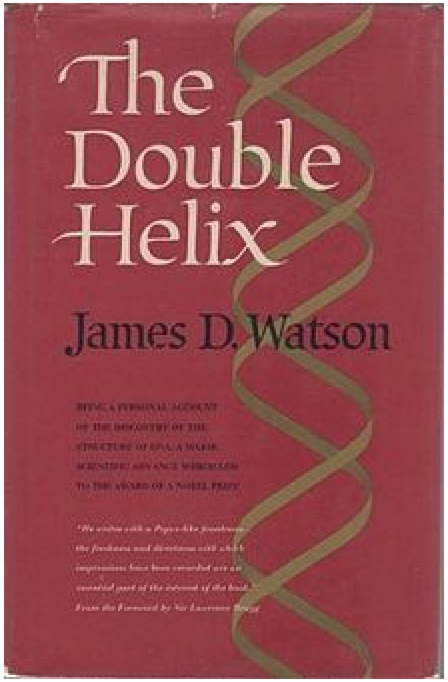
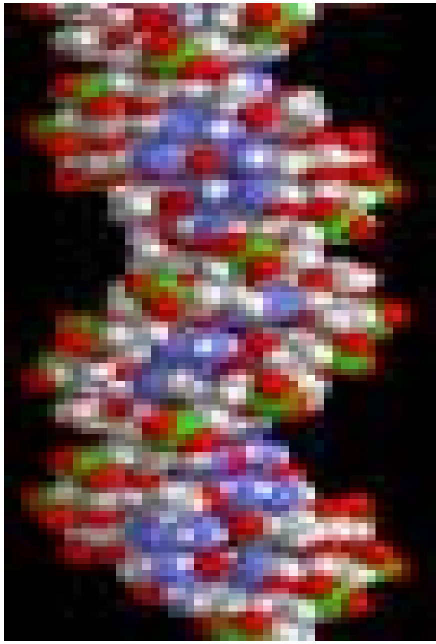
Kusuma School of Biological Sciences

SBL-100

Introductory Biology for Engineers
L-3

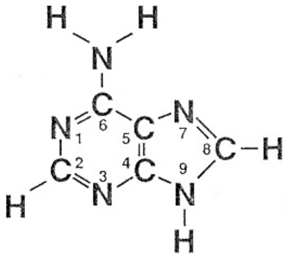


Nucleic Acids

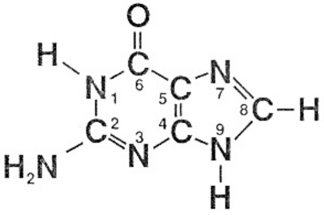


Disclaimer. These slides are compiled from diverse textbooks and resources in public domain, exclusively for teaching purposes. The author does not claim copyrights.

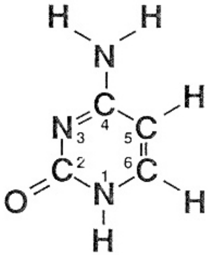
The nucleic acid bases, nucleosides & nucleotides



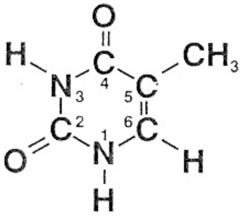
Adenine
(A)



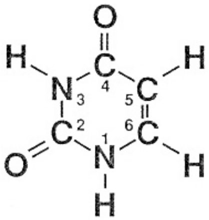
Guanine
(G)



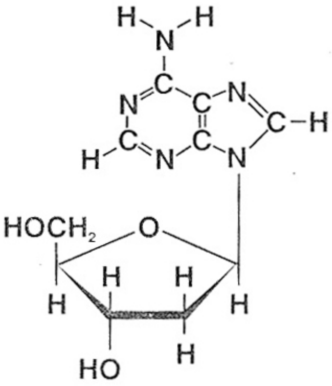
Cytosine
(C)



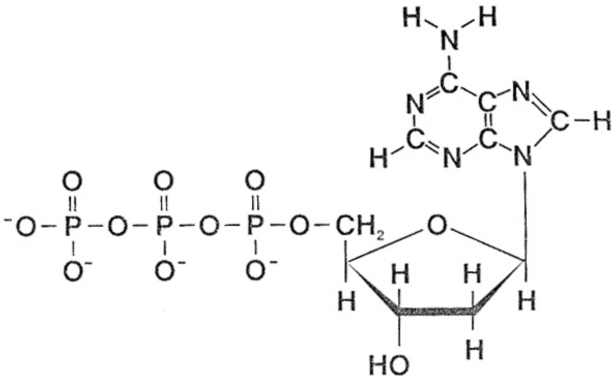
Thymine
(T)



Uracil
(U)

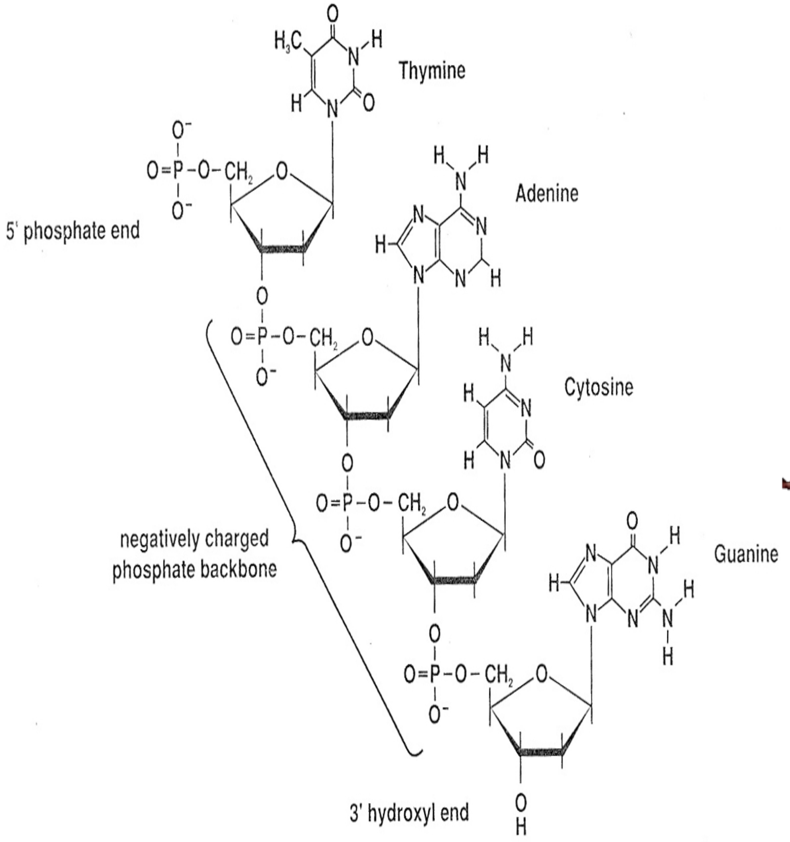


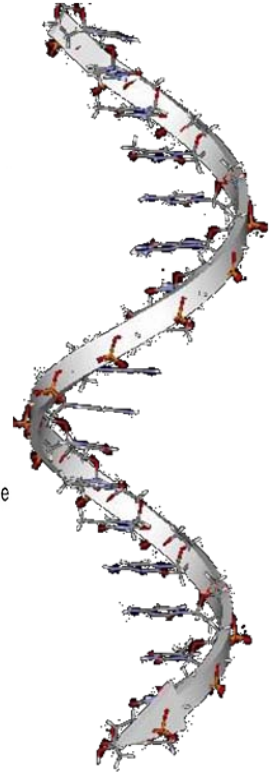
Deoxyadenosine
(a nucleoside)



Deoxyadenosine 5' -triphosphate
(dATP)
(a nucleotide)

The phosphodiester backbone, the bases and the sugars and their covalent connectivities





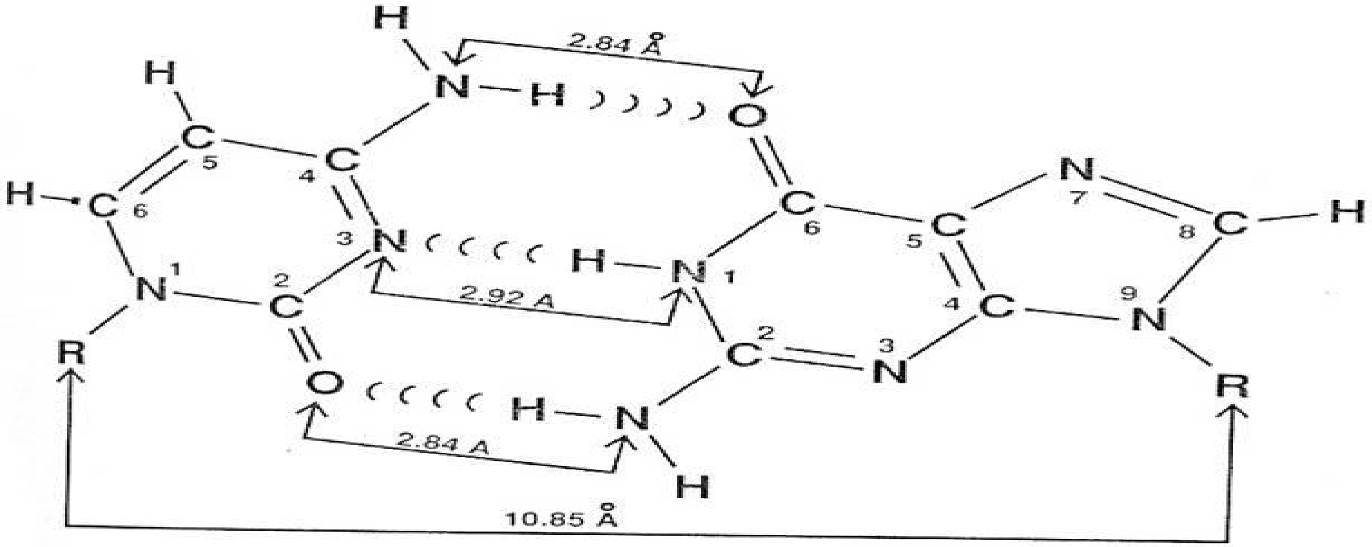
Rotational degrees of freedom in a nucleotide)

