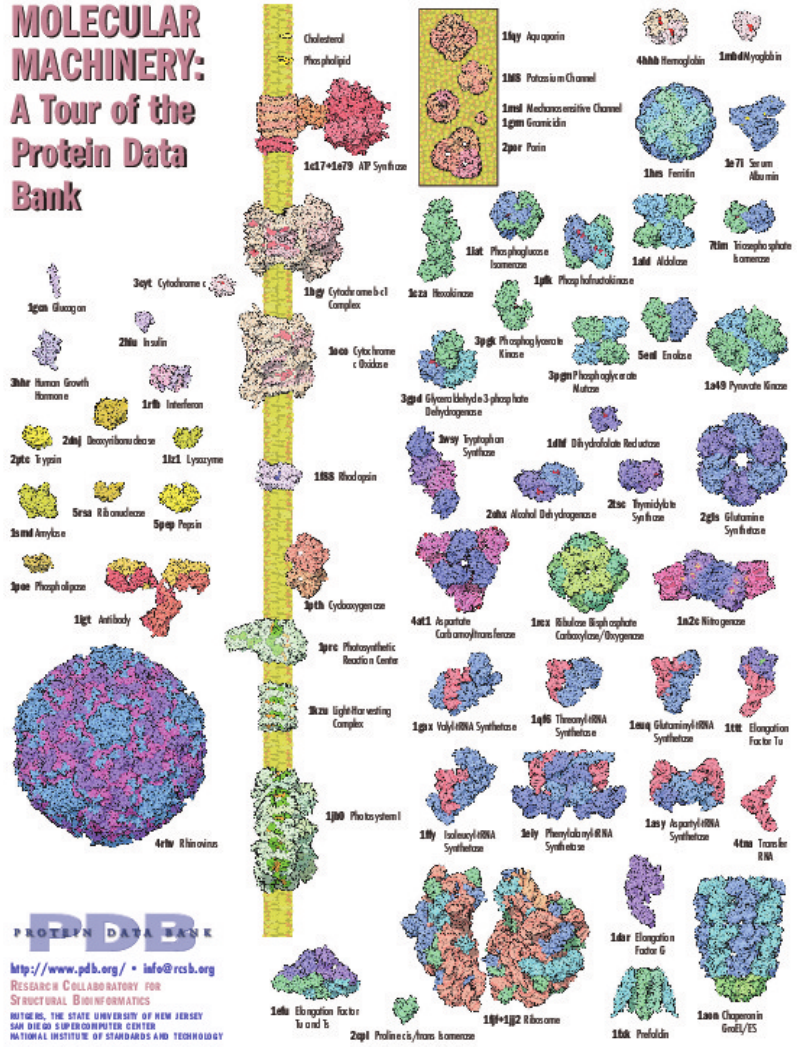
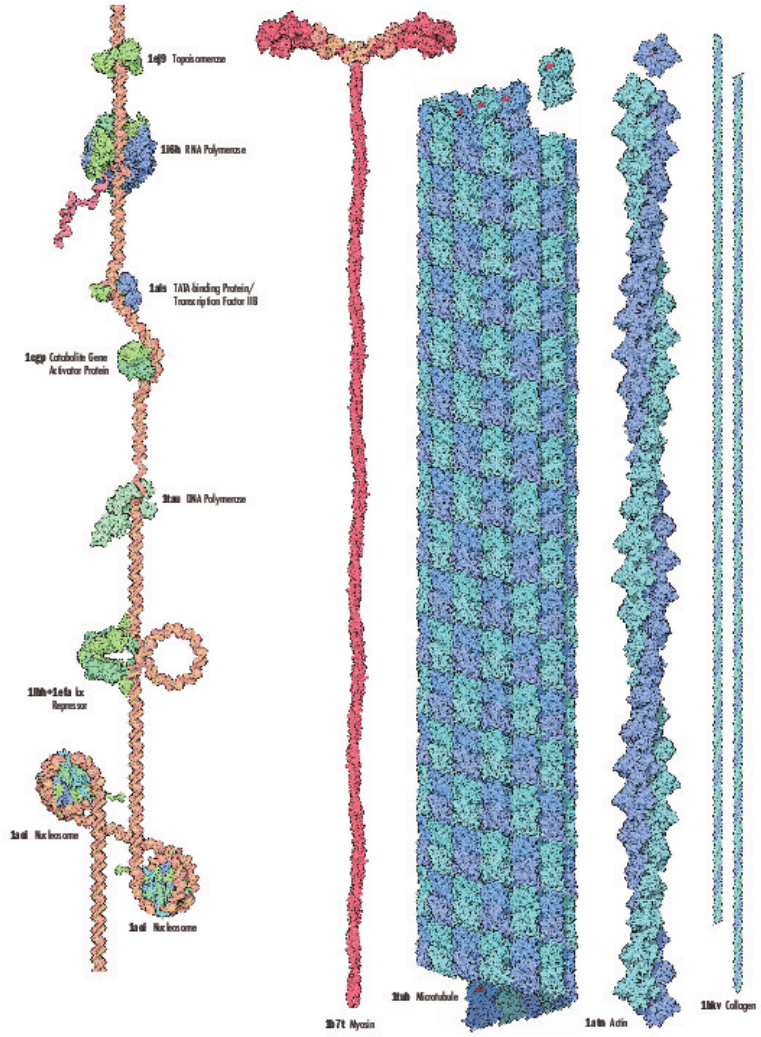
Treatment	Observations						Ave
nonsmokers	69	52	71	58	59	65	62.3
light smokers	55	60	78	58	62	66	63.2
moderate smokers	66	81	70	77	57	79	71.7
heavy smokers	91	72	81	67	95	84	81.7
							69.7

