1785

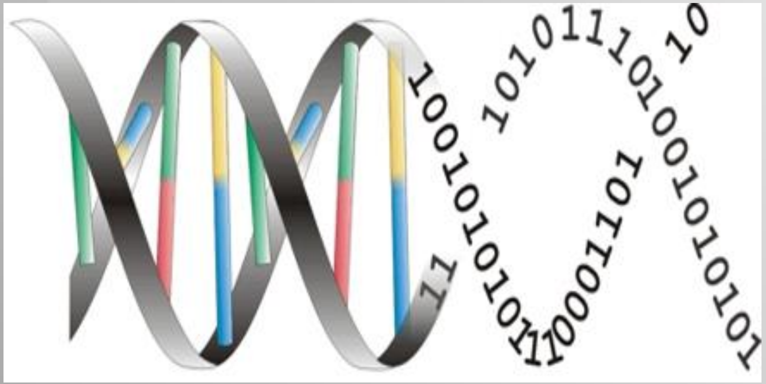
The University of Georgia

The University of Georgia

Programming for Computational & Systems Biology

Instructor: Paul Xie

Tue. & Thr. 9:35~10:50



Overview of Course

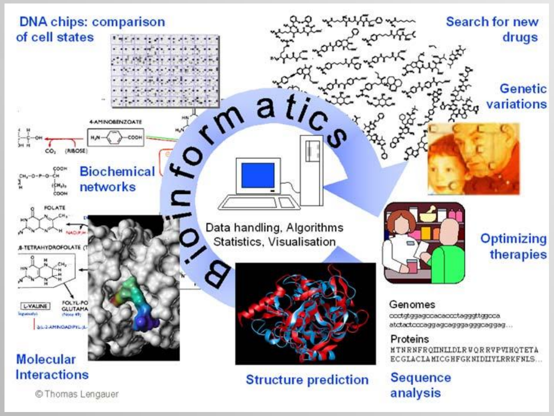
- Introduction to programming skills in computational and systems biology. Topics include real world examples, such as processing genome or proteome data, and analyzing large-scale data. The idea of “big data” will be emphasized to help students with their coding skills to discovering new knowledge in biomedical sciences and solving biomedical problems.

Course Format

- Computer labs + lectures (3 hours/week)
- Coding concepts (some knowledge about biological data)
- 6-8 assignments, published on eLC, (30-40%)
  - Please upload them by the due days
- Paper review (10-20%)
- 1 term paper (50-60%)

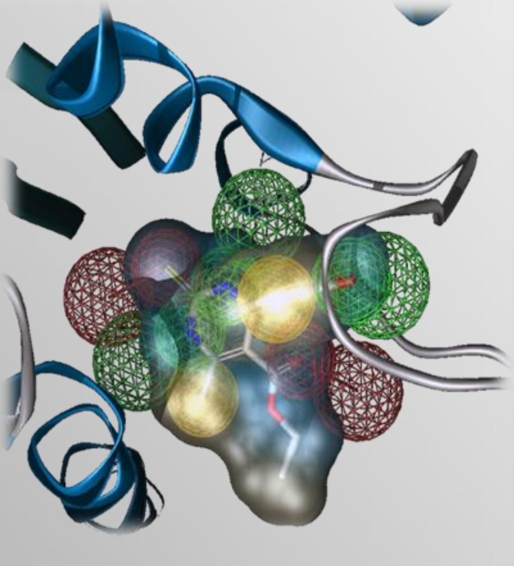
Topics

- -Omics data, e.g. Genomics, Proteomics…
- Basic Programming, e.g. I/O, variables, string, loop, regular express, array…etc.
- Data retrieval & processing
- Data analysis
- Model & building model
- Clustering and Classification
- Prediction



Logic & Loops
DNA, RNA, & Protein

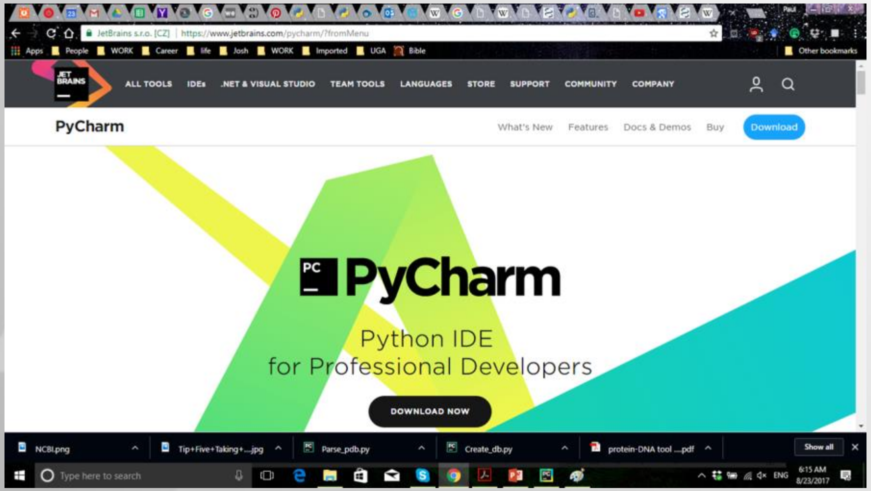
Instructor: Paul Xie (3)

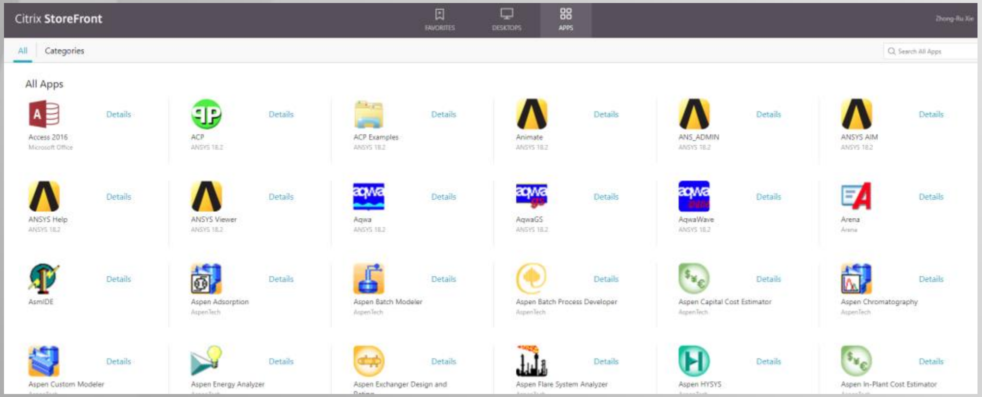


The University of Georgia

Open PyCharm

- Go to UGA CENGR Mylab
  - mylab.engr.uga.edu
- Login Citrix
  - Apps
  - PyCharm
  - New project





Code

```
def main():
        age = int(input("How old are you?"))
        if age < 16:
                print("You cannot drive")
        elif age < 21:
                print("You can drive but cannot drink")
        else:
                print("You can drink and drive")
        elif (DNA == 'G'):
                pair = 'C'
        print("%s\n|\n%s" %(DNA,pair))
main()
```

Code

```
def  main():
    a = 'Hello '
    b = 'World '
    print(a+b)
    DNA = input('please input a nucleotide: ')
    pair = ''
    if (DNA == 'A'):
        pair = 'T'
    elif (DNA == 'T'):
        pair = 'A'
    elif (DNA == 'C'):
        pair = 'G'
    elif (DNA == 'G'):
        pair = 'C'
    print("%s\n|\n%s" %(DNA,pair))
main()
```