	Angle	Rotation around	Common values for B-DNA	Comments
1	χ	C-N glycosidic bond	-123.7	Base orientation
2	Φ	sugar pucker	C2' Endo or C3' Endo	Phase
3	γ	C4'-C5'	56.0	Phosphodiester Backbone
4	β	C5'-O5'	169.6	--do--
5	δ	C3'-C4'	117.3	--do--
6	ε	C3'-O3'	-166.0	--do--
7	α	P-O5'	-74.5	--do--
8	ζ	P-O3'	-97.1	--do--

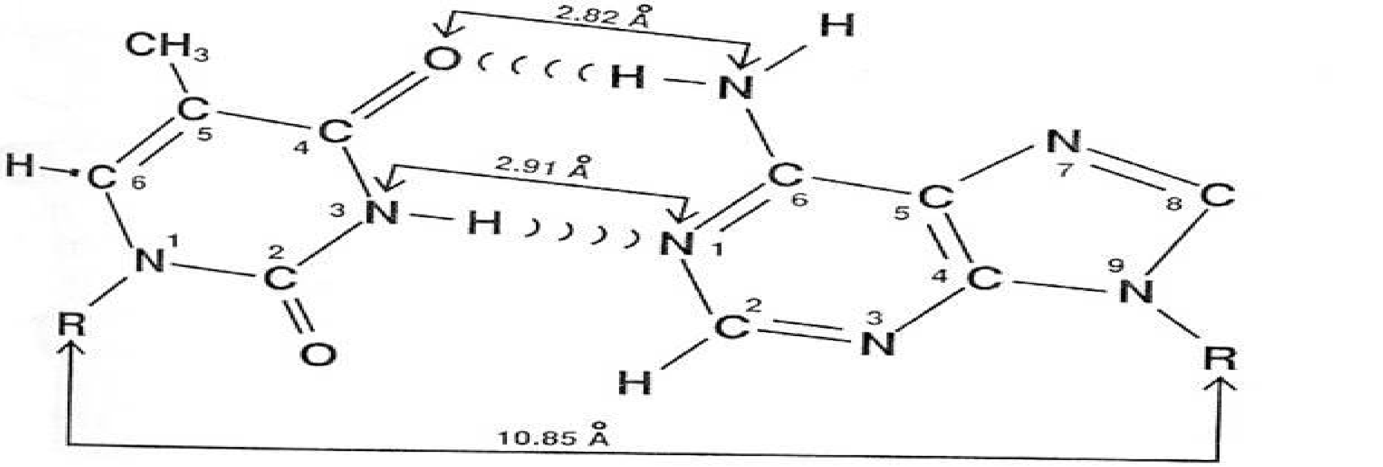
Single stranded nucleic acids

Watson-Crick (WC) pairing

C.G

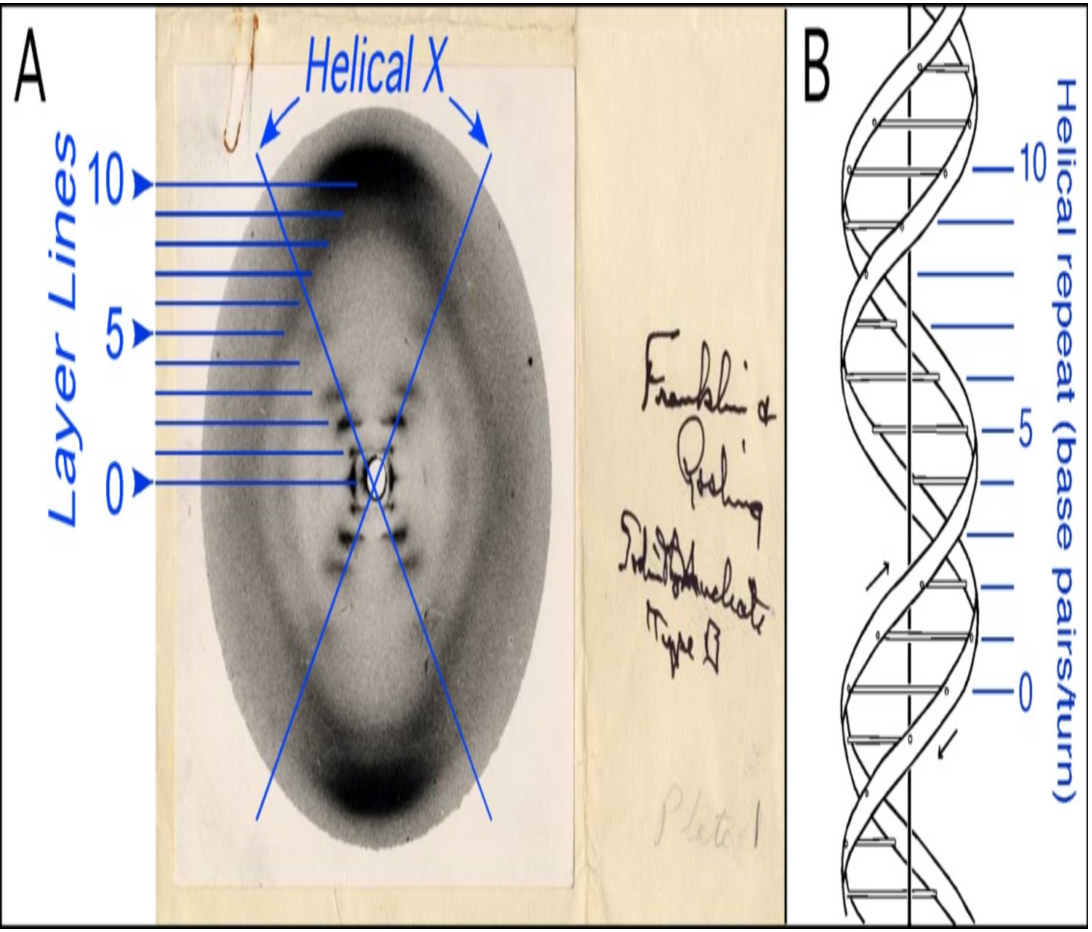


T.A



Double helical DNA

X-ray diffraction photograph of a DNA fiber at high humidity (Franklin and Gosling, 1953). Interpretation of the helical-X and layer lines added in blue.



The Nobel Prize in Physiology or Medicine 1962

"for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"

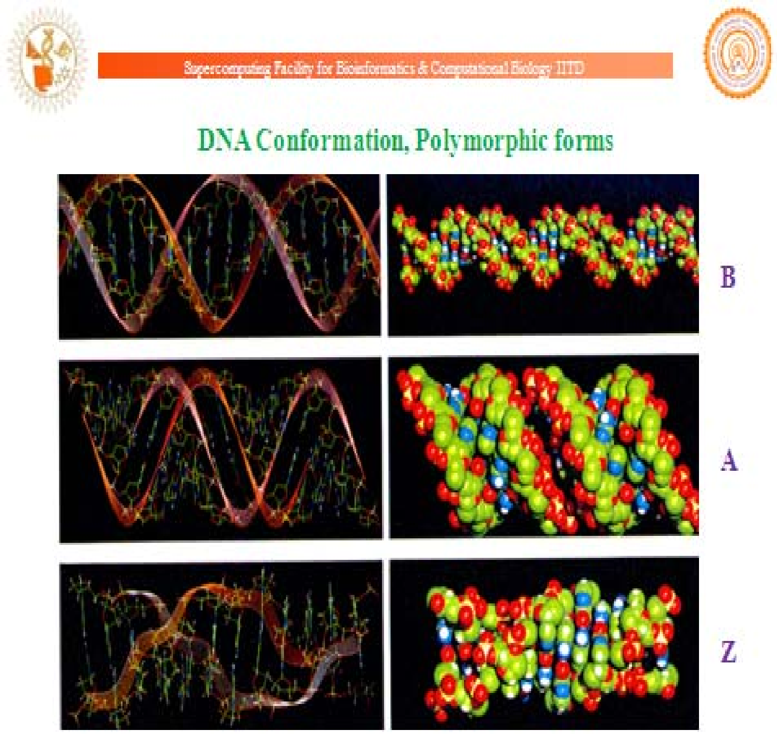
Francis Crick
MRC, UK
b. 1916 (UK)

James Watson
Harvard U., USA
b. 1928 (IUSA)

Maurice Wilkins
London U., UK
•b. 1916 (new Zealand)

