CH2: Genome Assembly and Waterman Statistics

CS 182/282 Spring 2024

Scribes (from past years): `kclark5`, `ewoo`, `adenadel`, `alee113`, `cmeyer5`, `amohan8`, `berdogdu`

Scribes (2022): `isu2`, `yguo62`, `achinta2`, `ypu7`, `nlee16`, `hvenkata`, `shong41`, `mlincol1`

Compiled & edited by `eyouth, smaffa, cbaker20`

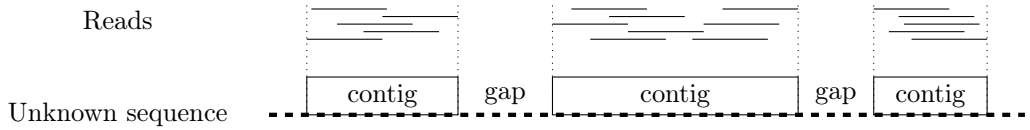
Reach out to the TAs on Ed for any clarifications or corrections.

Overview of Genome Assembly

Genome assembly is the process of inferring a sequence from many fragments. Sophisticated algorithms have been developed to align and merge large numbers of genomic fragments ("reads") obtained from shotgun sequencing or other experimental methods, with the goal of reconstructing the original genomic sequence which produced the fragments. Sequence assembly was the central challenge of the Human Genome Project, which succeeded in assembling the first reference human genome in 2001.

General Methods

A schematic for genome assembly is shown below:



Broadly, large numbers of *reads* (short genomic strands) are sequenced and aligned to each other to infer the unknown target sequence which produced them. *Contigs*, or contiguous regions, represent preliminary consensus sequences based on aligned reads; these are eventually merged to produce the final full-length alignment. *Gaps* are regions of the target sequence which cannot be inferred due to insufficient reads.

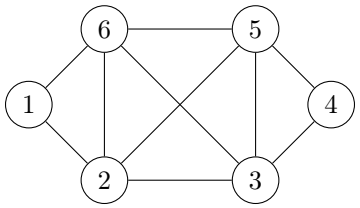
In addition to the difficulty inherent in determining the orientation and order of reads, there are several other complexities that arise in sequencing which complicate the process of genome assembly. Genomic sequences which are repeated will yield identical reads, which may obscure information about important genomic features. Low sequencing depth or insufficient coverage of the target genome can yield gaps in the inferred sequence, as shown in the schematic above. Read length is also a consideration; short reads may not provide enough information to fully reconstruct the target sequence, while long reads may include a higher proportion of errors due to missequencing of bases.

Sanger sequencing was the primary method of sequencing genomic material during the first several decades of the field. It was capable of producing reads of several hundred bases in length. *Next-generation sequencing* methods are the standard used today; these produce shorter reads of $100 - 200$ base pairs. The Human Genome Project involved a sequencing depth of 8x; human DNA was sheared and cloned into plasmids for amplification and extracted via restriction enzymes for sequencing and assembly. This process introduced issues of contamination (via plasmid DNA) and correction (via sequencing errors), both of which are significant considerations in any genome assembly workflow.

Graph Theory

Defn: An *Eulerian circuit (cycle)* is a path in a graph which visits edge exactly once and ends at the starting vertex.

Defn: An *Eulerian graph* is a graph which contains an Eulerian circuit.



Example of an Eulerian graph

Thm: A graph G is Eulerian if and only if it is a connected graph and all nodes have even degree. If G is connected and Eulerian, it has at least one Eulerian cycle. If G has r connected components and is Eulerian, then it has at least one Eulerian cycle for every component.

Sequencing Coverage

Defn: The sequencing *coverage* (also referred to as sequencing depth) of a genomic sequence is

$$a = \frac{NL}{G}$$

where G is the length of the target sequence, L is the read length, and N is the number of reads.

For the Human Genome Project, $a = 8$.

This equation can help inform the approximate number of reads necessary to achieve a desired coverage. For example, to achieve 10x coverage of a bacterial artificial chromosome (BAC) of length 125 kb, the number of 480 bp-long reads required is

$$N = a\left(\frac{G}{L}\right) = (10)\left(\frac{125,000}{480}\right) \approx 2{,}604.$$

However, since the orientation (i.e., strandedness) of each read is unknown, 10x coverage of both strands would require twice as many reads (in this case, about 5,200).

