Lecture 1

BNFO 136

Usman Roshan

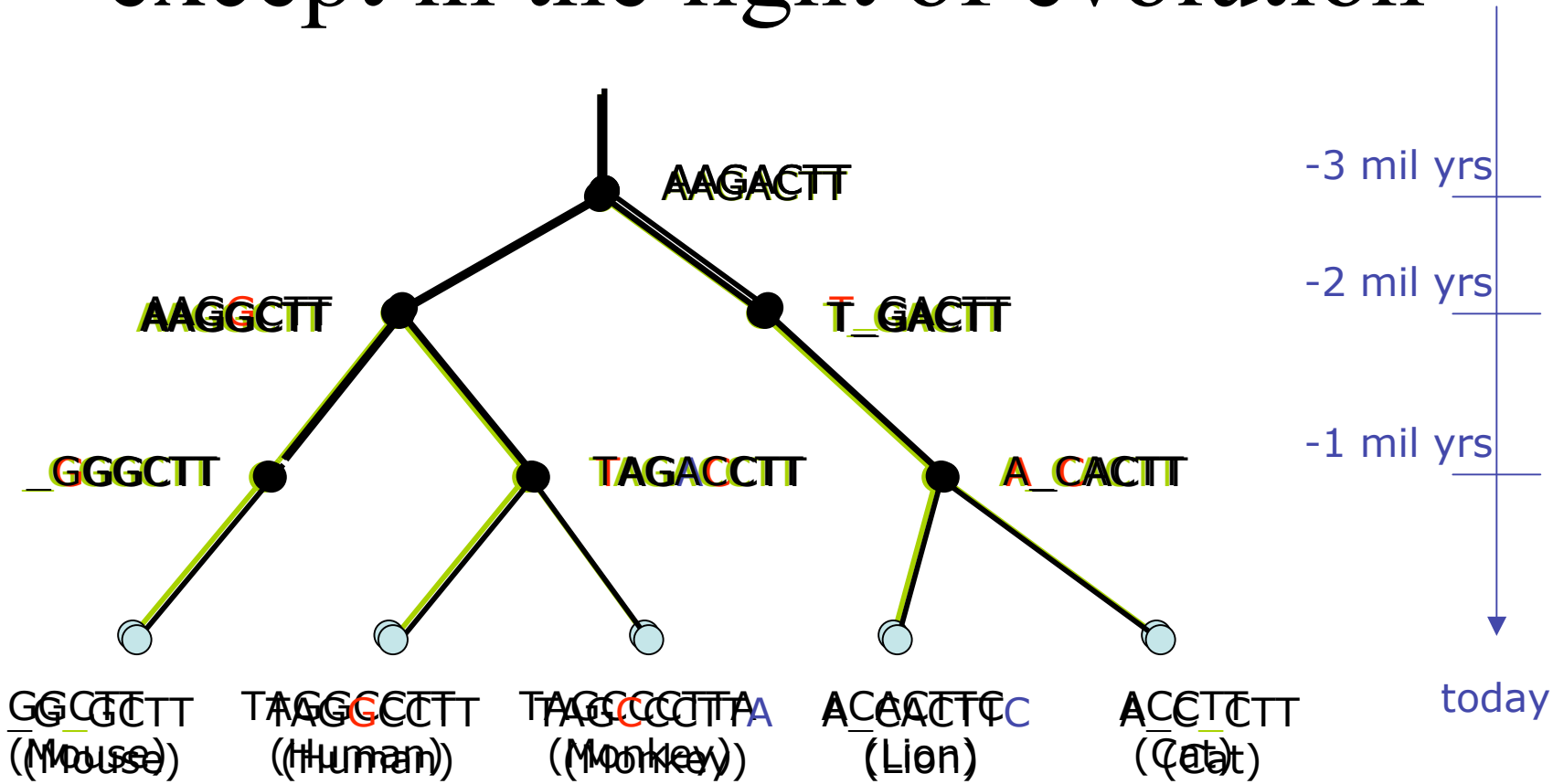
Course overview

- Pre-req: BNFO 135 or approval of instructor
- Python progamming language
  - Some unix basics
  - Input/output, lists, dictionaries, loops, if-else, counting
- Sequence analysis
  - Comparison of protein and DNA sequences
  - Scoring a sequence alignment
  - Picking the best alignment from a set
  - Computing a sequence alignment
  - Finding the most similar sequence to a query
- Phylogeny reconstruction
  - UPGMA and neighbor joining algorithms

Overview (contd)