7 Proteins





47

---

2 Homework (Infectious Disease Protein)
Select one protein that plays a key role in an important infectious disease. Prepare a brief presentation and one page report on this protein that includes the following:

(a) Cite and read what you believe are the three most important publications related to your protein.

(b) Complete a sequence search of Swissprot/Uniprot and comment on what you believe are the most important hits.

(c) Complete a multiple sequence alignment and comment on important segments of your protein's sequence.

(d) Complete a phylogenetic analysis of your protein.

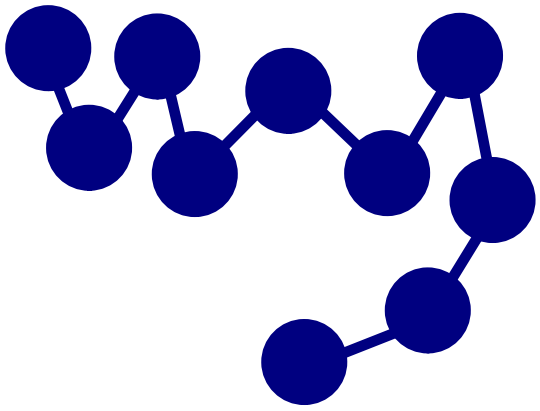
---

48 Definition (Protein Structure)

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

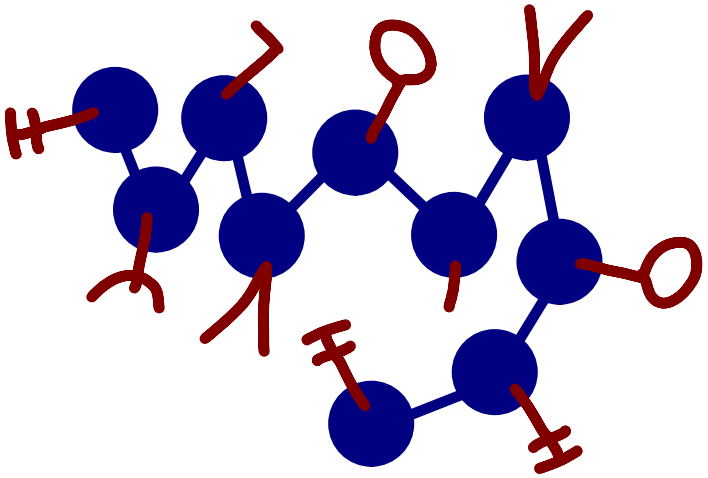
---

49 Definition (Primary Structure—Backbone)

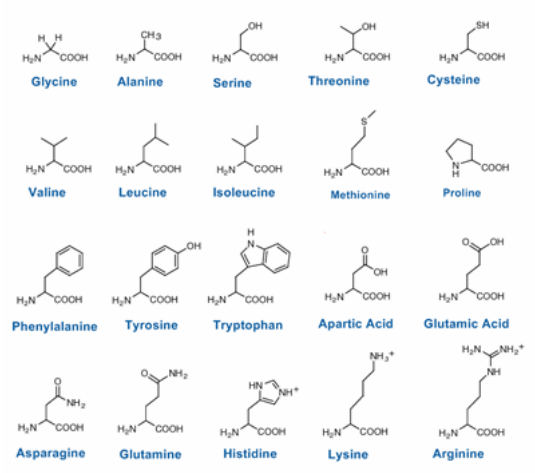


---

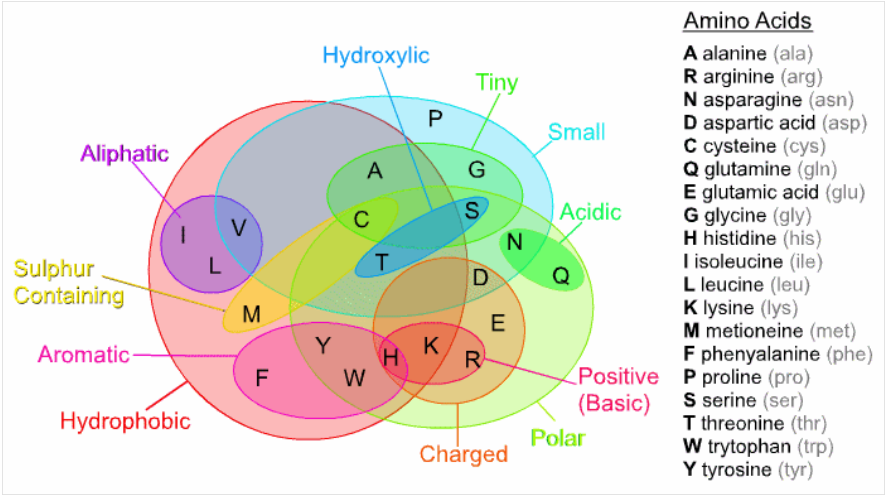
50 Definition (Primary Structure—Sidechains)



51 Example (Amino Acid Structures)



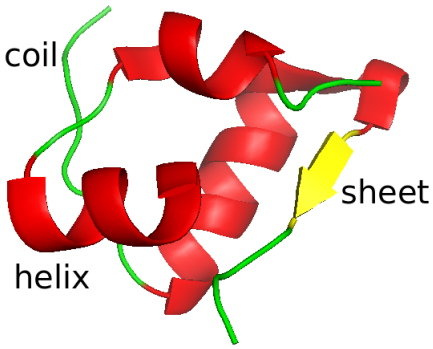
52 Example (Amino Acid Properties)



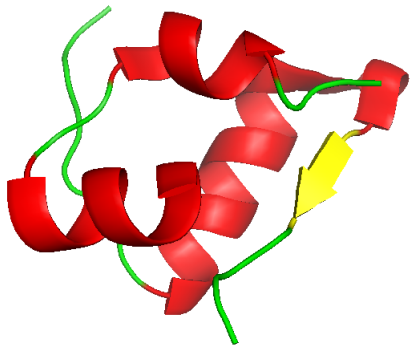
53 Example (Insulin Primary Sequence)

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

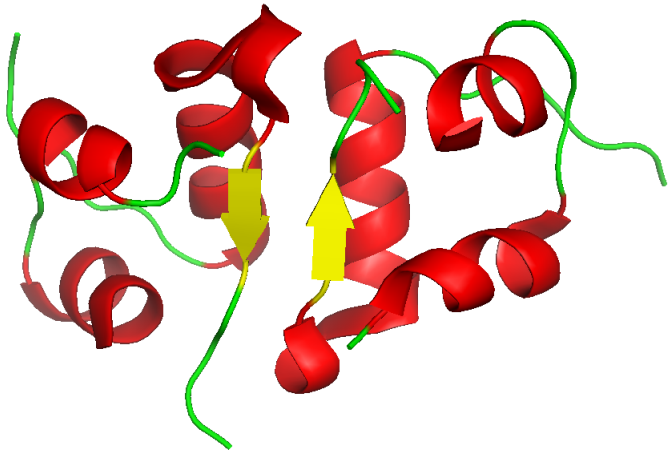
54 Example (Insulin Secondary Structure)



