Bio 208FS: Representing the Central Dogma of Molecular Biology in Python

Paul Magwene

Central Dogma of Molecular Biology

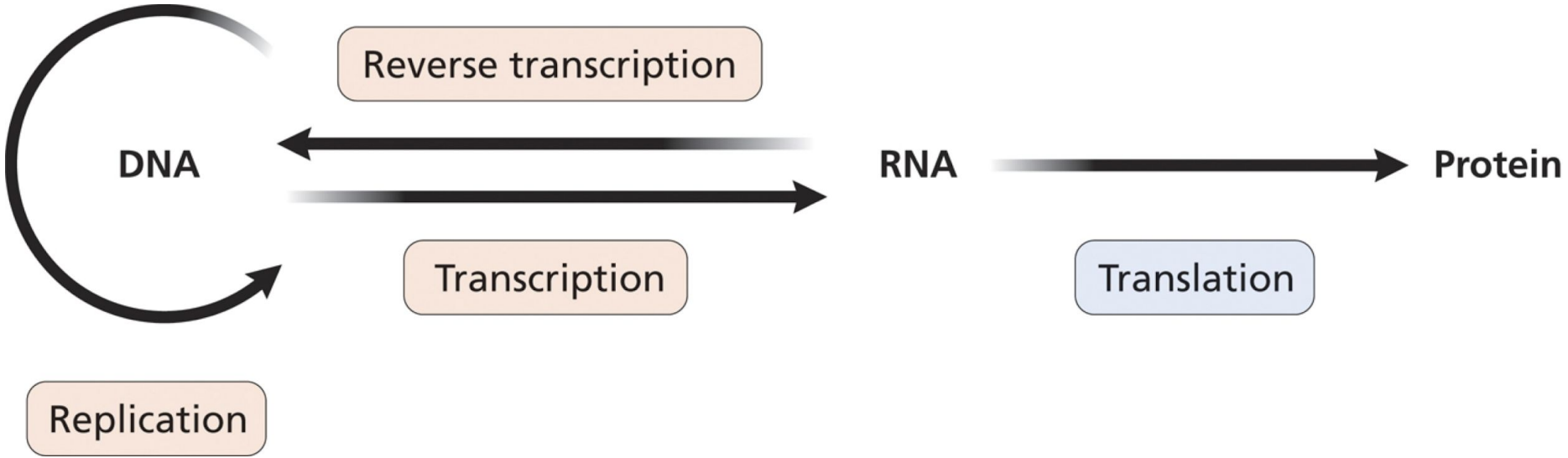

Image from Momand and McCurdy, 2016

DNA, RNA, and Proteins are biopolymers

- Polymer -- a material made of many repeating subunits (monomers)
- DNA and RNA (polynucleotides)
  - monomeric units are nucleotides
- Proteins (polypeptides)
  - monomeric units are amino acids
- While the molecules adopt 3D shapes that are important for their function, their basic structure can be represented as linear sequences of their respective monomers

Using Python strings to represent biopolymers

- Python strings are ordered sequences of characters (unicode in Python 3)
- Strings can be indexed and sliced like lists
- Unlike lists, strings are immutable
  - Efficiency reasons
  - Changing a string implies you're creating a new string
- Strings can be concatenated
- A variety of string methods (functions "attached to" an object) for manipulating strings are part of the Python standard implementation
  - see the Python Documentation on Strings

DNA monomers

- The size of the "alphabet" of DNA monomers (nucleotides) is 4
  - Adenine (A)
  - Guanine (G)
  - Thymine (T)
  - Cytosine (C)

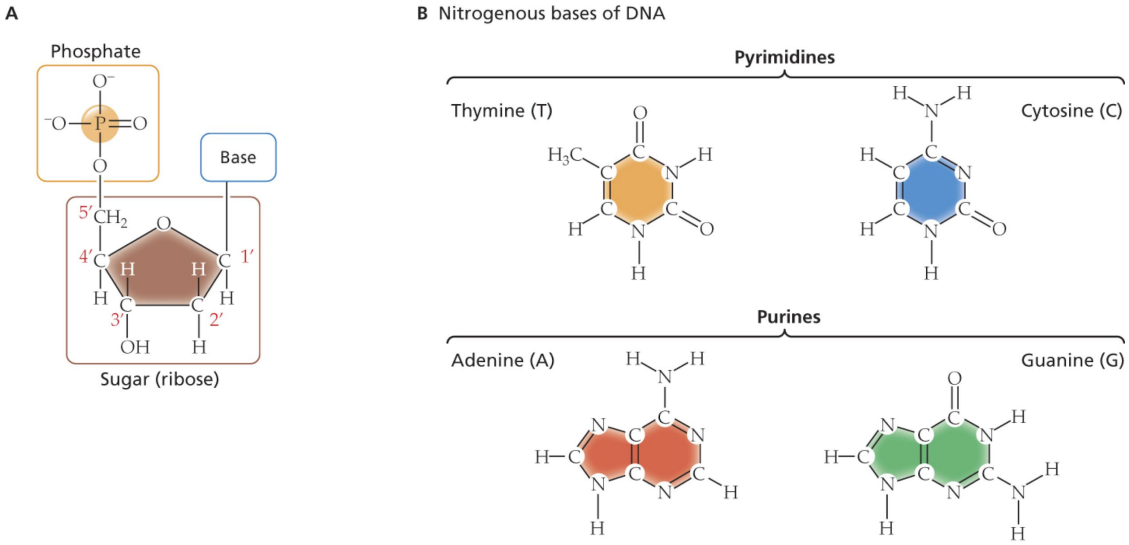

Image from Momand and McCurdy, 2016

DNA polymeric structure

- The backbone of the polymeric DNA structure is formed by chains of nucleotides joined by covalent bonds
  - Strings of characters from the DNA alphabet
- DNA molecules have a polarity: 5' ("5 prime") to 3' ("3 prime")
  - By convention we'll represent left (5') to right (3')
- DNA in the genome is usually in a "double stranded" state
  - The two strands are "antiparallel" -- complementary nucleotides from each strand for hydrogen bonds with each other (base pairing) -- A/T and G/C

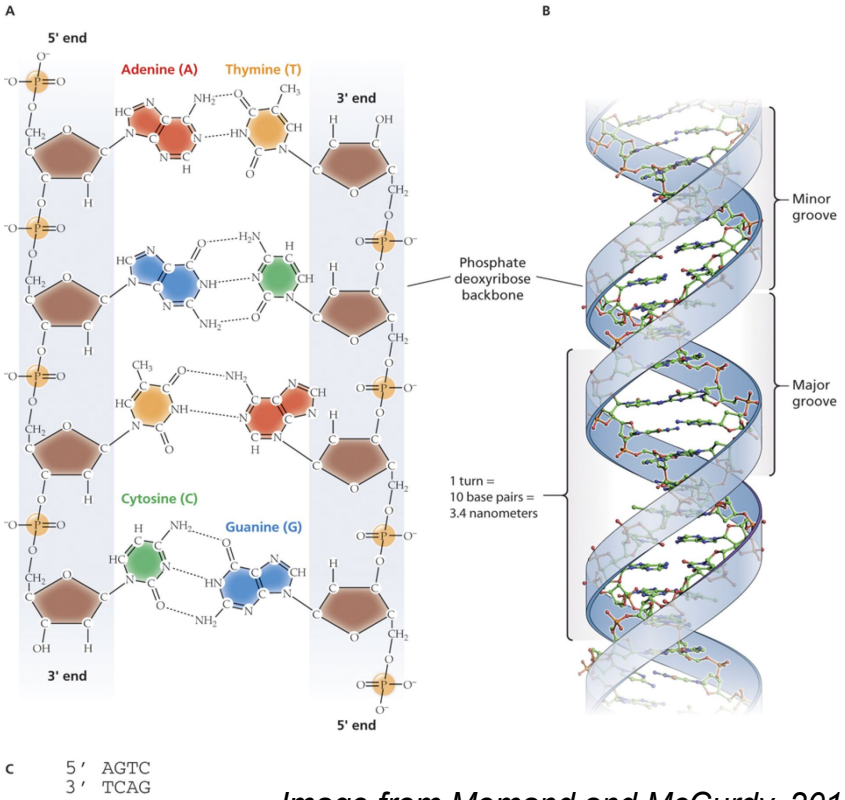

Image from Momand and McCurdy, 2016

Python string representations of DNA and functions on DNA

- See hands on
- Implement functions for:
  - Reverse
  - Complement
  - Reverse complement
- Programming concepts
  - control flow statements
    - if, if-else
    - for loops
    - while
    - list comprehensions
  - Python dictionaries

RNA

- Differences from DNA
  - Sugar is ribose instead of deoxyribose
  - Uracil is the nitrogenous base instead of thymine
- Often single stranded
  - But can form complementary interactions with DNA (or other RNAs)

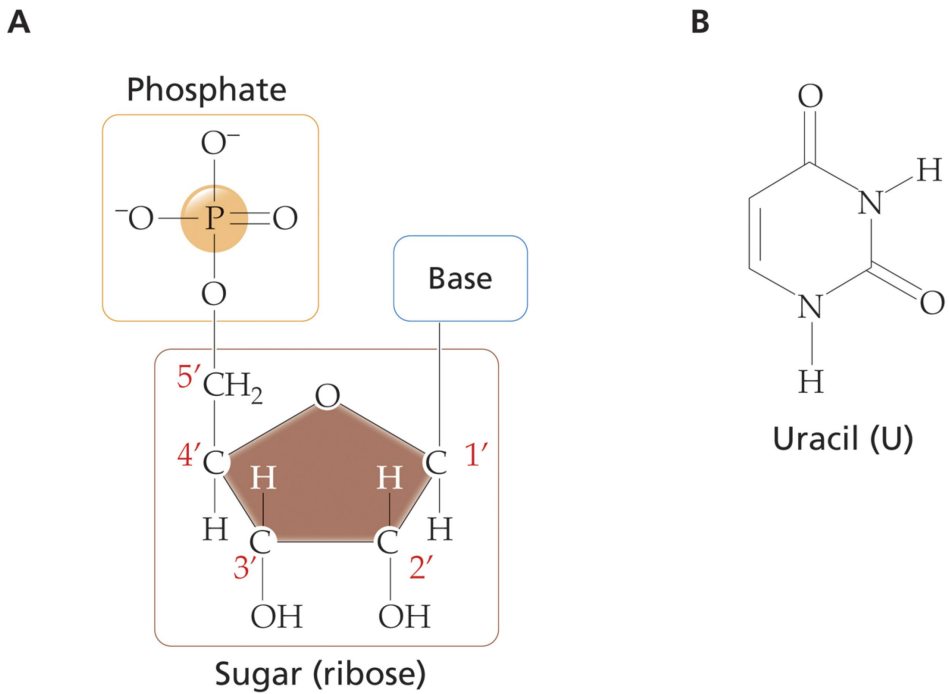

Image from Momand and McCurdy, 2016

Python string representations of RNA

- See hands on
  - RNA as strings
  - Implement function for simulating transcription

