Seeking a predictive theory of adaptation
Paul Joyce

Outline

- Experimental system
- Theory of adaptation
  - Description of the theory
  - Extensions of the theory
- Data versus theory
  - Testing the assumptions
  - Testing the predictions
- Beyond the first step

General Methods

- Evolved bacteriophage ID11 (family Microviridae) on cellular host *E. coli* C at elevated temp (37°C instead of optimal 33°C) for 20 flask passages.

- Passage bottleneck size: $10^4$ (SMALL) or $10^5$ (large)

General Methods

- Evolved bacteriophage ID11 (family Microviridae) on cellular host *E. coli* C at elevated temp (37°C instead of optimal 33°C) for 20 flask passages.

- Passage bottleneck size: $10^4$ (SMALL) or $10^5$ (large)
**Flask Passage Design**

Wildtype ancestor stock from single plaque

10^4 Transfer Size

10^6 Transfer Size

**Sampling for each replicate**

Passage:

<table>
<thead>
<tr>
<th>Plate &amp; pick individual plaques</th>
<th># plaques picked</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>32</td>
</tr>
</tbody>
</table>

Sequence whole genome of each plaque

• ID Beneficial mutations
• Fitness (growth rate) of each assayed

**DESCRIPTION OF THE THEORY**

Consider the set of all 3L +1 possible genotypes that differ from the wild-type by at most one basepair

\[
WT = AACGTAGCCCTATCGA TACCGATATCAAACTGCGCGAACAGACCAGTA
\]

\[
M_1 = AACGTAGCCCTATCGA TACCGGATATCAAACTGCGCGAACAGACCAGTA
\]

Fitness

\[
M_1 \quad WT
\]

Gillespie’s mutational landscape model

If we rank all of the 3L+1 sequences by fitness (fittest has rank 1), then the wild-type will have some rank i, indicating that i - 1 beneficial mutations are accessible.

\[
WT = AACGTAGCCCTATCGATTACCGATATCAAACTGCGCGAACAGACCAGTA
\]

\[
M_2 = AACGTAGCCCTATCGATTACCGATATCAAACTGCGCGAACAGACCAGTA
\]

Fitness

\[
M_2 \quad WT
\]

Gillespie's mutational landscape model
If we rank all of the 3L+1 sequences by fitness (fittest has rank 1), then the wild-type will have some rank i indicating that i-1 beneficial mutations are accessible.

\[ WT = AACGTAGCCTATCGATTACCGATATCAAACTGGCCGAAACAGCCAGTA \]

\[ M_3 = AACGTAGCCCTATCGATACCGATATCAAAACTGGCCGAAACAGCCAGTA \]

Testing the assumptions of the mutational landscape model: The Challenges

1. Identifying the appropriate alternative model
2. The inability to identify adaptive mutations of small effect
3. Low statistical power due to small number of beneficial mutations in each experiment

Extreme Value Theory has three types of tail distributions:

- Weibull (truncated)
- Gumbel (exponential)
- Fréchet (heavy tailed)

### Generalized Pareto Distributions

The probability density function for the Generalized Pareto Distribution is given by:

$$ f(x|\kappa, \tau) = \begin{cases} \frac{1}{\kappa}(1 + \frac{x\tau}{\kappa})^{-\frac{\kappa+1}{\kappa}}, & x \geq 0, \text{ if } \kappa > 0 \\ \frac{1}{\kappa}(1 + \frac{x\tau}{\kappa})^{-\frac{\kappa+1}{\kappa}}, & 0 < x < -\frac{\tau}{\kappa}, \text{ if } \kappa < 0 \\ \frac{\tau}{\kappa}e^{-x/\tau}, & x \geq 0, \text{ if } \kappa = 0. \end{cases} $$

### Testing the assumptions of the mutational landscape model: The Challenges

1. Identifying the appropriate alternative model
2. The inability to identify adaptive mutations of small effect
3. Low statistical power due to small number of beneficial mutations in each experiment

### Sensitivity Analysis

- **Probability of type I error**

### Adjusting for missing data

If the purpose of the study is to test the extreme value theory assumption then in most experimental evolution scenarios it is more appropriate to measure the selection coefficients relative to the smallest observed fitness rather than the fitness of the wild type. Let $y_k = s_k - s_j$

$$ f_{(y_1, y_2, \ldots, y_j)}(y_k) = (j-1)! \frac{f(y_1+s_j) f(y_2+s_j) \cdots f(y_{j-1}+s_j)}{1-F(s_j) 1-F(s_j) \cdots 1-F(s_j)} $$

### The effects of not shifting

- **Shift and rescale**

$$ f_{(y_1, y_2, \ldots, y_j)}(y_k) = (j-1)! \frac{f(y_1+s_j) f(y_2+s_j) \cdots f(y_{j-1}+s_j)}{1-F(s_j) 1-F(s_j) \cdots 1-F(s_j)} $$

$$ y_k = s_k - s_j $$

- **Probability of type I error**

- **Probability of type II error**
Problem 3: Low statistical power due to small number of beneficial mutations in each experiment

Solution: Pooling data across experiments.

However, one can distinguish the exponential distribution from the truncated alternatives with relatively few observations.

The effects of pooling data from non replicate experiments

Note that the likelihood ratio statistic generalizes to

\[
2 \ln(\Lambda) = \sum_{k=1}^{m} 2(\ell(X_k | \tilde{\kappa}_k, \tilde{\tau}_k) - \ell(X_k | 0, \tilde{\tau}_k)) \tag{3}
\]

where \( X_{ik} \) are the observed fitnesses, for the \( k \)th experiment, \( \tilde{\kappa}_k, \tilde{\tau}_k \) the parameter estimates under the GPD for the \( k \)th experiment and \( \tilde{\tau}_k \) is the estimate for \( \tau \) under the exponential model.

Testing the Assumptions

Null \( \kappa = 0 \); equivalent to the Gumbel (exponential) distribution

Alternative model \( \kappa > 0 \) or \( \kappa < 0 \)

Conclusion: Reject Gumbel (exponential) distribution in both cases.
Theoretical predictions under exponential distribution

Transition probabilities
\[
\frac{1}{i+1} = \frac{1}{i+1} \quad \frac{1}{i+1} \quad \frac{1}{i+1} \quad \frac{1}{i+1}
\]

Mean Change in rank
\[
\frac{i+2}{4}
\]

Parallel evolution
\[
\frac{2}{i}
\]

Mean fitness improvement
\[
2 \frac{i-1}{i} E(S)
\]

Orr, 2002

Testing the predictions

Mutation-adjusted Orr model performs well

EXTENDING THE THEORY

Testing the predictions

\[ P = 0.49 \]

Orr’s model explains the data poorly

See Rokyta et al. (2005) Nature Genetics 37:441-444

Theoretical predictions and extensions

<table>
<thead>
<tr>
<th></th>
<th>Exponential Tail</th>
<th>Generalized Pareto Tail</th>
<th>Uniform tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition probabilities</td>
<td>( \frac{1}{i+1} )</td>
<td>( \frac{k}{(i+1)(i+k)} )</td>
<td>( \frac{2(i-j)}{i(j-j)} )</td>
</tr>
<tr>
<td>Mean Change in rank ( E(P_i) )</td>
<td>( \frac{i+2}{4} )</td>
<td>( 1 + \frac{i-j}{2} \left( \frac{1}{i-j} \right) )</td>
<td>( \frac{i+1}{3} )</td>
</tr>
<tr>
<td>Parallel evolution</td>
<td>( \frac{2}{i} )</td>
<td>( \frac{2(1-k)}{2(1-k)(i-j)} )</td>
<td>( \frac{4}{3(i-j)} )</td>
</tr>
<tr>
<td>Mean fitness improvement</td>
<td>( 2 \frac{i-1}{i} E(S) )</td>
<td>( 2 \frac{i-1}{i} E(S) )</td>
<td>( \frac{4}{3} E(S) )</td>
</tr>
</tbody>
</table>

Joyce, Rokyta, Orr, and Beisel (2008)
Beyond the first step

To model adaptation, some common question must 1st be answered

Fitness Landscapes

Objectives

Using data from a virus adapting to lab conditions, we wish to know:

1. Are properties of a smooth landscape observed?
2. Are properties of a rugged landscape observed?

Is our fitness landscape smooth?

- Predictions:
  a. 1st and 2nd-steps should be same mutations.
  b. Fitness effects should be of similar magnitude.

Two well-sampled backgrounds
Is our fitness landscape smooth?

- Predictions: (a) 1st- and 2nd-steps should be same mutations. (b) Fitness effects should be of similar magnitude.

- Observations: (a) Only one 1st-step observed among 2nd-steps.

Conclusion: Effects are not additive; landscape is not smooth.

Is our fitness landscape smooth?

- Predictions: (a) 1st- and 2nd-steps should be same mutations. (b) Fitness effects should be of similar magnitude.

- Observations: (a) Only one 1st-step observed among 2nd-steps. (b) 2nd-steps have lower fitness effects than 1st-steps.

Is our landscape rugged (totally uncorrelated)?

- Prediction #1: Mutant 2534 is of high fitness rank (top 3) among observed 1st-steps.

# of beneficial mutations on 2534 background should be ~ Neg. binomial (0.5, rank=3). Expected # is 3. Is data consistent w/ this expectation?

Results:

```
# Negative Binomial Assuming 2534 is Rank 3

0.2
0.15
0.1
0.05

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

# Beneficial Mutations on 2534
```

- Prediction #1: Mutant 2534 is of high fitness rank (top 3) among observed 1st-steps.

# of beneficial mutations on 2534 background should be ~ Neg. binomial (0.5, rank=3). Expected # is 3. Is data consistent w/ this expectation?

Observation: Observe 9 beneficial mutations—all transitions—on 2534, and estimate (from ‘recapturing’ several) that 18 exist (95%CI: 10-41).
Results: Negative Binomial Assuming 2534 is Rank 3

- Conclusion: Too many beneficial mutations on 2534. Data inconsistent w/ uncorrelated landscape.

2nd Test of Uncorrelated Landscape

- Prediction #2: Mutations on 2534 background should come from upper tail of same distribution as 1st steps.

- Methods: Fit fitness estimates to:
  1) Single GPD (General Pareto Distribution) = NULL
  2) Separate GPD’s for each background = ALT

- Bayesian Analysis: Calculate Bayes’ Odds Ratio of 1 vs. 2 distribution models using MCMC to integrate out uncertainty in fitness & parameters values.

Results:

- P(2 Distribution Model) = 22
  P(1 Distribution Model)

Conclusion: Distribution of fitness effects is not constant between steps. Landscape is not uncorrelated.