6.047 Recitation 1 Notes

Samuel Kim (2019) *

September 6, 2018

(numbers) link text to the slides they correspond to

(5) Useful supplements to the biological material covered in this recitation include:

- the first chapter of the compiled scribe notes (especially Section 1.4)
- Chapter 3 of Jones and Pevzner
- Wikipedia - the sections on “The central dogma of molecular biology” and “Gene regulation” are of particular interest
- Khan Academy - Science - Biology for Transcription/Translation

1 Molecular biology review

(6) Like many organic molecules, DNA is a polymer, or long chain, made up of smaller building blocks. These blocks are referred to as nucleotides or bases. The nucleotides of DNA are adenine (A), thymine (T), guanine (G), and cytosine (C).

(7) The first major role of DNA is to pass on genetic information. DNA is replicated to create genetically identical daughter cells, and recombined to give offspring genetic information from their parents. We will discuss the molecular mechanisms of genetic replication and recombination in more detail another week.

(8) The second major role of DNA is to code for proteins, which carry out most cellular functions. The central dogma of molecular biology states DNA is transcribed to RNA which is translated into proteins. RNA is comprised of the same nucleotides as DNA, except uracil (U) is substituted for thymine (T).

DNA is a double-stranded molecule. The two strands are anti-parallel or reverse complements. We measure positions on the DNA molecule in units of base pairs (bp). The direction of each strand is 5′ to 3′ and denotes the direction in which transcription occurs. For this reason, by convention the 5′ end is upstream and the 3′ end is downstream. Transcription is performed by RNA polymerase, which reads the template strand and produces an RNA copy of the coding strand.

Genes encode the sequence which is to be transcribed/translated and can occur on either strand of DNA. The sequence to be transcribed is said to be on the coding or sense strand, and its complement is on the template or antisense strand.

Transcription occurs in three main stages:

- **Initiation** - RNA polymerase binds the the DNA strand, and opens the helix so that it may begin to read the template strand.
  
  Sometimes, transcription factors (TFs) may be required to be bound to promoter region in order to recruit other protein complexes before the RNA polymerase can be bound. We will learn more later about how these promoter regions can thus regulate gene expression.

- **Elongation** - RNA Polymerase grows the chain of nascent RNA in the 5′ to 3′ direction, by adding nucleotides (A,U,C,G) to the chain.

- **Termination** - At the transcription termination site, the RNA Polymerase falls off, leaving the RNA transcript.

The process by which RNA is converted into proteins is called translation. Proteins are also polymers, but of building blocks called amino acids. Each possible 3-mer of RNA bases, known as a codon, codes for a certain amino acid. There are 21 amino acids, but $4^3 = 64$ codons in the genetic code. This redundant code is important for tolerating noise in replication/transcription.

There are many types of RNA molecules which serve diverse functions. These include:

- **messenger RNA** (mRNA) code for proteins.

- **transfer RNA** (tRNA) help the translation machinery connect amino acids in the correct sequence to build a protein. Each type of tRNA binds to a specific codon on one side and a specific amino acid on the other side.

- **ribosomal RNA** (rRNA) is an integral part of the ribosome, which is the molecular machine that performs translation. Ribosomes draw in a mRNA molecule, bring in tRNA molecules that match the mRNA sequence, and join the amino acids on those tRNAs to the nascent protein.
• **long (intergenic) noncoding RNAs (lncRNAs)** are a diverse class of long RNA molecules that do not code for proteins. They may have a variety of roles in gene regulation, though the function of many is not yet well understood. One prominent example is the *XIST* lncRNA, which is responsible for silencing the extra copy of the X chromosome in females.

• **microRNAs (miRNAs)** are short non-coding RNA molecules that bind to mRNA. This can lead to the mRNA being destroyed or less efficient transcription, and thus miRNAs are a mechanism for post-transcriptional regulation of gene expression.

• **small nuclear RNAs (snRNAs)** are short non-coding RNAs that help guide the chemical modification of other RNAs.

(15) Genes begin at the **transcription start site** (TSS). The coding sequence begins and ends with **untranslated regions**, written 5′-UTR and 3′-UTR. The coding sequence is subdivided into **exons**, which will be translated into amino acids, and **introns** which are **spliced** out of the transcribed RNA molecule before translation.

(16) Transcripts produced from the same gene may have different combinations of introns removed, leading to multiple versions of spliced RNA. Different spliced transcripts from the same gene are called **alternate splice isoforms**.

