Lecture 3: Biology Basics Continued

Spring 2020

January 28, 2020

Genotype/Phenotype

Phenotype:

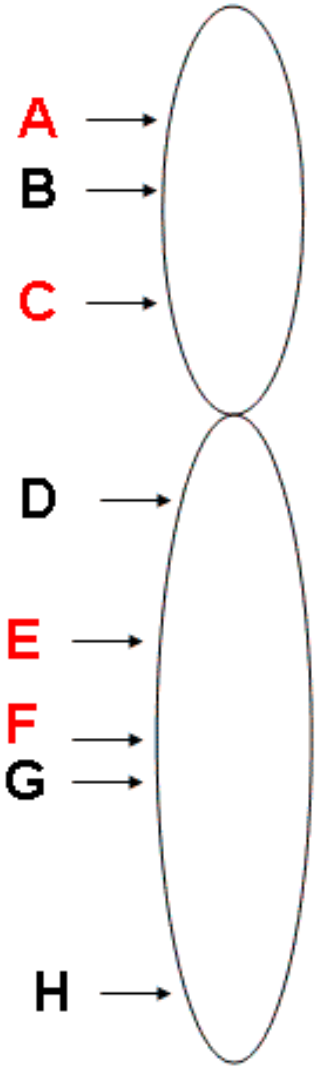
Blue eyes

Brown eyes

Genotype:

Recessive: bb

Dominant: Bb or BB



- Genes are shown in **relative order** and distance from each other based on pedigree studies.
- The chance of the chromosome breaking between A & C is higher than the chance of the chromosome breaking between A & B during meiosis
- Similarly, the chance of the chromosome breaking between E & F is higher than the chance of the chromosome breaking between F & G
- The closer two genes are, the more likely they are to be inherited together (co-occurrence)
- If pedigree studies show a high incidence of co-occurrence, those genes will be located close together on a genetic map

- **Pleiotropy**: when one gene affects many different traits.
- **Polygenic traits**: when one trait is governed by multiple genes, which maybe on the same chromosome or on different chromosomes.
  - The additive effects of numerous genes on a single phenotype create a continuum of possible outcomes.
  - Polygenic traits are also most susceptible to environmental influences.

Selection

- Some genes may be subject to **selection**, where individuals with advantages or “adaptive” traits tend to be more successful than their peers reproductively.
- When these traits have a genetic basis, selection can increase the prevalence of those traits, because the offspring will inherit those traits. This may correlate with the organism's ability to survive in its environment.
- Several different genotypes (and possibly phenotypes) may then coexist in a population. In this case, their genetic differences are called **polymorphisms.**

Genetic Mutation

- The simplest is the point mutation or substitution; here, a single nucleotide in the genome is changed **(single nucleotide polymorphisms (SNPs))**
- Other types of mutations include the following:
  - **Insertion.** A piece of DNA is inserted into the genome at a certain position
  - **Deletion.** A piece of DNA is cut from the genome at a certain position
  - **Inversion.** A piece of DNA is cut, flipped around and then re-inserted, thereby converting it into its complement
  - **Translocation.** A piece of DNA is moved to a different position.
  - **Duplication.** A copy of a piece of DNA is inserted into the genome

Mutations and Selection

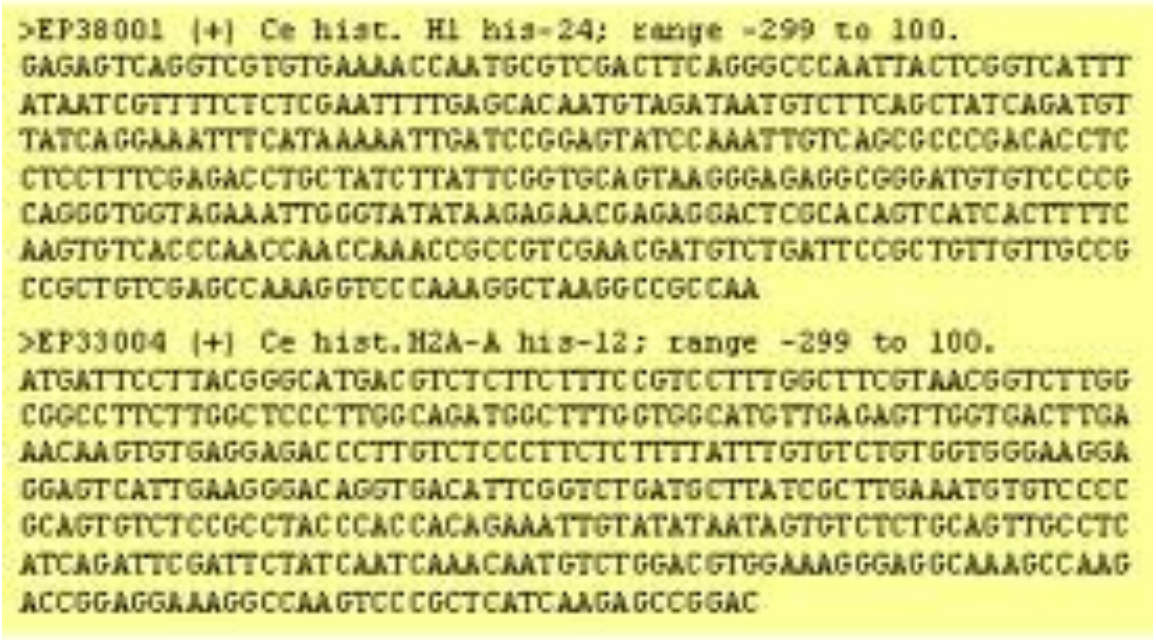
- While mutations can be detrimental to the affected individual, they can also, in rare cases, be beneficial; more frequently, neutral.
- Often mutations have no or negligible impact on survival and reproduction.
- Thereby mutations can increase the **genetic diversity** of a population, that is, the number of present polymorphisms.
- In combination with selection, this allow a species to adapt to changing environmental conditions and to survive in the long term.