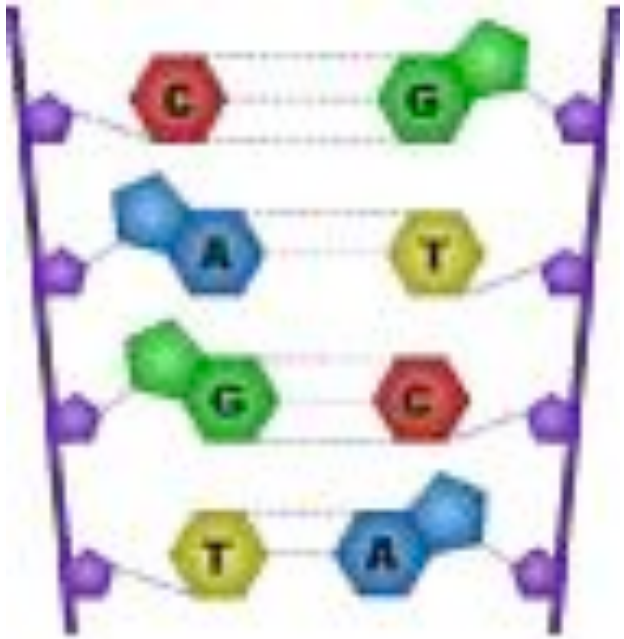
- Grade: 15% programming assignments, two 25% mid-terms and 35% final exam
- Recommended Texts:
  - Introduction to Bioinformatics by Arthur M. Lesk
  - Python Scripting for Computational Science by Hans Petter Langtangen

Nothing in biology makes sense, except in the light of evolution

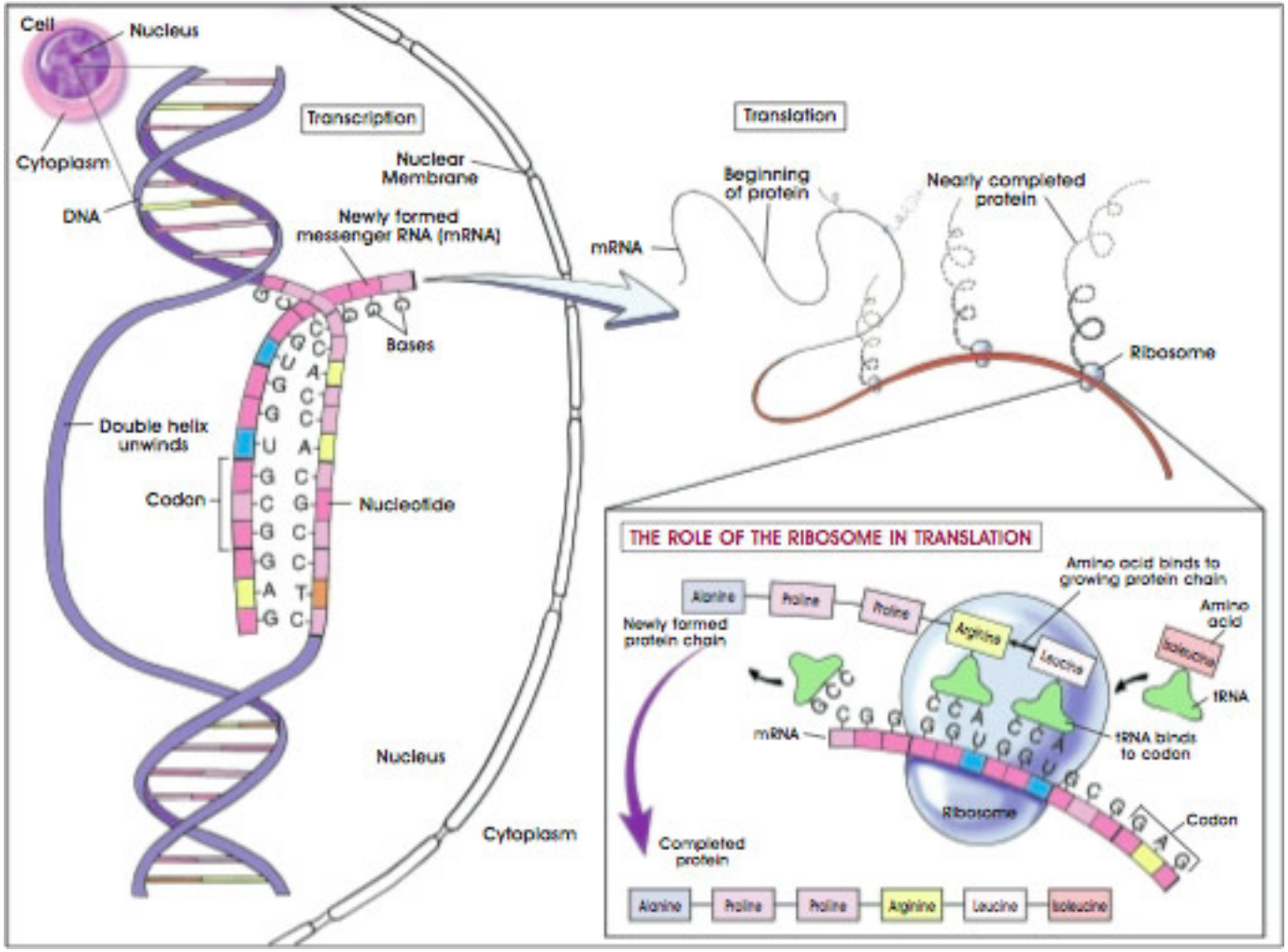


Representing DNA in a format manipulatable by computers

- DNA is a double-helix molecule made up of four nucleotides:
  - Adenosine (A)
  - Cytosine (C)
  - Thymine (T)
  - Guanine (G)
- Since A (adenosine) always pairs with T (thymine) and C (cytosine) always pairs with G (guanine) knowing only one side of the ladder is enough
- We represent DNA as a sequence of letters where each letter could be A,C,G, or T.
- For example, for the helix shown here we would represent this as CAGT.