Waterman Statistics

Waterman statistics provide a theoretical means of inferring information about the parameters involved in genome assembly. Proposed by Michael Waterman (who also developed the Smith-Waterman algorithm for local sequence alignment, along with Temple Smith) in the late 1980s, this statistical theory relies on properties of several well-established probability distributions, reviewed briefly below.

Review of Probability Distributions

The Bernoulli Distribution

Defn: A **Bernoulli trial** is a trial with two outcomes, "success" and "failure", where p is the probability of success and $q = 1 - p$ is the probability of failure. By definition, $p + q = 1$.

The random variable Y_1 can be used to model the number of successes for a single Bernoulli trial (i.e., either 0 or 1), with probability mass function

$$p_{Y_1}(y) = p^y(1-p)^{1-y}$$

$$= \begin{cases} p & \text{if } y = 1 \\ q & \text{if } y = 0 \end{cases}$$

The Binomial Distribution

Defn: A **binomial random variable** gives the number of successes in n independent Bernoulli trials with the same probability of success (p) in every trial.

The random variable Y_2 can be used to model the number of successes (ranging from 0 to n), with probability mass function

$$p_{Y_2}(y) = \binom{n}{y} p^y(1-p)^{n-y}$$

The Poisson Distribution

Defn: A **Poisson distribution** gives the probability of a given number of events happening within a fixed interval, provided that they occur with a constant mean rate $\lambda > 0$.

A random variable Y_3 has a Poisson distribution if its probability mass function is

$$p_{Y_3}(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

where the number of events is $y = 0, 1, 2, \ldots$. Calculus can be used to show that this is a probability mass function since

$$\sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{\lambda}$$

The Poisson distribution is a particularly relevant one for genome assembly statistics because it is a limiting form of the binomial distribution. If n (the number of trials) is large and p (the probability of success per trial) is small, then the quantity np is of moderate size and the binomial distribution can be approximated as a Poisson distribution with parameter $\lambda = np$. This leads to the following result:

Theorem: For any $k \in \{0, 1, \ldots, n\}$, the binomial distribution $Bin(k; n, p)$ approaches the Poisson distribution $Pois(k; \lambda)$ as $n \to \infty$ and $p \to 0$.

Proof. Manipulating the binomial probability distribution and recalling that $\lambda = np \implies p = \frac{\lambda}{n}$,

$$Bin(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k}$$

$$= \frac{np \cdot (n-1)p \cdots (n-k+1)p}{k!} (1-p)^{n-k}$$

$$= \frac{np \cdot (n-1)p \cdots (n-k+1)p}{k!} (1-p)^{n} (1-p)^{-k}$$

$$= \frac{n(\frac{\lambda}{n}) \cdot (n-1)(\frac{\lambda}{n}) \cdots (n-k+1)(\frac{\lambda}{n})}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Now, noting that

$$\lim_{n\to\infty} (n-i)\frac{\lambda}{n} = \lambda \qquad \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \qquad \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1$$

it follows that

$$\lim_{n\to\infty} Bin(k;n,p) = \frac{\lambda^k}{k!} e^{-\lambda}$$

□

Additional Distributions

The Geometric Distribution

The **geometric distribution** models the number of successes up until (but not including) the first failure in a series of Bernoulli trials with probability of success p. A geometrically-distributed random variable Y_4 has probability mass function

$$p_{Y_4}(y) = p^y(1-p)$$

for $y = 0, 1, 2, \ldots$.

The Exponential Distribution

The **exponential distribution** well-approximates the geometric distribution (note that the exponential distribution is continuous, while the geometric distribution is discrete). An exponentially-distributed random variable Y_5 has probability mass function

$$f_{Y_5}(y) = \lambda e^{-\lambda y}$$

for $y \geq 0$. Note that the mean of this distribution is $\frac{1}{\lambda}$ and that the cumulative distribution function is

$$F_{Y_5}(y) = 1 - e^{-\lambda y}$$

Poisson Processes

Defn: A sequence of events that occur within some interval (of time) is a **homogeneous Poisson process** if the following two conditions hold:

- The occurrence of any event in the time interval (a, b) is independent of the occurrence of any event in the time interval (c, d) if (a, b) and (c, d) are disjoint (i.e., $(a, b) \cap (c, d) = \varnothing$).
- There is a constant, $\lambda > 0$, such that for any sufficiently small time interval $(t, t + h)$, $h > 0$, the probability that a single event occurs within the interval $(t, t + h)$ is independent of t and has the formula $\lambda h + o(h)$, where $o(h)$ is a function that is sublinear with respect to h (i.e., $\lim_{h\to\infty} \frac{o(h)}{h} = 0$). Additionally, the probability that more than one event occurs in the interval $(t, t + h)$ must also be $o(h)$ (i.e., sublinear with respect to h). This property is called *time homogeneity*.