Raw Sequence Data

- 4 bases: A, C, G, T + other (i.e. N = any, etc.)
  - kb (= kbp) = kilo base pairs = 1,000 bp
  - Mb = mega base pairs = 1,000,000 bp
  - Gb = giga base pairs = 1,000,000,000 bp.
- Size:
  - E. Coli 4.6Mbp (4,600,000)
  - Fish 130 Gbp (130,000,000,000)
  - Paris japonica (Plant) 150 Gbp
  - Human 3.2Gbp

Fasta File

- A sequence in FASTA format begins with a single-line description, followed by lines of sequence data (file extension is .fa).
- It is recommended that all lines of text be shorter than 80 chara



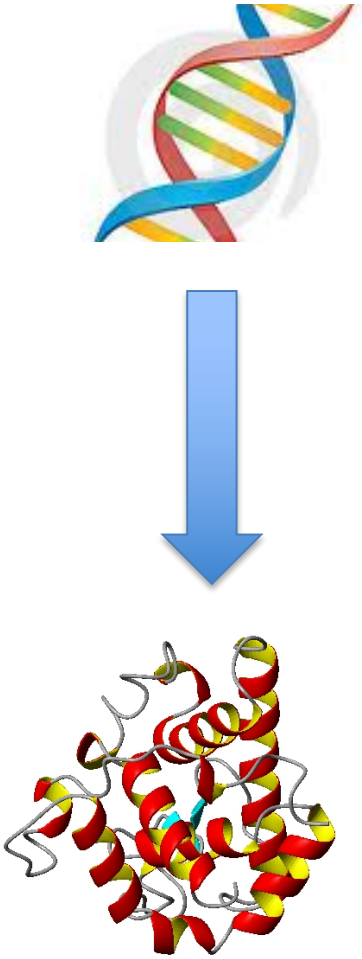
```
>EP38001 (+) Ce hist. H1 his-24; range -299 to 100.
GAGAGTCAGGTCGTGTGAAAACCAATGCGTCGACTTCAGGGCCCAATTACTCGGTCATTT
ATAATCGTTTTCTCTCGAATTTTGAGCACAATGTAGATAATGTCTTCAGCTATCAGATGT
TATCAGGAAATTTCATAAAAATTGATCCGGAGTATCCAAATTGTCAGCGCCCGACACCTC
CTCCTTTCGAGACCTGCTATCTTATTCGGTGCAGTAAGGGAGAGGCGGGATGTGTCCCCG
CAGGGTGGTAGAAATTGGGTATATAAGAGAACGAGAGGACTCGCACAGTCATCACTTTTC
AAGTGTCACCCAACCAACCAAACCGCCGTCGAACGATGTCTGATTCCGCTGTTGTTGCCG
CCGCTGTCGAGCCAAAGGTCCCAAAGGCTAAGGCCGCCAA

>EP33004 (+) Ce hist.H2A-A his-12; range -299 to 100.
ATGATTCCTTACGGGCATGACGTCTCTTCTTTCCGTCCTTTGGCTTCGTAACGGTCTTGG
CGGCCTTCTTGGCTCCCTTGGCAGATGGCTTTGGTGGCATGTTGACAGTTGGTGACTTGA
AACAAGTGTGAGGAGACCCTTGTCTCCCTTCTCTTTTATTTGTGTCTGTGGTGGGAAGGA
GGAGTCATTGAAGGGACAGGTGACATTCGGTCTGATGCTTATCGCTTGAAATGTGTCCCC
GCAGTGTCTCCGCCTACCCACCACAGAAATTGTATATAATAGTGTCTCTGCAGTTGCCTC
ATCAGATTCGATTCTATCAATCAAACAATGTCTGGACGTGGAAAGGGAGGCAAAGCCAAG
ACCGGAGGAAAGGCCAAGTCCCGCTCATCAAGAGCCGGAC
```

