Published in final edited form as:

Science. 2010 October 15; 330(6002): 376–379. doi:10.1126/science.1192001.

Rapid Construction of Empirical RNA Fitness Landscapes

Jason N. Pitt¹ and Adrian R. Ferré-D'Amaré^{1,*}

¹Howard Hughes Medical Institute and Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle WA 98109-1024, USA

Abstract

Evolution is an adaptive walk through a hypothetical fitness landscape, which depicts the relationship between genotypes and the fitness of each corresponding phenotype. We constructed an empirical fitness landscape for a catalytic RNA by combining next-generation sequencing, computational analysis and 'serial depletion', an *in vitro* selection protocol. By determining k_{obs} for every point mutant of a catalytic RNA, we demonstrated that abundance in serially depleted pools correlates with biochemical activity ($r = 0.67$, $Z = 7.4$). Therefore, enumeration of each genotype by deep sequencing yielded a fitness landscape containing $\sim 10^7$ unique sequences, without requiring measurement of the phenotypic fitness for each sequence. High-throughput mapping between genotype and phenotype may apply to artificial selections, host-pathogen interactions and other biomedically relevant evolutionary phenomena.

In vitro selection of RNA has led to the isolation of aptamers and ribozymes expanding our understanding of the biochemical capabilities of this nucleic acid (1). SELEX methodology also allows the study of the process of evolutionary adaptation (2), which has been conceptualized as an optimizing walk through a fitness landscape (3). In molecular evolution, such a landscape represents the sequence space of a macromolecule and the fitness associated with each genotype (4, 5). Two major difficulties have been identified in the empirical construction of macromolecular fitness landscapes. First, even for macromolecules of modest length, the sequence space is vast; a 20-mer RNA or protein has $\sim 10^{12}$ and $\sim 10^{26}$ possible sequences, respectively. Second, in order to characterize the landscape, the phenotypic fitness of each individual genotype needs to be measured, or an indirect measure of fitness needs to be validated. Therefore, although fitness landscapes have been constructed through computer simulations [*e.g.* (6-8)], experimental analyses of fitness landscapes have typically been limited to dozens to hundreds of genotypes [reviewed in (9)].

Advances in DNA sequencing methodology allow us to sequence $\sim 10^8$ individual molecules of lengths up to ~ 50 nucleotides (nt) with high accuracy, producing larger accessible experimental sequence spaces (10). During an *in vitro* selection experiment, the representation of RNA molecules that are more active should increase. Thus, we hypothesize that the abundance (or 'fecundity') of a particular sequence in an RNA pool undergoing selection, or its rate of increase, may serve as a surrogate fitness metric. Thus, deep sequencing of earlier stages of an *in vitro* RNA selection experiment (before the pool has become dominated by the most active species) and enumeration of the frequency of each genotype may directly provide an experimental fitness landscape.

*To whom correspondence should be addressed, aferre@fhcrc.org; telephone (206) 667-3622; facsimile (206) 667-3331.

Accession codes: Illumina sequence data have been deposited with the NCBI Sequence Read Archive with accession code SRA020870.1.

To establish the validity of this approach, we mutagenized a well-characterized *in vitro* selected RNA ligase ribozyme and subjected the resulting pool to reselection. We chose the 'class II' ligase (11) because its length (54 nt) is comparable to those of reliable individual reads by the Illumina genome analyzer, and because the 'wild-type' [isolate a4-11, hereafter 'master sequence', Fig. S1 and (11)] of this catalytic RNA is extremely active, likely being near-optimal. The mutagenized pool (45 nt at a degeneracy of 21% per position; expected frequency of master sequence = 0.0025%) can be analogous to a viral population arising due to error-prone replication (12). Our class II ligase construct catalyzes the formation of a 2'-5' phosphodiester bond between its 5' triphosphate and the 3' terminus of a 'substrate' RNA that is immobilized on a magnetic bead [Fig. S1 and (13)]. Additional sequence pools served as negative controls. One was comprised of $\sim 10^{13}$ random RNAs; the other of mutants of a sequence variant of the class II ligase that has been engineered to favor 3'-5' bond formation, but which is crippled in catalytic activity [(14) and Fig. S1].