Protein monomeric units are amino acids

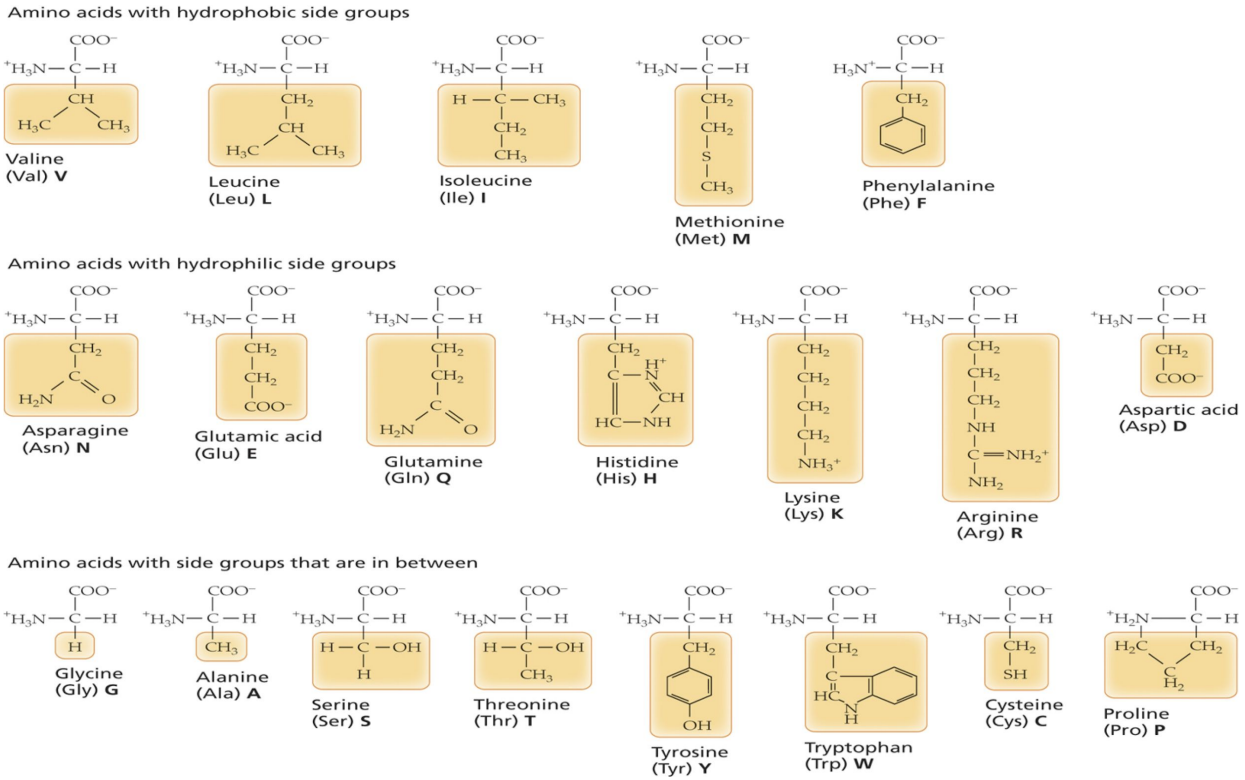

Image from Momand and McCurdy, 2016

Protein polymeric structure

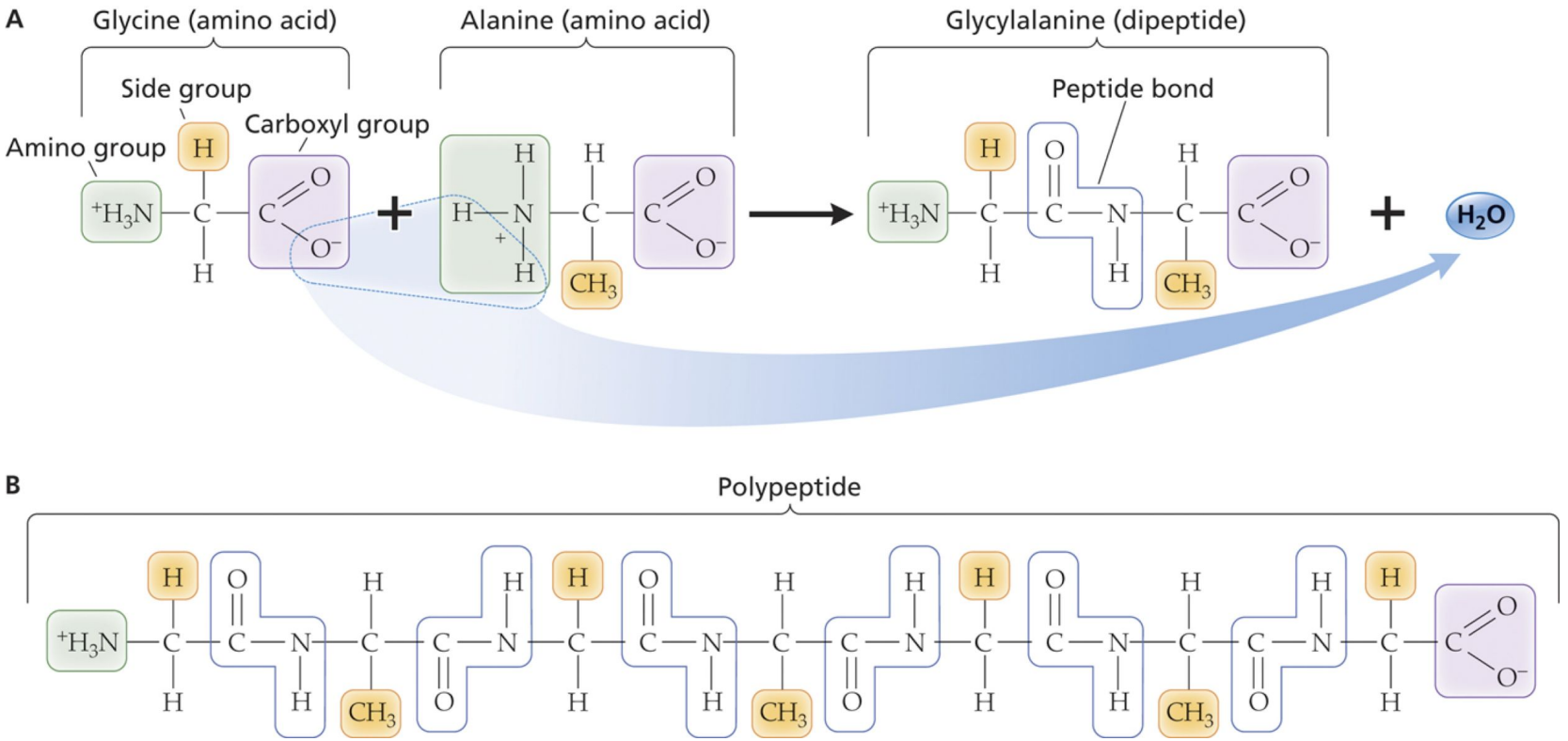

Image from Momand and McCurdy, 2016

Python string representation of polypeptides

- Alphabet size = 20
- By convention written from left to right (N to C)

Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid or aspartate	Asp	D
Glutamic acid or glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Standard Genetic Code

First position	Second position: T				Second position: C				Second position: A				Second position: G				Third position
	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	
T	TTT	phe	TCT	ser	TAT	tyr	TGT	cys	T
	TTC		TCC		TAC		TGC		C
	TTA	leu	TCA		TAA	STOP	TGA	STOP	A
	TTG		TCG		TAG	STOP	TGG	trp	G
C	CTT	leu	CCT	pro	CAT	his	CGT	arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	ile	ACT	thr	AAT	asn	AGT	ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	lys	AGA	arg	A
	ATG	met	ACG		AAG		AGG		G
G	GTT	val	GCT	ala	GAT	asp	GGT	gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	glu	GGA		A
	GTG		GCG		GAG		GGG		G

Image from Momand and McCurdy, 2016

Implement a translation function

- Input: string representing a DNA or RNA sequence
- Output: string representing a Protein sequence