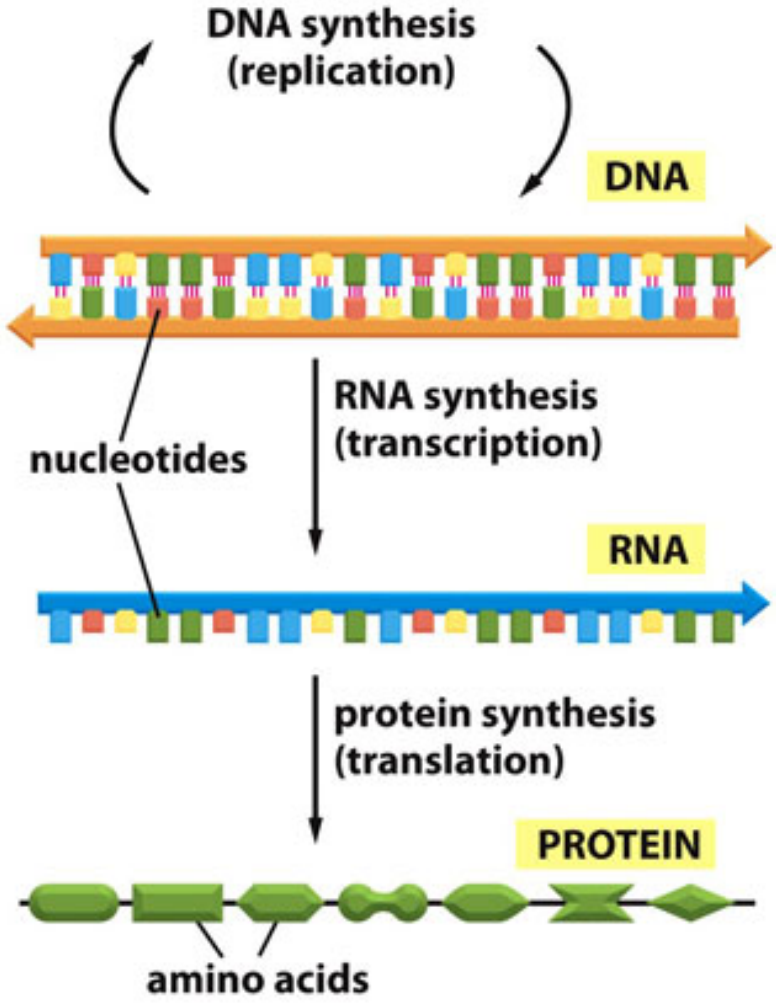
Fastq File

- Typically contain 4 lines:
  - Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description.
  - Line 2 is the sequence.
  - Line 3 is the delimiter '+', with an optional description.
  - Line 4 is the quality score.
  - file extension is .fq

```
@SEQ_ID
GATTTGGGGTTCAAAGCTTCAAAGCTTCAAAGC
+
!''*(((****+))%%%++++++++!!!!+++++
```

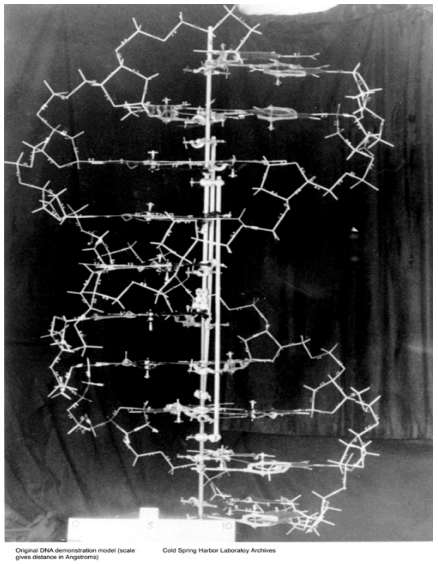
Central Dogma





Discovery of DNA

- DNA Sequences
  - Chargaff and Vischer, 1949
    - DNA consisting of A, T, G, C
      - Adenine, Guanine, Cytosine, Thymine
  - Chargaff Rule
    - Noticing #A≈#T and #G≈#C
      - A "strange but possibly meaningless" phenomenon.
- Wow!! A Double Helix
  - Watson and Crick, *Nature,* April 25, 1953
  - **1 Biologist**
    1 Physics Ph.D. Student
    + 900 words
    = Nobel Prize
  - Rich, 1973
    - Structural biologist at MIT.
    - DNA's structure in atomic resolution.



