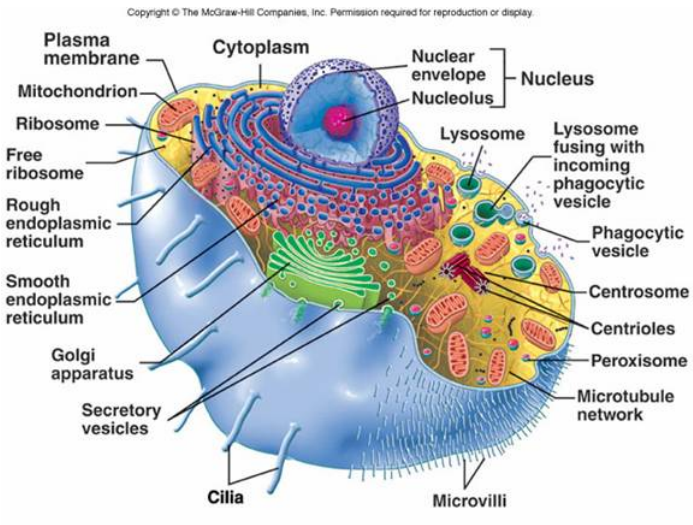
1 Central Dogma of Biology

1 Definition (The Cell)



2 Definition (The Cell)

- single cell organisms
- multi cell organisms

3 Definition (The Cell)

- prokaryotes
- eukaryotes

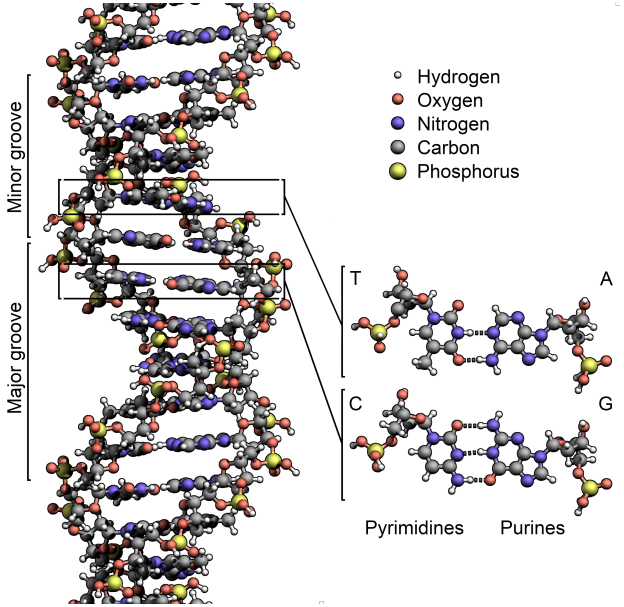
4 Definition (Nucleic Acids)

- DNA: A T G C
- RNA: A U G C

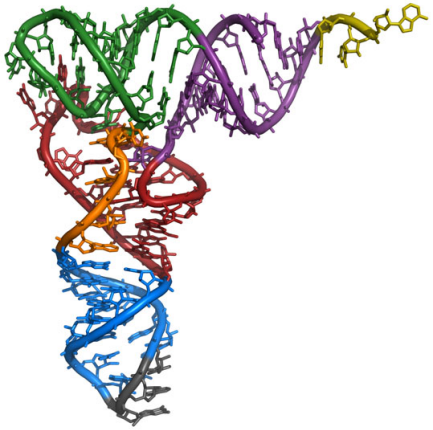
5 Definition (Amino Acids)

Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile
A	R	N	D	C	E	Q	G	H	I
Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
L	K	M	F	P	S	T	W	Y	V

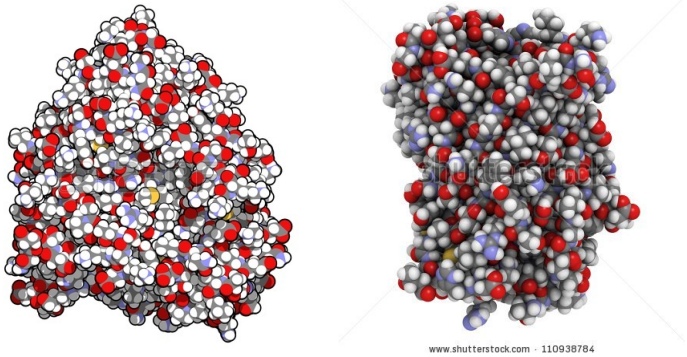
6 Definition (DNA Molecule)