Transcription and translation



Amino acids

Proteins are chains of amino acids. There are twenty different amino acids that chain in different ways to form different proteins.

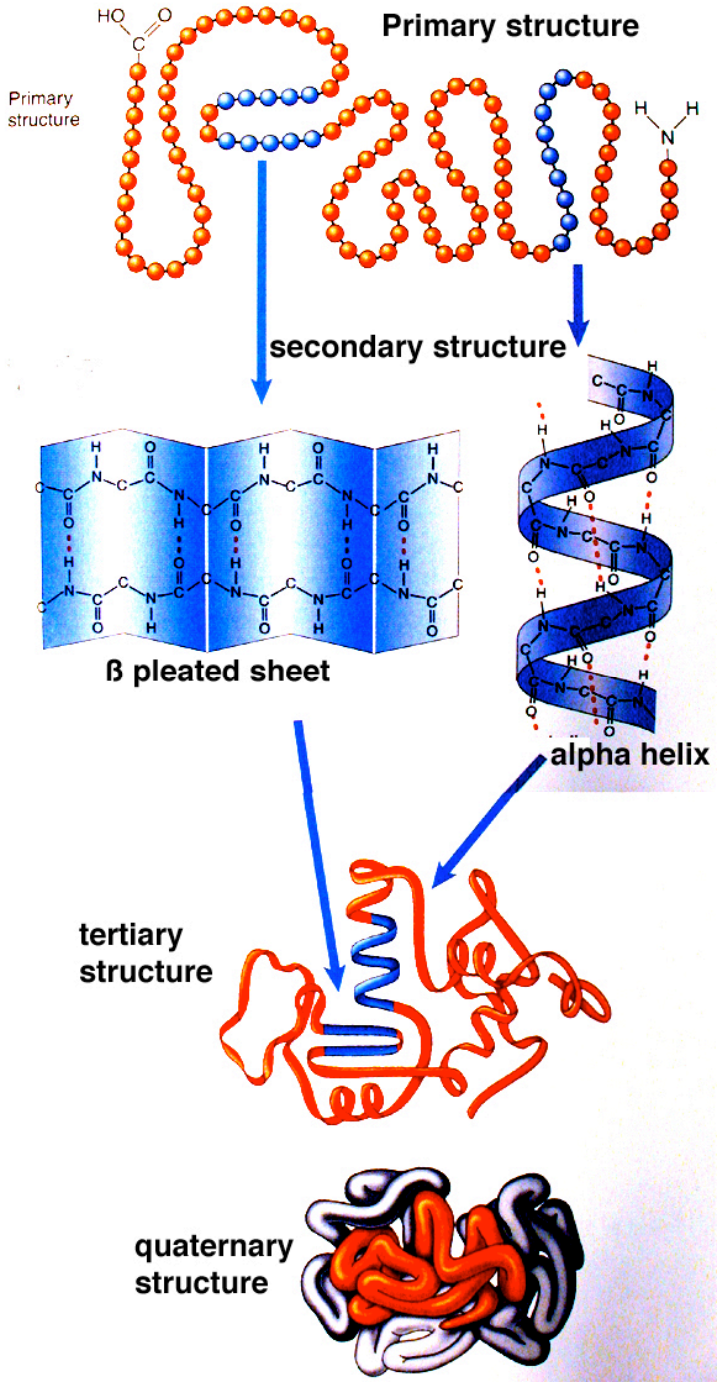
For example,
FLLVALCCRFGH
(this is how we could store it in a file)

This sequence of amino acids folds to form a 3-D structure

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

Protein structure

- Primary structure: sequence of amino acids.
- Secondary structure: parts of the chain organizes itself into alpha helices, beta sheets, and coils. Helices and sheets are usually evolutionarily conserved and can aid sequence alignment.
- Tertiary structure: 3-D structure of entire chain
- Quaternary structure: Complex of several chains



Getting started

- FASTA format:

```
>human
ACAGTAT
>mouse
ACGTA
>cat
AGGTGAAA
```

Python basics

- Basic types:
  - Scalar: number or a string,
  - Lists: collection of objects
  - Dictionaries: collection of (key,value) pairs
- Reading sequences from a file into a data structure
- Basic if-else conditionals and for loops

Some simple programs

- How many sequences in a FASTA file?
- What is the length and name of the longest and shortest sequence?
- What is the average sequence length?
- Verify that input contains DNA sequences.
- Compute the reverse complement of a DNA sequence

Two dimensional lists

- Represented as list of lists
- For example the matrix

23 4 1

12 5 12

1 6 20

would be stored as

[[23, 4, 1], [12, 5, 12], [1, 6, 20]]