Original DNA demonstration model (scale gives distance in Angstroms) Cold Spring Harbor Laboratory Archives

Watson and Crick walk along the Backs Cold Spring Harbor Laboratory Archives

Crick Watson

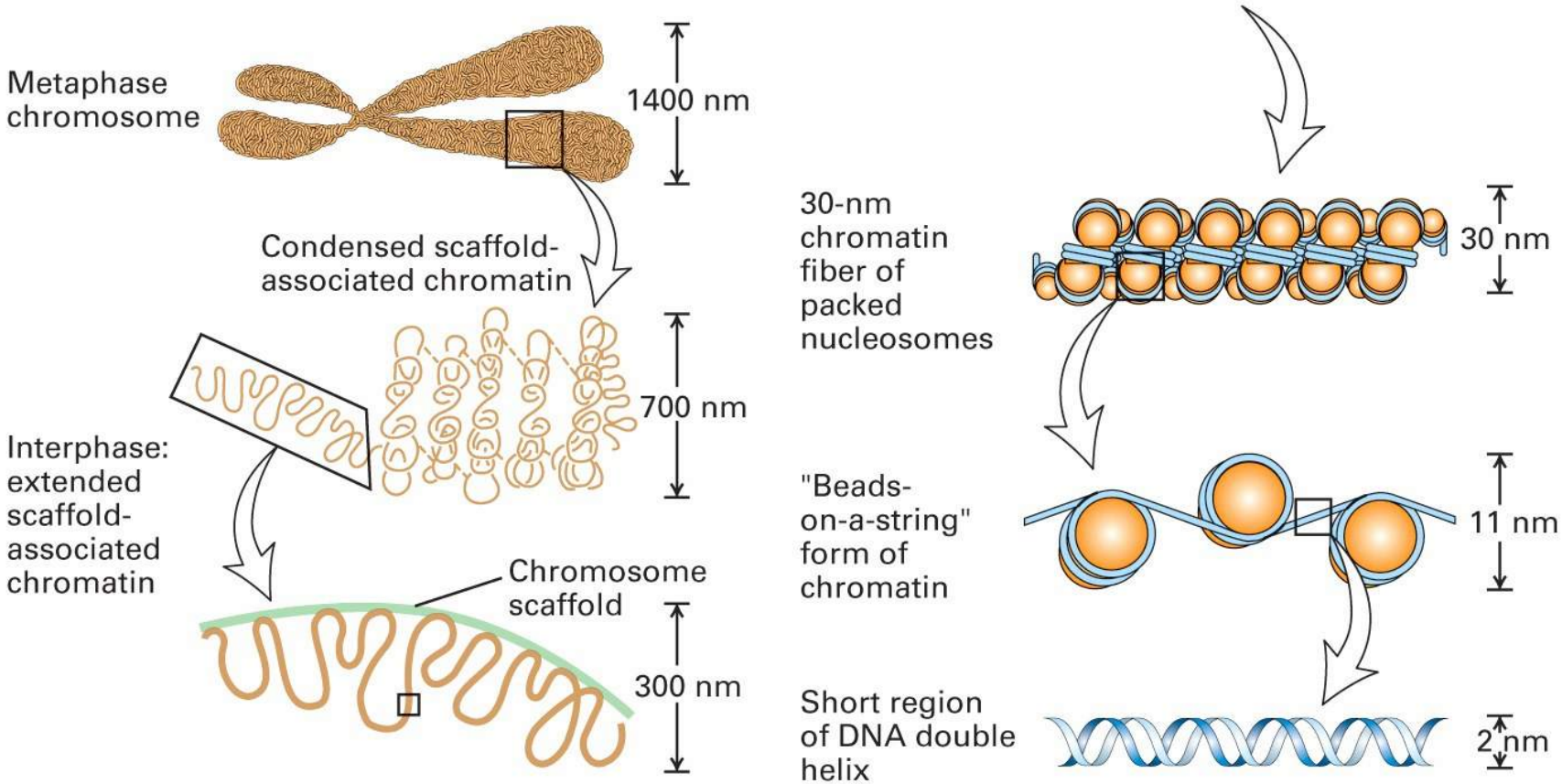
Watson & Crick – *"…the secret of life"*

- Watson: a zoologist, Crick: a physicist
- "*In 1947 Crick knew no biology and practically no organic chemistry or crystallography..*" – www.nobel.se
- Applying Chagraff's rules and the X-ray image from Rosalind Franklin, they constructed a "tinkertoy" model showing the double helix.
- Their 1953 *Nature* paper: *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."*

