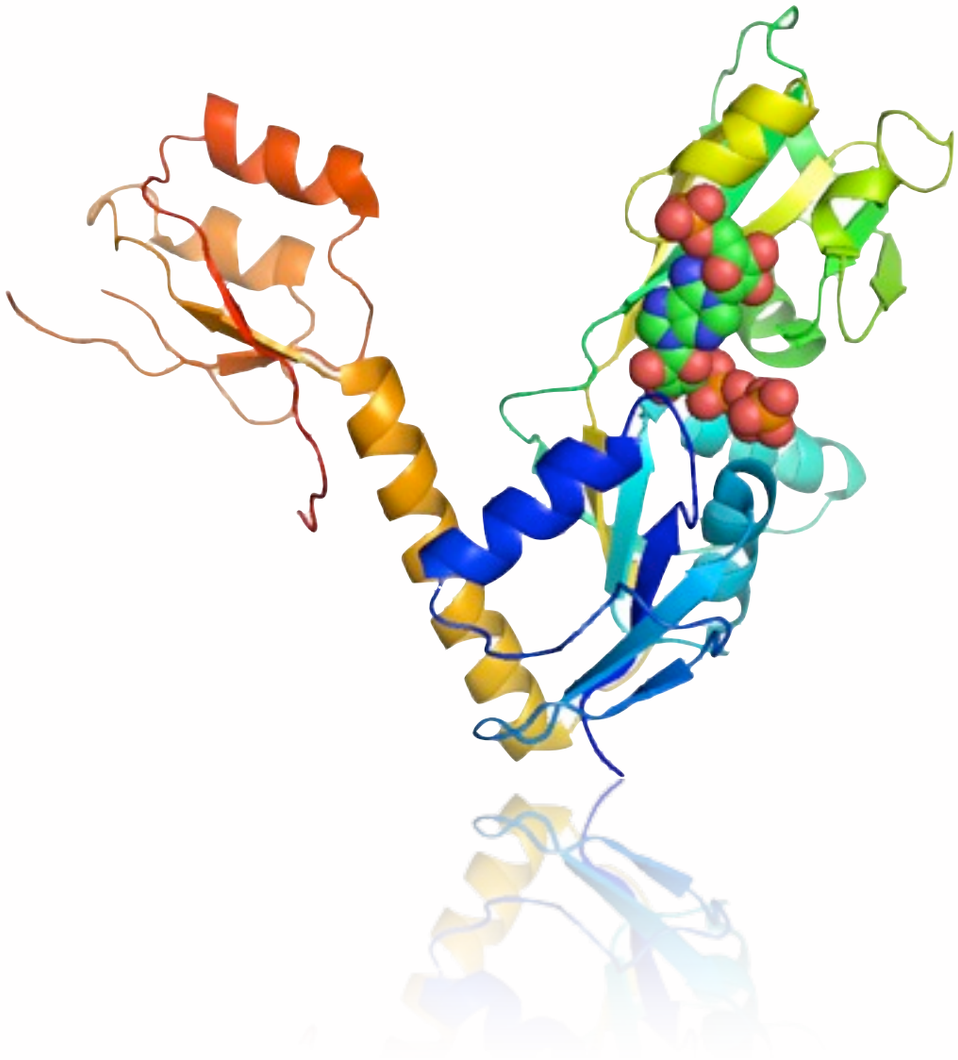
Structural Bioinformatics

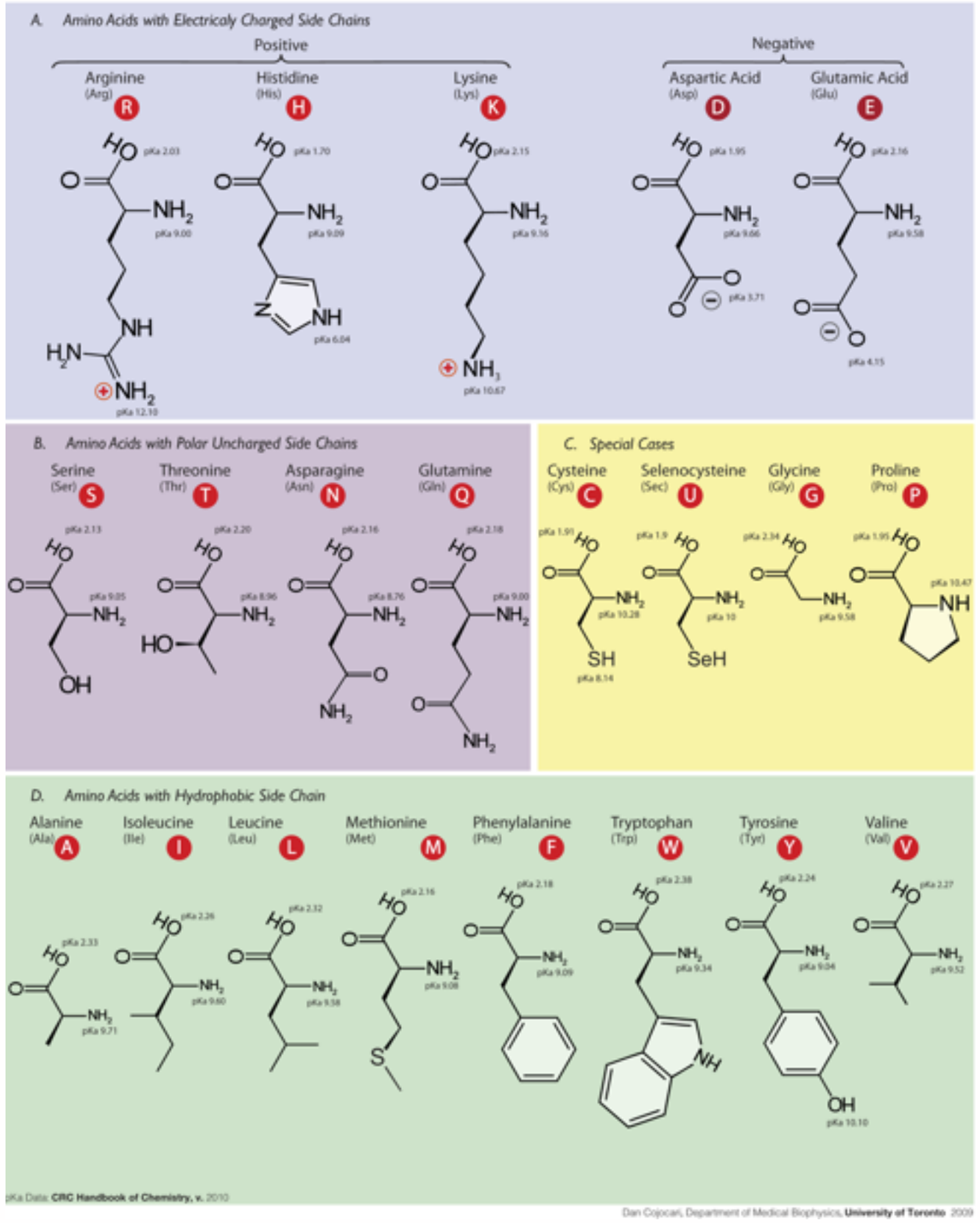
Davide Baù
Staff Scientist
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