Our control pools were flanked by constant sequences allowing their members to be distinguished from those of the master sequence pool, and neither was expected to be significantly enriched during the course of selection. The three RNA pools were mixed to generate a starting pool with a complexity of $\sim 6 \times 10^{13}$ (13), with the genotypes of members of each pool centered on a characteristic sequence distance (number of mutations) from the master sequence (Fig. 1A).

To separate individual sequences by their reactivity, we incubated the RNA pool with substrate RNA linked to beads, depleting the most active species from the population. RNAs were allowed to react for 24 hours. Then, species capable of ligation to the beads were sequenced, and their genotypes and frequencies analyzed [Fig. 1, Table S1, and (13)]. This one-step selection experiment resulted in an enrichment of the master sequence pool at the detriment of the random and engineered pools. The most abundant sequences in the starting pool are PCR artifacts from the random pool (Fig. 1B). These disappear in the 24 hour selected pool (Fig. 1D). Moreover, the representation of the master sequence (Hamming distance = 0) increased markedly, from 0.0015% to 0.28% of the total population (Fig. 1C,E).

Since the enrichment of a sequence in the selection experiment should be proportional to its catalytic activity, we successively depleted an RNA pool with a stoichiometric excess of immobilized substrate for increasing time intervals in a 'serial depletion' experiment. Comparison of the population captured after a one minute incubation with that captured after 24 hours shows that the former is composed of sequences that are markedly closer to the master sequence (Fig. 2A). If the fecundity of genotypes is proportional to their activity, then the rate of change of their frequency over the time-course, plotted as a function of genotype space, would represent a fitness landscape (Figs. 2B, S2, and S3). Indeed, populations selected early were more active than those selected at the longest time-point, as evidenced by their bulk k_{obs} (Fig. 3A). The populations selected at the three shortest time-points had comparable activity to each other, but the serially depleted population that was allowed to react for 24 hours was only one third as active. Consistent with early selection of the most active sequences, the information content (15, 16) of the early time-points is also highest [Fig. 3A and (13)].

To demonstrate experimentally that the frequency of each sequence is indicative of its catalytic activity and not a due to interactions within the RNA population, all 162 possible single point mutations of the master sequence were synthesized, and their k_{obs} measured [Fig. S4, Table S3, and (13)]. The k_{obs} of point mutants in all positions varied in the library ($n = 135$) correlated positively ($r = 0.67$) with the frequency of sequences containing the mutations in the 24 hour incubation (Fig. 3B). We tested the significance of this correlation

through computation of the distribution of the correlation coefficients of randomly assorted experimental k_{obs} and observed mutation frequencies [mean = 0.0088, s.d. = 0.089, Fig. 3B, Table S4, (17)]. This analysis indicates a significant regression coefficient ($Z = 7.4$, $p \ll 0.0001$, t -test). Therefore, the fecundity of any individual sequence provides a metric of its fitness, and we can create an experimental fitness landscape comprised of $\sim 10^7$ different RNA genotypes in a single experiment (Fig. 2B)

The empirical fitness landscape we generated is a high-dimensional object. We visualized it by computing the information content per residue of the master sequence, in essence projecting the landscape onto the ribozyme sequence. We plotted the information content for sequences most abundant in the 1 minute time-point ($D_{\text{max}} = 1$ minute), and those most abundant in the 24 hour time-point ($D_{\text{max}} = 24$ hours) on a revised secondary structure of the ribozyme generated from sequence covariation information implicit in our data [Figs. 4A,B, Tables S5–S9, Fig S5, (13)]. This analysis reveals that, in addition to the previously characterized key residues surrounding the ligation junction (11, 14), sequences in the central bulge of the RNA and the distal end of P3 are important for maximal activity. The location of these residues under purifying selection suggests that the class II ligase may fold into a Y-shaped structure (18) that brings the ligation junction and L3 into close proximity. The difference in information content between members of the slow- and fast-reacting populations (Fig. 4C) provided a measure of the slope of the landscape that the evolving population must traverse to achieve maximum fitness. As expected, the catalytic rates measured for each point mutation (Fig. S5D) support the functional importance of P1, P3 and L3.