Iterations

- Loop
  - While loop
  - For loop

Exercise

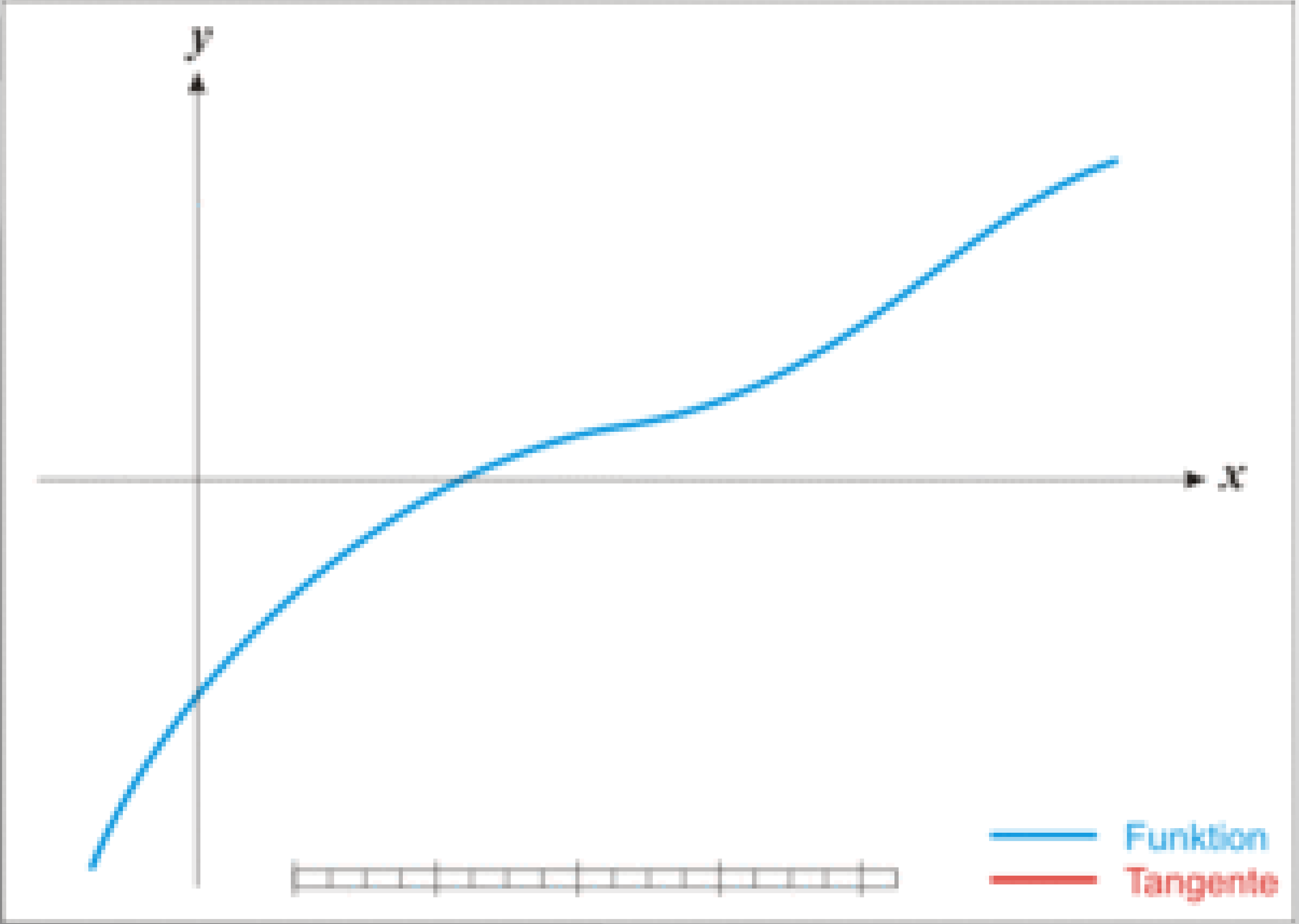
- While loop
  - Count or count down
  - Factorial
  - e.g. 3!, 5!, 12!, x!....

Iterations

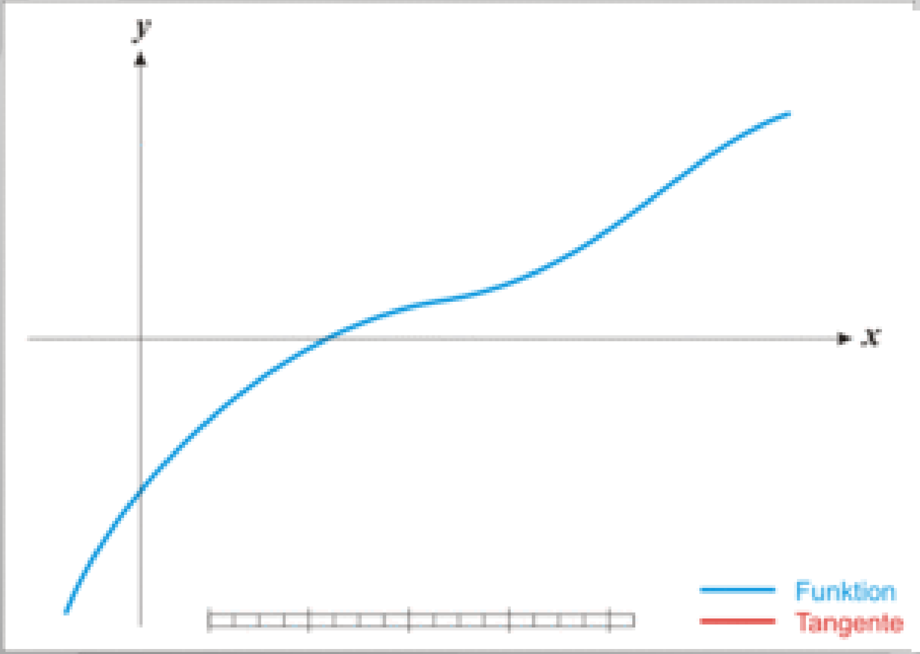
The University of Georgia

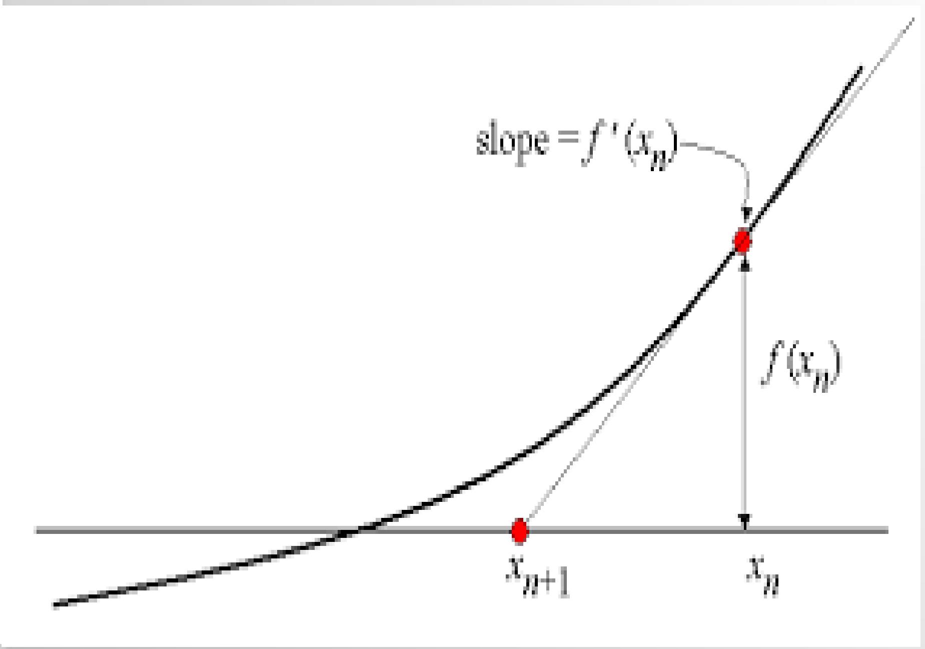
- Loop
  - While loop
  - Do-while loop
  - For loop
- Fibonacci number
- $(f_0 = 0,)\ f_1 = 1,\ f_2 = 1,\ f_n = f_{n-1} + f_{n-2}$
- 1,1,2,3,5,8,13,21,34,55,….
- Given a number N, how can we find all the factors (e.g. n = 120) f = 2,3,4,6,12,15,
- $y = ax^2+bx+c$; y = x1, x2 (y = -b+/-