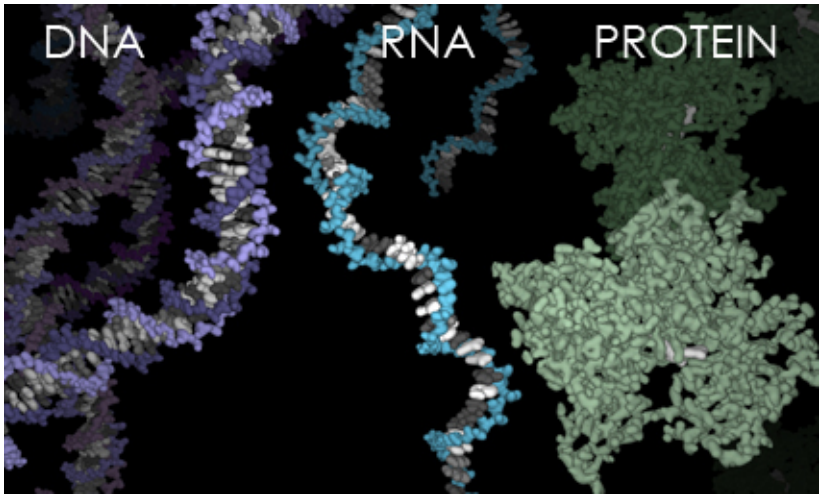
7 Definition (tRNA Molecule)



8 Definition (Protein Molecules)



9 Definition (DNA vs RNA vs Protein Molecule)



1 Lesson (Molecules of Life)
Discuss how the four types of molecules, DNA, mRNA, tRNA and proteins are:

(a) similar.

(b) different.

2 Lesson (Genetic Code)
How can the 20 different amino acides of protein sequences be coded using only the four different nucleic acids of DNA?

10 Definition (Genetic Code)

Genetic Code

First Position	Second Position T		C		A		G		Third Position
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop	A
	TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

11 Definition (Genes)
Genes are segments of DNA that are transcribed and translated into a protein sequence. Splicing may be required.

- codon: group of three nucleic acids that code for a single amino acid.
- introns: portions of a gene that are removed before translation.
- exons: portions of a gene that are spliced before translation.

3 Lesson (cDNA)
What is cDNA?

4 Lesson (DNA Transcription and Translation)
The following cDNA sequence contains the beginning of the gene that codes for the human insulin protein.

```
TGCCTGTCTCCCAGATCACT
GTCCTTCTGCCATGGCCCTG
TGGATGCGCCTCCTGCCCC
```

(a) Determine the sequence of cDNA that codes for the first nine letters of the insulin protein sequence.

Solution:

(b) Determine the sequence of mRNA that codes for the first nine letters of the insulin protein sequence.

Solution:

(c) Determine the first nine letters of the insulin protein sequence.

Solution:

(d) Use the Uniprot database to check your answer to part (c) by looking up the protein sequence for the human insulin protein.

Solution:

5 Lesson (Transcription and Translation)
Use the link below to look at an animation of the transcription and translation of DNA and discuss what you see.

`http://www.dnalc.org/resources/3d/central-dogma.htmli`

Because it takes three letters of a DNA sequence to translate to a single letter of a protein sequence, where we start the translation in the DNA sequence has a big effect on the resulting protein sequence. In other words, we have to choose the correct **reading frame** before we attempt to translate a DNA sequence into a protein sequence. Reading frames which start with the start codon `ATG` are called **open reading frames.**

6 Lesson (Jemboss (Translation))
Install the free bioinformatics software package `Jemboss` and use it to check your answers to Lesson 4 part (c). Use the commands

`NUCLEIC, TRANSLATION, transeq`

7 Lesson (Jemboss (Open Reading Frames))

(a) Determine three of the six possible translations of the following segment of DNA.