If the conditions above hold, the number of events (N) that occur up to time t has a Poisson distribution with parameter λt (it is a *homogeneous Poisson process*). These two conditions taken together are often referred to as the "randomness assumption" in other contexts.

Relevance to Genome Assembly

There are three major questions that Waterman statistics can help to answer:

1. What is the percentage of the target DNA region covered by the assembly?
2. What is the mean number of contigs in the assembly?
3. What is the mean length of a given contig in the assembly?

Solution to Question 1

Recall that coverage is defined as $a = \frac{NL}{G}$. Assuming that all reads are sampled uniformly at random from the region of interest and that $G \gg L$ (so that "end effects" at the boundaries may be disregarded), the left-hand ends of the reads are independently and uniformly distributed over the interval $[0, G]$. Then each left-hand end falls within an arbitrary interval $[x, x + h)$ with probability $\frac{h}{G}$, and the distribution of the number of left-hand ends falling in this interval has a binomial distribution with (Bernoulli) probability of success $p = \frac{h}{G}$ and mean $\frac{Nh}{G}$. If N is large and h is small, this distribution can be approximated by a Poisson distribution with parameter $\lambda = Np = \frac{Nh}{G}$ (the mean of the aforementioned binomial distribution).

Let Y be a random variable equal to the number of fragments whose left-hand ends fall within a particular interval of length L to the left of a randomly-chosen base x^*; in other words, Y will count the number of fragments that cover x^*, which lies at position $x + h - 1$ in the context of the above paragraph. Then Y has a Poisson distribution with mean $\lambda = a$ (the coverage). This means that the probability of at least one read covering x^* is

$$1 - \mathbb{P}(Y = 0) = 1 - e^{-a} \cdot \frac{a^0}{0!} = \mathbf{1 - e^{-a}}$$

Another way to think about this question is to note that, if we ignore end effects, the probability of a read covering a given base x^* is $\frac{L}{G}$. If we consider each read as an independent trial where success is defined by whether the read covers x^*, we get that the number of reads covering x^* follows a binomial distribution with $n = N$, $p = \frac{L}{G}$. When L is relatively small with respect to N, this distribution can be approximated by a Poisson distribution with the same parameter as a above, $\lambda = np = \frac{NL}{G} = a$. From here, we get the same probability as found above $(1 - e^{-a})$ that a given base is covered.

The table below summarizes the relationship between coverage (sequencing depth) and percent of target sequence covered with the technology available at the time of the original Human Genome Project:

Coverage (a)	% target region
2	86%
4	98%
6	99.8%
8	99.97%
10	99.995%
12	99.999%

Solution to Question 2

Note that the probability that no reads cover an arbitrary position x^* is e^{-a} as shown above. Consider $p = e^{-a}$ to be the probability that a given base immediately following one of our N reads is uncovered. Each time this event occurs, we will observe a new contig in our final assembly; whenever a read has a base uncovered to its right, there must be another read (in the limit) starting a new contig somewhere to the right of this uncovered base. As a result, the number of reads with a following uncovered base will, in the limit, fully describe the number of contigs in our assembly. If we consider that each read is an independent Bernoulli trial of this property with $p = e^{-a}$, we get that the mean number of contigs is

$$Np = \boldsymbol{Ne^{-a}}$$

Note that the assumption of independence here requires us to be working in the limit. A summary of values is given in the table below (note that $N \propto a$):

Coverage (a)	mean # contigs
0.5	60.7
0.7	69.5
1.0	73.6
1.5	66.9
2.0	54.1
3.0	29.9
4.0	14.7
5.0	6.7
6.0	3.0
7.0	1.3

Note that as coverage increases, the mean number of contigs increases to a point and then begins to decrease. This is because at low coverage additional reads will cover "new" areas of the region (thus generating additional contigs), while at higher coverage additional reads will cause existing contigs to merge together.

