CSE280A Class Projects

1 Primer design for RNA

As discussed in class, some genes may come together due to a structural variation and produce fusion transcripts. In this project we look to amplify transcripts (transcribed sequences or mRNA) of fusion proteins directly. Thus, our experiment will be to take total RNA, drop primers in it, and only a fusion protein would end up being amplified. However, we’d like to avoid the cost of sequencing, as we have the technology to measure the length of the amplified product. In this project, we will try and do a primer design so that every possible amplification is of a different length if possible.

Primer design for RNA:

Input: The input is the following

1. A collection of exons given as genomic coordinates.
2. A collection $P$ of locations on the exons signifying candidate primer end-points. Each exon either has forward candidate primers or reverse candidate primers.
3. A collection $E \subseteq P \times P$ of primer pairs that dimerize so that both primers cannot be part of the same solution.

Output: A collection of primers $P \subseteq P$ s.t. no primer pair in $P$ dimerizes, all pairs of forward reverse exon fusions are captured, and each primer pair has a distinct amplicon lengths separated by at least $10$ bp. If it is not possible to find a perfect solution, you should formulate an optimization problem that minimizes the number of distinct pairs which have similar lengths.

Project goals: You should come up with a formulation that adequately models the problem, and use any of the optimization techniques available to solve it. Ideally, we would also like to see your insights into the problem. For example, most coding exons are very small and it may be that there simply aren’t enough primer locations to allow for a good design.

Primer design papers: See Bashir et al. [1], Patel et al. [3]

Data: Anand Patel can supply a collection of exons, candidate primers, and dimerizing pairs.
Project 1: Primer design for fused transcripts

- Filled primers represent selection from the candidate pool.
- Edges indicate dimerization.
- All possible sums $F_i + R_j$ give the possible lengths, and each length should be distinct.

Figure 1: Primer design with RNA
2 Primer design with multiplexing

Recall that in primer design of a genomic region, our goal is to find a collection of non-dimerizing primers such that each breakpoint is amplified by at least one primer pair. Here, we pursue primer designs for large instances where it is simply not possible to get a good design. Either the set of primers will have a dimerizing pair, or some breakpoint will not be covered. In this case, we increase the number of reactions. Instead of forcing that all reactions be done in a single experiment, we ask for the smallest number of experiments such that the other constraints are maintained. If the primers are nodes in a graph, and edges represent dimerization, then a ‘graph coloring’ represents a solution.

**Input:** Forward and Reverse genomic regions; Candidate primer locations $\mathcal{P}$; dimerizing pairs of primers

**Output:** A collection of primer subsets $P_1, P_2, \ldots, P_k$, where $P_i \subseteq \mathcal{P}$ such that no $P_i$ contains a dimerizing set, each breakpoint is amplified by at least one forward-reverse primer pair in one of the subsets $P_i$, and that the total number of experiments is minimized. Note that $P_i$ can reuse primers.

**Project Goals:** You should come up with a problem formulation that models the constraints, and provide a practical but good performing solution on the available data-sets.

**Primer design papers:** See Bashir et al. [1], Patel et al. [3].

**Data:** Anand Patel (adp@ucsd.edu) can supply a collection of exons, candidate primers, and dimerizing pairs.

---

**Project 2: Primer multiplexing for large regions**

![Diagram of multiplexed primer design](image)

Figure 2: Multiplexed Primer design
3 Time to Fixation (somewhat theoretical)

Consider a fixed haploid population of size $N$ with one locus and two alleles. Thus, each individual is of type $A$ or $a$. In each generation, an individual picks a parent at random. Due to selection, allele $A$ is picked with probability $\approx 1 + s$, while $a$ is chosen with probability $\approx 1$. Starting with $A$ being in exactly one copy, and assuming that it goes to fixation, what is the time in number of generations for fixation to occur. Your goal is to experiment on this for $N = 10^3, 10^4, 10^5, 10^6$, and $s \in \{0.01, 0.02, 0.04, 0.1\}$, and come up with a plausible formula. Next, justify your formula theoretically, or comment on the different results in the two attached manuscripts.

Reading: See attached manuscripts in the appendix and Campbell [2].

Project goals: Caution: The math in both manuscripts might be a little off, and is not the critical requirement of this project. Careful experimentation on a very large number of examples should clarify what the actual fixation time should be. You can show it by providing a scatter plot of actual and theoretical times.