Watson & Crick with DNA model

Rosalind Franklin with X-ray image of DNA

Superstructure



Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

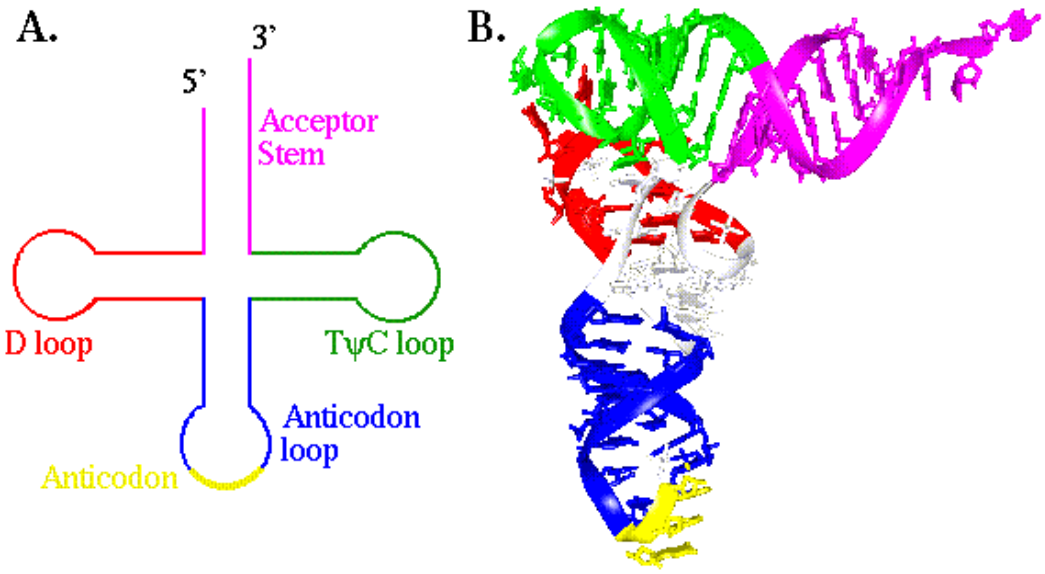
Superstructure implications

- DNA in a living cell is in a highly compacted and structured state.
- Transcription factors and RNA polymerase need ACCESS to do their work.
- Transcription is dependent on the structural state – SEQUENCE alone does not tell the whole story.

RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)

- RNA can form secondary structures by "pairing up"



tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

RNA, continued

- Several types exist, classified by function
- mRNA – carries a gene's *m*essage out of the nucleus.
- tRNA – *t*ransfers genetic information from mRNA to an amino acid sequence
- rRNA – *r*ibosomal RNA. Part of the ribosome machine.

Protein

- A polymer composed of amino acids.
- There are 20 naturally occurring amino acids.

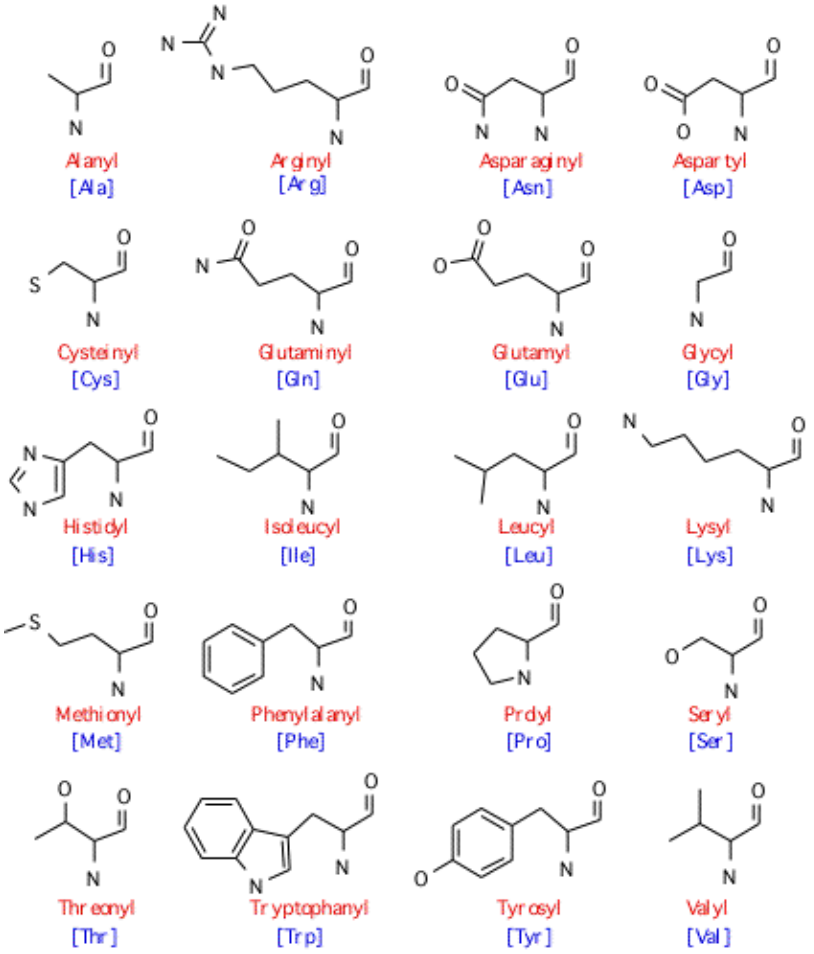
- Usually functions through molecular motion or binding with other molecules.