Solution to Question 3

Consider the left-hand ends of a succession of reads beginning with the leftmost read of a given contig. Assume that the distance from the leftmost read to the next read in the contig can be modeled with a geometric distribution, approximated by an exponential distribution with $\lambda = \frac{N}{G}$. Under this model, the second read will overlap the first read if this distance is less than L. This occurs with probability

$$\int_0^L f(x)\,dx = \int_0^L \lambda e^{-\lambda x}\,dx = 1 - e^{-\frac{NL}{G}} = 1 - e^{-a}$$

A contig is defined as a sequence of reads which overlap; i.e., a contig composed of n reads could be thought of as a sequence of $n-1$ Bernoulli "successes" (i.e., overlaps) followed by a single "failure" (i.e., no overlap

between the rightmost read of the contig and any other read to the right). The mean of this geometric distribution (with probability of success p as derived above) is

$$\frac{p}{1-p} = \frac{1-e^{-a}}{1-(1-e^{-a})} = \frac{1-e^{-a}}{e^{-a}} = e^a - 1$$

The total length of a contig is the sum of the distances between adjacent reads, where the distances L are distributed randomly as described above. The conditional distribution of the exponential random variable x given that $0 \leq x < L$ is

$$\frac{\lambda e^{-\lambda x}}{1-e^{-\lambda L}}$$

which implies that the expected (mean) length of the random distances is

$$\int_0^L x \left(\frac{\lambda e^{-\lambda x}}{1-e^{-\lambda L}} \right) dx = \frac{1}{\lambda} - \frac{L}{e^{\lambda L} - 1}$$

The mean of a sum of a random number of these random distances is then

$$\text{(mean number of random distances)} \times \text{(mean random distance)} = (e^a - 1)\left(\frac{1}{\lambda} - \frac{L}{e^{\lambda L} - 1} \right)$$

which implies that the mean contig length (accounting for the last, non-overlapped read) is

$$\begin{aligned}(e^a - 1)\left(\frac{1}{\lambda} - \frac{L}{e^{\lambda L} - 1} \right) + L &= \frac{e^a - 1}{\lambda} - L + L \\ &= \frac{e^a - 1}{\lambda} \\ &= \boldsymbol{\frac{L}{a}(e^a - 1)}\end{aligned}$$

Approximate examples values are given in the table below for reads of length $L = 500$ bases:

Coverage (a)	2	4	6	8	10
mean contig length	1,600	6,700	33,500	186,000	1,100,000

Genome Assembly

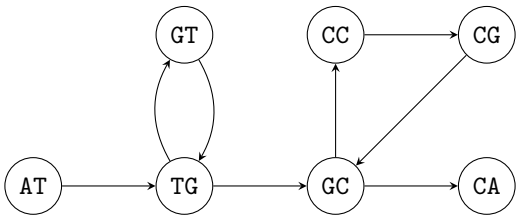
A robust field of theory underlies the process of assembling a genome from vast numbers of reads. Central to this algorithmic process are *de Bruijn graphs*.

de Bruijn Graphs

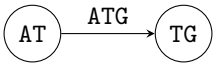
Consider the following sequence:

`ATGTGCCGCA`

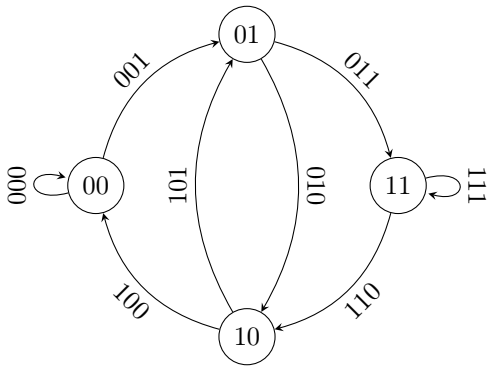
This sequence can be represented as a directed graph whose nodes are labeled with k-tuples, where edges are drawn between all pairs of k-tuples which "overlap" in the sequence. For $k = 2$, such a graph is



Edges can then be labeled with the appropriate $(k+1)$-tuple; for example,



The following is a *de Bruijn graph* for the set $\{000, 001, 010, 011, 100, 101, 110, 111\}$ (with $k = 2$):



The Idury-Waterman Algorithm

In 1995, Ramana Idury and Michael Waterman introduced an algorithm for assembling DNA sequences from reads obtained via shotgun sequencing. Their algorithm assumes *ideal data*; i.e., no sequencing errors, full coverage of the genome, and overlap between adjacent reads of at least k for some constant $k > 0$.