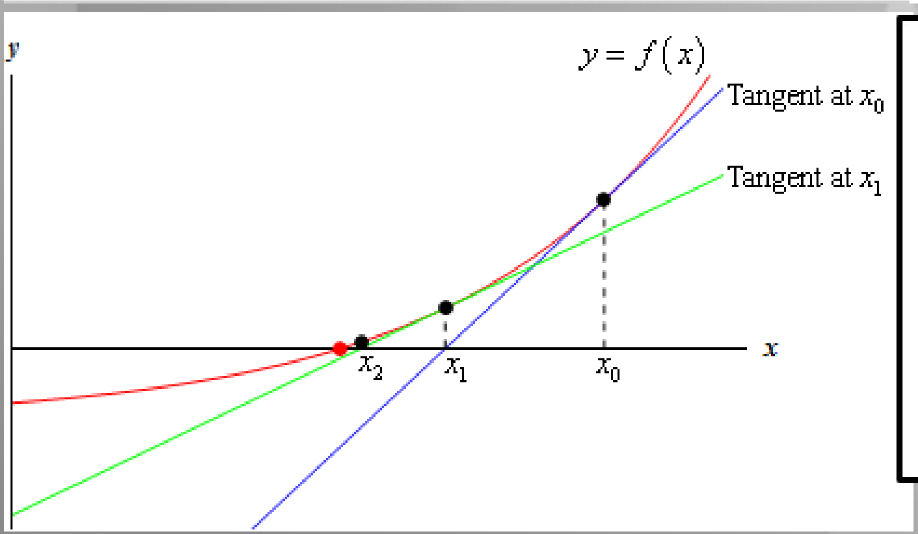
Newton’s Method

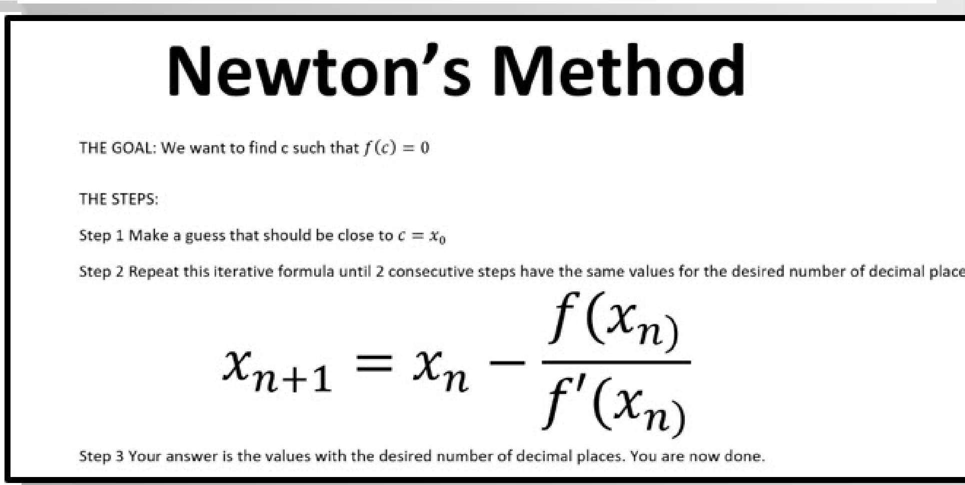


Newton's Method









Newton's Method

THE GOAL: We want to find c such that $f(c) = 0$

THE STEPS:

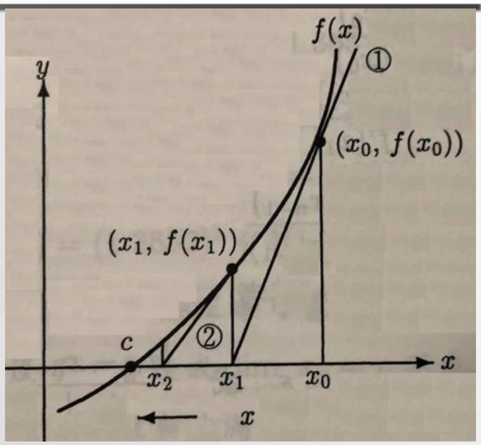
Step 1 Make a guess that should be close to $c = x_0$

Step 2 Repeat this iterative formula until 2 consecutive steps have the same values for the desired number of decimal place

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Step 3 Your answer is the values with the desired number of decimal places. You are now done.

Newton-Raphson Method



- Taylor expansion
- $f(x) = f(c) + f'(c)\,(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^n(c)}{n!}(x-c)^n$
- The first-degree Taylor Polynomial
- $f(x) = f(c) + f'(c)\,(x-c) + \cancel{O(h^2)}$
- When $f(x) = 0$,
  $f(x) = f(x_0) + f'(x_0)\,(x-x_0) = 0$

$$\Rightarrow x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Random Number

- Loop
  - While loop
  - For loop

Iterations

- Loop
  - While loop
  - For loop
- Another input method – open file
  - F = open(“file path”,”r”)
  - Read and write
  - F.readline()
- Function : len()

Iterations

- String
  - – String is a series of characters

Random Number

- Import random
- random.randint()
- random.randrange()
- random.random()
- random.uniform
- random.choice()
- random.choices
- random.sample

Syntax

- random.randint(begin,end) ➔ to generate a random integer begin ≤ v ≤ end
- random.randrange(begin,end(,skip)) ➔ to generate a random integer begin ≤ v < end (and skip by)
- random.sample(range(begin,end),n)
- random.random() ➔ to generate a random float
- random.uniform(begin,end) ➔ float begin ≤ v < end
- random.choice() ➔ randomly choose a character from a list

Exercise

- Guess number
  1. Generate random number
  2. Input a guess number
  3. Check if the guess is correct
  4. iterate

Exercise

- Turtle and Rabbit race
  1. Dice (random number generator: 1~6)
  2. Record (the updated positions of turtle and rabbit)
  3. demonstrate track and current positions (visualized positions)
  4. The unfinished track (visualized track)
  5. The finish line
  6. The loop (iteration of dice rolling and run forward)
  7. Clean the previous line (system(“cls”)

Exercise

- Secret message (encryption) t➔ 7➔#
  1. Input a message to deliver
  2. (secret code)
  3. Split the message and put into a random code
  4. Decoder
     ex： I/|u;#\+|DNgi+
     9(5uF!Ivsncof#O_w{JJaF6A6imX<h-D(
     7 NgL#*|Fr5p{(3\LQd-0Pfv8
     +s\T

Assignment

- Self-assessment
  - Do and submit your assignment onto eLC
  - Grade your own assignment based on the rubric
  - I will grade your assignment too
  - Download and compare the two assessment
- Peer-assessment
  - Download the assignment done by classmates
  - Grade those assignments based on the rubric
  - Submit your grading