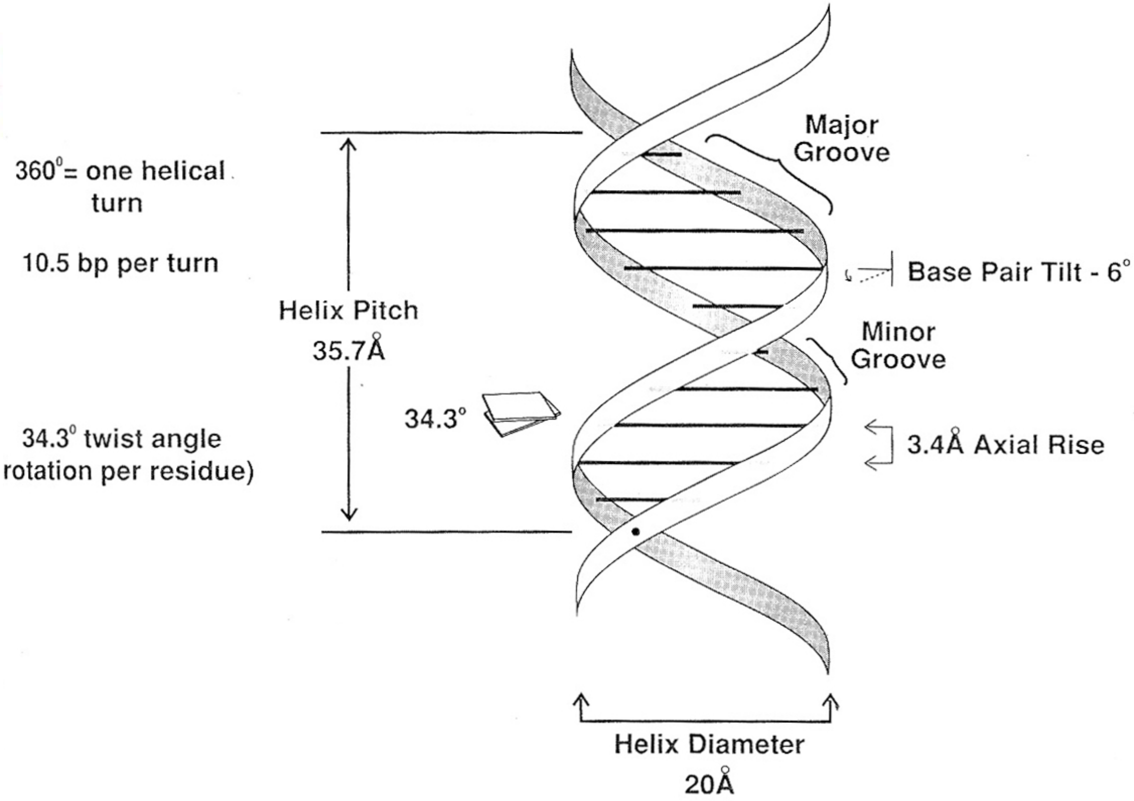
Structure types and helical parameters*

Structure type	Pitch	Helical Symmetry	Axial rise	Turn angle	Minor groove(w)	Major groove(w)	Minor depth (d)	Major groove(d)
A DNA	28.2	11	2.56	32.7	11.0	2.7	2.8	13.5
B DNA	**33.8**	**10**	**3.38**	**36.0**	**5.7**	**11.7**	**7.5**	**8.5**
Z DNA	45.0	6	3.70	-30.0	8.8	2.0	3.7	13.8

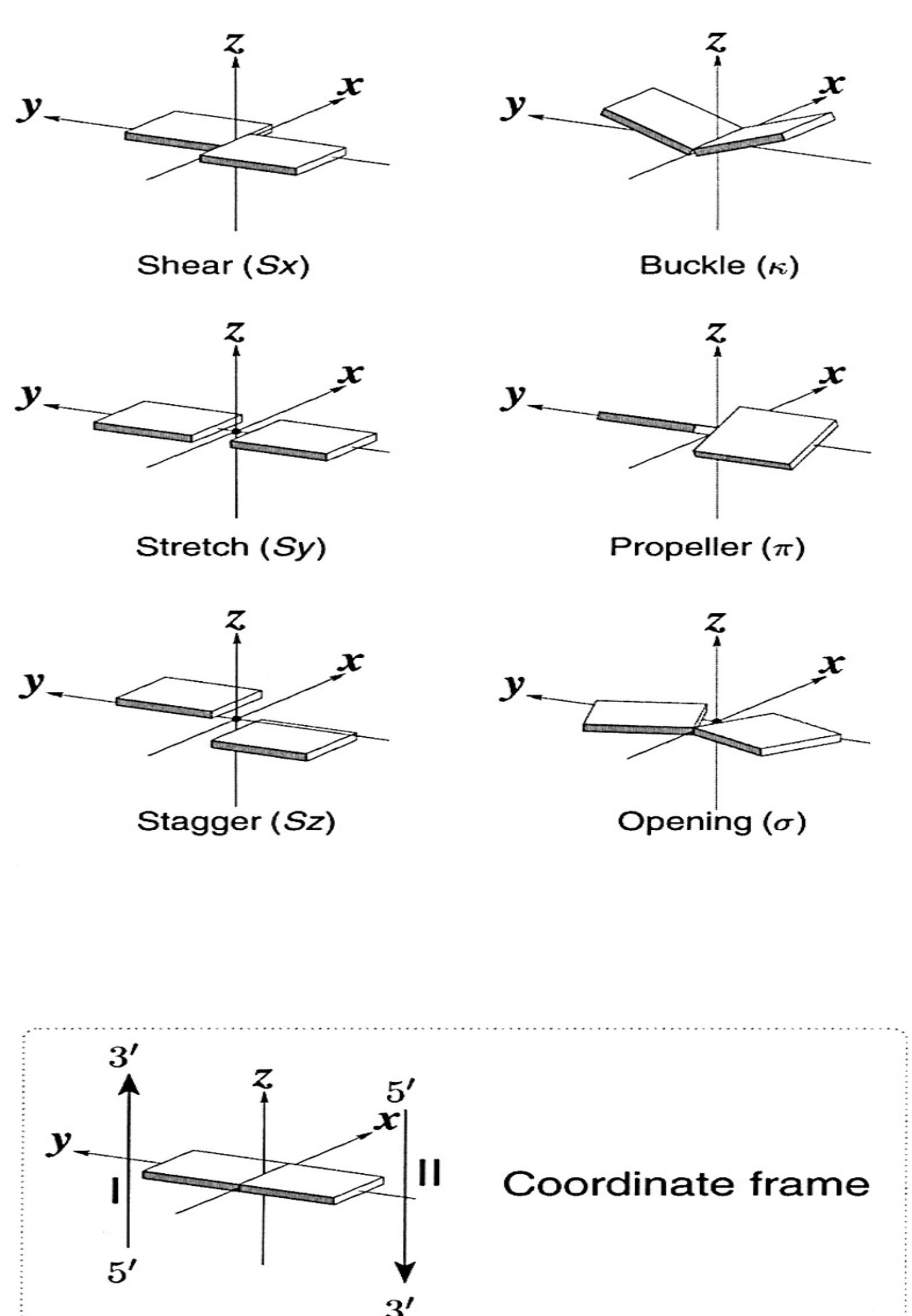


Supercomputing Facility for Bioinformatics & Computational Biology IITD

DNA Conformation, Polymorphic forms



B DNA is physiologically the most relevant form



Shear (Sx)

Buckle (κ)

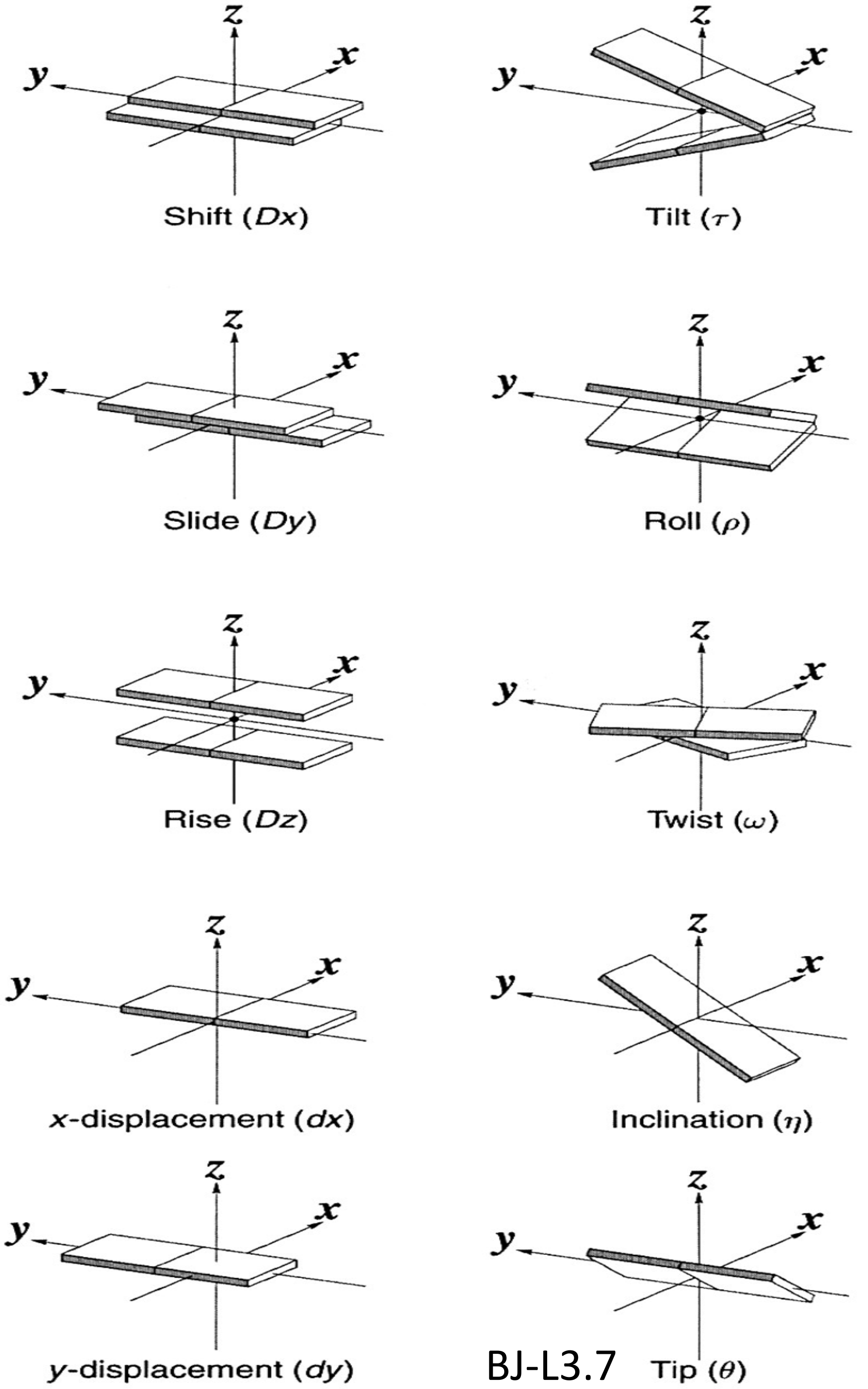
Stretch (Sy)

Propeller (π)

Stagger (Sz)

Opening (σ)

Coordinate frame



Shift (Dx)

Tilt (τ)

Slide (Dy)

Roll (ρ)

Rise (Dz)

Twist (ω)

x-displacement (dx)

Inclination (η)

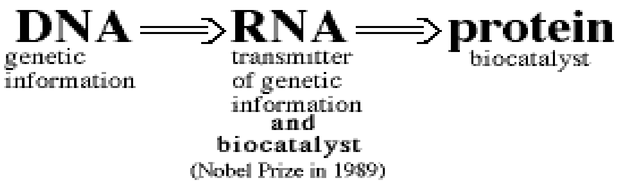
y-displacement (dy)

Tip (θ)

Intra-base pair & inter-base pair parameters

DNA is a dynamic molecule!!

