6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

# 6.047/6.878 Fall 2008 Recitation 1 Notes

Pouya Kheradpour, Matt Rasmussen

September 5, 2008

## 1 Python

Python is popular a 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:

- Official tutorial, `http://docs.python.org/tut/tut.html`
- Library reference, `http://docs.python.org/lib/lib.html`
- WikiBooks, `http://en.wikibooks.org/wiki/Programming: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.

### 1.1 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.

```
# 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
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
from math import cos
```

```
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
```

# 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:

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 0.6 | 0.25 | 0.10 | 1.0 |
| G | 0.4 | 0.25 | 0.10 | 0.0 |
| T | 0.0 | 0.25 | 0.40 | 0.0 |
| C | 0.0 | 0.25 | 0.40 | 0.0 |

   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):

| | |
|---|---|
| A | 0.1 |
| G | 0.4 |
| T | 0.1 |
| C | 0.4 |

   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 $S =$ 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:

$$\begin{aligned} P(M|S) &= \frac{P(S|M)P(M)}{P(S)} \\ &= \frac{P(S|M)P(M)}{P(S|B)P(B) + P(S|M)P(M)} \\ &= \frac{0.01 \times 0.1}{0.0016 \times 0.9 + 0.01 \times 0.1} = 0.40984 \end{aligned}$$

# 3 Basic definitions in molecular biology

1. The fundamental building blocks of DNA are A, T, G, C. RNA has the same *nucleotides* except for T which is replaced by U.

2. The central dogma of molecular biology states that DNA is *transcribed* to mRNA which is *translated* to proteins. 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.

3. Genes in DNA are **int**errupted by **int**rons that do not code for proteins but often play an important role in regulation. mRNA has these introns stripped away and only contains **ex**ons or regions that are **ex**pressed.

4. Many organisms have their DNA broken into several chromosomes. Each chromosome contains two strands of DNA, which are complementary to each other but are read in opposite directions. 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".