dbau@pcb.ub.cat

Proteins



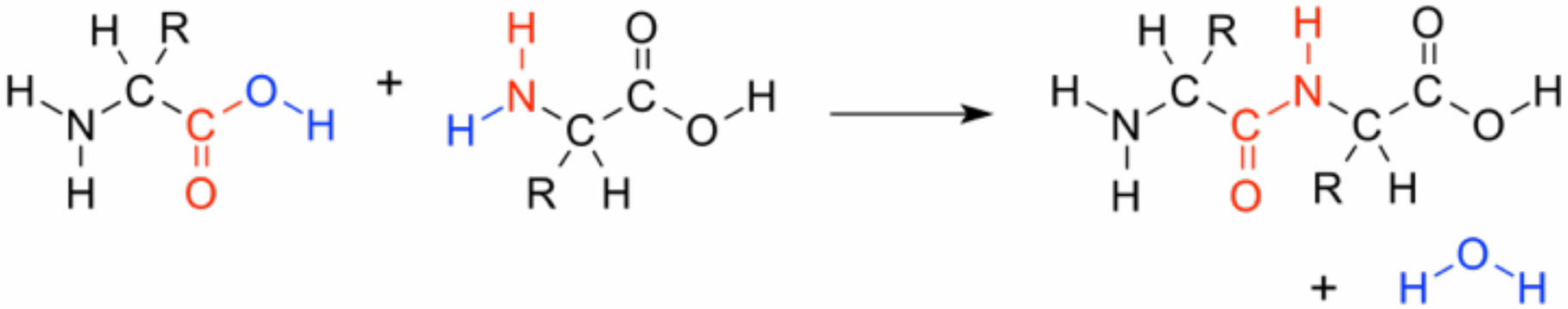
Amino Acids


A. Amino Acids with Electricaly Charged Side Chains
Positive
Arginine (Arg) R
Histidine (His) H
Lysine (Lys) K
Negative
Aspartic Acid (Asp) D
Glutamic Acid (Glu) E
B. Amino Acids with Polar Uncharged Side Chains
Serine (Ser) S
Threonine (Thr) T
Asparagine (Asn) N
Glutamine (Gln) Q
C. Special Cases
Cysteine (Cys) C
Selenocysteine (Sec) U
Glycine (Gly) G
Proline (Pro) P
D. Amino Acids with Hydrophobic Side Chain
Alanine (Ala) A
Isoleucine (Ile) I
Leucine (Leu) L
Methionine (Met) M
Phenylalanine (Phe) F
Tryptophan (Trp) W
Tyrosine (Tyr) Y
Valine (Val) V
pKa Data: CRC Handbook of Chemistry, v. 2010
Dan Cojocari, Department of Medical Biophysics, University of Toronto 2009

The peptide bond

Properties

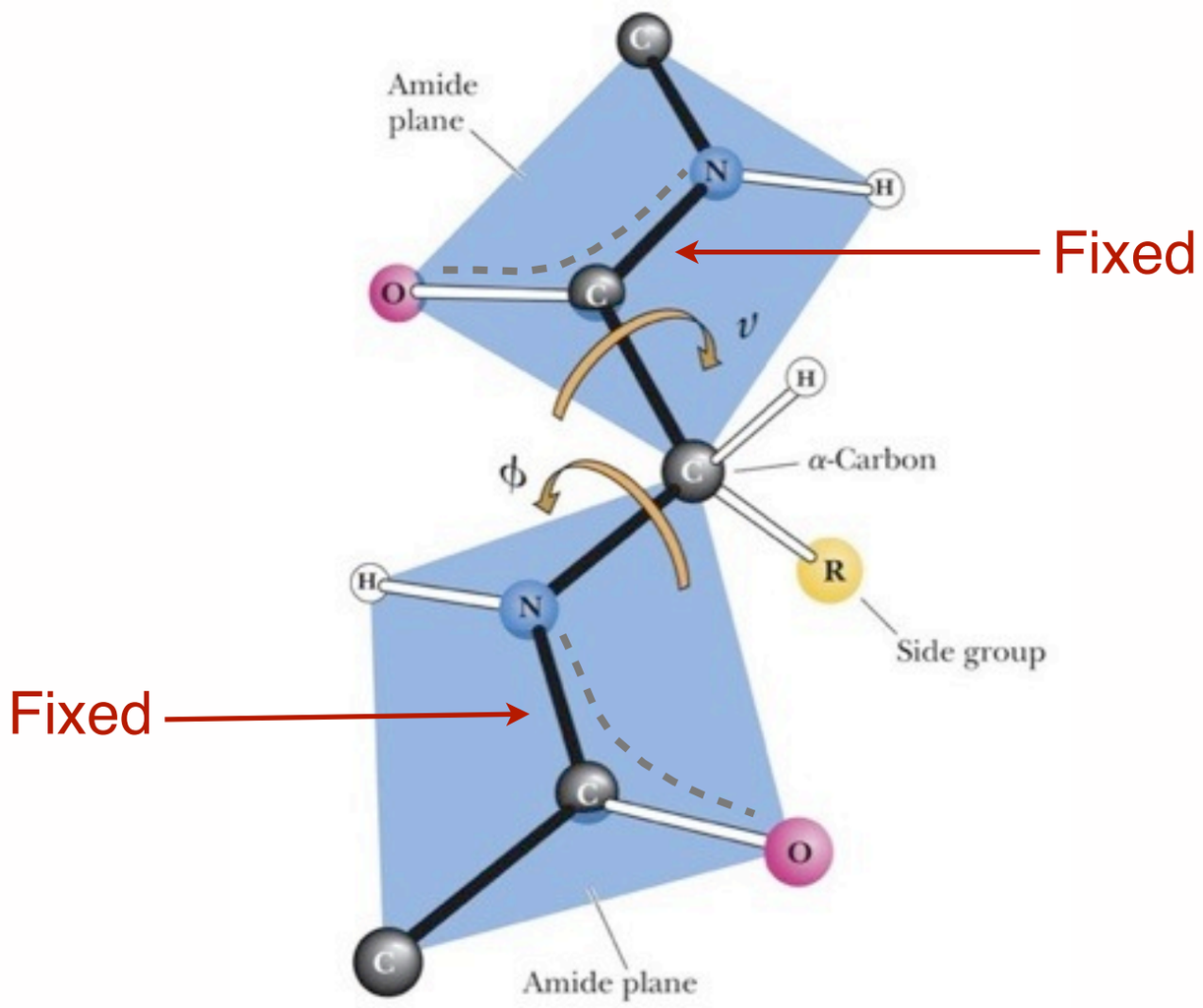
A peptide bond is a **covalent bond** formed between two molecules when the carboxyl group of one molecule reacts with the amino group of the other molecule, causing the release of a molecule of water (H_2O).



Polypeptides and proteins are chains of amino acids held together by peptide bonds.

The peptide bond

The peptide bond is planar

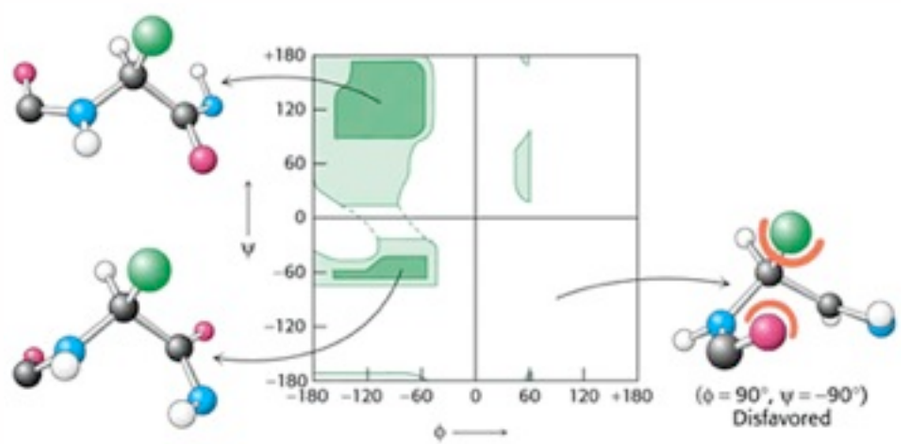


Only 2 bonds can freely rotate: C_α–N and C_α-C(O)

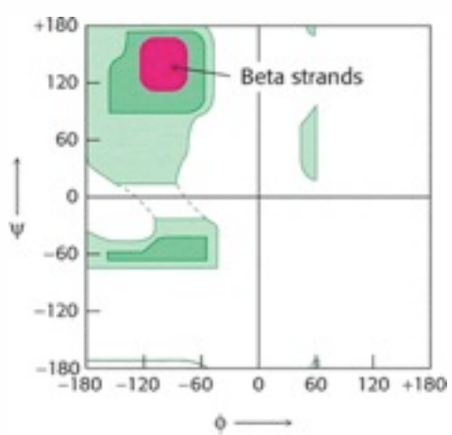
Adapted from http://oregonstate.edu

Ramachandran plots

Protein structures Φ and Ψ angles fall within allowed regions (displayed in green and red).