BJ-L3.7



DNA ⟹ RNA ⟹ protein

genetic information

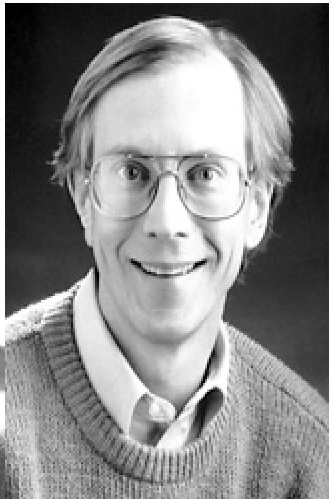
transmitter of genetic information and biocatalyst (Nobel Prize in 1989)

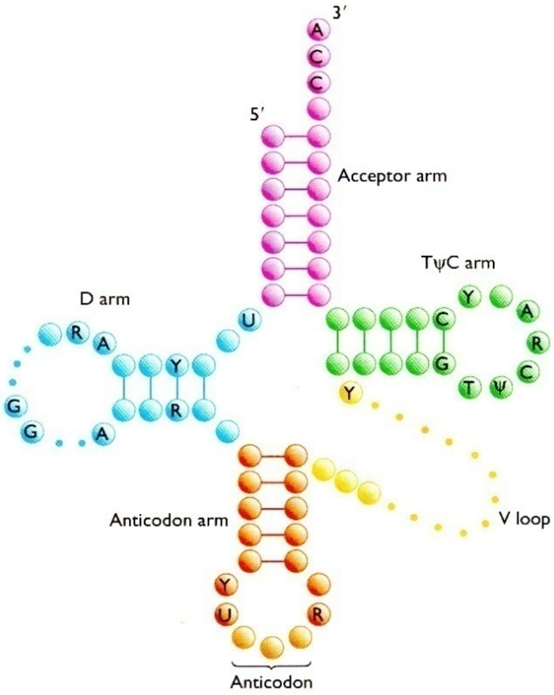
biocatalyst

Ribozymes

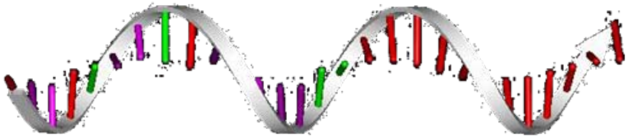
Self splicing of RNA

The Nobel Prize in Chemistry 1989 was awarded jointly to Sidney Altman and Thomas R. Cech "for their discovery of catalytic properties of RNA"

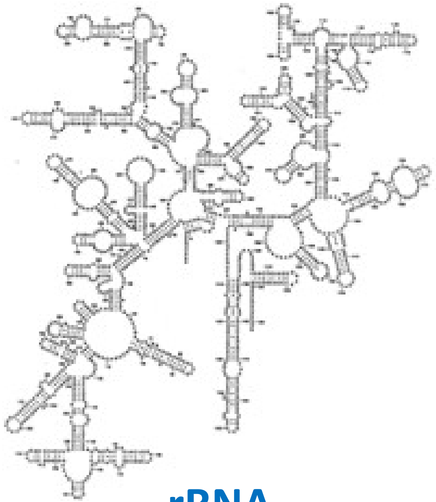




tRNA



mRNA

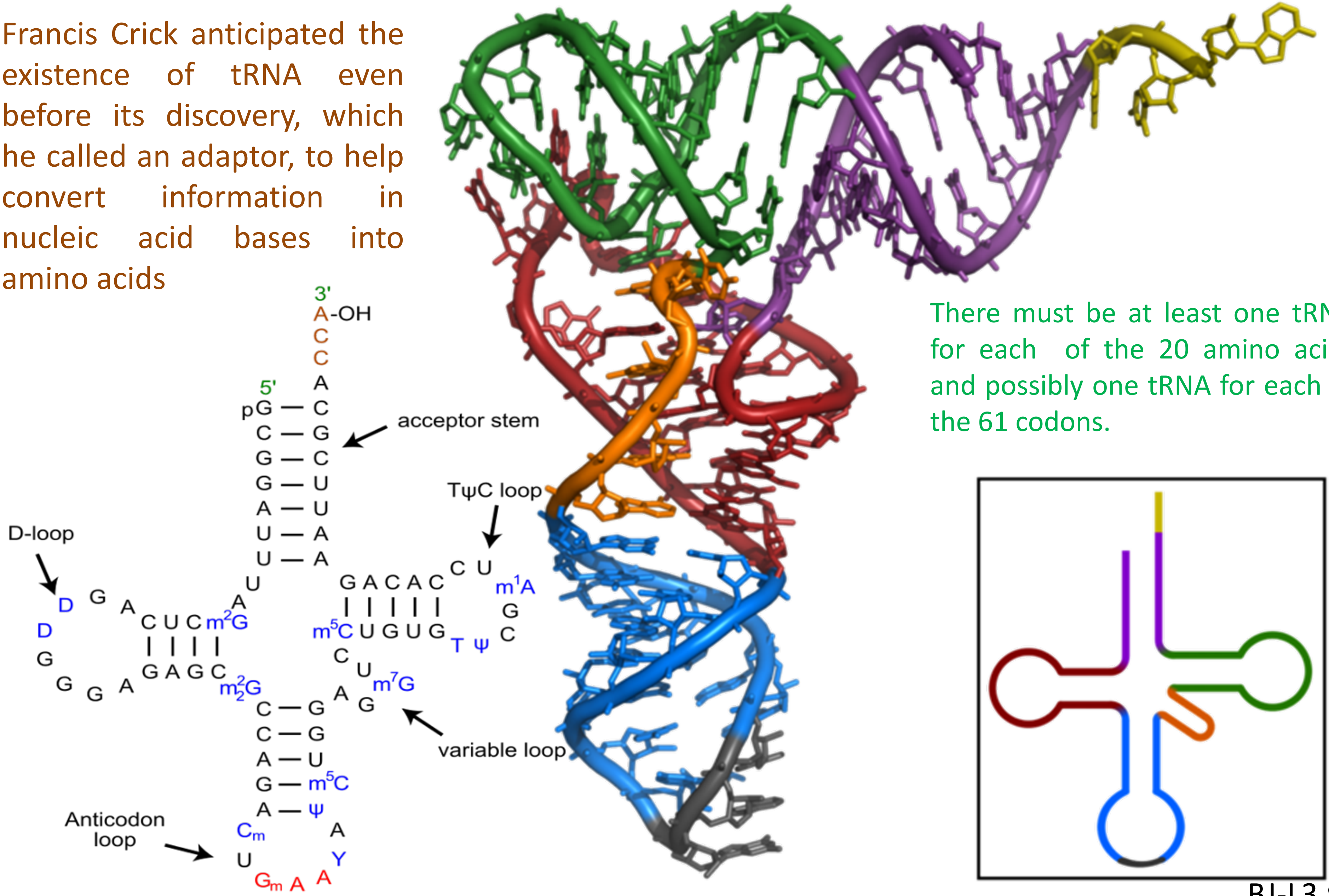


rRNA

Genes of 16SrRNA, a constituent of small subunit of ribosome, are used in Phylogeny since they are more conserved

More of t-RNA

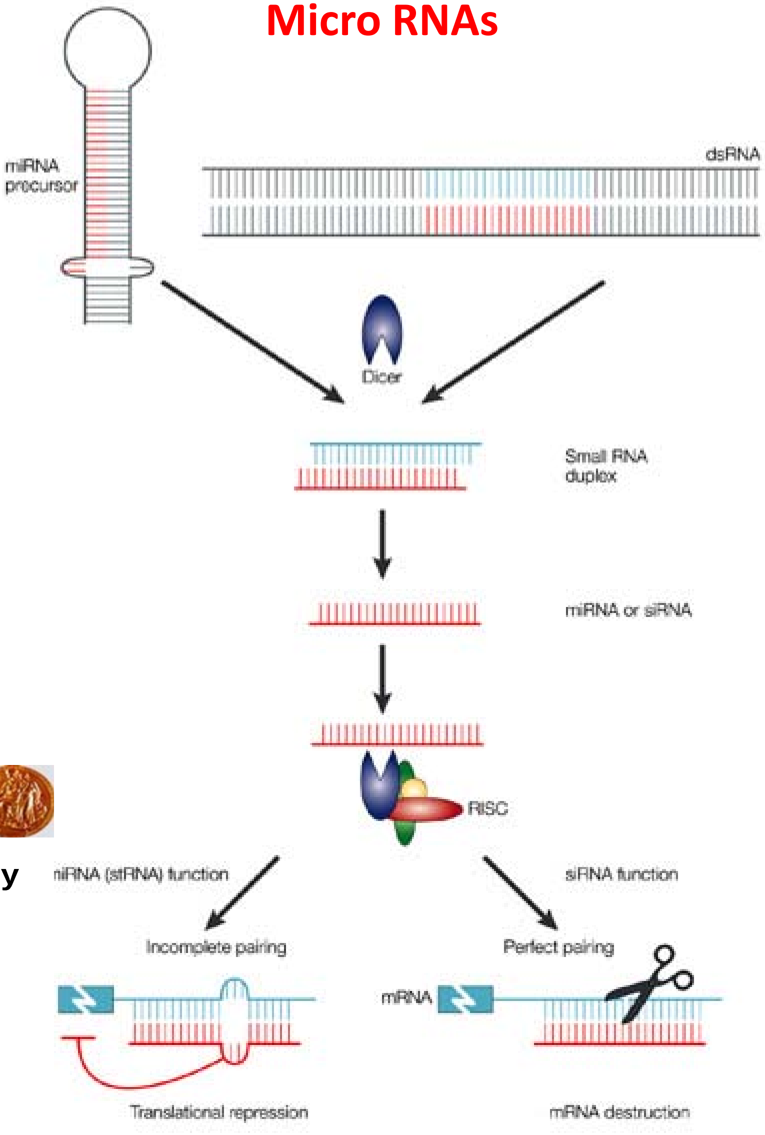
Francis Crick anticipated the existence of tRNA even before its discovery, which he called an adaptor, to help convert information in nucleic acid bases into amino acids



There must be at least one tRNA for each of the 20 amino acids and possibly one tRNA for each of the 61 codons.

Micro RNAs

MicroRNAs and short interfering RNAs might use the same RNA-processing complex to direct silencing. Evidence is accumulating that microRNAs (miRNAs) and the short interfering (siRNAs) involved in RNA interference are generated using the same pathway. Processing of the miRNA hairpin precursor or long double-stranded RNA (dsRNA) uses an enzyme called Dicer, and produces a single-stranded 21-23-nucleotide RNA. This small RNA attaches to an RNA interference silencing complex (RISC) and is directed to the messenger RNA (mRNA) of interest. Here, the mechanisms diverge. miRNA attaches itself to the target mRNA, but slight imperfections in the match between the miRNA and its recognition site mean that the miRNA forms a bulge, which blocks the mRNA from being transcribed into protein. siRNA, however, binds perfectly with its taget mRNA and tags the mRNA for destruction.



The Nobel Prize in Physiology or Medicine 2006

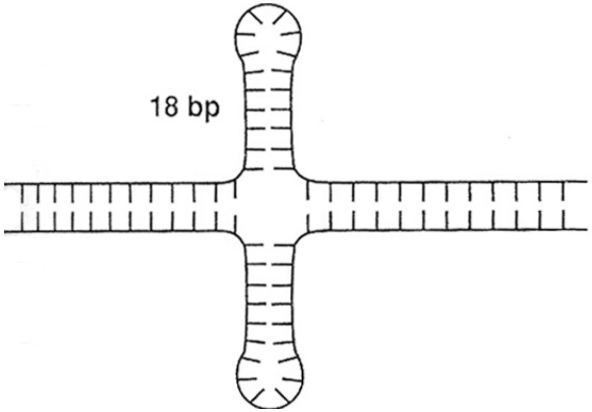
for their discovery of RNA interference – gene silencing by double-stranded RNA

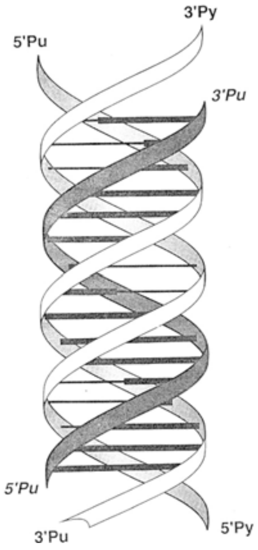
Andrew Z. Fire
Stanford U, USA
b. 1959, (USA)

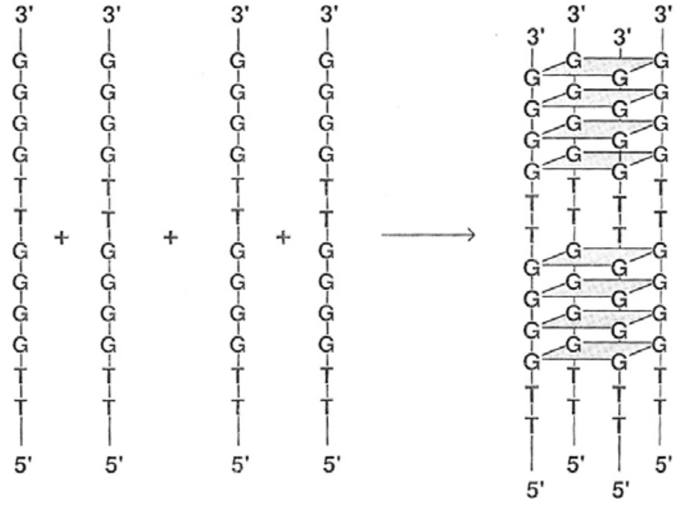
Craig C. Mello
U.Mass USA
b. 1960 (USA)

Regulation of protein synthesis through RNAs

Unusual structures of Nucleic Acids: Cruciforms, junctions, triplexes, quadruplexes





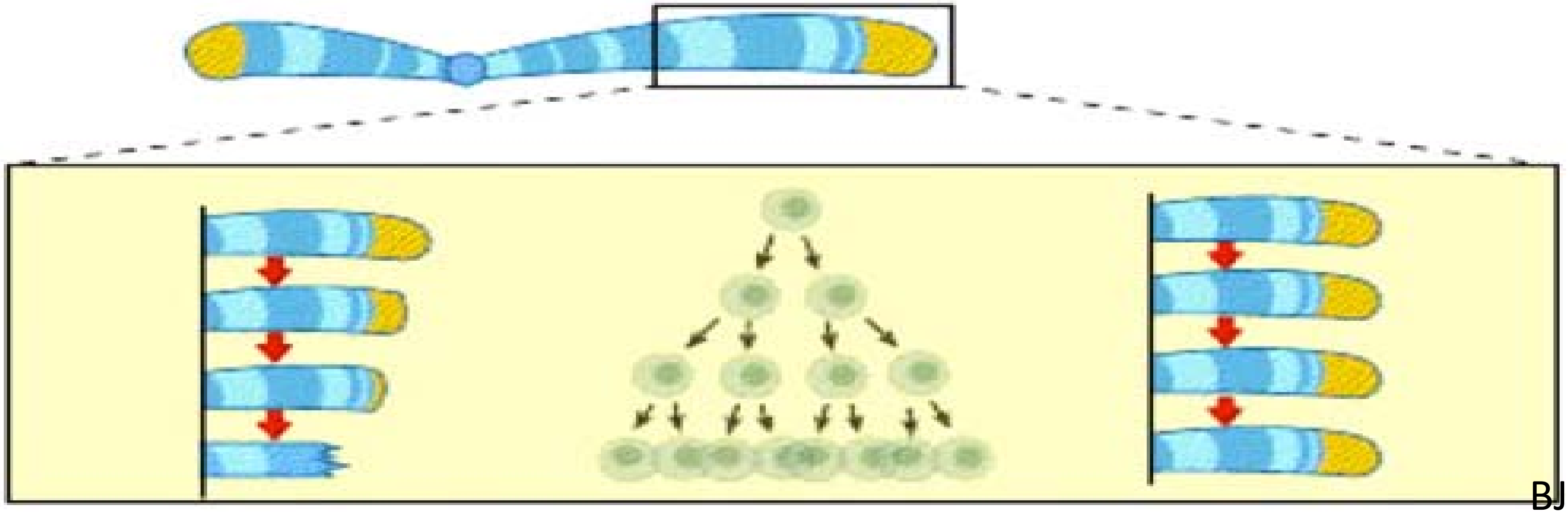


Inside the nucleus of a cell, our genes are arranged along twisted, double-stranded molecules of DNA called chromosomes.

At the ends of the chromosomes are stretches of DNA called **telomeres**, which protect our genetic data, make it possible for cells to divide, and hold some secrets to how we age and get cancer.

Telomeres have been compared with the plastic tips on shoelaces, because they keep chromosome ends from fraying and sticking to each other, which would destroy or scramble an organism's genetic information.

Yet, **each time a cell divides, the telomeres get shorter**. When they get too short, the cell can no longer divide; it becomes inactive or "senescent" or it dies. **This shortening process is associated with aging, cancer, and a higher risk of death.** So telomeres also have been compared with a bomb fuse.



Structure of a G-quadruplex.

Left: a G-tetrad.
Right: an intramolecular G-quadruplex

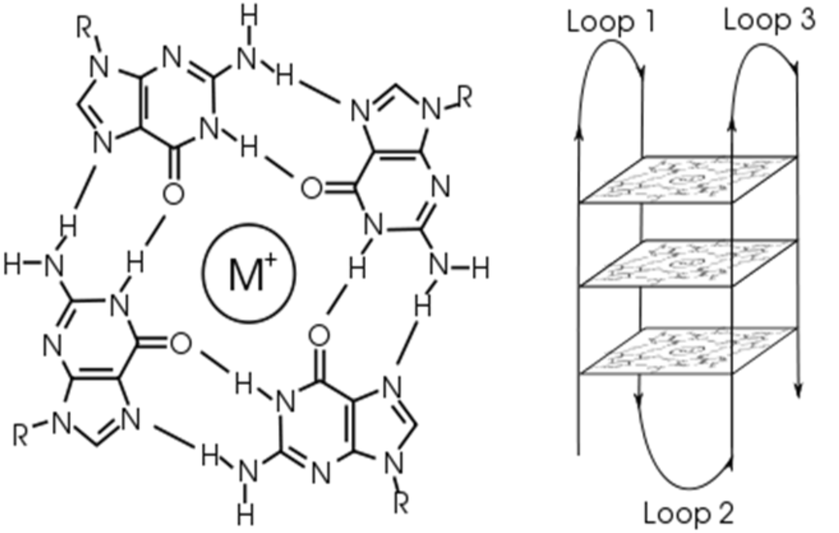
The Nobel Prize in Physiology or Medicine 2009

for their discovery of how chromosomes are protected by telomeres and the enzyme telomerase

Elizabeth H Blackburn
UCSF, USA
b. 1948, (USA)

Carol W Greider
Johns Hopkins, USA
b. 1961 (USA)

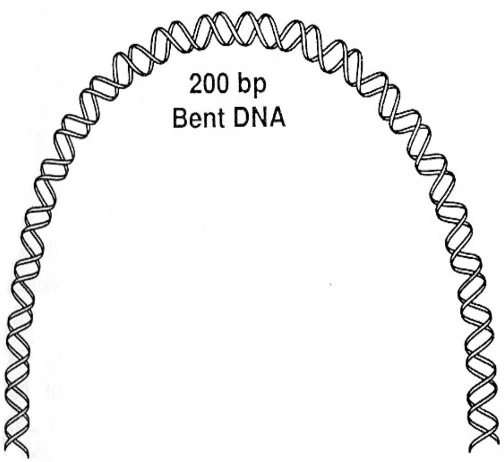
Jack W Szostak
Harvard Med School, USA
b. 1952 (UK)

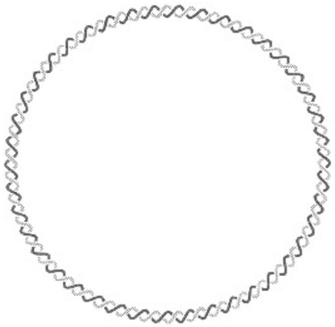


Telomeric quadruplexes

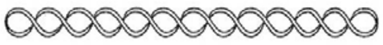
Telomeric repeats in a variety of organisms have been shown to form these structures *in vitro*, and they have also been shown to form *in vivo* in some cases. The human telomeric repeat (which is the same for all vertebrates) consists of many repeats of the sequence d(GGTTAG), and the quadruplexes formed by this structure have been well studied by NMR and X-ray crystal structure determination. The formation of these quadruplexes in telomeres has been shown to decrease the activity of the **enzyme telomerase**, which is responsible for maintaining length of telomeres and is involved in around 85% of all cancers. This is an active target of drug discovery.

Higher order structures of DNA: Packaging DNA into chromsomes

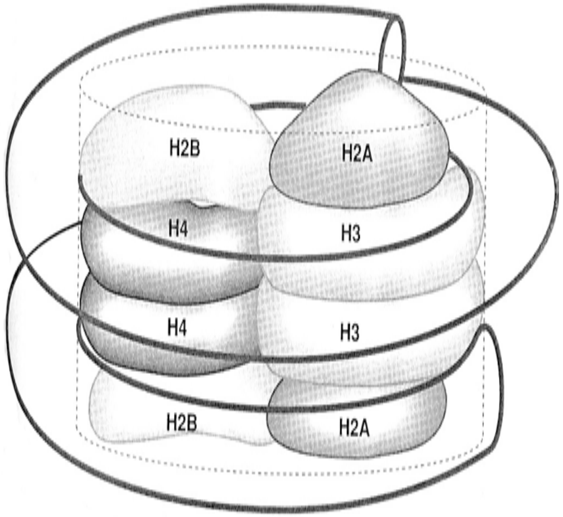




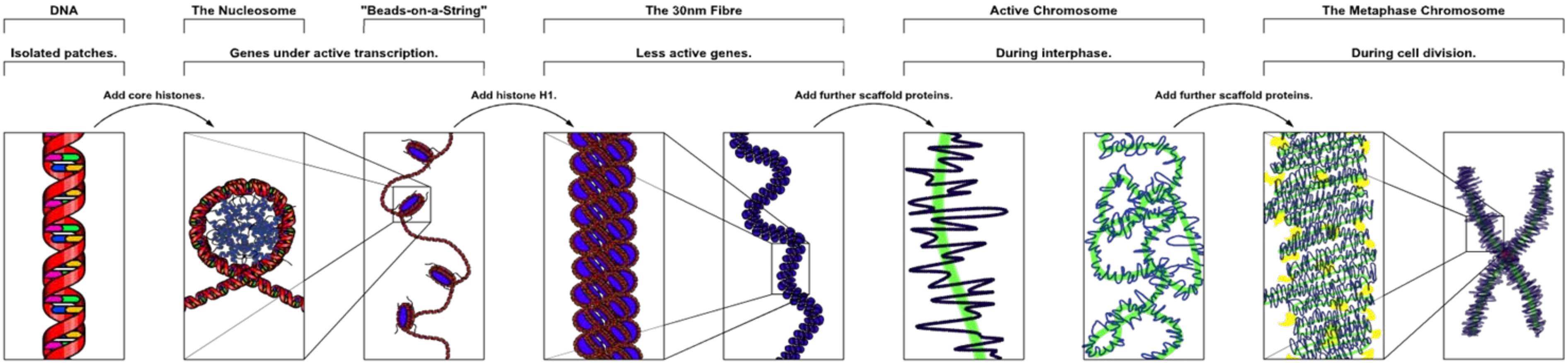
Circular DNA



Supercoiled DNA



A schematic of nucleosomes



DNA
The Nucleosome
"Beads-on-a-String"
The 30nm Fibre
Active Chromosome
The Metaphase Chromosome

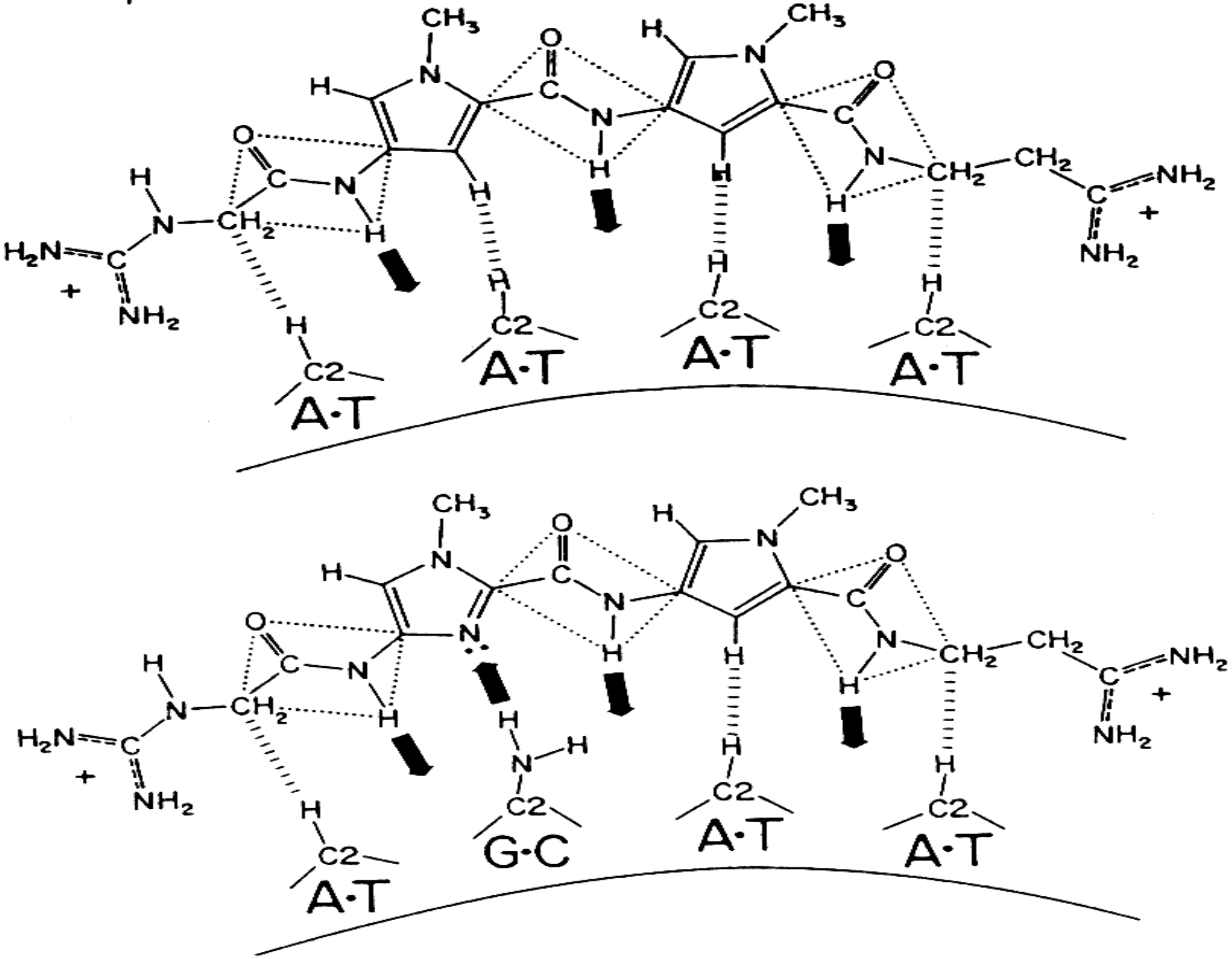
Isolated patches.
Genes under active transcription.
Less active genes.
During interphase.
During cell division.