T A T A G G G A C T C A

(b) Check your answer with `Jemboss`. Use the commands `NUCLEIC`, `TRANSLATION`, `sixpack`

Solution:

8 Lesson (Jemboss — Open Reading Frame)
Use Jemboss to translate the human insulin gene DNA sequence into the protein sequence for insulin using:

(a) the GenBank entry J00265.1. (Hint: Use Google to search for GenBank and select `Nucleotide` and search for J00265.1).

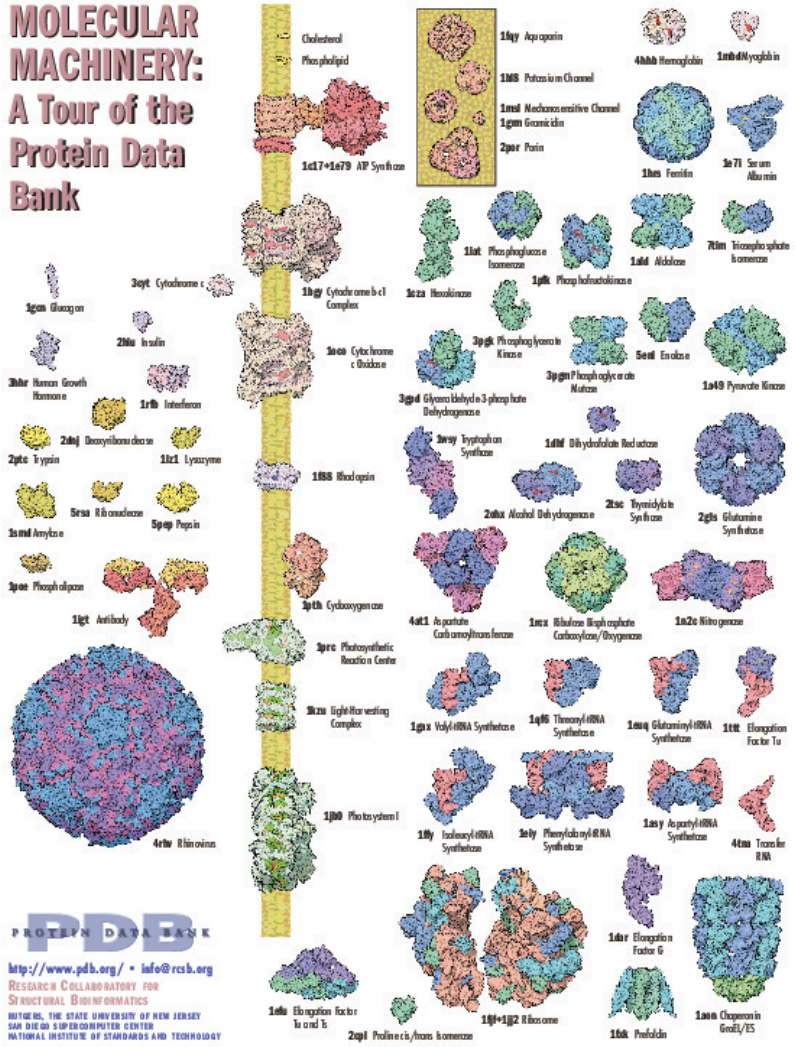
(b) the file `insulin_human_cDNA.txt` which contains the sequence for the cDNA for human insulin.