– Getting scores on doing and grading the assignment

Weakness & Solutions

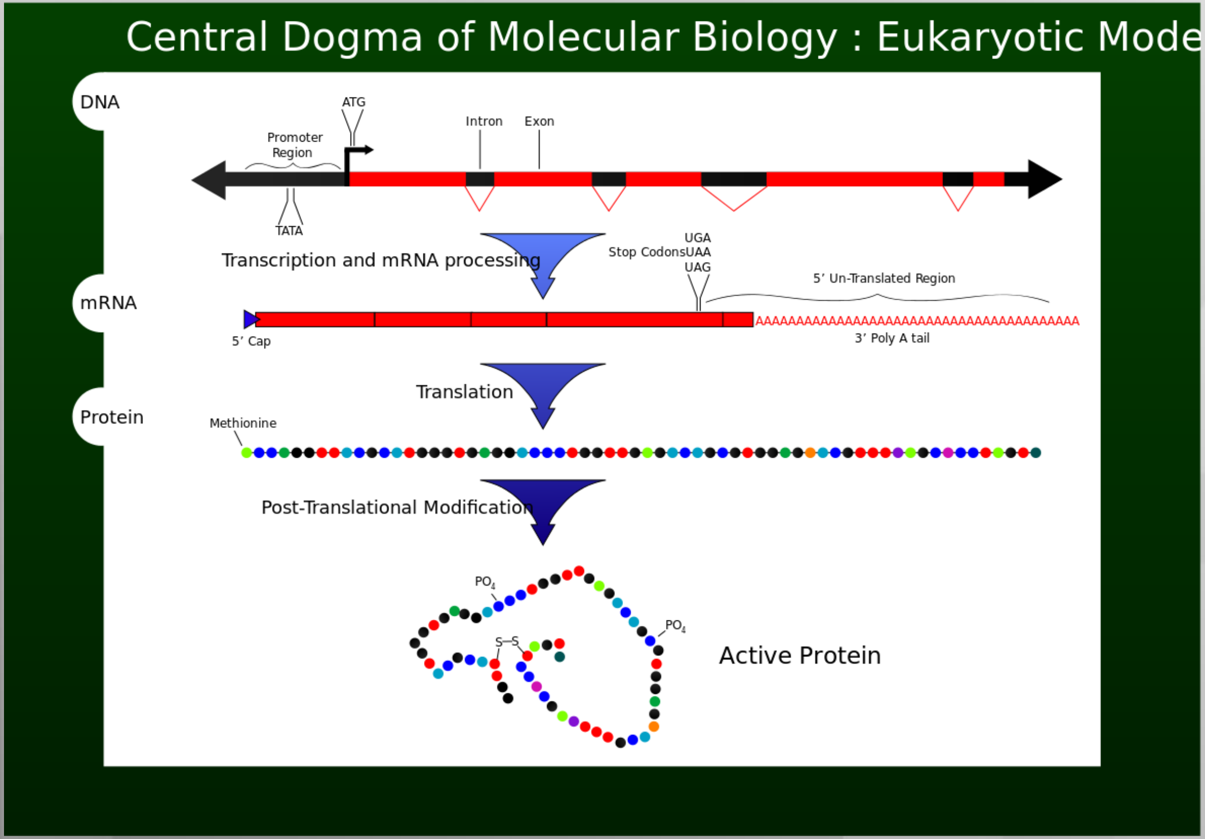
- Q: How can we get all the factors (prime) of an integer?
  - Exercise
- Possible weakness:
  1. Empty spaces ➔ leaking key1
  2. Total length ➔ leaking key2
- Possible solutions?
  - Exercise

Exercise

- Random sequences/proteins?
  1. Randomly generate sequences
  2. Find (and record) the starting code ‘ATG’
  3. Count/Translate each 3 codes
  4. Calculate the length
  5. Find the stop code ‘TGA’
  6. Record the length

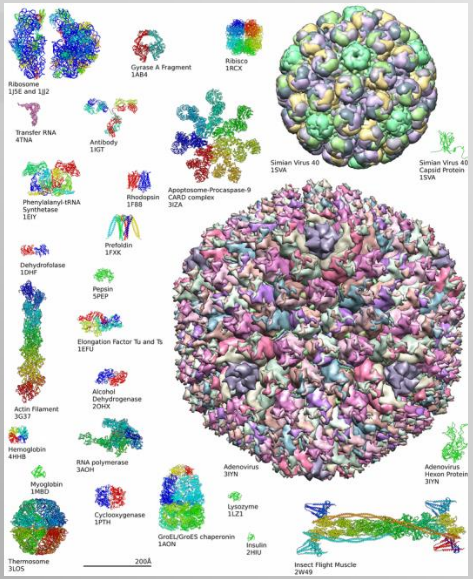
TIME
FOR A
BREAK

Central Dogma



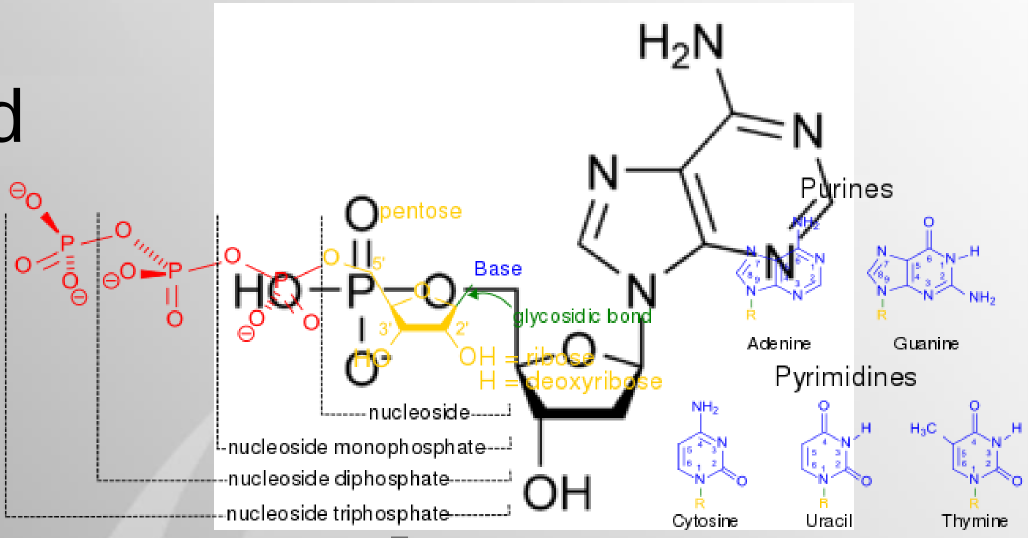
Central Dogma of Molecular Biology : Eukaryotic Mode
DNA
ATG
Promoter Region
Intron
Exon
TATA
Transcription and mRNA processing
Stop Codons UGA UAA UAG
5′ Un-Translated Region
mRNA
5′ Cap
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
3′ Poly A tail
Translation
Protein
Methionine
Post-Translational Modification
PO4
PO4
S–S
Active Protein

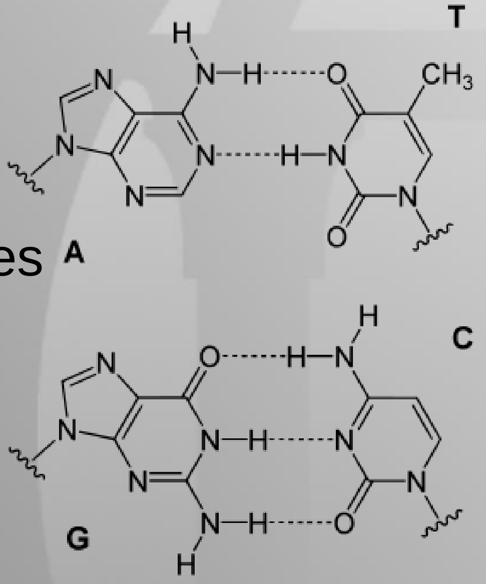
- DNA→RNA→ Protein
- Transcription & Translation (&Replication)

