MATE 6685 - Bioinformatics

Humberto Ortiz Zuazaga

humberto@hpcf.upr.edu

http://www.hpcf.upr.edu/~humberto/

Outline

- Programming languages
- Why program?
- What is Python?
- Why Python?
- How?

Programming languages

A formal language in which to direct the computer. It's like a very specific vocabulary for writing laboratory protocols. A program is like a lab protocol even work-study students could follow.

Why program?

If you have ever created a large excell (or Lotus, or ...) spreadsheet, you have programmed. Real programming languages give you more control, more expressive power. Even if you are “just a biologist” eventually you will want to do something no program has been written to do. You can either explain the science to a programmer, or do it yourself!

Finally, a program is both a very specific model of a biological system, and a method for simulating those models, testing them, and making predictions.

What is python?

1. Free!
2. Interpreted language
3. Dynamic typing
4. Free!
5. Object oriented
6. Very high level
7. Free!

Simple Sequence Analysis

- %GC content
- Tm $= 64.9 + 41 \times (G + C - 16.4)/(A + T + G + C)$
- MW $= (A \times 313.21) + (T \times 288.20) + (G \times 329.21) + (C \times 289.19) + 18.02$

Counting bases

```
sequence = "ATCGGACCTACGCCTCAAGCACCTACATCCCGATAGAAGACCCTTTT"

total = len(sequence)
gccount = 0

for base in sequence:
    if 'G' == base or 'C' == base:
        gccount = gccount + 1

print 100.0 * gccount / total
```

Functions

```
def gccontent(sequence):
    """Return the percent GC content of an
       unambiguous nucleotide SEQUENCE"""
    total = len(sequence)
    gccount = 0

    for base in sequence:
        if 'G' == base or 'C' == base:
            gccount = gccount + 1

    return 100.0 * gccount / total
```

Calling a function

```
seq1 = "ATCGGACCTACGCCTCAAGCACCTACATCCCGATAGAAGACCCTTTT"
seq2 = "ATCGGACCCCGTAGACAATTCAAGCACCTACATCCCGATAGAAGACCCTTTT"
seq3 = "GACCATACTGCCTCCAAGCAGCCTAACAATCCCCGTATGAGGAAGAACCC"

print seq1, gccontent(seq1)
print seq2, gccontent(seq2)
print seq3, gccontent(seq3)
```

Running python code

```
$ python gccontent.py
ATCGGACCTACGCCTCAAGCACCTACATCCCGATAGAAGACCCTTTT 51.0638297872
ATCGGACCCCGTAGACAATTCAAGCACCTACATCCCGATAGAAGACCCTTTT 48.0769230769
GACCATACTGCCTCCAAGCAGCCTAACAATCCCCGTATGAGGAAGAACCC 54.0
```

Homework

I have put the `gccontent.py` program on the course page. Using this as a model, write a function to compute the melting point of a sequence, using the formula:

$$\text{Tm} = 64.9 + 41 \times (G + C - 16.4)/(A + T + G + C)$$

Print the %GC and melting points of `seq1 seq2 seq3`.