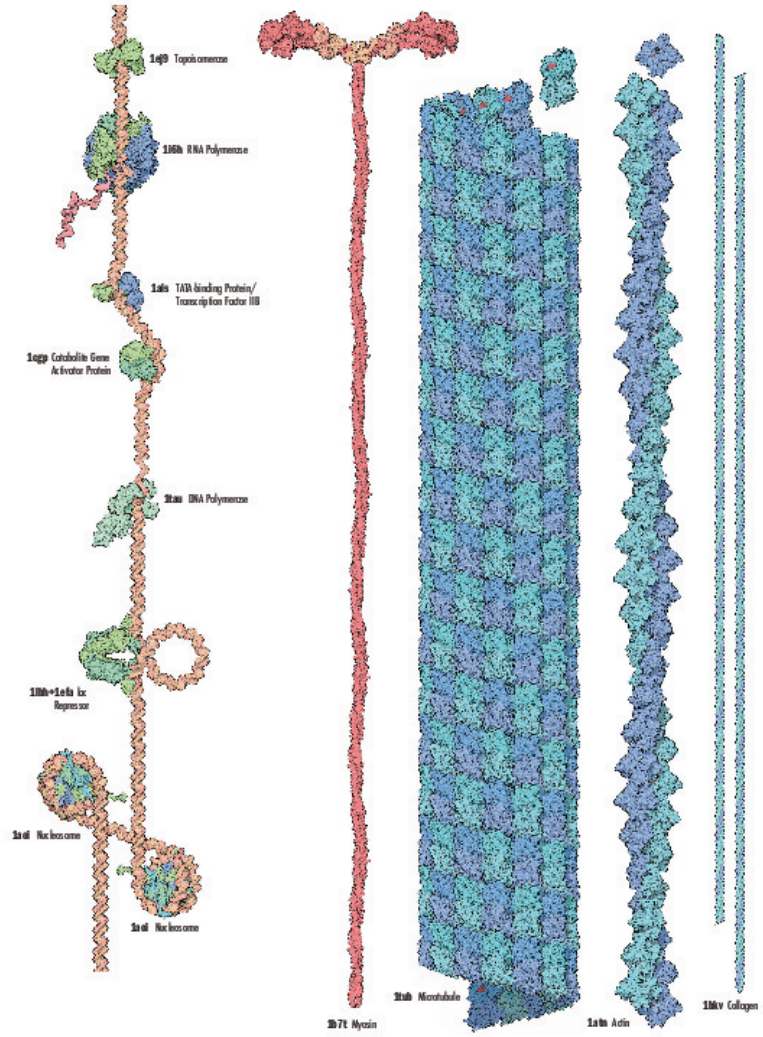
(c) Compare the protein sequence from part (a) and part (b). The first 62 letters of the sequences should be the same. Why are the rest different?

(d) The section `CDS` in the GenBank insulin entry J00265.1 describes special aspects of the insulin gene, e.g. where the exons are. Use this section to determine the length of the first protein segment corresponding to the first exon of the insulin gene. Does this length agree with part (c)?

12

2 Proteins





13

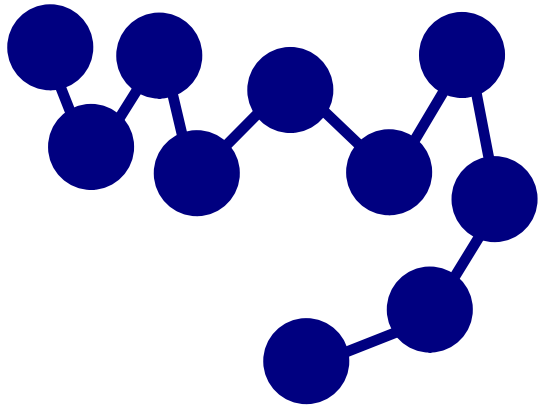
---

14 Definition (Protein Structure)

- primary structure
- secondary structure
- tertiary structure
- quaternary structure

