6.047/6.878 Fall 2012 Recitation 1 Notes

September 8, 2012

Author: Melissa Gymrek (2011)
Revised by Rachel Sealfon (2012)

1 Basic definitions in molecular biology

As may be obvious, there are many concepts and new terms in molecular biology. Encountering all of this terminology can be bewildering at first. During recitation, we will discuss (at least at the level of a computational practitioner) many of these concepts. The first chapter of the course textbook has a nice summary of the crash course in molecular biology from yesterday’s class (section 1.4). Other resources for broadening your biological background include Chapter 3 of the Jones and Pevzner book, “Molecular biology for computer scientists” by Lawrence Hunter (www.biostat.wisc.edu/~craven/hunter.pdf; start with pages 23 to 29) and of course Wikipedia. The sections on “The central dogma of molecular biology” and “Gene regulation” are relevant and well done.

To review from lecture:

1. The central dogma of molecular biology states that DNA is transcribed to mRNA which is translated to protein. Notice that because the nucleotide difference between DNA and mRNA is minimal, it is called transcription whereas the reading of mRNA to construct proteins is called translation.

2. The fundamental building blocks of DNA are A, T, G, C. RNA has the same nucleotides except for T which is replaced by U. The basic building block of proteins are amino acids.

3. The mechanism of DNA replication is implied by the structure of DNA. Two antiparallel strands are held together by weak hydrogen bonds, so that the strands can be pulled apart for replication or transcription. A pairs with T and G pairs with C, so the sequence of one strand is completely implied by the sequence of its complement. Nucleotides in individual strands are held together by strong covalent bonds along the phosphate and sugar backbone. Strands are directional, and are written by convention in the 5’ to 3’ direction. (5’ and 3’ refer to the positions on the pentose ring where the phosphate backbone connects).

4. Genes can occur on either strand of DNA. The DNA before a gene (in the 5’ region) is considered “upstream” whereas the DNA after a gene (in the 3’ region) is considered “downstream”. RNA polymerase synthesizes a new RNA molecule by adding bases complementary to the template strand of DNA (antisense strand). The new RNA molecule grows in the 5’ to 3’ direction and looks like the coding strand of DNA (sense strand).

5. Genes in DNA are interrupted by introns that do not code for proteins but often play an important role in regulation. mRNA has these introns stripped away and only contains exons or regions that are expressed.

6. Genes are composed of sequences of nucleotide triplets called codons. Each codon specifies an amino acid. There are 64 possible codons (4 possible nucleotides in each of 3 positions), but only 20 amino acids are encoded. Therefore, the codon code is redundant, with some amino acids encoded by multiple codons. In particular, for many codons, changing the base at the third position does not alter the amino acid encoded.

7. DNA is coiled around histone proteins, forming nucleosomes. These are further packed and condensed into chromatin fiber and chromosomes. Many organisms have their DNA broken into several chromosomes.
8. Epigenetics is a term that has different meaning to different people, but most broadly means any inherited phenotypic change that is not due to a change in DNA sequence. This often refers to DNA methylation (chemical modifications to the nucleotide bases) or chemical modifications to the histones that make up the packaging of DNA. *Epigenomics* refers to the large-scale study of epigenetic mechanisms using modern genomics tools.

9. *Transcription factors* are proteins that bind DNA and affect (promote or repress) gene expression. *MicroRNAs* are small (surprise!) RNA molecules that can bind to messenger RNAs and can induce the degradation of the bound mRNA. Both items fall under the generic category of *regulators of gene expression*.

10. A *motif* is a pattern that has some biological significance. Most commonly, we will talk about DNA sequence motifs. Some examples of motifs are sequences where specific transcription factors bind or sequences that can promote mRNA splicing.

## 2 Probability

1. We will quickly review some basic probability by considering an alternate way to represent motifs: a *position weight matrix* (PWM). We would like to model the fact that proteins may bind to motifs that are not fully specified. That is, some positions may require a certain nucleotide (e.g. A), while others positions are free to be a subset of the 4 nucleotides (e.g. A or C). A PWM represents the set of all DNA sequences that belong to the motif by using a matrix that stores the probability of finding each of the 4 nucleotides in each position in the motif. For example, consider the following PWM for a motif with length 4:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
<td>0.25</td>
<td>0.10</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.4</td>
<td>0.25</td>
<td>0.10</td>
<td>0.0</td>
</tr>
<tr>
<td>T</td>
<td>0.0</td>
<td>0.25</td>
<td>0.40</td>
<td>0.0</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>0.25</td>
<td>0.40</td>
<td>0.0</td>
</tr>
</tbody>
</table>

We say that this motif can generate sequences of length 4. PWMs typically assume that the distribution of one position is not influenced by the base of another position. Notice that each position is associated with a probability distribution over the nucleotides (they sum to 1 and are nonnegative).