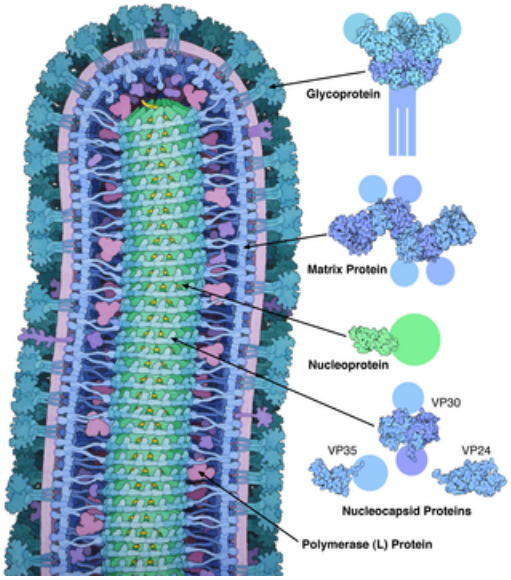
55 Example (Insulin Tertiary Structure)



56 Example (Insulin Quaternary Structure)



57 Example (Proteins in Ebola)



6 Lesson (Uniprot)
Go to the UniProt protein sequence database (http://www.uniprot.org/) and search for *human insulin*.

(a) Which entry corresponds to the entry for human insulin? Specify the **primary accession number**. (This is a citable number.) Hint: Next to Display click none and then select ENTRY INFORMATION. *Solution:*

(b) What is the title of the most recent article cited? *Solution:*

(c) Write down the first 10 (one-letter) amino acids codes for human insulin. *Solution:*

---

58 Example (Uniprot)

- UniProt stands for the Universal Protein Resource and is one of the first and most popular protein databases.
- The goal of UniProt is to provide a one-stop-shop which links information on proteins.
- UniProt has two parts:
  - SwissProt: Approximately half a million manually curated proteins.
  - TrEMBL: Approximately 88 million computer curated proteins.

---

7 Lesson (Sequence Search)
Go to www.uniprot.org and search for human insulin (P01308). Click the BLAST button to search for sequence similar to human insulin. Comment on your results.

---

59 Definition (Mutations)
DNA is subject to mutations. We will only consider insertions, deletions and substitutions.

```
orignal sequence   ATTGCTCC

orignal sequence   ATTG_CTCC
       insertion   ATTGGCTCC

orignal sequence   ATTGCTCC
        deletion   ATT_CTCC

orignal sequence   ATTGCTCC
    substitution   ATTTCTCC
```

---

60 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 are:

```
TAGTA   _TAGTA
_A_TAT  ATA_T_
```

---

61 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 homologous sequences, i.e. sequences that have evolved from a distant ancestor, may no longer be very similar to each other. Likewise, non-homologous sequences, can, throught the process of evolution, converge to similar sequences.

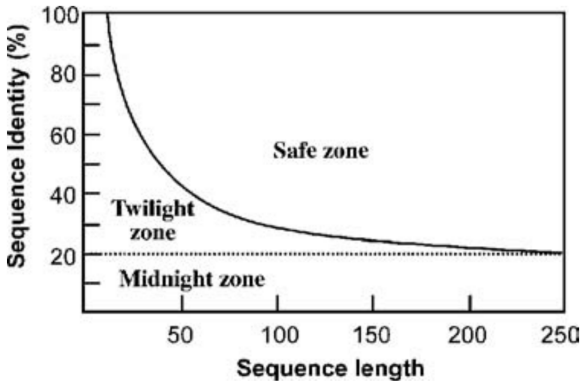
62 Definition (Evolution)

Divergent Evolution:
Identical sequences which diverge to different sequences due to random mutations.

Convergent Evolution:
Different sequences which converge to similar sequences due to similar structural or functional evolutionary forces.

---

63 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.

---

64 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

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

---

8 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:

---

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