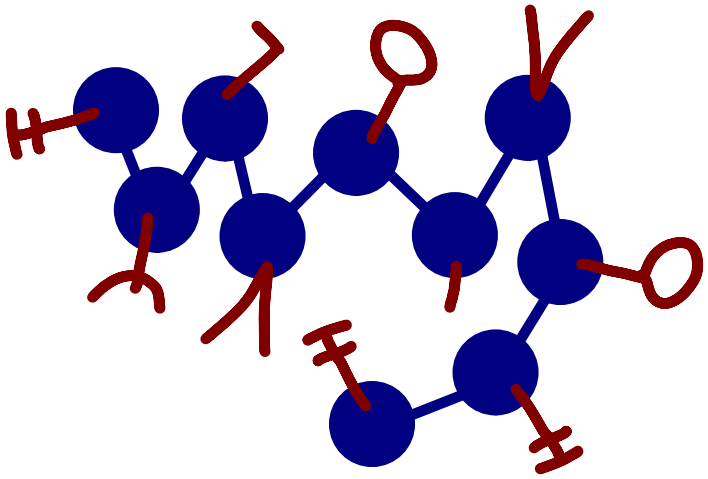
---

15 Definition (Primary Structure—Backbone)



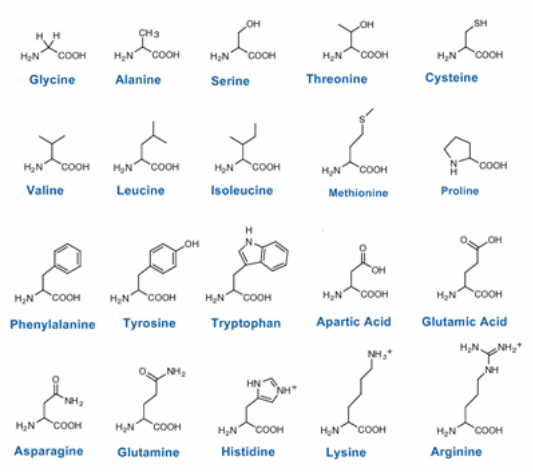
---

16 Definition (Primary Structure—Sidechains)

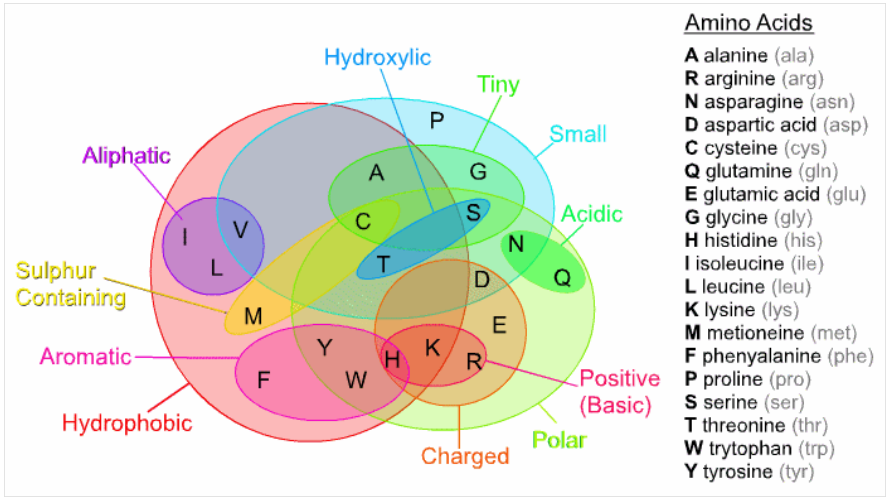


---

17 Example (Amino Acid Structures)



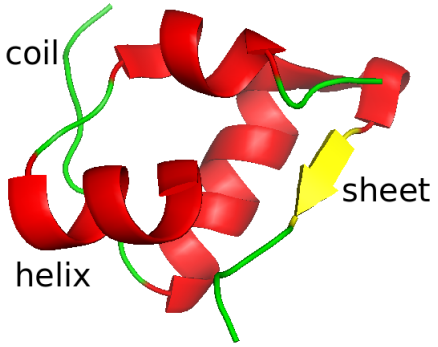
18 Example (Amino Acid Properties)



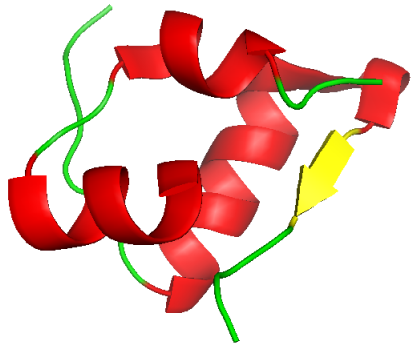
19 Example (Insulin Primary Sequence)

```
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY    50
TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC   100
SLYQLENYCN                                               110
```