![Project 3: Fixation time](image)

Figure 3: Number of generations to fixation.
4 Tests of selection in the post selection regime

Ronen et al. show that there is great discriminative power in the allele frequency spectrum of regions under selection and those evolving neutrally. However, key parameters (time since selection and selection coefficient) are not known. Given a training collection of data (neutrally evolving regions and regions under selection in the post-fixation regime), develop a classifier that can test for selection.

**Input:** A collection of SNP matrices corresponding to neutral evolution, and those under selection. All we know is that we are in the post-selection regime but the selection coefficient and time since selection are not known.

**Output:** Classification of the data into ‘under selection’ and ‘neutral’ labels.

**Project goals:** We spent a fair amount of time thinking about selection class. Any extension of those ideas or your own novel ideas are welcome.

**Reading:** Class notes. Also see Ronen [4].

**Data:** Contact Roy Ronen (rronen@ucsd.edu).

Figure 4: Power analysis shows the gap between what is achievable and what is currently known for tests of selection in the post-fixation regime.
5 Identification of introgressed sequences

Two recent papers [5, 6] suggest that modern humans and Neanderthals interbred. In this project, you should critically examine the techniques used to arrive at this conclusion. Please try and download the data sets and reproduce the figures in the paper.

**Reading:** Read the papers by Sakararaman and Vernot, 2014. Also look at the methods papers
References


Appendix: Fixation time for alleles under selection
1 Wright Fisher (WF) or Coalescent model

We have a fixed population of $N$ individuals going from generation to generation. The individuals have allele $A$ or $a$. In each generation, individual chooses a single parent from the previous generation, and adopts their allele.

**Discrete time modeling.** Let $N_t$ denote the number of copies of allele $A$ at time $t$, so that $N_0 = 1$. Under selection with parameter $s$,

$$\Pr(A) = c(1 + s) \frac{N_{t-1}}{N}$$
$$\Pr(a) = c \left(1 - \frac{N_{t-1}}{N}\right)$$

Where the normalizing constant

$$c = \frac{1}{(1 + s)^{\frac{N_{t-1}}{N}} + \left(1 - \frac{N_{t-1}}{N}\right)} = \frac{1}{1 + \frac{sN_{t-1}}{N}}$$

Therefore,

$$E(N_t) = N \Pr(A) = \frac{(1 + s)}{1 + \frac{sN_{t-1}}{N}} N_{t-1}$$

**Continuous time.** The growth follows a logistic curve. Our goal is to get the rate of growth. In continuous time, suppose that we are sampling individuals at time $t + \delta t$

$$\Pr(A) = c(1 + s)^{\delta t} \frac{N_t}{N}$$
$$\Pr(a) = c \left(1 - \frac{N_t}{N}\right)$$

Where the normalizing constant

$$c = \frac{1}{(1 + s)^{\delta t} \frac{N_t}{N} + \left(1 - \frac{N_t}{N}\right)} \simeq \frac{1}{1 + \frac{s\delta t N_t}{N}}$$

Therefore,

$$E(N_{t+\delta t}) = N \Pr(A) = \frac{(1 + s\delta t)}{1 + \frac{s\delta t N_t}{N}} E(N_t)$$

$$E(N_{t+\delta t}) + \frac{s\delta t N_t}{N} E(N_{t+\delta t}) = E(N_t) + s\delta t E(N_t)$$

$$E(N_{t+\delta t}) - E(N_t) = s\delta t \left(E(N_t) - \frac{N_t}{N} E(N_{t+\delta t})\right)$$

Or,

$$\frac{E(N_{t+\delta t}) - E(N_t)}{\delta t} = s \left(N_t - \frac{N_t N_{t+\delta t}}{N}\right)$$
Note that

$$\lim_{\delta t \to 0} N_{t+\delta t} = N_t$$

Therefore,

$$\frac{dE(N_t)}{dt} = sN_t(N - N_t) \frac{N}{N}$$

Suppose the population goes from $N_i$ to $N_j$ in time $\tau$. Then,

$$\int_{t=0}^{\tau} s dt = \int_{N_i=N_i}^{N_j} dE(N_t) \left( \frac{1}{N_t} + \frac{1}{N-N_t} \right)$$

$$s\tau = \ln \frac{N_j}{N_i - N_i} \frac{N_j}{N_i}$$

This gives us the basic logistic regression equation for time as:

$$\tau = \frac{1}{s} \ln \left( \frac{N_j(N - N_i)}{N_i(N - N_j)} \right)$$

(1)

and for population growth as

$$\tau s = \ln \left( \frac{N_j(N - N_i)}{N_i(N - N_j)} \right)$$

(2)

$$e^{\tau s} = \frac{(N/N_i - 1)}{(N/N_j - 1)}$$

(3)

$$N_t \simeq \frac{N}{1 + N e^{-\tau s}}$$

(4)

where Eqn. 4 is obtained by choosing $N_i = 1$ (population at time 0), and $N_j = N_t$. We can use Eqn. 1 to compute times to fixation, as well as the time to reach different population milestones.