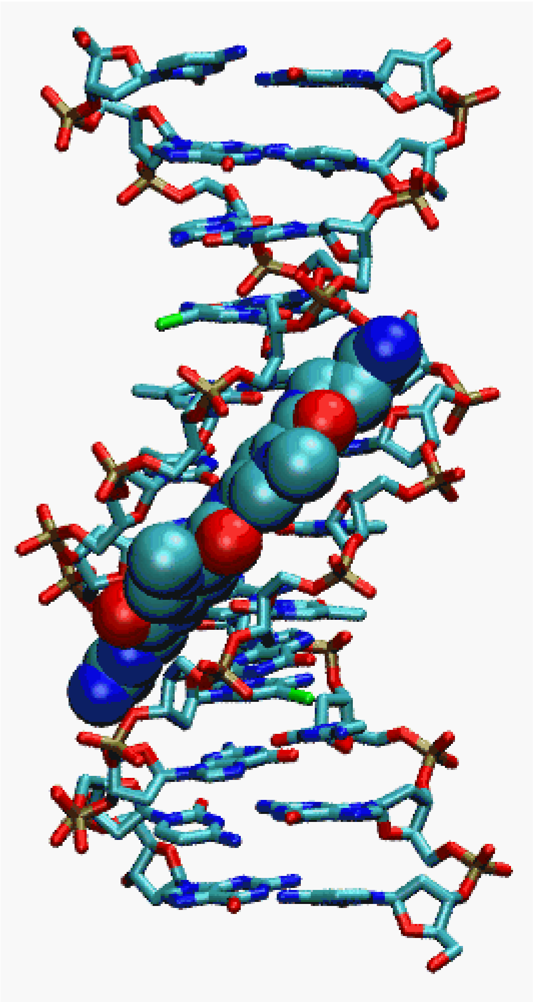
Add core histones.
Add histone H1.
Add further scaffold proteins.
Add further scaffold proteins.

Making Drugs against DNA: DNA-Drug: Minor groove interactions

Make a molecule that fits well in the grooves of DNA (Steric complementarity) and additionally makes hydrogen bonds with the base pairs (electrostatic complementarity) to cure cancer etc. or to control gene expression..

Netropsin



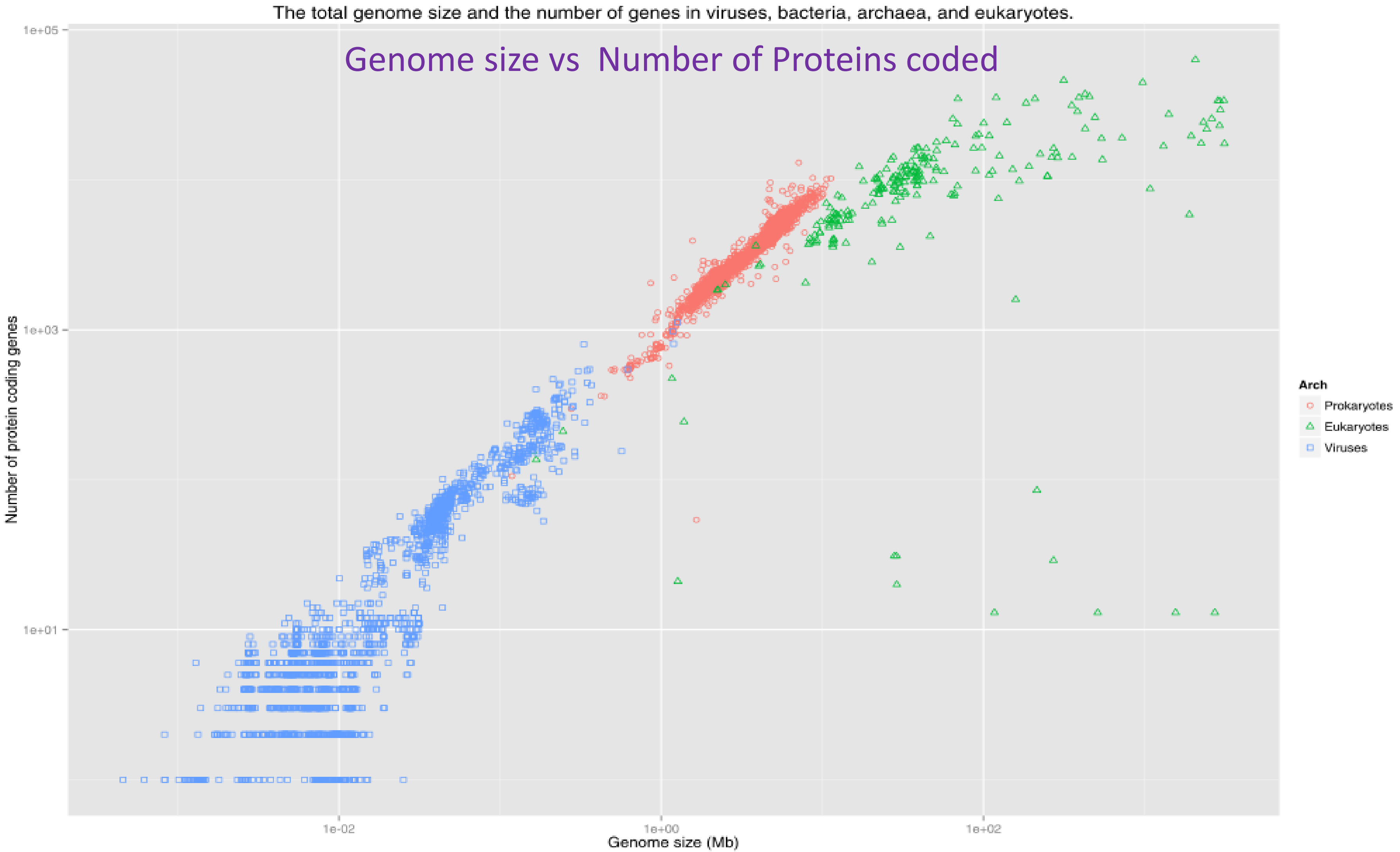


Genome sizes of some organisms

Organism	***Genome size*** ***((Mb)** (Mb=Mega base)*
Eschericia coli	4.6
Sacchromyces cerevisiae (Yeast)	15
M tuberculosis	4.4
H.Influenza	1.83
C. elegans (Nematode)	100
Drosophila melanogaster (Fruit fly)	120
Gallus gallus (Chicken)	120
Homo sapiens (humans)	3300
Mouse	**3000**
Rice	**430**
Wheat	13500

(source: www.wormlab.caltech.edu/briggsae/genomeSize.html)

Why do some organisms including plants have so much more DNA than us? The C-Value paradox!


The total genome size and the number of genes in viruses, bacteria, archaea, and eukaryotes.
1e+05
1e+03
1e+01
Number of protein coding genes
1e-02
1e+00
1e+02
Genome size (Mb)
Arch
Prokaryotes
Eukaryotes
Viruses

Genome size vs Number of Proteins coded

Log-log plot of the total number of annotated proteins in genomes submitted to GenBank as a function of genome size. Based on data from NCBI genome reports.

Genomics and Proteomics

The Nucleotide sequence and the corresponding amino acid sequence of Human Insulin (which participates in metabolism of fat and proteins).

atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgac

M A L W M R L L P L L A L L A L W G P D

ccagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctac

P A A A F V N Q H L C G S H L V E A L Y

ctagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggac

L V C G E R G F F Y T P K T R R E A E D

ctgcaggtggggcaggtggagctgggcgggggccctggtgcaggcagcctgcagcccttg

L Q V G Q V E L G G G P G A G S L Q P L

gccctggaggggtccctgcagaagcgtggcattgtggaacaatgctgtaccagcatctgc

A L E G S L Q K R G I V E Q C C T S I C

tccctctaccagctggagaactactgcaactag

S L Y Q L E N Y C N -

A base ‘A’ is inserted in the above nucleotide sequence as shown below. The whole protein sequence changes.

A mutation causing frame shift

atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgac

M A L W M R L L P L L A L L A L W G P D

ccagccgcagAcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctcta

P A A D L C E P T P V R L T P G G S S L

cctagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagagga

P S V R G T R L L L H T Q D P P G G R G

cctgcaggtggggcaggtggagctgggcgggggccctggtgcaggcagcctgcagccctt

P A G G A G G A G R G P W C R Q P A A L

ggccctggaggggtccctgcagaagcgtggcattgtggaacaatgctgtaccagcatctg

G P G G V P A E A W H C G T M L Y Q H L

ctccctctaccagctggagaactactgcaactag

L P L P A G E L L Q L…….

ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/CHR_X/

```
DEFINITION  Homo sapiens chromosome X
ORIGIN
      1  ccaggatggt ccttctcctg aaggttaatc cataggcaga tgaatcggat attgattcct
     61  gttcttggaa taatctagag gatctttaga atccattggg attcataatc acagctatgc
    121 cgatgccatc atcaccggct tagccctttc tgaaaacaca gtcatcatct acccccattg
    181 gaatcacgat gcaaaaaacc tgtcccaaag cggtggtttc ctatgtgatt cttgcatcca
    241 ggacaaatga cagtcagcag agaggcgccc tgttccatct tttggtttga tccagttaaa
    301 ggcacacacg tgagcaccca acgtttgcca actcagcact gggcagagcc tggcctctga
    361 ggaaattggc atcttcgtaa tcaatatatt attatgtttt attgaaatgt aagtcattgc…..
```

```
DEFINITION  Homo sapiens chromosome Y
ORIGIN
      1   ggtttcacca agttggccag gctggtctcg aactcctgac ctcaggtgat ctgtccacct
     61  cggtgtccca aagtgctggg attacaggtg tgaaccacca cacccagcct catgtaatac
    121 ttaaaaatga actacaggtg gattacaaac ctgaatatca aagaaaactt ttttttttga
    181 aaaatagagg gaaatgtctt ataacctcag agttaggagg tttttcttag atacaataca
    241 aaaagcataa ccacgcccat agtcccagct actcaggagg ctgaggcata agaatcactt
    301 gagctcgaga ggtggaggtt gcagtgagcc gagatcctgc cattgcactc cagctgaggc
    361 tacagagtga gagtataaaa aaaaaaaaaa aagcataacc tttaaaaatg ggttagccta…..
```

Question: Can you infer the meaning of the sequences on the left just by reading without looking at definitions or using some software?