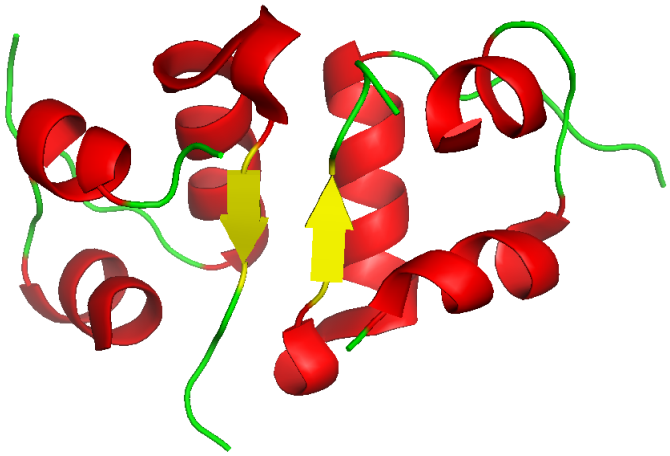
20 Example (Insulin Secondary Structure)



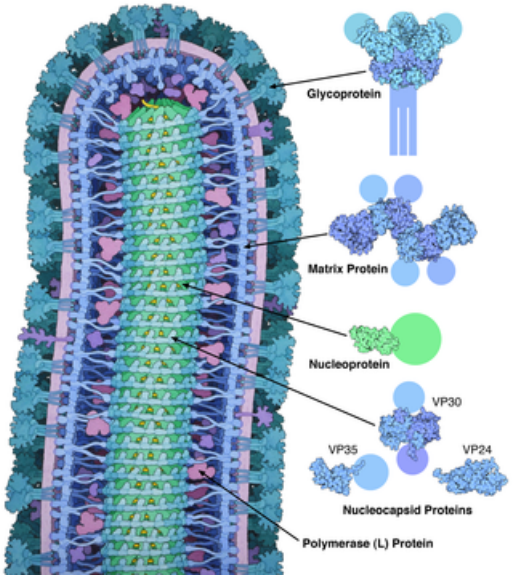
21 Example (Insulin Tertiary Structure)



22 Example (Insulin Quaternary Structure)



23 Example (Proteins in Ebola)



3 Biopython

Use the following instructions to install the Anaconda Python package:

- Download the appropriate package from `http://continuum.io/downloads`
- Open a terminal using the `ALT-CLTR-t` keyboard short-cut.
- Use the linux command `cd` to change to the directory containing the downloaded file.
- Type `bash <downloaded file name>` to install the package. (You can use the tap key to complete the name once you have typed the first few characters of the file name.)
- Type `conda install Biopython` to install the python bioinformatics package.

Startup the IPython notebook. If you are using linux, open a terminal and type

```
ipython notebook -pylab inline
```

If you are using Windows enter the command `%pylab inline` at the beginning of your notebook.

Use the following commands to plot a sine curve. To create 100 points equally spaced between 0 and 2π type

```
t = linspace(0,2*pi,100)
print t
```

Plot the sine curve

```
plot(sin(t))
```

The ipython notebook combines the best features of programs like Matlab and Maple.

24 Example (IPython)
Download the `insulin_cDNA.txt` file. Explain what each of the following commands do:

```
file = open('insulin_cDNA.txt')
file.  (press tab key)
seq = file.readline().strip()
print seq
len(seq)
seq[0]
seq[1]
```

```
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:

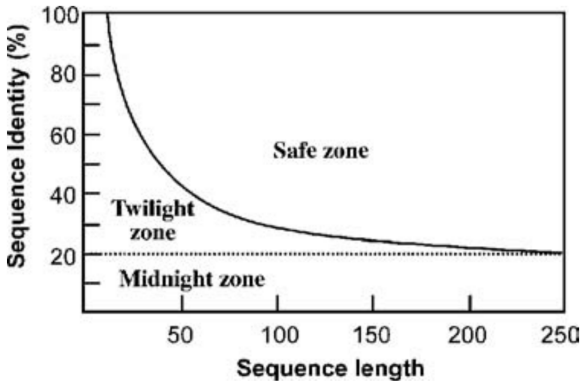
```
P01319 INS_CAPHI (Goat)
P01317 INS_BOVIN (Cow)
P01318 INS_SHEEP
```

(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 insulin 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:
