6.047 Recitation 1 Notes

Sitara Persad (2017) *

September 8, 2017

(numbers) link text to the slides they correspond to

(4) 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

(5) 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).

(6) DNA plays two central roles in cell functions: it is replicated to pass genetic information on to daughter cells and it codes for proteins. We leave replication aside until we discuss population genetics in Module IV.

(7) The central dogma of molecular biology states DNA is transcribed to RNA which is translated into proteins. Like DNA, RNA is also comprised of nucleotides. The nucleotides of RNA are the same except uracil (U) is substituted for thymine (T). Because the difference between DNA and mRNA is minimal, it is called transcription. You can remember it as re-writing (transcribing) notes in the same language (the language of nucleotides).

(8-9) 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 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.

(10-11) The process by which RNA is converted into proteins is called translation. This is because proteins are considerably different from DNA and RNA, so the process of turning RNA into proteins, amino acid by amino acid is akin to translating notes into a different language.

Proteins are also polymers, but of building blocks called amino acids. 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.

(12-13) There are many types of RNA molecules which serve diverse functions:

- messenger RNA (mRNA) is produced by transcription and is the mechanism by which the code for proteins is moved from within the nucleus to the translation machinery outside the nucleus

- transfer RNA (tRNA) brings amino acids to the translation machinery and are the mechanism by which the translation machinery constructs proteins

- ribosomal RNA (rRNA) makes up the ribosome which is the driver of translation. Ribosomes draw the mRNA through them, join it to tRNA which recognize a particular word of the sequence, and bind the attached amino acid to the protein being constructed.
- **microRNAs** (miRNAs) are small RNA molecules which silence genes by binding to mRNA.

- **long (intergenic) noncoding RNAs** (lncRNAs) interact with the structure of DNA molecules. A prominent example is the **XIST lncRNA**, which is responsible for silencing the extra copy of the X chromosome in females and preventing double expression of X chromosome genes (called dosage compensation) and has been recently engineered to silence other chromosomes.

- **small nuclear RNAs** (snRNAs) modify other RNAs, for example to process pre-mRNA.

- **piwi-interacting RNAs** (piRNAs) are involved in gene silencing.

(14) 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 are expressed (translated into amino acids) and **introns** which are spliced out of the transcribed RNA molecule before translation.

(15) The spliced RNA is not necessarily unique for each gene; that is, different introns may be removed to give multiple versions of spliced RNA. Different RNA transcripts coming from the same gene are called **alternate splice isoforms**.

Variation between these transcripts plays significant roles, and the disruption of the splicing mechanism can have major phenotypic consequences. For example, a mutation in exon 7 of the SMN gene does not affect which amino acid is produced by the DNA sequence. However, it does affect splicing, which leads to a biologically useless protein and spinal muscular atrophy in affected individuals.

(16-17)

The region upstream of the TSS is called the **promoter** and is responsible for recruiting the various proteins required for transcription to start. In addition to RNA polymerase itself, proteins called **transcription factors** (TF) 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.

Transcription factors may also bind to regions further from the TSS, and be brought closer to the start of transcription through conformational changes in DNA. We call promoters **proximal regulators** as they are in proximity to the gene they target.

' We are also interested in **distal regulators** which are distant from their target gene. **Enhancers** are similar to promoters in that proteins bind to motifs contained within. However, they interact with the transcriptional machinery through conformational changes in the DNA molecule which bring distant parts of a single chromosome or even multiple chromosomes together. **Insulators** are similar to enhancers but interfere with interactions required for transcription. Enhancers are said to **upregulate** or increase gene expression where insulators **downregulate**.
18) Epigenetics literally means "above genetics". However, a precise definition is still controversial as we learn more about the biology. One potential definition is heritable modifications to the DNA molecule itself. However, some epigenetic marks we will consider are not yet known to be heritable.

19) DNA methylation is the addition of a methyl group directly to the C nucleotide in a CpG. 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. Methylation is known to be heritable; however, there are many open questions about the biology of methylation.

20) 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. We will learn about the various modifications which we know about and how they form a sort of epigenetic code in tandem with the genetic code. There are many open questions about histone modifications, such as whether the epigenetic code specifies the genetic code or vice versa.

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. By way of introduction to the probability theory we will be using, we introduce a basic model for simulating transcription factor binding.

A motif is a pattern that has some biological significance. Here we will consider motifs in DNA sequence that describe where transcription factors are likely to bind. Such motifs are usually represented by position weight matrices (PWMs). Each position is represented by a discrete random variable.

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 density function. We observe a particular realization $X = x$ with probability $P(X = x)$. We will refer to the probability density function as $P(X)$.

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

$$0 \leq P(X = x) \leq 1$$
If we sum the probability of each realization over all possible realizations, we get 1

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

Discrete random variables are thus used in PWMs, which take into account the fact that proteins may bind to sequences that are not fully-defined. As opposed to fully-defined sequences (i.e. AACGTAA), these sequences may allow some positions to be occupied by any one of multiple nucleotides.

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

**Conditional probability**

$$P(A | B) = \frac{P(A, B)}{P(B)}$$

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

$$P(A, B, C, \ldots) = P(A)P(B | A)P(C | A, B)\ldots$$

**Bayes’ rule**

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

**Independence**

$$P(A | B) = P(A)$$

**Conditional independence**

$$P(A | B, C) = P(A | B)$$

(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 $P(X | Y)$. From Bayes rule, we know that we need to compute $P(Y | X)$. 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 $\arg \max_x P(X | Y)$. 
To exercise some of these tools, we consider the problem of distinguishing between real DNA sequence motifs and those which occur just by chance. Suppose you are given a background or null model of nucleotide frequencies in the genome:

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

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.

Now suppose we have a sequence \( S = GCAA \). 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 \( P(S \mid B) = 0.4 \times 0.4 \times 0.1 \times 0.1 = 1.6 \times 10^{-3} \).

Given that the human genome has roughly \( 3 \times 10^9 \) bp and assuming that each 4-mer in the human genome is independent (note: this is not true), we would expect to find \( (1.6 \times 10^{-3})(3 \times 10^9) = 4.8 \times 10^6 \) occurrences of \( GCAA \) in the genome just by chance. This fact motivates much of the innovation in motif discovery which we will learn about.

Now, suppose I have a model of a DNA sequence binding motif. This model can be interpreted in two ways: it gives the probability of seeing a particular nucleotide in a particular position in a real motif, or it is a set of parameters from which we could generate motifs. This representation is our PWM:

<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>.6</td>
<td>.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>.1</td>
<td>.1</td>
<td>.4</td>
<td>.4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

From this PWM, we can compute \( P(S \mid M) = 0.4 \times 0.25 \times 0.1 \times 1 = 0.01 \). However, a more interesting quantity is \( P(M \mid S) \), i.e. how likely is it that the observed sequence came from a real motif? Given priors \( P(B) = 0.9, P(M) = 0.1 \), we can compute this posterior probability using Bayes’ Rule:

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

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

\[
= \frac{1.6 \times 10^{-3} \times 0.9 + 0.01 \times 0.1}{0.01 \times 0.1}
\]

\[
= 0.41
\]
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.