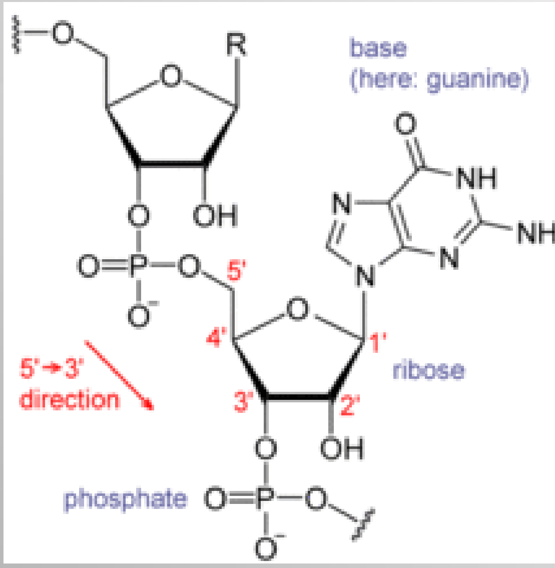


DNA & RNA

- Deoxyribonucleic acid
- Ribonucleic acid
  - Base: A, T/U, C, G
    - Adenine
    - Thymine
    - Uracil
    - Guanine
    - Cytosine
    - Purines & Pyrimidines
  - Ribose
  - Phosphate

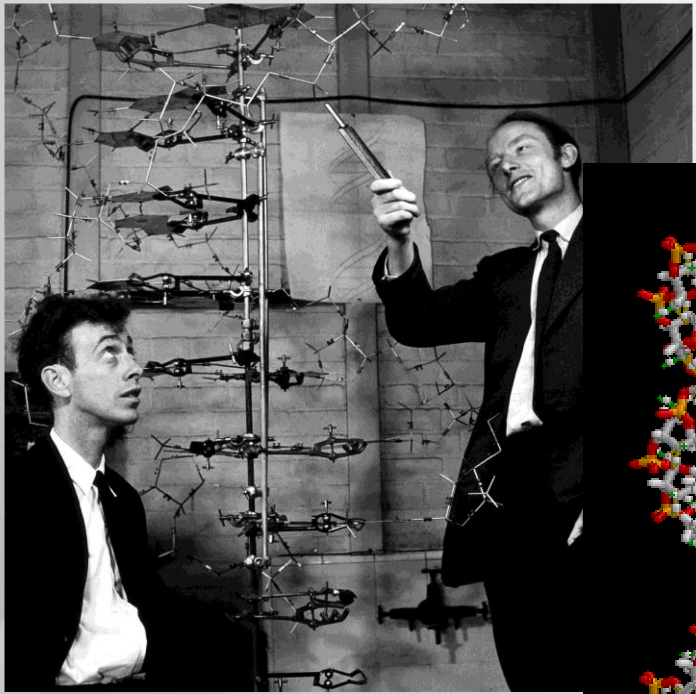


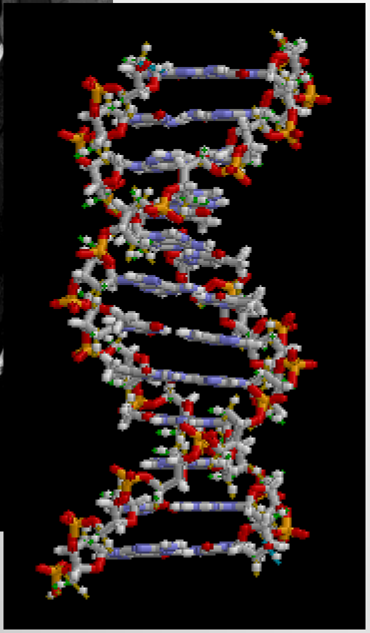


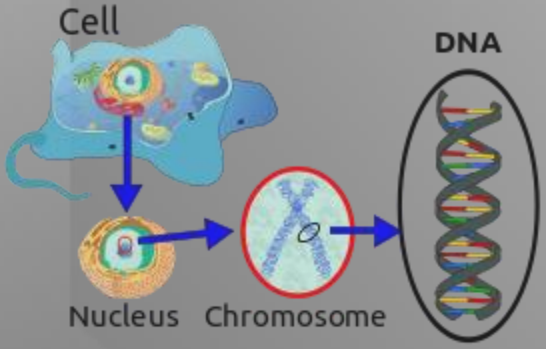


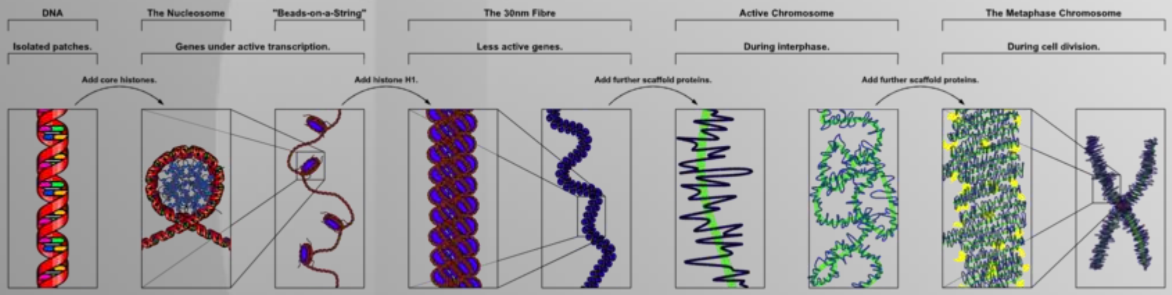
Double Helix!

- Double helix
  - Chromosomes
- Replication
  - DNA-polymerase
- Transcription
- Translation