Proteins: Primary Structure

- Peptide sequence:
  - Sequence of amino acids = sequences from a 20 letter alphabet (i.e. `ACDEFGHIKLMNPQRSTVWY`)
  - Average protein has ~300 amino acids
  - Typically stored as fasta files

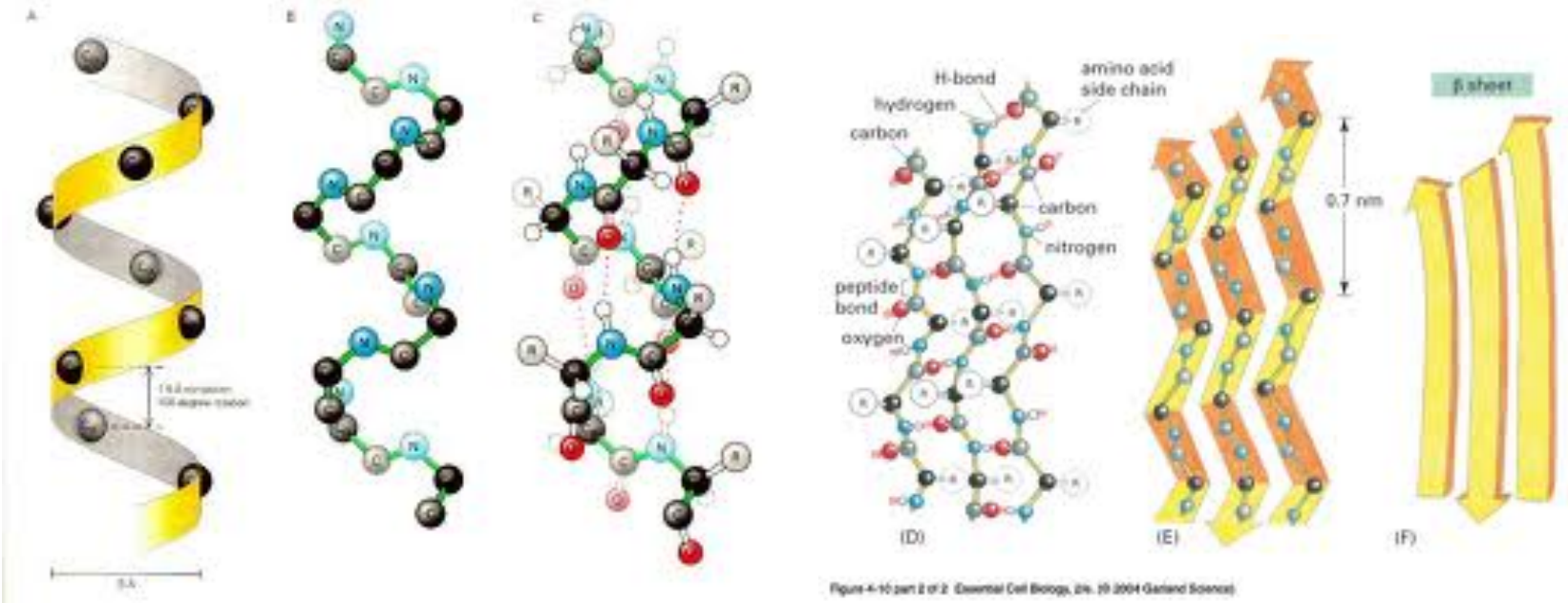
```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Naturally Occurring Amino Acids



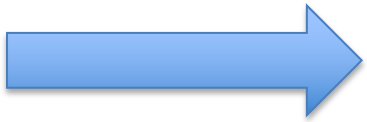
Proteins: Secondary Structure

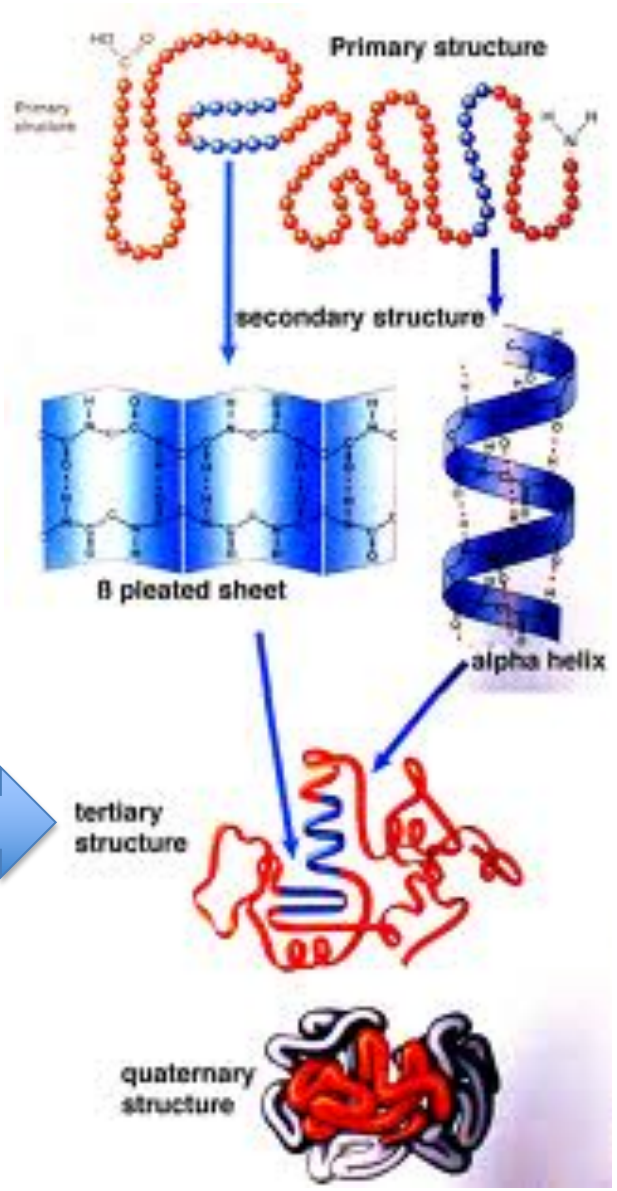
- Polypeptide chains fold into regular local structures
  - Common types: alpha helix, beta sheet, turn, loop
  - Defined by the creation of hydrogen bonds



Proteins: Tertiary Structure

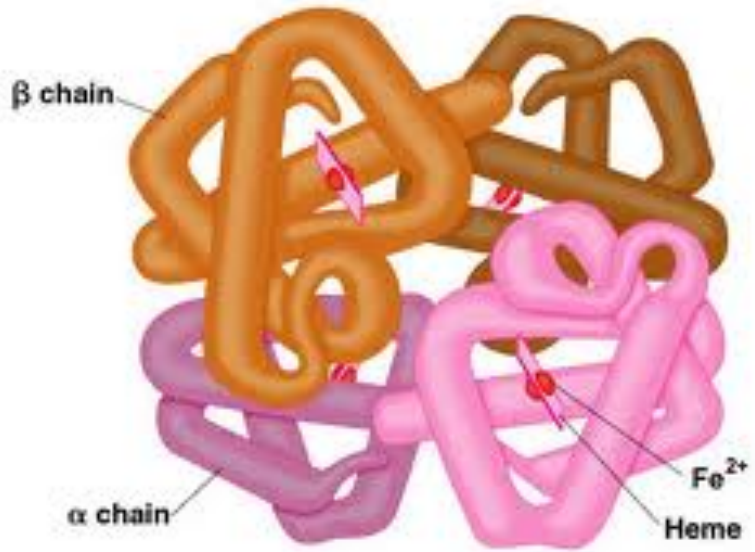