Proteins produced from alternate splice isoforms often have distinct functions, so alternative splicing increases the functional diversity of the proteome that can be generated from the genome. Further, mutations that lead to incorrect splicing often lead to dysfunctional proteins that are associated with disease.

(17) **Regulatory regions** are sections of the genome that may bind proteins required for transcription. In general, the transcription level of a gene is determined by whether its associated regulatory regions are active.

The region upstream of the TSS is called the **promoter** and is responsible for recruiting the various proteins required for transcription to start, particularly RNA polymerase. In addition, proteins called **transcription factors** (TFs) may have to bind to **transcription factor binding motifs** or **DNA sequence motifs** in the promoter region in order for transcription to begin, or indeed for polymerase to bind to the TSS. We call promoters **proximal regulatory regions** as they are in proximity to the gene they target.

We are also interested in **distal regulatory regions** which are distant from their target gene. They may interact with the target gene through conformational changes in the DNA molecule which bring distant parts of a single chromosome or even multiple chromosomes together. **Enhancers** are distal regulatory elements where protein binding tends to enhance the expression of their target genes. **Insulators** are distal regulatory elements that may block the activity of enhancers or certain epigenetic modification mechanisms. Not all genes have distal regulatory regions.

The activity of regulatory regions may be controlled through many mechanisms. For exam-
ple, the TFs that bind to a certain regulatory region must be sufficiently expressed, DNA sequence variation may inhibit or enhance protein binding to a given region, and epigenetic modifications may influence chromatin accessibility.

(18) Epigenetics literally means “above genetics”, and generally refers to modifications to DNA or chromatin. However, a precise definition is still controversial as we learn more about the biology. Some definitions require that epigenetic modifications be heritable, but it is not yet known how some of the modifications we will consider are inherited.

(19) DNA methylation is the addition of a methyl group directly to the DNA molecule. In mammals, this nearly always only occurs on C nucleotides that are followed by a G. We often write “CpG”, where the p denotes the phosphate bond between adjacent bases on the same strand, to distinguish it from the complementary base pair C—G. DNA methylation in promoters is associated with repression of gene expression, by interfering with protein binding, though it may have other functions in other contexts. DNA methylation is known to be heritable.

(21) Histone modification is the addition of small molecules to histone proteins around which DNA is wrapped. Histone proteins form complexes called nucleosomes which are spaced every 200 bp along the genome. Histone modifications are also called chromatin marks, where chromatin refers to the DNA and nucleosomes together.

Many chemical modifications can be found on histones, including but not limited to methylation and acetylation. Histone modifications may help recruit specific proteins to the DNA, or regulate how tightly nucleosomes are bound and therefore how accessible a regulatory region is to transcription factor binding. There are many open questions about histone modifications, including what the specific functions of various modifications are, whether histone modifications themselves are causal or simply a result of processes causing transcriptional activation and silencing, and which histone modifications may be inherited.

**Probability review**

(21) For much of the course, we will be concerned with formulating probabilistic models of biology and using them to infer quantities of interest. Here, we will introduce some concepts of probability theory, and then apply them to solve the problem of finding transcription factor binding sites using position weight matrices.

(22) A discrete random variable $X$ can take multiple values from a sample space $S$, but each with a specific probability specified by a probability mass function. We observe a particular realization $X = x$ with probability $P(X = x)$. We will refer to the probability mass function as $P(X)$.

We need two axioms. First, the probability of any particular realization is between 0 and 1.

\[ 0 \leq P(X = x) \leq 1 \]
Next, the probabilities of all possible realizations sum to 1.

\[
\sum_{x \in S} P(X = x) = \sum_x P(X) = 1
\]

(23) There are other probability theorems that we will be using in this course, the most important of which are:

**Conditional probability**

\[
P(A \mid B) = \frac{P(A, B)}{P(B)}
\]

**Chain rule** For any ordering of \(A, B, C, \ldots:\)

\[
P(A, B, C, \ldots) = P(A)P(B \mid A)P(C \mid A, B) \cdots
\]

**Bayes’ rule**

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}
\]

**Independence** \(A\) and \(B\) are independent if

\[
P(A \mid B) = P(A)
\]

**Conditional independence** \(A\) and \(B\) are conditionally independent given \(C\) if

\[
P(A \mid B, C) = P(A \mid B)
\]