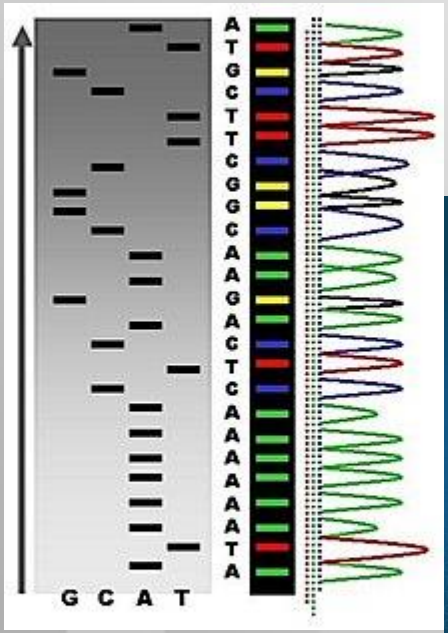


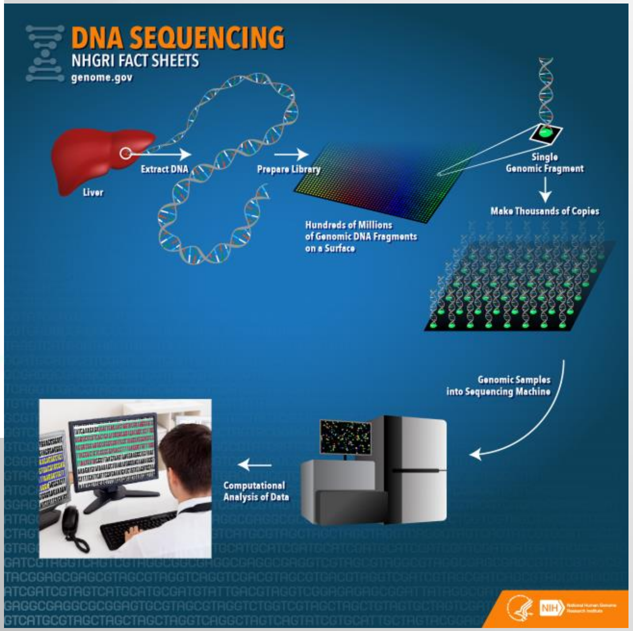


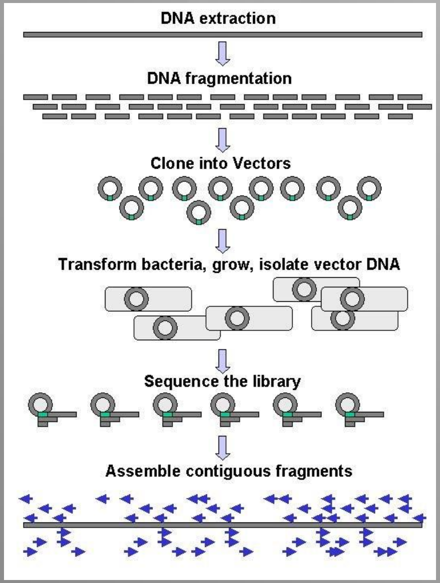


Sequencing

- Restriction Enzyme
- Gel & Electrophoresis

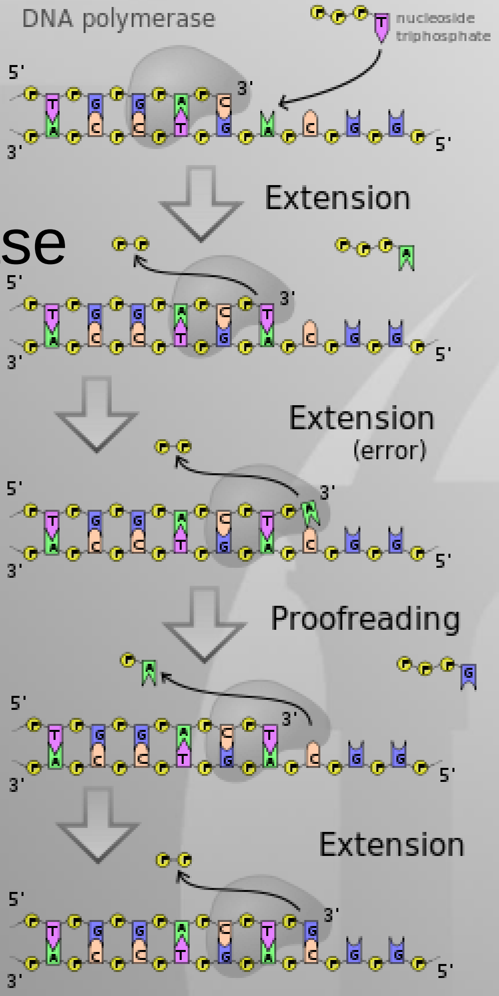


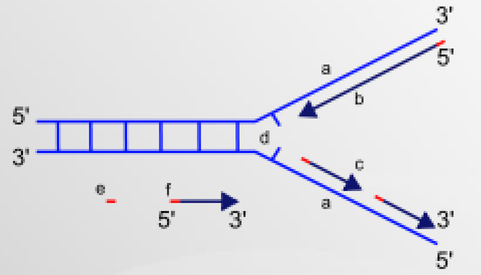

DNA SEQUENCING
NHGRI FACT SHEETS
genome.gov
Liver
Extract DNA
Prepare Library
Hundreds of Millions of Genomic DNA Fragments on a Surface
Single Genomic Fragment
Make Thousands of Copies
Genomic Samples into Sequencing Machine
Computational Analysis of Data

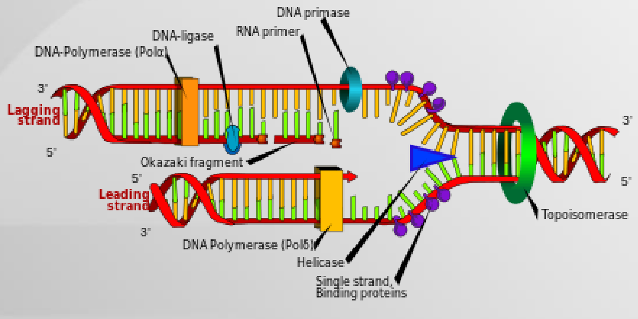

DNA extraction
DNA fragmentation
Clone into Vectors
Transform bacteria, grow, isolate vector DNA
Sequence the library
Assemble contiguous fragments

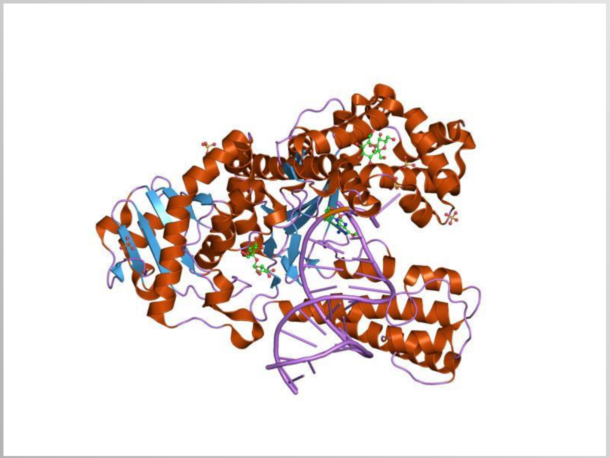
Replication

- Double helix
- Replication
  - DNA-polymerase
- Transcription
- Translation



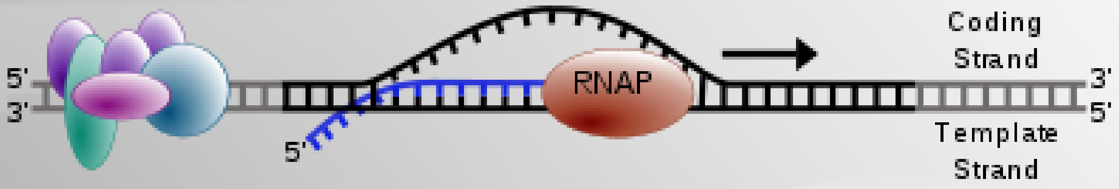


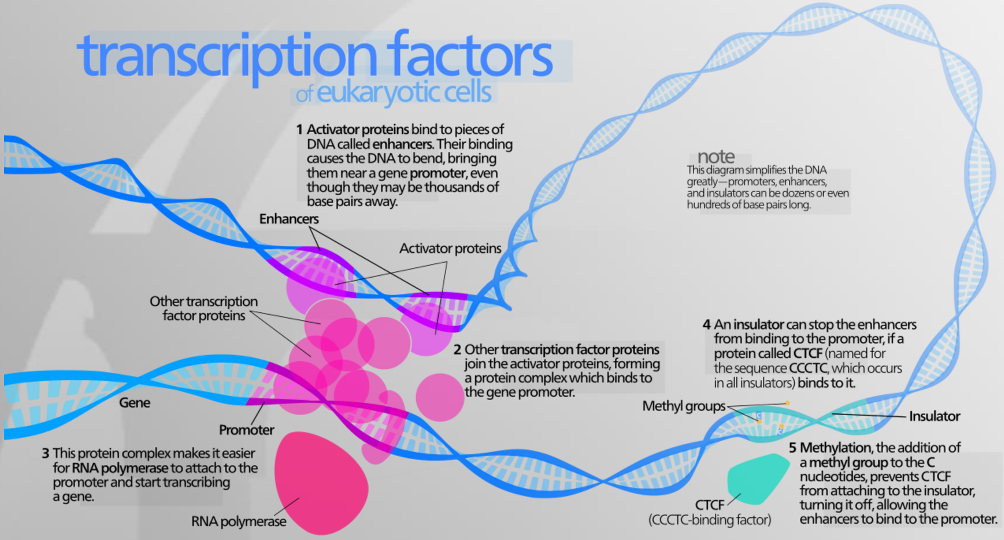




Transcription

- Double helix
- Replication
- Transcription
  - RNA Polymerase
  - Transcription Factors
  - Expression control
- Translation





1 **Activator proteins** bind to pieces of DNA called **enhancers.** Their binding causes the DNA to bend, bringing them near a gene **promoter**, even though they may be thousands of base pairs away.