Secondary structure elements are defined by specific pairs of Φ and Ψ angles:



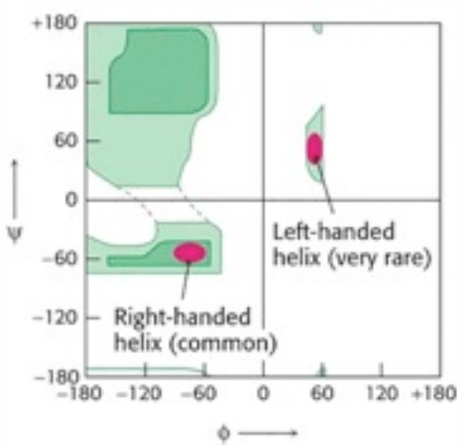


Image credits: http://www.imb-jena.de/~rake

Take home message

Proteins
Chains of amino acids held together by the peptide bond

Configuration
Defined by limited pairs of **Φ** and **Ψ** angles

Role
Fundamental constituents of the cell

Summary

Protein structural levels

Primary Secondary Tertiary Quaternary

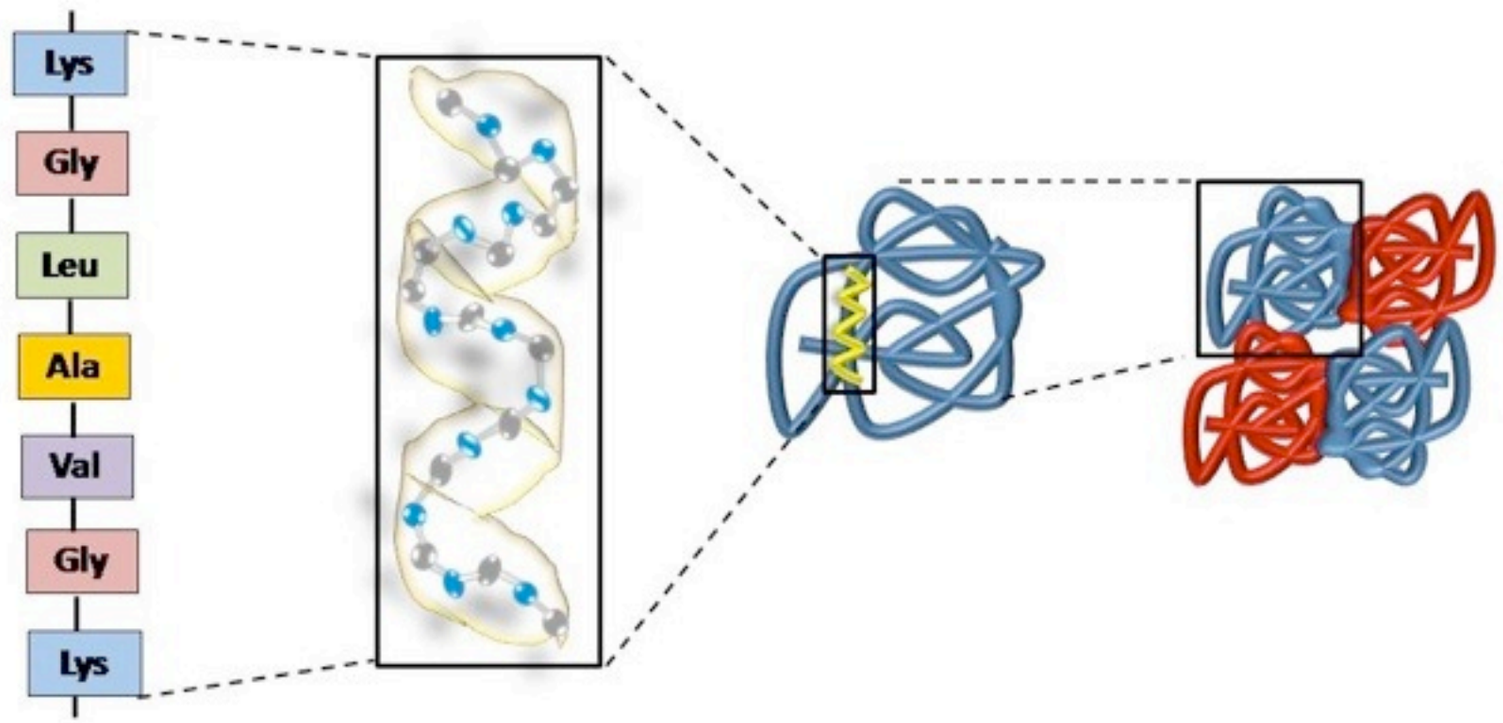


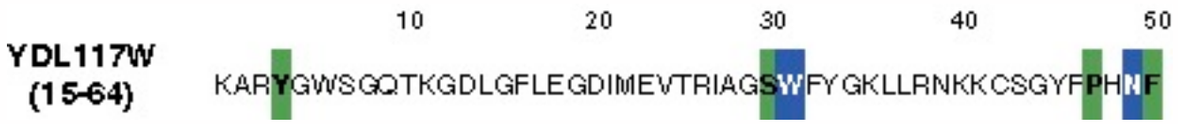
Image credits: http://iitb.vlab.co.in/

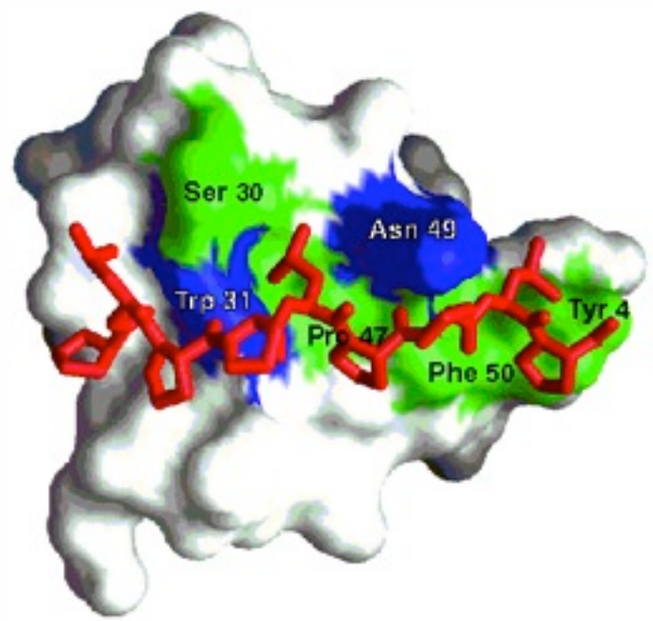
Protein structure relevance

The **biochemical function** (activity) of a protein is defined by its interactions with other molecules.