- 3D structure of a polypeptide sequence
  - interactions between non-local and foreign atoms

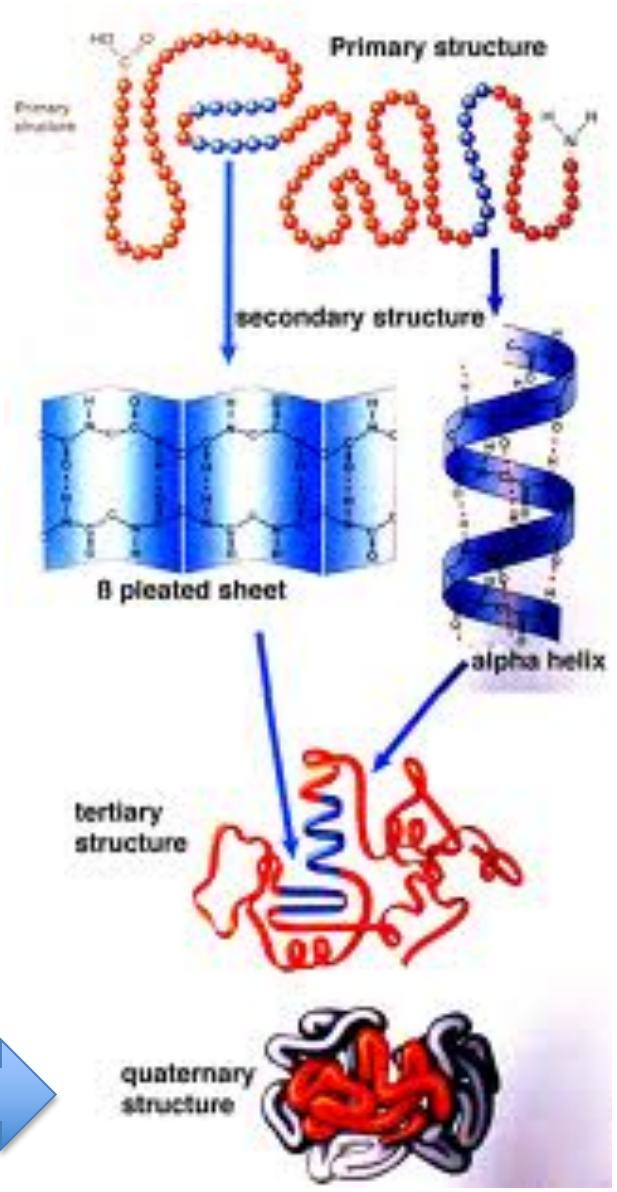




Proteins: Quaternary Structure

- Arrangement of protein subunits





Conclusions

Challenges in Bioinformatics

- Need to feel comfortable in interdisciplinary area
- Depend on others for primary data
- Need to address important biological and computer science problems

Basic Steps in Bioinformatics Research

1. Data management problem: storage, transfer, transformation (Information Technology)
2. Data analysis problem: mapping, assembly
   – algorithm scaling (Computer Science)