Input: N reads ($\{f_1, f_2, \ldots, f_N\}$); positive integer k

Output: Full genome sequence which produced the reads $f_{1:N}$

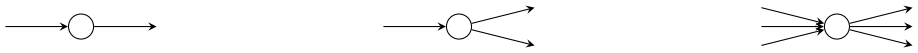
Overview of the Idury-Waterman Algorithm

1. Construct the union of the *spectra* of all the reads and their reverse complements where the spectrum of a given read f_i is the set of all k-tuples in f_i
2. From the union spectra created in step 1, construct the de Bruijn graph with each distinct $(k-1)$-tuple as a node and each k-tuple as an edge
3. Perform a variant of an Eulerian tour through the de Bruijn graph and infer the assembled sequence(s)
4. Align all reads to the assembled sequence(s) inferred from step 3

Refer to the original paper and algorithm slides on the website for a more detailed procedure.

Complexities of de Bruijn Graphs

Ideally, de Bruijn graphs enable unambiguous reconstruction of a sequence of interest. However, this is rarely possible from the outset, due to sequencing errors and the presence of repeated subsequences. In general, there are 3 main types of nodes within an unsimplified de Bruijn graph: *singletons*, *forks* and *crosses*.



singleton fork cross

Singletons account for approximately 90% of nodes, and enable straightforward simplification of the de Bruijn graph around them. Many heuristics involved in de Bruijn graph-based genome assembly are therefore focused on reducing forks and crosses to singletons, by eliminating erroneous or extraneous edges and resolving cycles induced by repeated subsequences.

Statistics of de Bruijn Graphs

de Bruijn graph-based genome assembly relies upon several assumptions:

- Errors are uniformly distributed, both over all reads and across the length of each read
- The global error rate is small (i.e., no more than $1 - 2\%$)
- The number of reads covering a given position of the sequence is a Poisson random variable with mean $a = \frac{NL}{G}$, so that $P(\text{base } t \text{ is covered by } k \text{ reads}) = \frac{e^{-a}a^k}{k!}$
- There are no repeats of length k or greater
- The only sequencing errors are substitution errors (i.e., a single base substituted for another)

Using G, N, and k as previously introduced, the following additional variables can be defined:

$$\begin{aligned} L &= \text{length of a read} \\ U &= \#\text{ of } k-1 \text{ base pair regions in all the reads} = N(L-k+2) \\ r &= \text{error rate of a single DNA base} \\ G' &= G-k+2 \\ R &= 1-(1-r)^{k-1} \end{aligned}$$

Denoting V, S and F as the sets of vertices, singletons, and forks, respectively, the following holds:

Theorem: The de Bruijn graph obtained from applying the Idury-Waterman algorithm has

$$\begin{aligned} \mathbb{E}(|V|) &= RU + [1-e^{-a(1-R)}]G' \\ \mathbb{E}(|S|) &= RU + e^{-a(1-R)}[e^{a(1-R)(1-r)^2} + a(1-R)r(2-r) - 1]G' \\ \mathbb{E}(|F|) &= 2e^{-a(1-R)}[e^{a(1-R)(1-r)} - e^{a(1-R)(1-r)^2} - a(1-R)r(2-r)]G' \end{aligned}$$

Note that the first term in the formula for $\mathbb{E}(|V|)$ is the expected number of "false" $k-1$-mers, while the second term is the expected number of "true" $k-1$-mers.

Slides 78-82 of the algorithm slides outline the proof of $\mathbb{E}(|V|)$.

The Ham Smith Problem

The Ham Smith Problem originates from the problem of non-uniform read sampling. In particular, when a DNA sequence (or corresponding mRNA subsequences) is read in practice, some regions/snippets may be

more accessible or abundant than others, leading to a biased sample of reads (that may leave some parts of the genome uncovered or inaccurate, once assembled).

Suppose there were 10000 different species of mRNA in a cell, all of which fall into an abundance class proportional to their concentration: $5, 50, 200, 1000$. Then, we can examine the following example table:

Copies per cell	Number of distinct mRNA species	Number of mRNAs per abundance level
5	4000	20000
50	3250	162500
200	2500	500000
1000	250	250000
TOTAL	10000	932500

Suppose S mRNAs form our experimental database; i.e., we randomly sample S strands from the cell. We can then define $J_L(S)$ as the number of different mRNA species with abundance L that we observe in our database. We want to compute $E[J_L(S)]$, since from this value we can assess how large S needs to be to ensure that some abundance class has sufficient representation in our database, according to some threshold. For notational purposes, let n_L be the number of distinct mRNA species at abundance level L, and $|L|$ be the total number of mRNAs at abundance level L. Note that $n_L L = |L|$ by this construction. We will define the set of distinct species at level L to be s_L.