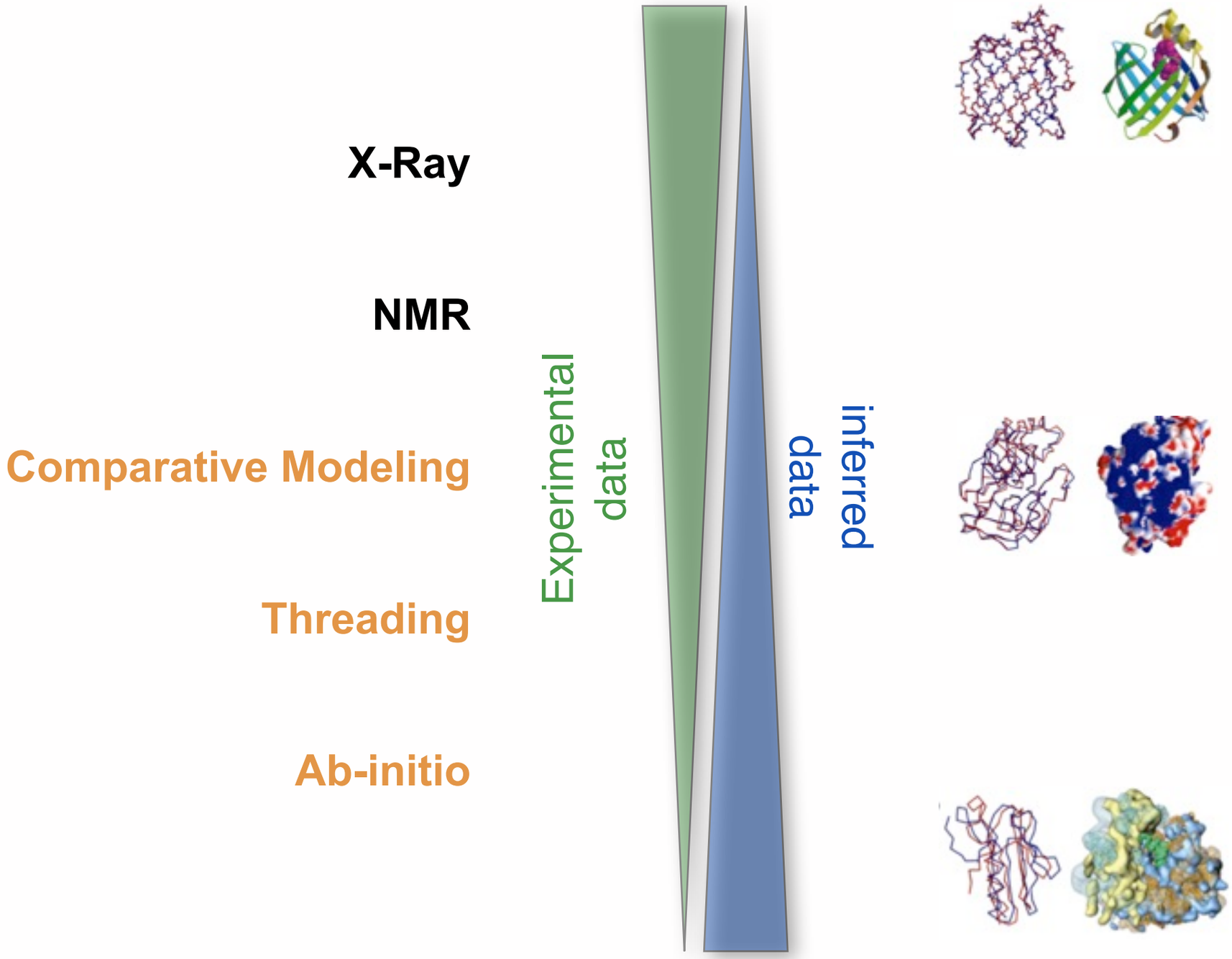
The biological function is in large part a consequence of these **interactions**.

The 3D structure is **more informative** than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.

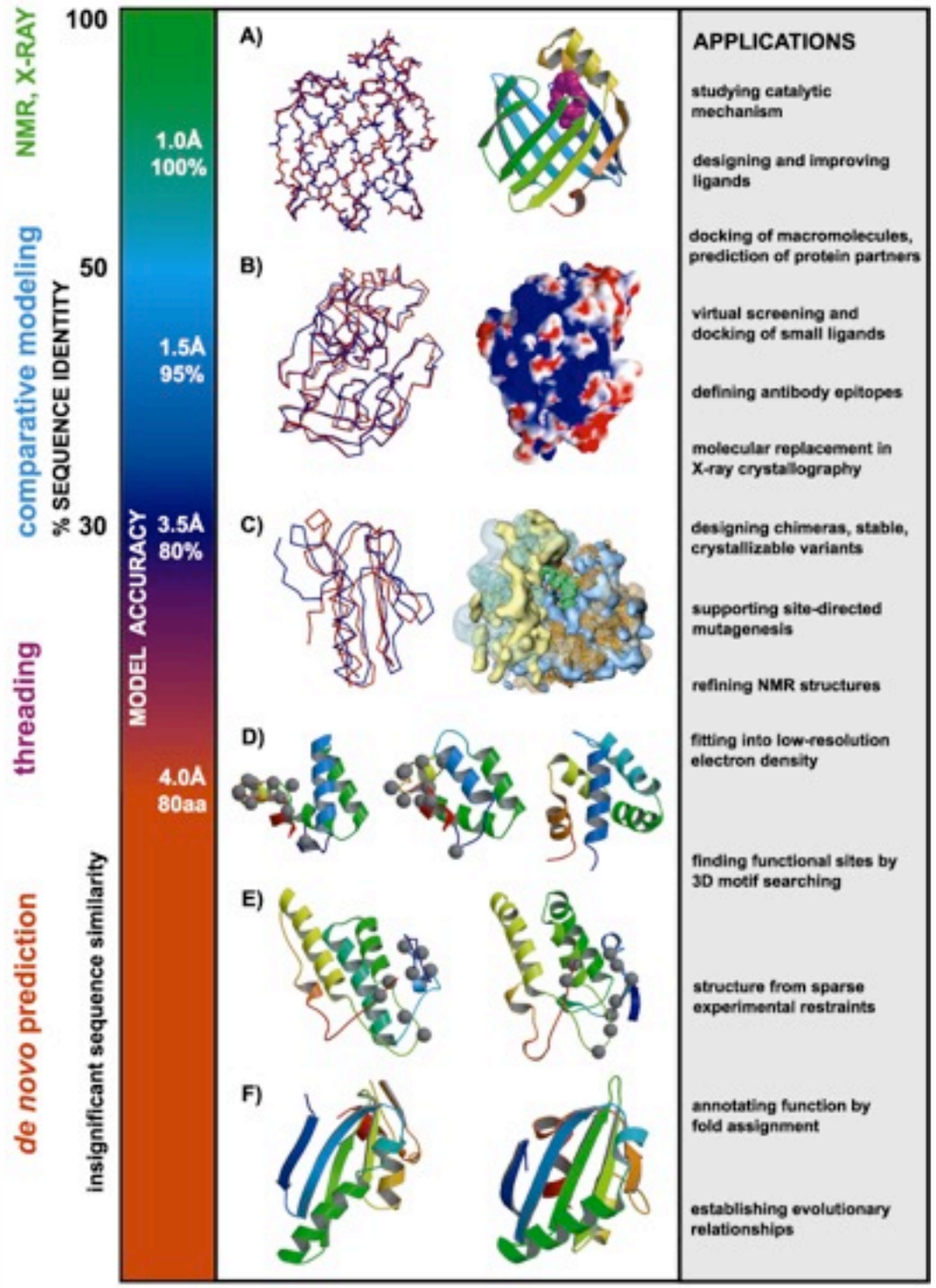




Protein prediction vs protein determination



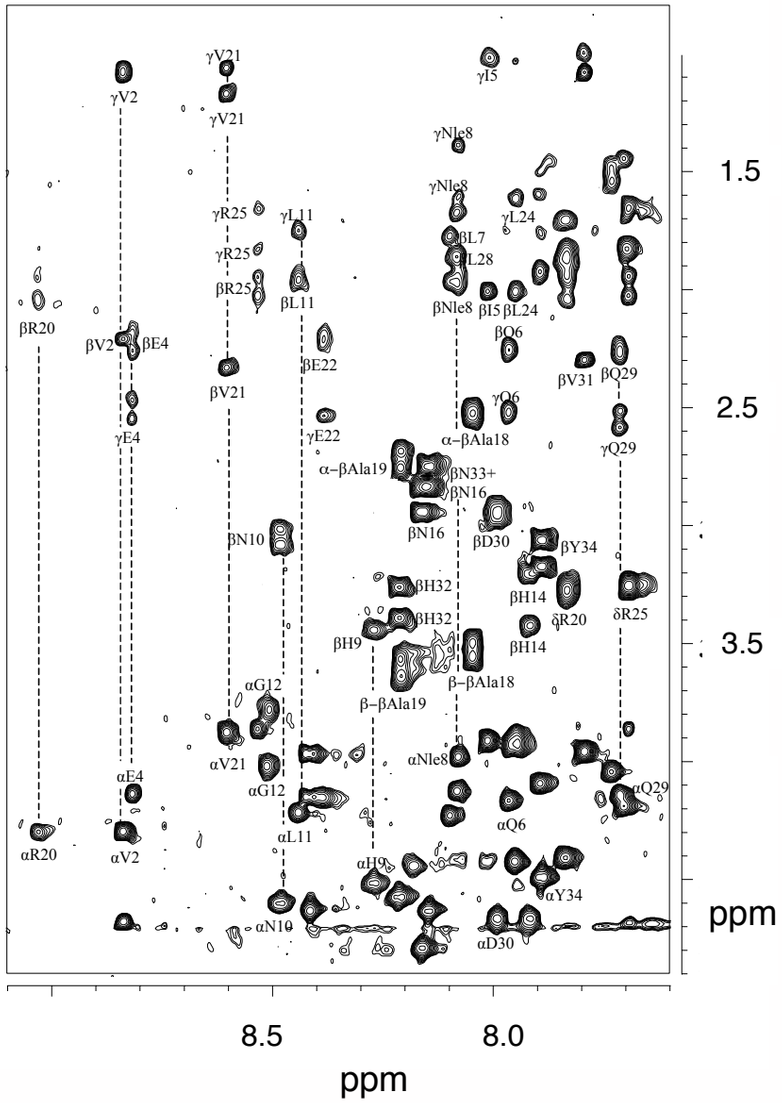
Utility of protein structure models, despite errors