(24) We will build probabilistic models consisting of many random variables \(X = \{X_1, X_2, \ldots, X_n\}\) which are not observed and \(\{Y_1, Y_2, \ldots, Y_m\}\) which are. In this setting, we will study two inference problems:

**Learning** Also known as *updating posterior probabilities* or beliefs, we want to compute \(P(X \mid Y)\). From Bayes rule, we know that we need to compute \(P(Y \mid X)\). From the chain rule, we know that we can factor the joint distribution of \(X, Y\) into smaller conditional distributions; however, we need to be careful about how we do so in order to efficiently compute the desired quantities. The key insight is that we can use conditional independence to simplify the factorization.

**Estimating the most likely configuration** Also known as the *maximum a posteriori* estimate of the parameters \(X\), we want to find the realizations of \(X\) which maximize the posterior probability above, i.e. compute \(\arg \max_X P(X \mid Y)\).
We will now explain a biological example we can apply these tools to. Many transcription factors prefer to bind to specific DNA sequences, often referred to as motifs. The motif of a given transcription factor can be described using a position weight matrix (PWM), which represents a sequence of discrete random variables in the sample space of DNA bases.

To generate a PWM, we might perform ChIP-seq to get a list of sequences that a given transcription factor binds to in a particular cell type. Then, we can calculate the proportion of A’s, C’s, G’s, and T’s at each position in the observed binding sequences.

PWMs are often visually represented using motif logos, where the total height $R_i$ of the letters in a column is the information content in bits at position $i$. If $P(X_{bi})$ is the probability of base $b$ at position $i$, then we can approximate

$$R_i = 2 + \sum_{b \in \text{bases}} P(X_{bi}) \log_2 P(X_{bi})$$

The height of an individual letter in column $i$ is equal to $P(X_{bi})R_i$.

PWMs can be used to find locations in the genomic sequence where the transcription factor might bind. Why would we need to do this if already did ChIP-seq? First, both accessible chromatin and a motif match are often required for a TF to bind successfully, and since chromatin accessibility varies across cell types, the binding locations observed experimentally may not translate to other cell types. With PWMs, which are widely available online, we can measure chromatin accessibility in a new cell type and then easily predict binding sites for as many factors as we have PWMs.

We now consider the problem of determining whether we believe a given sequence in the genome is a match to the motif or is part of the general genomic background.

In other words, we would like to calculate $P(M \mid S)$, or the probability that we have a motif instance given the sequence we have observed.

Our prior belief is that $P(B)$, the probability of a sequence belonging to the background, is equal to 0.9, while $P(M)$, the probability of a sequence matching the motif, is equal to 0.1.

Our background or null model of nucleotide frequencies in the genome is as follows:

$$\begin{array}{cccc}
A & G & T & C \\
0.1 & 0.4 & 0.1 & 0.4 \\
\end{array}$$

We can interpret this in two ways: each entry represents the relative frequency of that nucleotide in the genome, or this model gives a set of parameters from which we could generate a sequence.

Finally, we have a PWM representing the DNA sequence binding motif of interest:
<table>
<thead>
<tr>
<th>Position</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Solution**

Using Bayes' Rule:

\[
P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}
\]

Since \(S\) must either be a motif or background, \(P(S) = P(S, M) + P(S, B)\). So,

\[
P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S, M) + P(S, B)}
\]

By the chain rule,

\[
P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S \mid M)P(M) + P(S \mid B)P(B)}
\]

To evaluate this expression, we need the values of \(P(S \mid B)\) and \(P(S \mid M)\).

Assuming that each nucleotide in the genome is independent (note: this is not true), the probability that this sequence comes from the null model is \(P(S \mid B) = 0.4 \times 0.4 \times 0.1 \times 0.1 = 1.6 \times 10^{-3}\).

Similarly, the probability that this sequence came from the motif model is \(P(S \mid M) = 0.4 \times 0.25 \times 0.1 \times 1 = 0.01\).

Plugging these values into the expression we derived for \(P(M \mid S)\) above, we can calculate the *posterior probability*

\[
P(M \mid S) = \frac{0.01 \times 0.1}{0.01 \times 0.1 + 1.6 \times 10^{-3} \times 0.9} = 0.41
\]

Interpreting this result, the sequence we observed is more likely to have come from the null model than from the motif model, suggesting it is not a real motif.