This notation will allow us to assess the proportion of rarer mRNA transcripts in our database given some choice of S. For example, we might ask how many mRNAs must be read and recorded in our database in order to expect to have seen at least 50% of the different mRNA species that are expressed at the abundance level L.

To begin to answer this kind of question, we define I_a, which will be an indicator variable for the presence of a given *unique* mRNA species a in our database. Then,

$$I_a = \begin{cases} 1 \text{ if } a \text{ was seen in the sample} \\ 0 \text{ otherwise} \end{cases}$$

The representation (number) of different species of abundance class L^* in our database is thus

$$J_L(S) = \sum_{a \in s_{L^*}} I_a$$

Let's first consider a simple evaluation of this statement, in the case that $S = 1$. In this case (single draw of an mRNA transcript), we know that $E[J_L(1)] = E[\sum_{a \in s_{L^*}} I_a] = n_{L^*} \frac{L^*}{\sum |L|} = \frac{|L^*|}{\sum |L|}$. In other words, the expected number of distinct species in s_{L^*} we see when drawing a single read for our database is the mean fraction of non-unique species in s_{L^*} in the overall pool of non-unique species, $\bigcup_i s_{L_i}$. This result makes intuitive sense – the number of unique species in s_{L^*} we see in a single draw will be 1 if we draw a species in s_{L^*} and 0 otherwise, such that its expectation is proportional to the fraction of total available mRNA fragments whose sequence is in s_{L^*}.

We can generalize this for a generic S. Define $E[I_a] = p_{L^*}$ for all $a \in L^*$. Then, by linearity of expectations over n_{L^*} species, we have:

$$E[J_L(S)] = E[\sum_{a \in s_{L^*}} I_a] = \sum_{a \in L^*} E[I_a] = n_{L^*} p_{L^*}$$

Note that this assumes uniformity in expression of species in the same abundance class L^*. Now, all we need is to find p_{L^*}. Note that $I_a = 1$ when we find *any* instance of a, such that p_{L^*} is one minus the probability that we get 0 reads of a. Let r_{L^*} be the probability that some arbitrary species $a \in s_{L^*}$ is read 0 times in S samples from the mRNA species pool described above. We can model the number of

reads of a we collect over S samples as a binomial random variable N_a, which has parameters $n = S$, $p = \frac{L^*}{\sum |L|}$ (assuming both replacement and uniform probability of selecting a given non-unique strand). By our construction, we have that $r_{L^*} = P(N_a = 0) = \binom{n}{0} p^0 (1-p)^{n-0} = (1 - \frac{L^*}{\sum |L|})^S$. As a result, we have that $p_{L^*} = 1 - r_{L^*} = 1 - (1 - \frac{L^*}{\sum |L|})^S$, such that by plugging back into our original equation, we get:

$$E[J_L(S)] = n_{L^*}\big(1 - (1 - \frac{L^*}{\sum |L|})^S\big)$$

This expression allows us to calculate the mean number of unique species at a given abundance level L^* for a database of size S. For example, consider our original table. Then, the mean number of species at abundance level 50 expected in a sample of 10000 molecules is:

$$\begin{aligned} n_{50}(1 - (1 - \frac{50}{932500})^{10000}) &= 3250(1 - (1 - \frac{50}{932500})^{10000}) \\ &= 1348.87 \end{aligned}$$

These 1348.87 species make up $\frac{1349}{3250} \approx 41.5\%$ of the species at abundance level 50.

From here, we can compute a full table using these values; i.e., the expected percentage of mRNAs in each abundance class for a sample of size S:

Abundance Level	S = 50	5K	10K	50K	250K	1M
5	0.53	2.65	5.22	23.52	73.83	99.53
50	5.22	23.52	41.52	93.15	100	100
200	19.31	65.78	88.29	100	100	100
1K	65.80	99.53	100	100	100	100

Thinking about the distribution of available reads is critical to assembly – our estimates of coverage from Waterman statistics are a wonderful approximation in theory, but break down when we have non-uniformity. The statistical evaluation we have done in examining the Ham Smith problem helps demonstrate that we might have more gaps than anticipated by Waterman statistics in a real world setting.