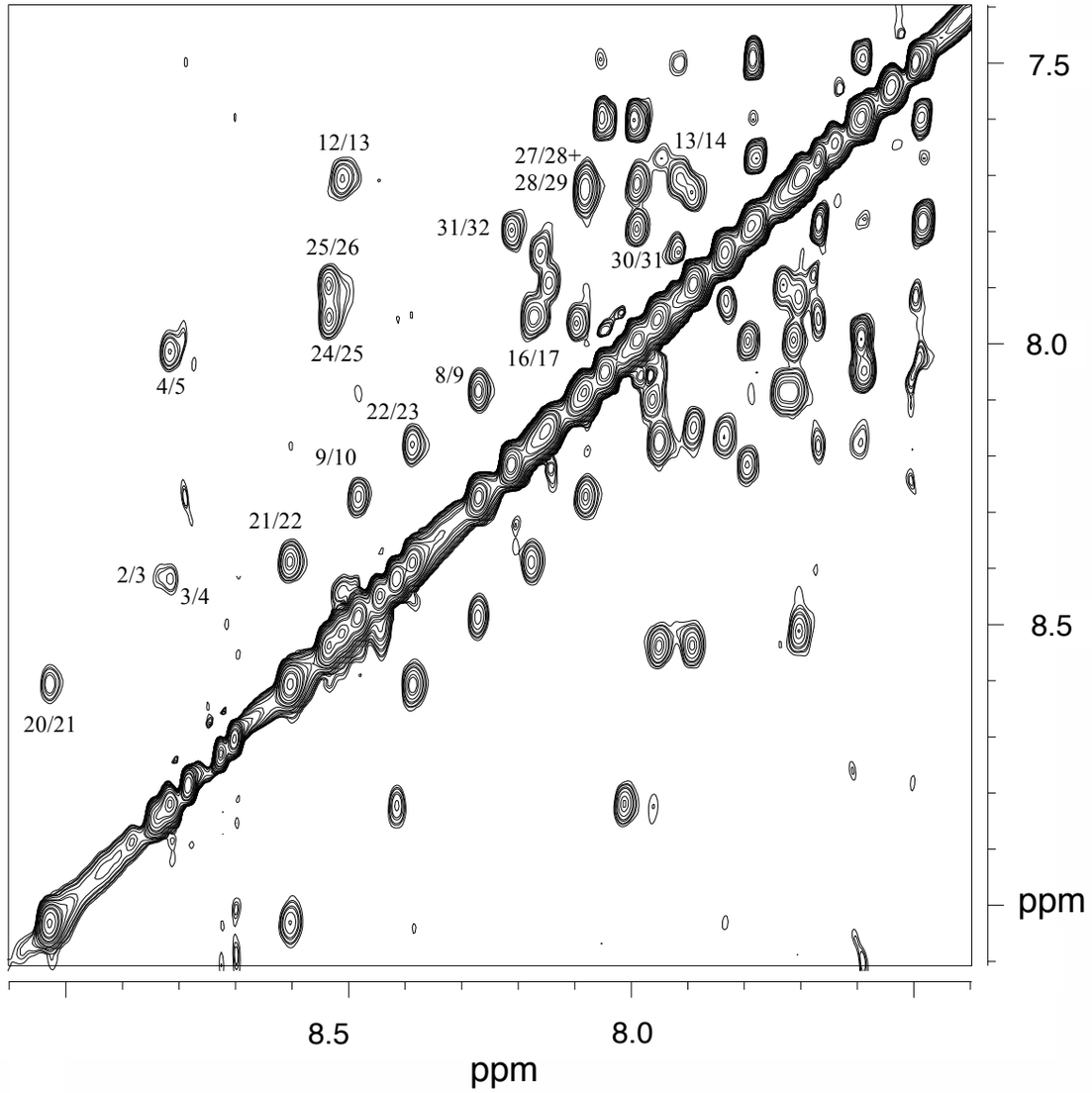
D. Baker & A. Sali. Science 294, 93, 2001.

NMR spectroscopy

Nuclear magnetic resonance



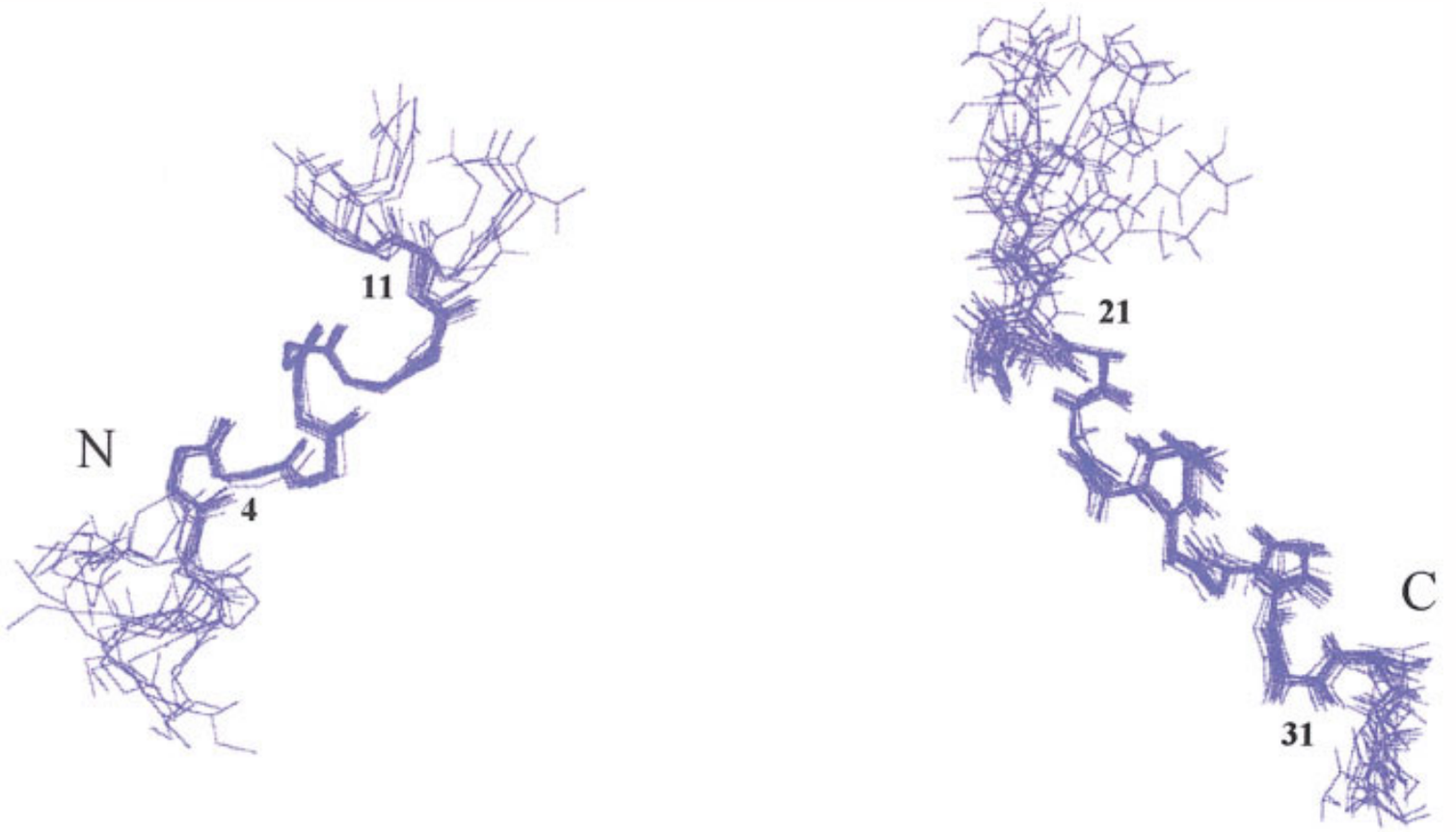
TOCSY



NOESY

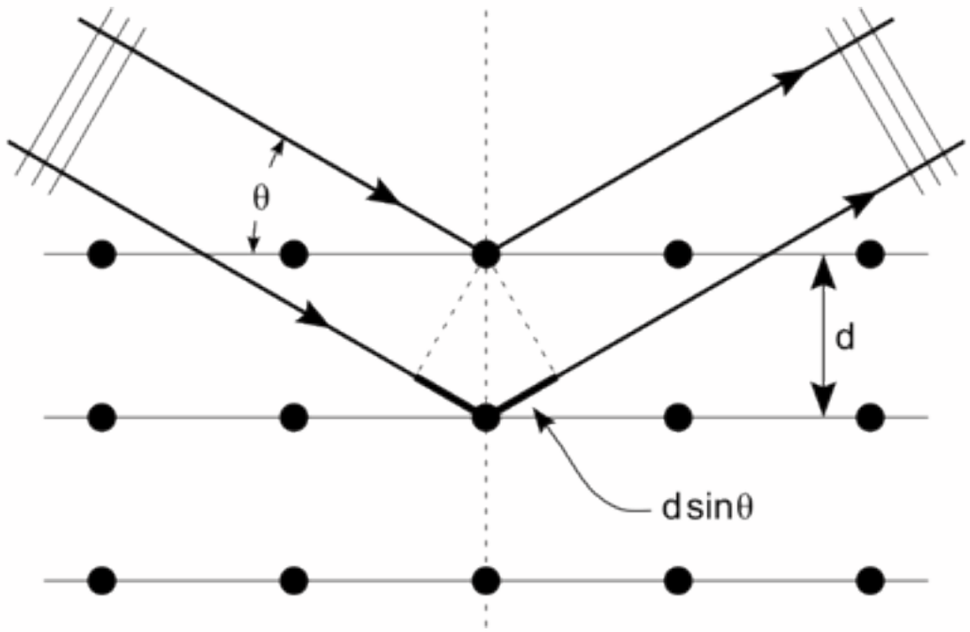
NMR spectroscopy

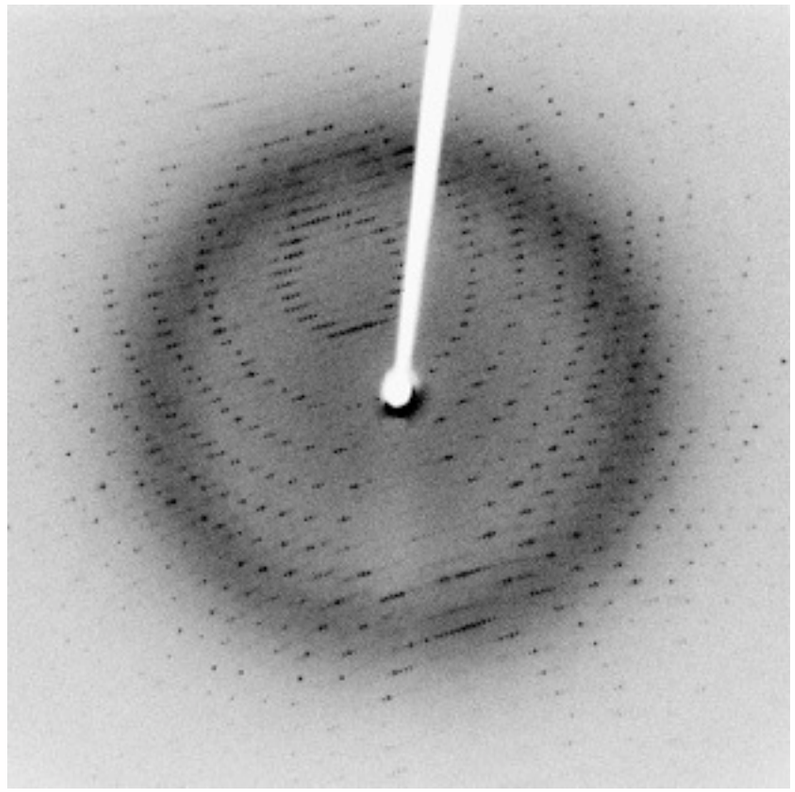
Nuclear magnetic resonance



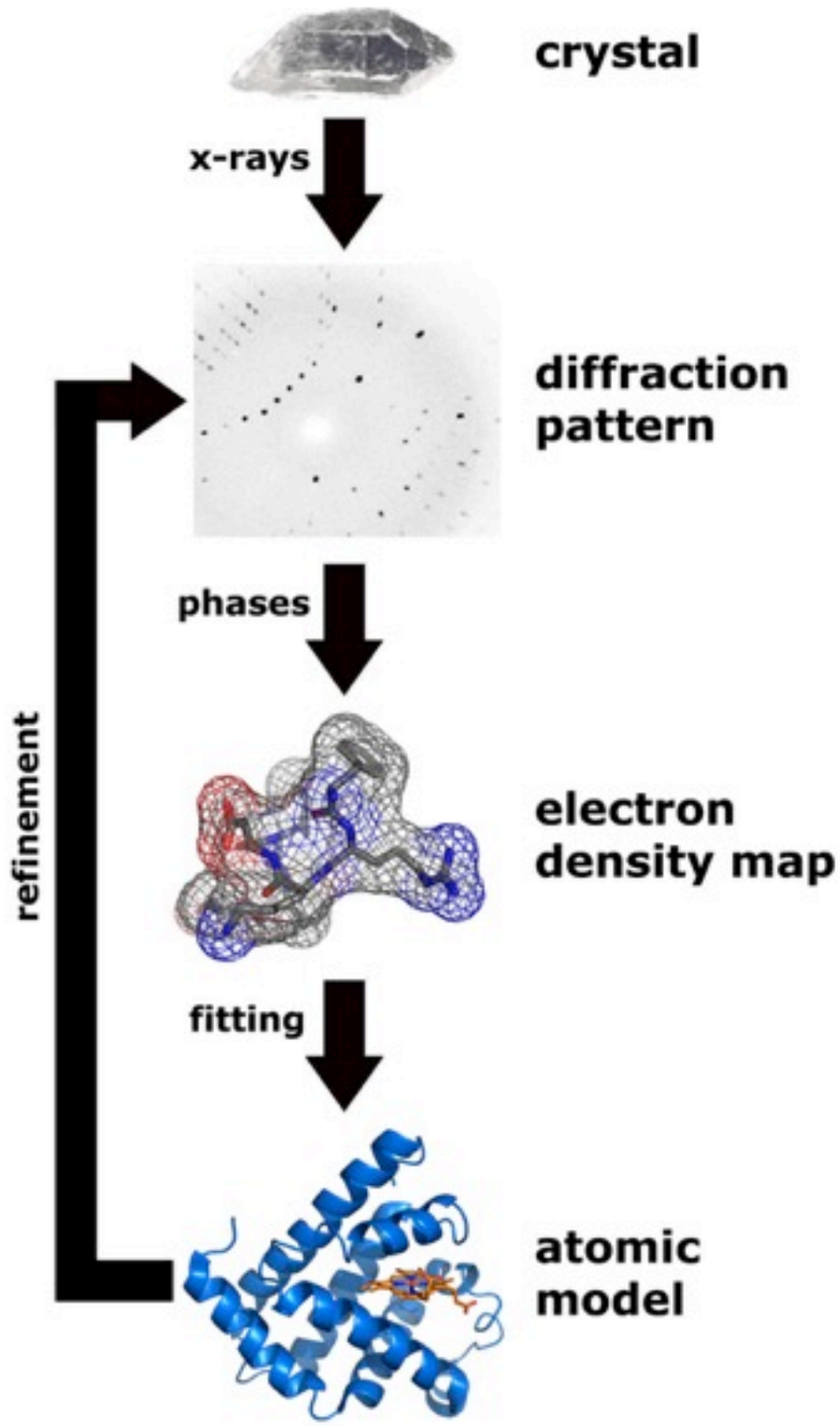
Superimposition of the ensemble of lowest energy structures of a peptide.