$$N_0 = 1$$

$$N_1 = \ln n$$

$$N_2 = \frac{N}{\ln N}$$

$$N_3 = \frac{N}{2}$$

$$N_4 = N - \frac{N}{\ln N}$$

$$N_5 = N - \ln N$$

$$N_f = N - 1$$

**Fixation, and other times.** To go from $N_0$ to $N_f$, we have

$$\tau = \frac{1}{s} \ln(N(N - 1)) = \frac{2}{s} \ln N$$

\[ \Box \]
This time can be split into 3 different epochs: \( N_0 \) to \( N_2 \), \( N_2 \) to \( N_4 \), and \( N_4 \) to \( N_f \). Let the respective times be \( \tau_1, \tau_2, \tau_3 \). Then,

\[
\tau_1 = \frac{1}{s} \ln \left( \frac{N}{mN} \left( \frac{N}{N - \frac{N}{mN}} \right) \right) = \frac{1}{s} \ln \left( \frac{N - 1}{\ln N - 1} \right) \simeq \frac{1}{s} \left( \ln N - \ln \ln N \right)
\]

\[
\tau_2 = \frac{1}{s} \ln \left( \frac{(N - \frac{N}{mN})(N - \frac{N}{mN})}{\frac{N}{mN} \frac{N}{mN}} \right) = \frac{2}{s} \ln \ln N
\]

\[
\tau_3 = \frac{1}{s} \ln \left( \frac{(N - 1) \frac{N}{mN}}{(N - \frac{N}{mN})} \right) = \frac{1}{s} \ln \left( \frac{N - 1}{\ln N - 1} \right) \simeq \frac{1}{s} \left( \ln N - \ln \ln N \right)
\]

### 2 Sampling from a population where the beneficial allele has reached fixation

Consider the case when the beneficial allele is just driven to fixation, so that \( N_0 = N \). Sample a pair of individuals. **Q: What is the expected time to coalescence for the pair?** Recall from standard coalescent theory that if the population remains constant at \( N \), the time to coalescent is distributed exponentially, with success probability \( \frac{1}{N} \). Thus, probability that we coalesce exactly at \( t \) generations (assuming continuous time) is

\[
P_t = \frac{1}{N} \left( 1 - \frac{1}{N} \right)^t \simeq \frac{1}{N} e^{-\frac{t}{N}}
\]  

Also, the expected time to coalesce is

\[
\int_{t=1}^{t=\infty} \frac{t}{N} e^{-\frac{t}{N}} dt = N
\]

Similarly, if there are \( k \) individuals, the expected time for a coalescence to occur is \( \frac{N}{k(2)} \), and the number of mutations created in the time that \( k \) lineages become \( k - 1 \) is thus

\[
\mu \frac{N}{k(2)} = \frac{2N\mu}{k - 1}
\]

This allow us to calculate the total number of polymorphic sites we expect to see as

\[
\sum_{k=n}^{2} \frac{2N\mu}{k - 1} \simeq 2N\mu \ln n
\]  

We want to do a similar calculation for the case when the individuals are sampled from the process under selection at a time when the beneficial allele has just fixed in the population. In this case, the population is decreasing as we go back in time. Specifically, using an approach similar to Eqn. 4, if a population of \( N_i \) individuals becomes \( N_j \) \( t \) generations later, then

\[
e^{st} = \frac{\frac{N}{N_j} - 1}{\frac{N}{N_i} - 1}
\]

\[
N_t \simeq \frac{N}{1 + e^{st} \frac{N}{N}}
\]
The last equation is derived choosing \( i = 0, N_i = N, \) and \( N_j = N_t. \) Now, the probability that two individuals coalesce at time \( \tau \) is

\[
P_\tau = \frac{1}{N_\tau} e^{-\int_{t=1}^\tau \frac{1}{N_t} dt} \tag{11}
\]

Q: Can we solve this expression using Eqn. 10, and then find the expected time to coalesce for (a) 2 individuals, and (b) \( k \) individuals?