In addition to studying the phenotypic fitness and positional information content of populations comprised of closely related sequences, our methodology (which produces orders of magnitude more sequence information than was practical before) could also be used to query the early rounds of SELEX experiments starting with random pools. Indeed, the locality-sensitive hash algorithm we implemented is more efficient in analyzing libraries with greater sequence diversity (19). A typical SELEX experiment employs multiple rounds of purifying selection and amplification to separate highly active sequences from other members of the starting population, assumed to be inactive (20–22). However, this assumption is only rarely tested (2). While aggregate activity of the population in earlier rounds of selection is low, there may be many sequences present that represent suboptimal solutions. Recent experiments demonstrated that minute RNAs, working alone or in ensembles, can be highly active catalysts (23, 24). For such small functional nucleic acids (13 nt) our methodology can determine the fitness of every possible RNA (4^{13} or 6.7×10^7 possible genotypes).

One application of empirically generated fitness landscapes is the analysis of systems in genetic conflict (25); for example, between a host immune response and an infecting viral quasispecies (26, 27). In such systems, positive selection drives phenotypic variation that maximizes sampling of their respective fitness landscapes. However, these systems are dynamic, with changes in one population leading to changes in the other. By analyzing regions of the fitness landscapes that remain fixed between multiple conditions we predict that critical regions of the molecules in question can be identified for vaccine or drug development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank R. Basom, A. Marty, and the staff of the Fred Hutchison Cancer Research Center Genomics Shared Resource for performing the Illumina sequencing, and D. Bartel, B. Fritz, R. Knight, H. Malik, W. Noble, H. Robins, A. Roll-Mecak, and M. Tewari, for discussions. This work was supported by the W.M. Keck Foundation and the Howard Hughes Medical Institute (HHMI). A.R.F. is an Investigator of the HHMI.

References

1. Wilson DS, Szostak JW. *Annu Rev Biochem.* 1999; 68:611. [PubMed: 10872462]
2. Lehman N, Joyce GF. *Curr Biol.* 1993; 3:723. [PubMed: 11539560]
3. Wright S. *Proc Sixth International Congress Genet.* 1932; 1:355.
4. Maynard Smith J. *Nature.* 1970; 225:563. [PubMed: 5411867]
5. Schuster P. *European Rev.* 2009; 17:281.
6. Schuster P. *Biol Chem.* 2001; 382:1301. [PubMed: 11688713]
7. Huynen MA, Stadler PF, Fontana W. *Proc Natl Acad Sci USA.* 1996; 93:397. [PubMed: 8552647]
8. Stich M, Lázaro E, Manrubia SC. *BMC Evol Biol.* 2010; 10:46. [PubMed: 20163698]
9. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. *Nature.* 2007; 445:383. [PubMed: 17251971]
10. Bentley DR. *Curr Opin Genet Dev.* 2006; 16:545. [PubMed: 17055251]
11. Eklund EH, Szostak JW, Bartel DP. *Science.* 1995; 269:364. [PubMed: 7618102]
12. Domingo E, Wain-Hobson S. *EMBO Rep.* 2009; 10:444. [PubMed: 19343047]
13. Materials and methods are available as supporting material on Science Online.
14. Pitt J, Ferré-D'Amaré AR. *J Am Chem Soc.* 2009; 131:3532. [PubMed: 19220054]
15. Carothers JM, Oestreich SC, Davis JH, Szostak JW. *J Am Chem Soc.* 2004; 126:5130. [PubMed: 15099096]
16. Schneider TD, Stromo GD, Gold L, Ehrenfeucht A. *J Mol Biol.* 1986; 188:415. [PubMed: 3525846]
17. The correlation was calculated using the frequencies of all haplotypes containing up to 8 mutations (projection 1 hash score = 800; 706,899 reads of 106,012 unique genotypes). The value of r is statistically unchanged ($p = 5\%$, two-tailed t -test) if the analysis is limited to sequences with 1, 2, ..., 6 mutations, indicating that higher-order effects are not detectable (Table S2).
18. de la Peña M, Dufour D, Gallego J. *RNA.* 2009; 15:1949. [PubMed: 19741022]
19. Pitt JN, Rajapakse I, Ferré-D'Amaré AR. *Nucleic Acids Res.* (in press).
20. Joyce GF. *Gene.* 1989; 82:83. [PubMed: 2684778]
21. Ellington AD, Szostak JW. *Nature.* 1990; 346:818. [PubMed: 1697402]
22. Tuerk C, Gold L. *Science.* 1990; 249:505. [PubMed: 2200121]
23. Turk RM, Chumachenko NV, Yarus M. *Proc Natl Acad Sci USA.* 2010; 107:4585. [PubMed: 20176971]
24. Vlassov A, Khvorova A, Yarus M. *Proc Natl Acad Sci USA.* 2001; 98:7706. [PubMed: 11427715]
25. Hurst LD, Atlan A, Bengtsson BO. *Q Rev Biol.* 1996; 71:317. [PubMed: 8828237]
26. Fernández G, Clotet B, Martínez MA. *J Virol.* 2007; 81:2485. [PubMed: 17151104]
27. Robins HS, et al. *Blood.* 2009; 114:4099. [PubMed: 19706884]
28. Hamming R. *Bell Systems Technical J.* 1950; 29:147.
29. Kimura M. *Nature.* 1968; 217:624. [PubMed: 5637732]