X-RAY crystallography





X-RAY crystallography



Take home message

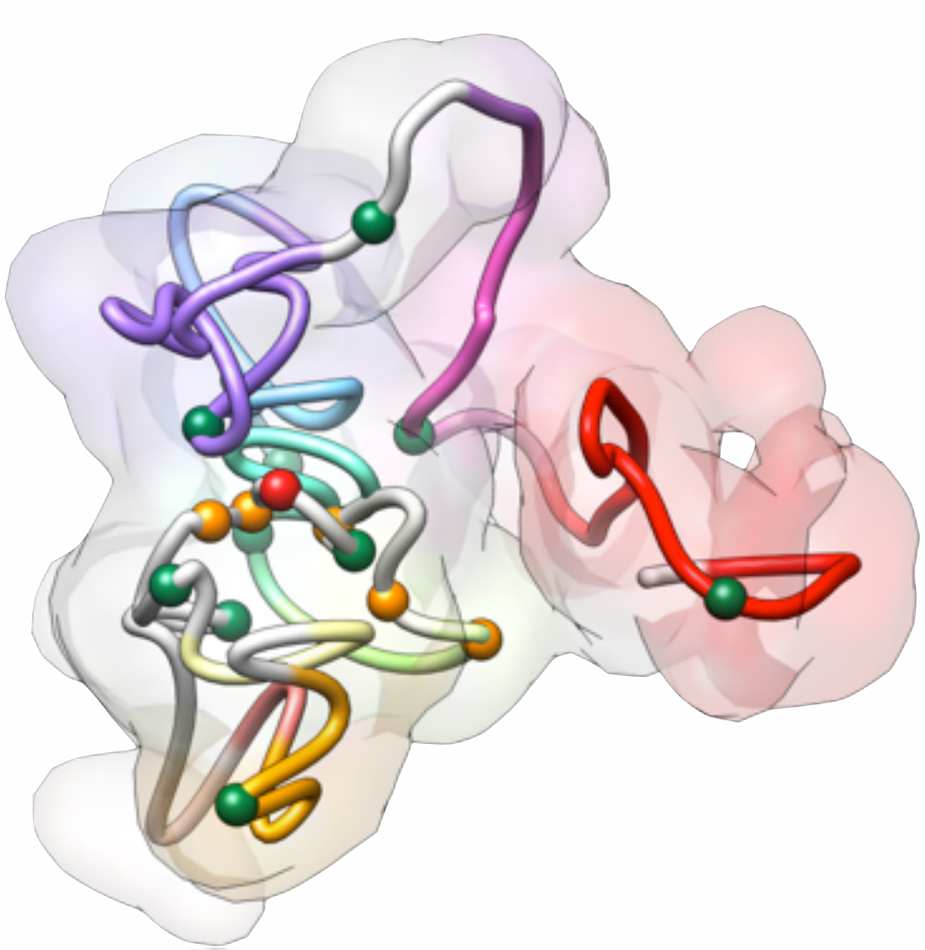
Biochemical function
Activity depends on the 3D structure

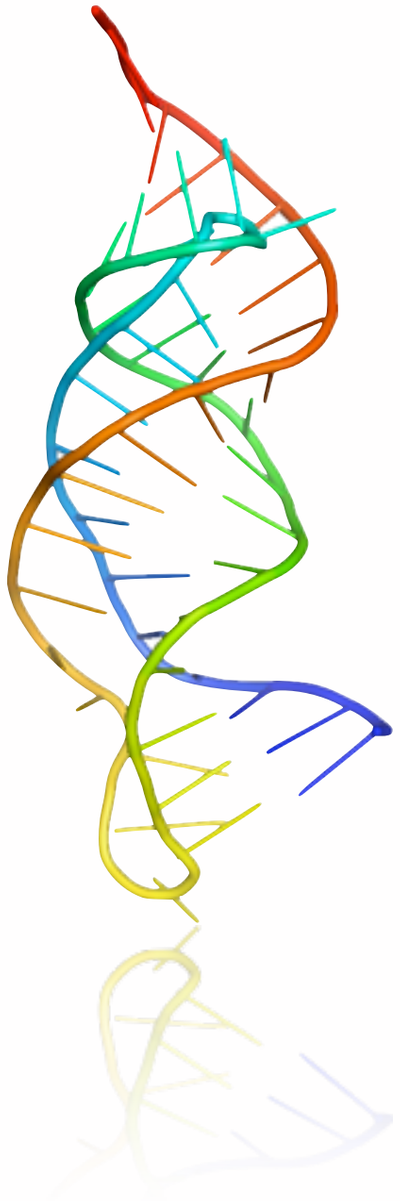
Evolution conserve
Structure is more conserved than sequence

Protein types
Fibrous
Membrane
Globular

Nucleic acids

DNA and RNA





Nucleic acids

DNA and RNA

DNA and RNA are polymers made up of repeating units called **nucleotides**.

Each nucleotide is composed of a nitrogen-containing **nucleobase**, a monosaccharide **sugar** and a **phosphate** group.

The nucleotides are joined to one another in a chain by **sugar-nucleobase** covalent bonds.

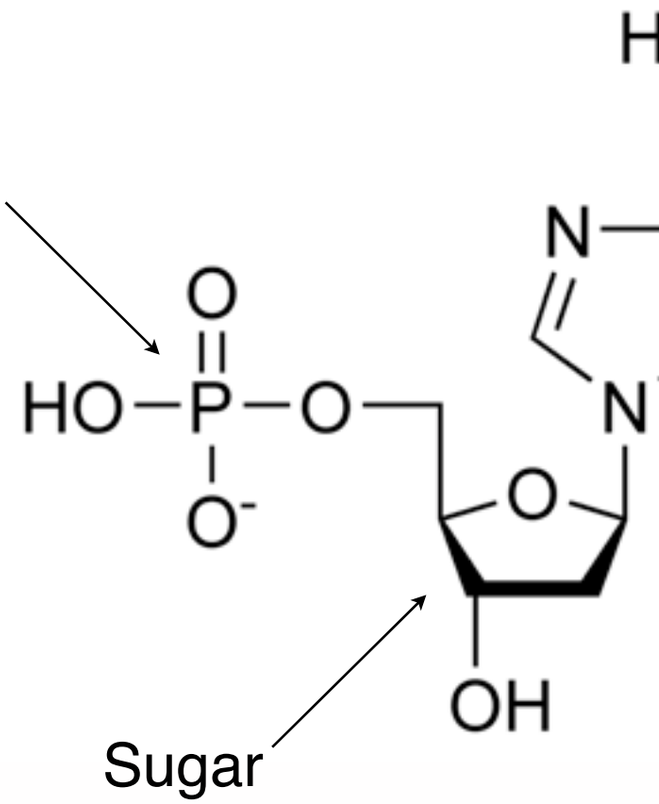
DNA (Deoxyribonucleic acid) encodes the genetic information.