2. We can also model the *background distribution* of nucleotides (the distribution found across the genome):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.4</td>
</tr>
<tr>
<td>T</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Notice how the probabilities for A and T are the same and the probabilities of G and C are the same. This is a consequence of the complementarity DNA which ensures that the overall composition of A and T, G and C is the same overall in the genome.

3. Consider the sequence $S = \text{GCAA}$.

   The probability of the motif generating this sequence is $P(S|M) = 0.4 \times 0.25 \times 0.1 \times 1.0 = 0.01$.

   The probability of the background generating this sequence $P(S|B) = 0.4 \times 0.4 \times 0.1 \times 0.1 = 0.0016$.

4. Alone this isn’t particularly interesting. However, given fraction of sequences that are generated by the motif, e.g. $P(M) = 0.1$, and assuming all other sequences are generated by the background ($P(B) = 0.9$) we can compute the probability that the motif generated the sequence using Bayes’ Rule:
\[
P(M|S) = \frac{P(S|M)P(M)}{P(S)}
\]
\[
= \frac{P(S|M)P(M)}{P(S|B)P(B) + P(S|M)P(M)}
\]
\[
= \frac{0.0016 \times 0.9 + 0.01 \times 0.1}{0.00984} = 0.40984
\]

3 Python

Python is a popular programming language that is frequently used in computational biology. Its main features are simple syntax, dynamic typing, and a large number of supporting libraries. Python is available within all MIT Athena accounts and is also available for download at http://python.org for all platforms (Windows, Linux, Mac OS X).

There are several tutorials and documentation sites for Python:


We won’t be using any complex or exotic features of Python, so it is probably not necessary to buy a Python book just for this course. If you would like one for the future, however, Learning Python and Programming Python by Mark Lutz are both excellent. You can access both of these books online at Safari Books using your MIT Library account (go to: http://proquest.safaribooksonline.com.libproxy.mit.edu/).

3.1 Brief summary of Python commands to learn

Here is a brief tour of basic Python language features, including print, variables, functions, lists, list comprehensions, loops, tuples, dictionaries, import, dir, and help. You should type these commands into the interactive interpreter to learn what each command does.

```python
# hello world
print 'hello, world!
print "hello, world!"

# functions & variables
def fact(n):
    if n == 0 or n == 1:
        return 1
    else:
        return n*fact(n-1)

print fact(8)
x = fact(8)

# print with formatting syntax
# In the C programming language this would be:
# printf("x = %d", x);
print 'x = %d' % x
```
# lists & loops

```python
lst = [1, 2, 3, 4]
print lst
print lst[2]
lst[2] = 0
lst
lst.append(5)
lst
del lst[0]
lst

lst = range(1,5)
lst

print len(lst)
for i in range(1,5):
    print i

# list comprehension (advanced feature)
print [x*x for x in [1,2,3,4]]

# equivalent to
lst = []
for x in [1,2,3,4]:
    lst.append(x*x)
print lst

# filtering with list comprehension (advanced feature)
print [x*x for x in range(1,11) if x % 2 == 0]

# equivalent to
lst = []
for x in range(1,11):
    if x % 2 == 0:  # only append even numbers, x / 2 has remainder 0
        lst.append(x*x)
print lst

# tuples
from math import sqrt

def csqrt(n):
    if n >= 0:
        return (sqrt(n),0)
    else:
        return (0,sqrt(-n))

real, imag = csqrt(-16)
print '%d+%di' % (real, imag)

# dictionaries (hash tables)
profs = {}
profs['6.047'] = 'kellis'
profs['7.012'] = 'lander'
```
profs['6.047']
profs['bogus']

# import, dir, help
cos(0)
import math
dir(math)
help(math.hypot)

# matrices (lists of lists)
m = [[0 for j in range(10)] for i in range(10)]
m
m[3][4]
m[4][5] = 6
m

4 Announcements

- Sign up to scribe on the course wiki: [https://wikis.mit.edu/confluence/display/6DOT047fa12/Scribing+sign-up](https://wikis.mit.edu/confluence/display/6DOT047fa12/Scribing+sign-up)

- First problem set has been posted, due 9/24.

- Start thinking about project ideas: browse recent publications, look at lecture topics, go to interesting computational biology talks at Broad/Whitehead/bio department/CS department/Harvard. A list of some large project databases is posted for your reference on the course wiki.

- TA office hours on Mondays from 6-8 pm, couches area outside 32D-507.