values).

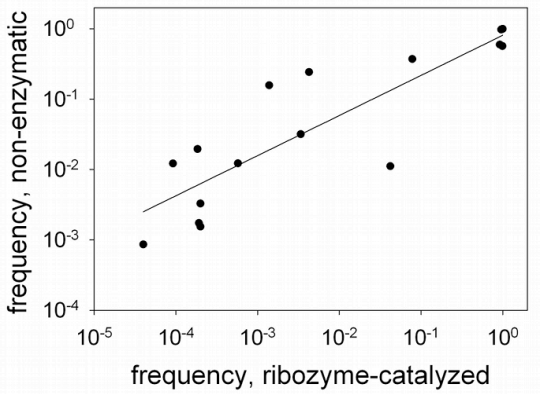


Figure S3. Phenotypic look-ahead effect from relatively high transcriptional error. For a given genotypic fitness landscape (green lines), a high transcriptional error rate (A) effectively smoothens the fitness landscape (red line) for a given genotype (green dot) producing phenotypic variants (red dots) compared to a low error rate (B).

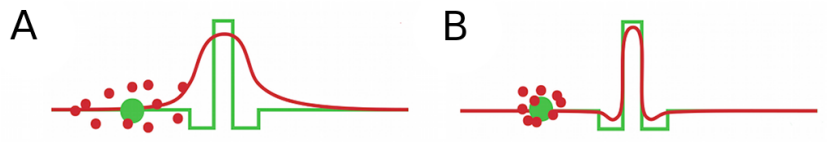
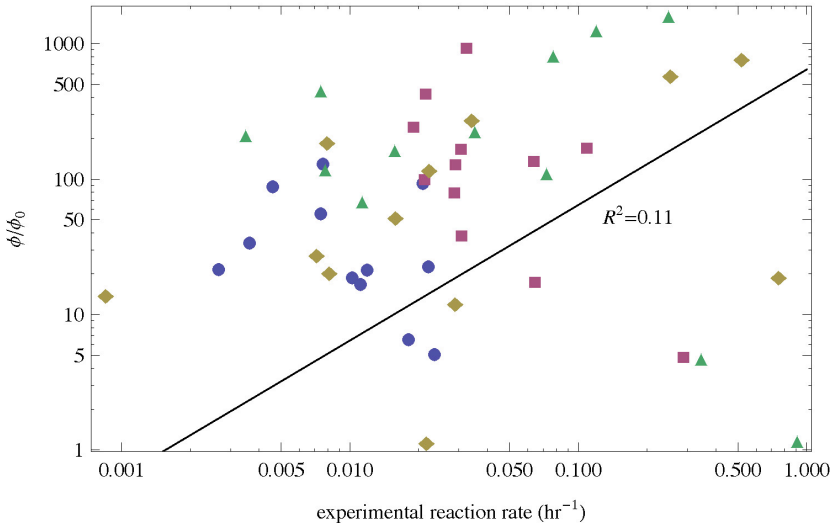


Figure S4. Relative error ratio vs. experimental reaction rates for mis-incorporation. Blue dots are DNA replication, green triangles are RNA replication, yellow diamonds are RNAt/DNAp and red squares are DNAt/RNAp. Line is from linear regression.



Supporting Information References

(1) Jost, D.; Everaers, R. *J Chem Phys* **2010**, *132*, 095101.

(2) Ohmichi, T.; Nakano, S.; Miyoshi, D.; Sugimoto, N. *J Am Chem Soc* **2002**, *124*, 10367.

(3) Lee, H. R.; Johnson, K. A. *J Biol Chem* **2006**, *281*, 36236.

(4) Minetti, C. A.; Remeta, D. P.; Miller, H.; Gelfand, C. A.; Plum, G. E.; Grollman, A. P.; Breslauer, K. J. *Proc Natl Acad Sci USA* **2003**, *100*, 14719.

(5) Hopfield, J. J. *Proc Natl Acad Sci USA* **1974**, *71*, 4135.

(6) Carlon, E.; Heim, T. *Physica A* **2006**, *362*, 433.

(7) Sugimoto, N.; Nakano, M.; Nakano, S. *Biochemistry* **2000**, *39*, 11270.

(8) Johnston, W. K.; Unrau, P. J.; Lawrence, M. S.; Glasner, M. E.; Bartel, D. P. *Science* **2001**, *292*, 1319.