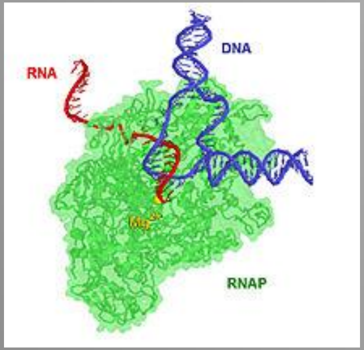
2 Other **transcription factor proteins** join the activator proteins, forming a protein complex which binds to the gene promoter.

3 This protein complex makes it easier for **RNA polymerase** to attach to the promoter and start transcribing a gene.

4 An **insulator** can stop the enhancers from binding to the promoter, if a **protein called CTCF** (named for the sequence **CCCTC**, which occurs in all insulators) **binds to it.**

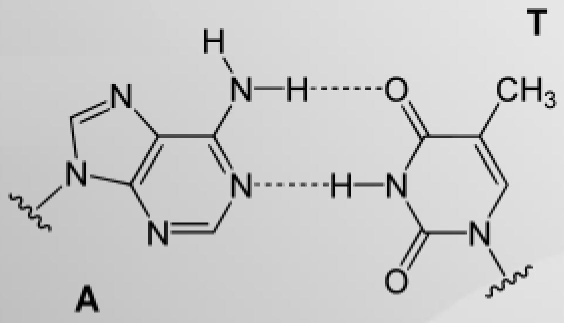
5 **Methylation**, the addition of a **methyl group** to the **C** nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.

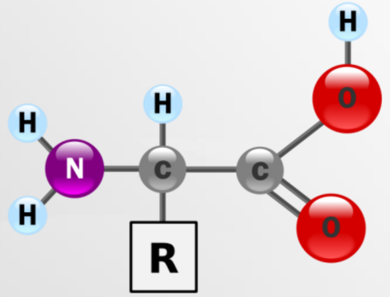
This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be dozens or even hundreds of base pairs long.

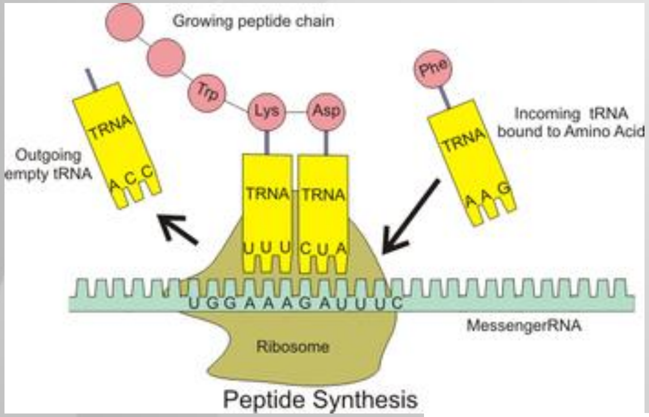


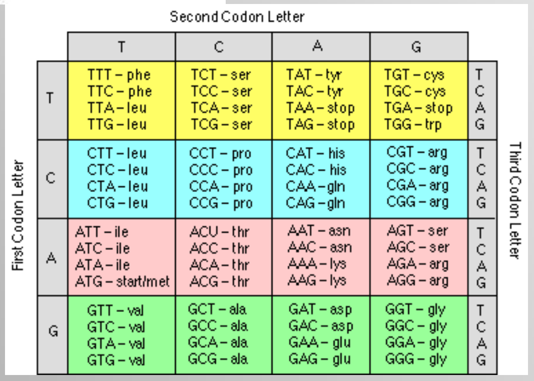
Translation

- Double helix
- Replication
- Transcription
- Translation
  - Ribosome
  - mRNA
  - tRNA
  - Nucleic acid to Amino acid







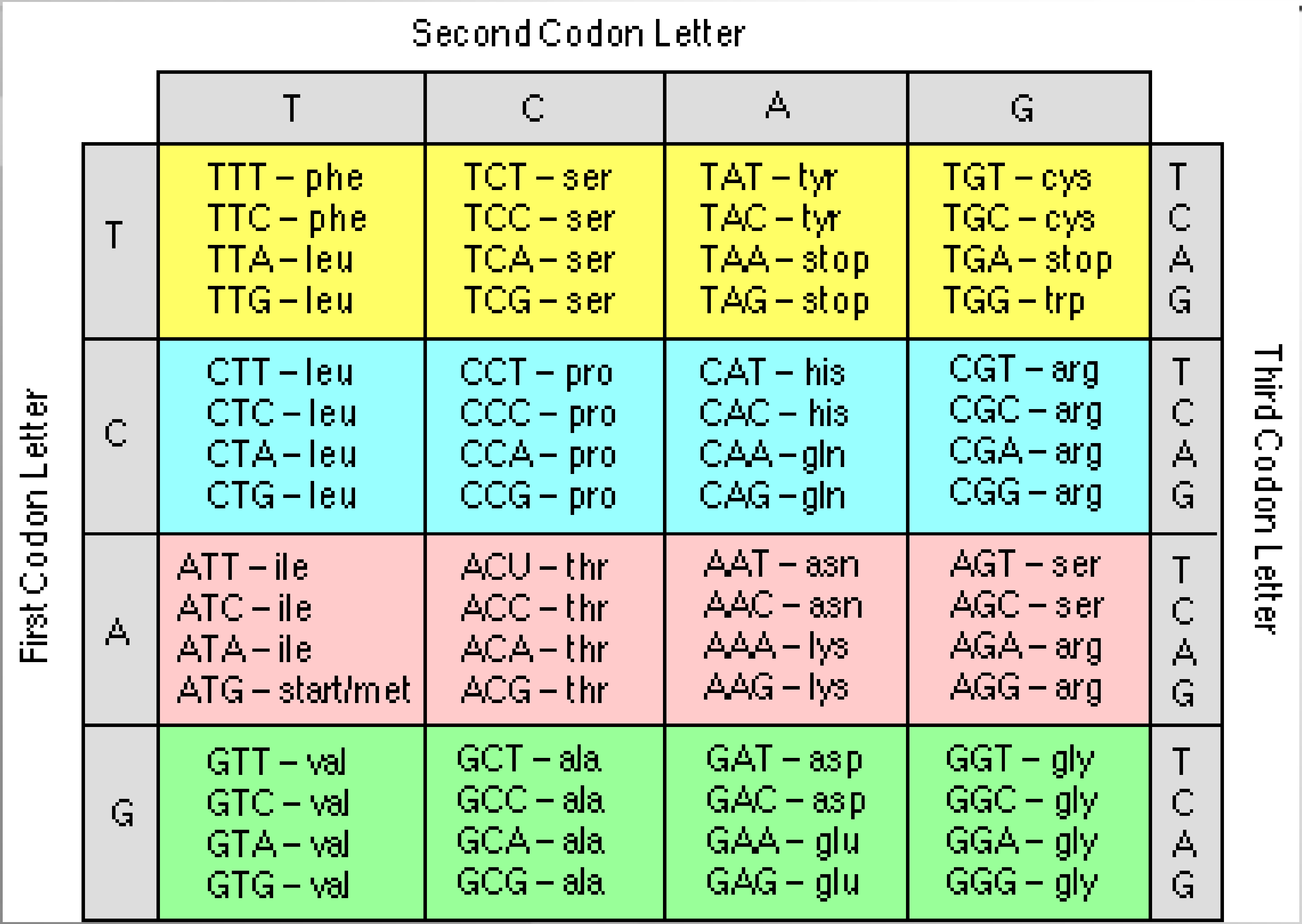


		Second Codon Letter				
		T	C	A	G	
First Codon Letter	T	TTT – phe TTC – phe TTA – leu TTG – leu	TCT – ser TCC – ser TCA – ser TCG – ser	TAT – tyr TAC – tyr TAA – stop TAG – stop	TGT – cys TGC – cys TGA – stop TGG – trp	T C A G
	C	CTT – leu CTC – leu CTA – leu CTG – leu	CCT – pro CCC – pro CCA – pro CCG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

