```
seq[-1]
seq. (press tab key)
seq.find?
seq.find('ATG')
seq[42]
seq[42:45]
seq.count('ATG')
seq[::-1]
```

Python uses object oriented programming.

25 Example (Object Oriented Programming)
For example, consider the python command:

```
marker = MARKER(color = blue)
```

The function `MARKER(color = blue)` is a factory function which manufactures objects, in this case markers. The `color=blue` argument specifies that a blue marker should be manufactured.

Objects have attributes. For example, the color attribute `marker.color` should equal blue.

Objects also have methods. For example, the marker method `marker.change_color(red)` changes the color attribute of the marker from blue to red.

In IPython, typing `marker.` followed by the tab key will list all the attributes and methods associated with the marker object. Typing `marker?` will provide information about the marker object.

---

26 Example (Biopython)
Explain what the following Biopython commands do:

```
file = open('insulin_cDNA.txt')
seq = handle.readline().strip()
print seq

from Bio.Seq import Seq
from Bio.Alphabet import IUPAC
DNA = Seq(seq,IUPAC.unambiguous_dna)
DNA
print DNA
DNA?
DNA. (press tab key)
print DNA.reverse_complement()
mRNA = DNA.transcribe()
print protein
protein.find('M')
protein.find('*')
print protein[14:125]
```

---

9 Lesson (Translating DNA)
Download the following files:

```
insulin_human_DNA.txt
insulin_human_cDNA.txt
```

(a) How many start codons are there in a the complete gene for human DNA? Make sure you check all six reading frames.

Solution:

(b) Translate the cDNA sequence for insulin to a protein sequence. Check your answer using Uniprot.

---

4 Sequence Alignment

DNA is subject to mutations. We will only consider insertions, deletions and substitutions.

27 Definition (Mutations)

```
orignal sequence   ATTGCTCC

orignal sequence   ATTG_CTCC
       insertion   ATTGGCTCC

orignal sequence   ATTGCTCC
        deletion   ATT_CTCC

orignal sequence   ATTGCTCC
    substitution   ATTTCTCC
```

28 Example (Sequence Alignment)
Consider the sequences:

```
TAGTA
ATAT
```

Before we can determine how similar the sequences are to each other, we must first align the sequences. Two optimal alignments obtained using *dynamic programming* are:

```
TAGTA   _TAGTA
_A_TAT  ATA_T_
```

29 Example (Dot Plots)
Use a dot plot to compare the following sequences:

```
TAGTA
ATAT
```

	T	A	G	T	A
A		○			○
T	○			○	
A		○			○
T	○			○	

10 Lesson (Dot Plots)
How similar are human, horse and chicken insulin? Use Jemboss to create dot plots comparing the insulin sequence for each.

- Go to www.uniprot.org.
- In the search field click on advanced.
- Select Gene name [GN] and type INS (for the insulin gene).
- Scroll down the results and click on the check box in the left column for human, horse and chicken insulin.
- Select download and a new window will appear containing the insulin sequences for human, horse and chicken in `fasta` format.
- Open Jemboss.
- Select ALIGNMENT, Dot Plots, polyplots.
- Cut and paste the `fasta` sequence data into Jemboss.
- Select pdf format for the output.
- Go to the Jemboss folder to retrieve the results.
- Interpret the plots.

11 Lesson (Dot Plots)
Repeat the previous lesson except compare the following insulin sequences:

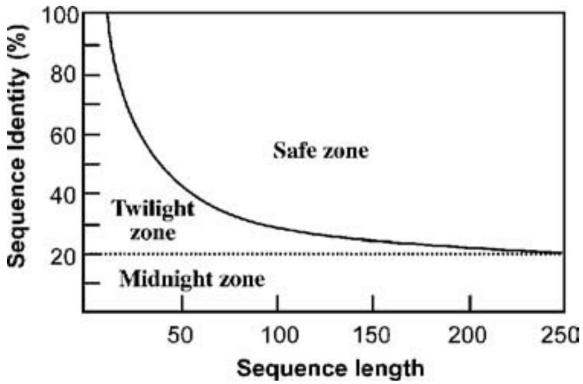
```
P01319 INS_CAPHI (Goat)
P01317 INS_BOVIN (Cow)
P01318 INS_SHEEP
```

(You may need to click in bottom right corner to display all insulin sequences at once.)

30 Definition (Homology)
Sequences which have evolved from a common ancestor are called **homologous**.

Similar sequences are likely to be homologous. However, we should keep in mind that sequences that have evolved from a distant ancestor may no longer be very similar to each other.

31 Definition (Sequence Alignment Zones)



Jin Xiong, Essential Bioinformatics, p. 33.

- safe zone: sequences are very likely to be homologous.
- twilight zone: sequences may be homologous.
- midnight zone: no reliable conclusion possible.

---

32 Definition (Percent Sequence Identity and Similarity)
After two sequences have been aligned, sequence identity and similarity is computed in one of two possible ways:

L_a is the length of the shorter sequence.

L_b is the length of the longer sequence.

N is either the number of identical or the number of similar letters in the alignment.

Sequence identity/similarity is computing using one of the two following formulas:

Formula 1

$$I = 100\,\frac{N}{L_a}$$

Formula 2

$$I = 100\,\frac{N}{\frac{L_a+L_b}{2}}$$

---

12 Lesson (Sequence Identity and Similarity)
Use `uniprot.org` to align cow insulin `P01317`, sheep insulin `P01318` and goat insulin `P01319`.

(a) In the `uniprot.org` search box type

```
P01317 or P01318 or P01319
```

Select the check boxes for these Ainsulin sequences and then select the alignment button. Wait a few seconds for the alignment to be computed by `uniprot.org`.

(b) Which sequences have a signal peptide attached? (Hint: check the box `signal peptide` in left column.)

(c) Which sequences have the propeptide attached? (Hint: check the box `propeptide` in left column.)

(d) Which sequences have the peptide segment? (Hint: check the box `peptide` in left column.)

(e) Complete the following tables *using only the peptide segment of each sequence*.

Sequence Identity:

	cow	sheep	goat
cow	100%		
sheep		100%	
goat			100%

Sequence Similarity:

	cow	sheep	goat
cow	100%		
sheep		100%	
goat			100%

Solution:

---

33 Definition (Paralogs)
If two sequences *from the same organism* are homologous, then the sequences are **paralogs**.

---

13 Lesson (Paralogs)
Use `uniprot.org` to align the insulin protein sequences: `P01325, P01326, P01322, P01323`.

(a) Which pairs of sequences are homologs and which are paralogs? Explain.

(b) Look at just the peptide segment of each sequence. (Check the box `peptide` in the left column.) Did the insulin gene duplicate before or after mouse and rat become separate species? Justify your answer.

---

34 Definition (Local vs Global Alignment)
Two basic types of sequence alignments are possible:

- local alignment (also called Smith-Waterman alignment)
- global alignment (also called Needleman-Wunsch alignment)

If sequence lengths are very different, we should consider using a local alignment. If the sequences are of similar lengths and likely to be closely related, we should use a global alignment. (Local alignments are used more often than global alignments.)

---

14 Lesson (Local Alignment)
Use Jemboss to align the cDNA and DNA sequences for human insulin. Identify where the exons and introns in the DNA sequence for human insulin are located.