Answer: No body can today…(Correction.. Rajni can do that! Let us first see if, as the hoarding near the Library says "Lipton can do that", before we go to Rajni sir!!)

What is the language of DNA that proteins understand and we don't.

Specific genetic disorders

Genetic Disorder	***Reason***
Huntington’s Disease	Excessive repeats of a three-base sequence, "CAG" on chromosome
Parkinson’s Disease	Variations in genes on chromosomes 4,6.
Sickle Cell	Disease Mutation in hemoglobin-b gene on chromosome 11
Tay-Sachs Disease	Controlled by a pair of genes on chromosome 15
Cystic Fibrosis	Mutations in a single (CFTR) gene
Breast Cancer	Mutation on genes found on chromosomes 13 & 17
Leukemia	Exchange of genetic material between the long arms of chromosome 6 & 22.
Colon cancer	Proteins MSH2, MSH6 on chromosome 2 & MLH1 on chromosome 3 are mutated.
Asthma	Disfunctioning of genes on chromosome 5, 6, 11, 14&12.
Rett Syndrome	Disfunctioning of a gene on the X chromosome.
Brukitt lymphoma	Translocations on chromosome 8
Alzheimer disease	Mutations on four genes located on chromosome 1, 14, 19 & 21.
Werner Syndrome	Mutations on genes located on chromosome 8.
Angelman Syndrome	Deletion of a segment on maternally derived chromosome 15.

Several reasons for making genome cards for every individual / organism except that we need to work out the science for making sense out of the information on the cards.

Eukaryotic Gene Prediction Accuracies

Today's Computational Challenge!

Genome assembly and genome annotation (understanding what each base pair does after correctly assembling the genome)

Intra- and inter-species gene prediction accuracy Intra-species performance figures derived from 5-fold cross-validation are along the diagonal in bold. (Korf, 2004)

Genomic DNA									
		At		**Ce**		**Dm**		**Os**	
Parameters	**Measure**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**
At	Nuc	**97.1**	**95.2**	78.7	91.3	77.7	68.0	90.7	71.8
	Exon	**82.9**	**81.2**	44.3	52.8	38.6	24.0	57.1	42.3
	Gene	**54.3**	**46.8**	20.9	11.3	18.8	5.7	20.5	9.7
Ce	Nuc	83.5	91.5	**97.6**	**94.2**	81.3	73.6	79.7	74.5
	Exon	40.5	49.9	**85.5**	**79.3**	42.2	29.8	27.5	26.0
	Gene	25.7	18.1	**46.0**	**32.5**	21.9	8.8	13.9	7.3
Dm	Nuc	30.0	95.3	45.9	95.0	**94.3**	**86.5**	78.4	89.8
	Exon	16.5	41.3	29.9	47.2	**78.6**	**67.2**	50.0	58.4
	Gene	3.2	4.3	7.8	6.9	**50.8**	**37.5**	36.3	28.9
Os	Nuc	39.3	96.3	24.9	95.5	79.8	88.7	**86.2**	**94.0**
	Exon	30.7	47.6	11.1	36.6	47.4	44.4	**70.2**	**72.4**
	Gene	5.1	6.1	5.3	7.8	27.2	17.2	**51.2**	**37.0**

Most methods today are based on sophisticated mathematical and statistical techniques but rely heavily on sparse experimental data for training the models to do predictions. These methods are typically organism specific .

There is no universally applicable model!

First step in (mRNA) gene prediction is to find ORFs on the genome.

ORF (Open reading frame) : A potential protein coding region..It starts with ATG, the start codon in most cases and, stops at one of the three stop codons viz. TAA, TAG, TGA. For an ORF to become a protein, it also needs control regions upstream (5’ side), which can initiate transcription. Amino acid sequence is inferred from ORF via genetic code.

Question.

How many ORFs are there in the following DNA sequence?

5’ TCACCTAATGCGTGCGCAATGCATGACTTAATATAA 3’
3’ AGTGGATTACGCACGCGTTACGTACTGAATTATATT 5’

Answer. Four

Frame 1. one. (ATG CAT GAC TTA ATA) Amino acid sequence (M H D L I)

Frame 2. Two: (i) (ATG CGT GCG CAA TGC ATG ACT) Amino acid sequence (M R A Q C M T) &
(ii) (ATG ACT): Amino acid sequence (M T)..The longer ORF embeds the shorter one.

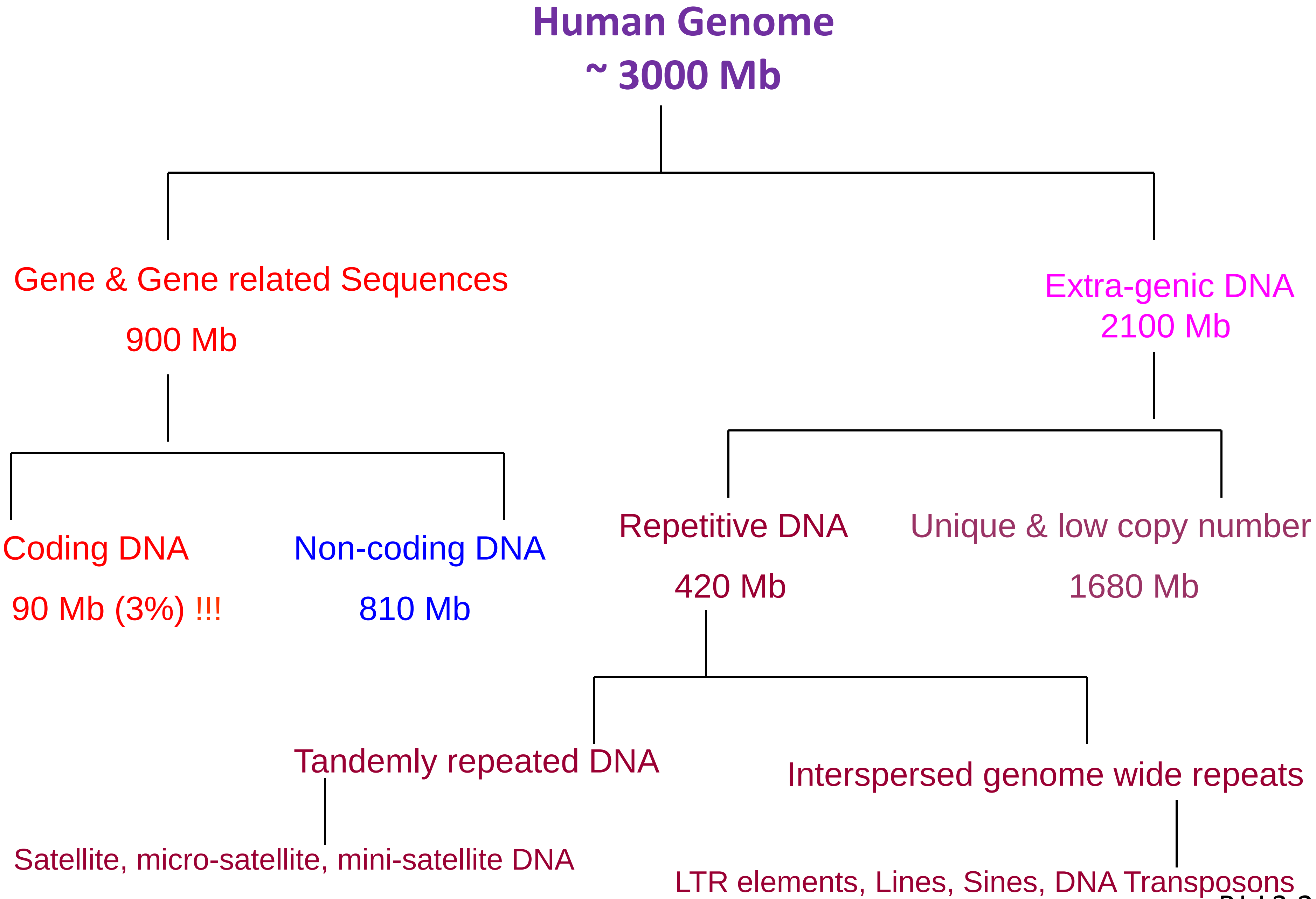
Frame 3: None

Frame 4: One (ATG CAT TGC GCA CGC ATT AGG) Amino acid sequence (M H C A R I R)

Frame 5: None

Frame 6 : None

Let us read the book of Human Genome soon like a Harry Potter novel !



Human Genome

~ 3000 Mb

Gene & Gene related Sequences

900 Mb

Coding DNA

90 Mb (3%) !!!

Non-coding DNA

810 Mb

Extra-genic DNA

2100 Mb

Repetitive DNA

420 Mb

Tandemly repeated DNA

Satellite, micro-satellite, mini-satellite DNA

Interspersed genome wide repeats

LTR elements, Lines, Sines, DNA Transposons

Unique & low copy number

1680 Mb

Assignment 7 (2020). Develop a universally applicable model for genome annotation so that one can read the books of genomes of all organisms like novels. (Human genome: Novel-1; Buffalo genome: Novel -2; Rice genome: Novel-3……..)