Third Codon Letter

DNA → AA

The University of Georgia



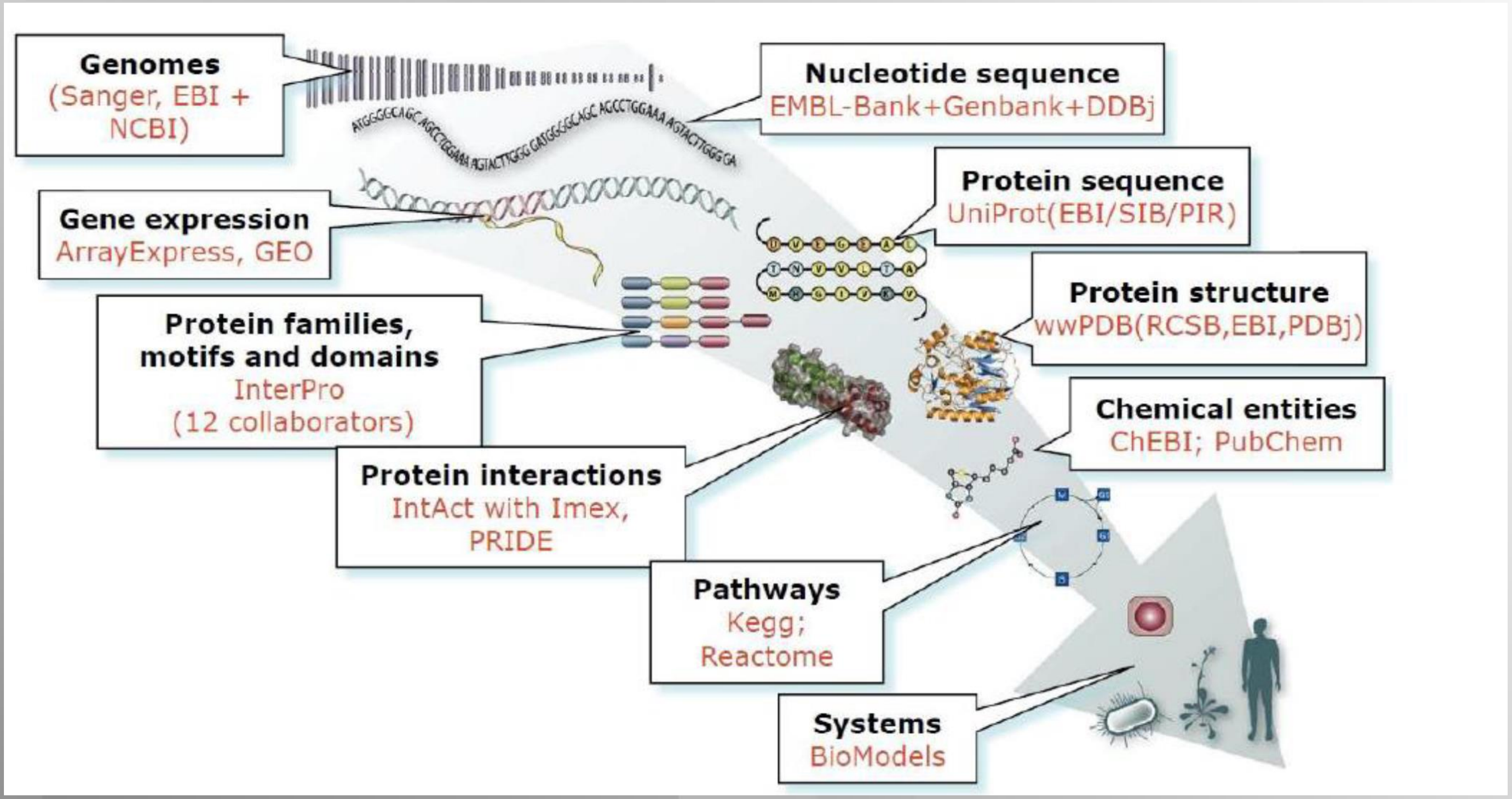
Second Codon Letter

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

20 AAs

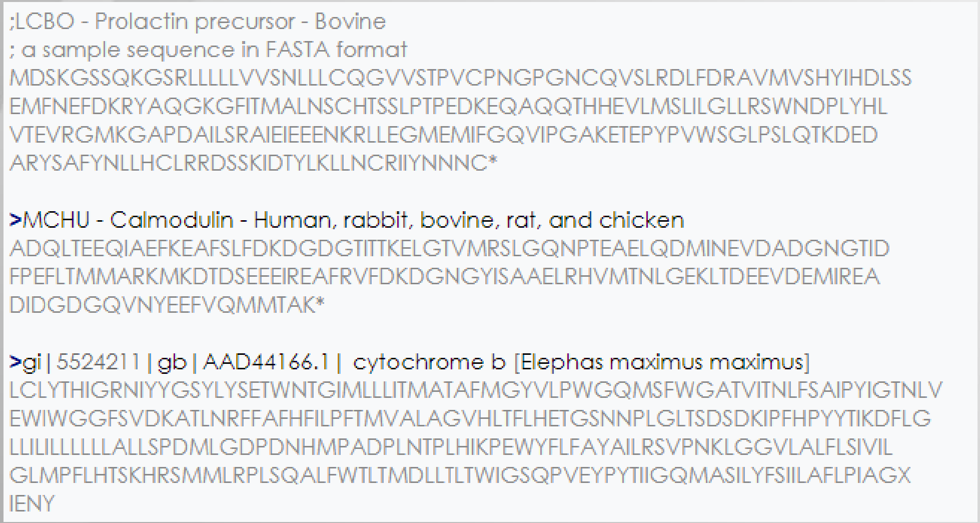
Amino Acid	3 letter code	1 letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine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

-Omics & Online Databases



Sequence Data

- NCBI
- RefSeq
- FASTA format
  - Single letter code
  - Sequence names and comments preceding the sequences
  - FASTA



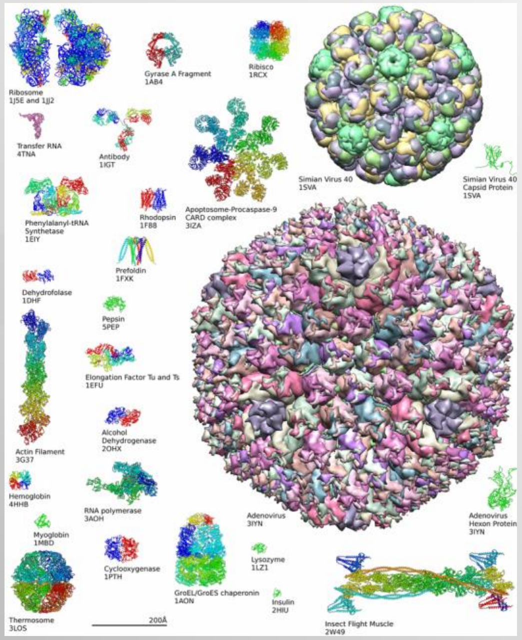
```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

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

Summary

- DNA→RNA→Protein
- NCBI & RefSeq
- FASTA format
- Pycharm
- The first program of Python
- Input/Output
- Variables



The University of Georgia