Fig. 1. Population structure before and after one round of *in vitro* selection

(A) Histograms of RNA ligase ribozyme populations before (blue) and after (red) *in vitro* selection (6.7×10^6 sequences each). Sequences are binned according to their Hamming distance (28) from the 'a4-11' (11) master sequence (MS) (13). (B, C) Pre-selection mutant spectrum (13, 19). Each spot is a unique species. Projection axes 1 and 2 are hash scores to the master sequence and an arbitrary string, respectively. Genotype frequency is the number of times a sequence was observed (13). (D, E) Mutant spectrum after one 24-hour *in vitro* selection step.

Fig. 2. Changes in population structure during serial depletion

(A) Hamming distance histograms from serial depletion, showing genotype frequencies from the pre-selection (blue), and one minute (green) and 24 hour (pink) time points. Frequencies of the master sequence in the three populations are indicated. Asterisk denotes a subpopulation that is dominated by the parental sequence (14) of the engineered pool. (B) Rates of depletion of genotypes most abundant in the one minute time-point (green) and those most abundant in the 24 hour time-point (magenta) as a function of their similarity to the master sequence.

Fig. 3. Genotype frequency correlates positively with experimental rate constants (A) k_{obs} (green, measured in triplicate, error bars represent SOM) and information content (black) for the entire populations from each serial depletion time-point. (B) Correlation between biochemically measured k_{obs} of individual point mutants (Table S3) and observed frequencies of the mutations (Table S4) in all sequence variants with Hamming distance 8 from the master sequence (red line, $r = 0.67$) (17). Green line is the k_{obs} of the master sequence (14). Dashed lines denote two independent estimates of the lower detection limit of the biochemical assay (13). (C) Histogram of correlation coefficients of k_{obs} ($n = 135$) with randomly reassorted mutation frequencies. The real correlation ($r = 0.67$) between the mutant frequencies in the selection and the experimental k_{obs} is 7.4 standard deviations from the mean.

Fig. 4. Analysis of the experimentally constructed fitness landscape as information content per position

Information content of a position in bits (15, 16) of genotypes with a projection 1 hash score 800 (Hamming distance of 8 from the master sequence) and $D_{\max} = 1$ minute (**A**), and $D_{\max} = 24$ hours (**B**) depicted as a heat map. Analyses based on 4,485,943 reads of 311,869 unique sequences, and 586,606 reads of 117,507 unique sequences for (**A**) and (**B**), respectively (Fig S5, Tables S5–S8). P1, P2, P3 denote helices, L2 and L3 loops. (**C**) Change in information content between $D_{\max} = 1$ minute and $D_{\max} = 24$ hours. Positions 12, 22, 33, 38, 39 appear to be selectively neutral (29). Black and red base-pair symbols indicate pairing predicted from aligning the 256 most common sequences, and from analysis of the Watson-Crick covariation of all sequences with a projection 1 hash score 800 (13), respectively.