3. Statistical challenges: traditional statistics is not well suited for modeling systematic errors over large number of observations (Biostatistics)
4. Biological hypothesis testing
   – data interpretation (Life Science)

Basic Skills

- Artificial intelligence and machine learning
- Statistics and probability
- Algorithms
- Databases
- Programming
- Biology/Chemistry knowledge

Genomics:
- Assembly
- Detection of variation
- GWAS

RNA:
- Gene expression
- Transcriptome assembly
- Pathway analysis
- RNA-RNA interaction

Protein:
- Mass spectrometry
- Structure prediction
- Protein-Protein interaction

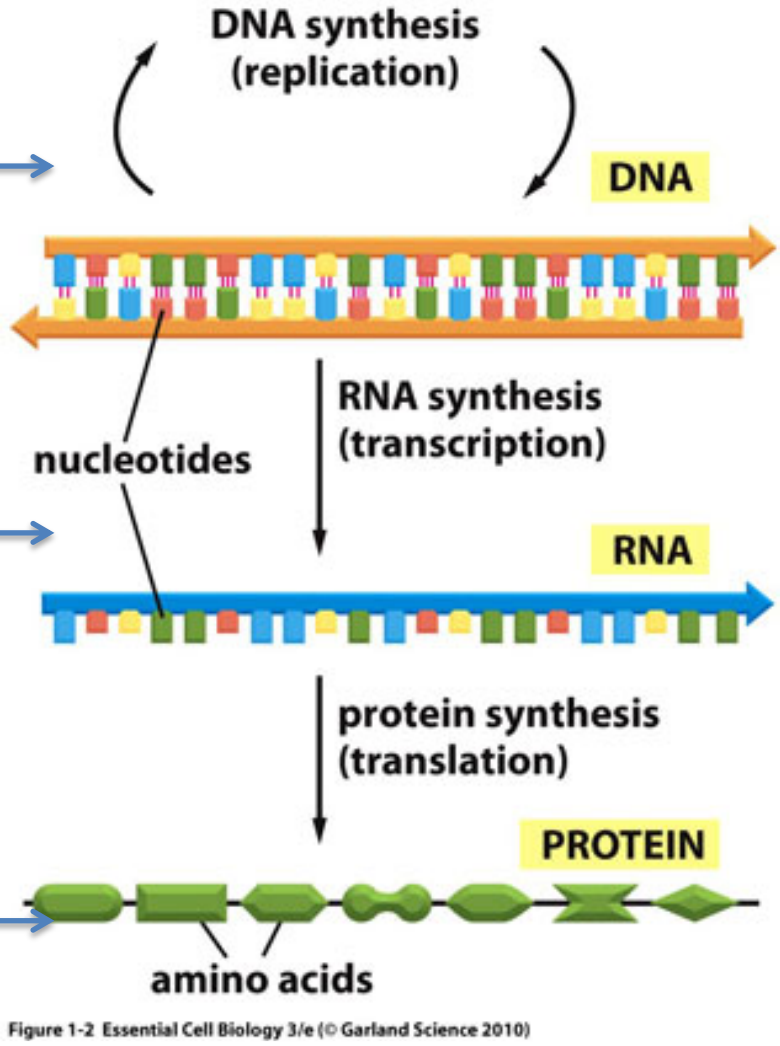


Figure 1-2 Essential Cell Biology 3/e (© Garland Science 2010)