RNA (Ribonucleic acid) is implicated in various biological roles including coding, decoding, regulation, and expression of genes.

The nucleotides

DNA

Phosphate group



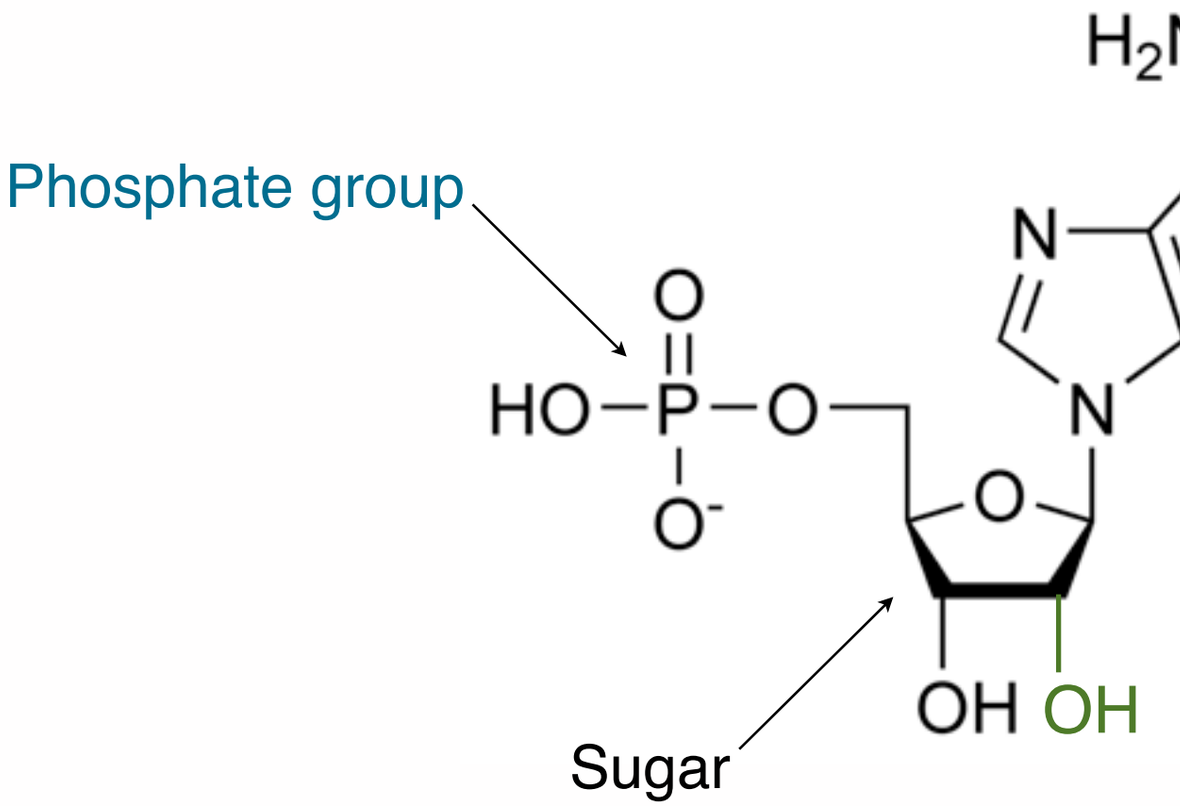
Sugar

Nitrogenous base

Guanine (G), Adenine (A), Thymine (T), or Cytosine (C)

The nucleotides

DNA



Nitrogenous base

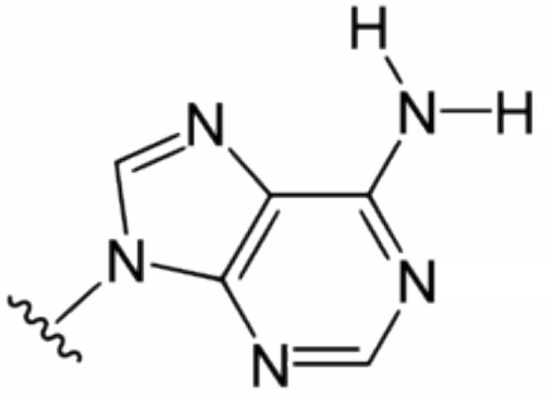
Guanine (G), Adenine (A), Thymine (T), or Cytosine (C)

Uracil (U)

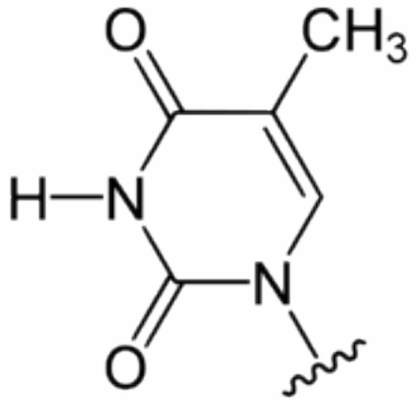
RNA

Nitrogens bases

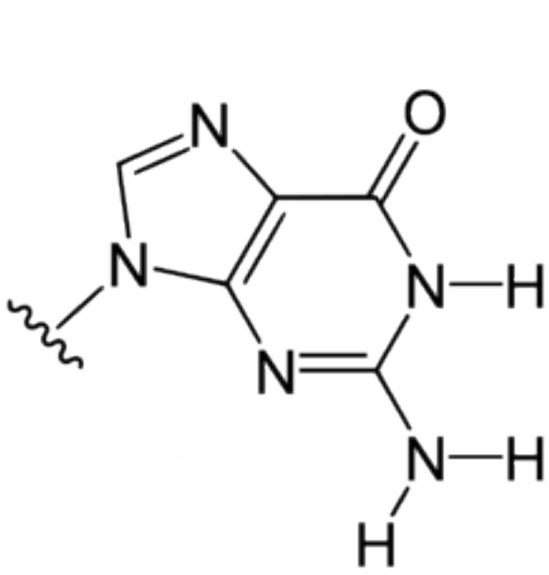
DNA



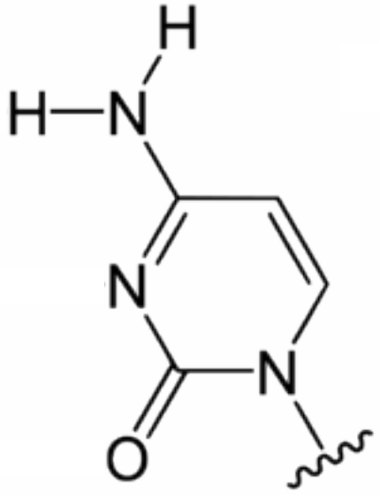
Adenine (**A**)



Thymine (**T**)

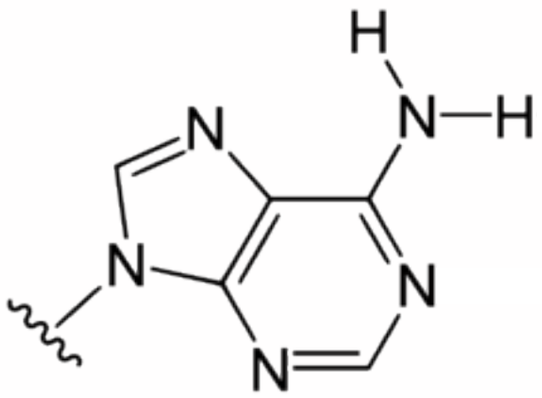


Guanine (**G**)



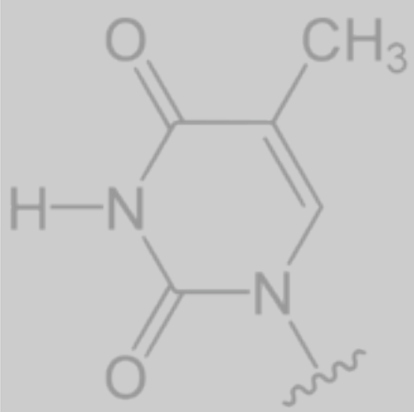
Cytosine (**C**)

Nitrogens bases



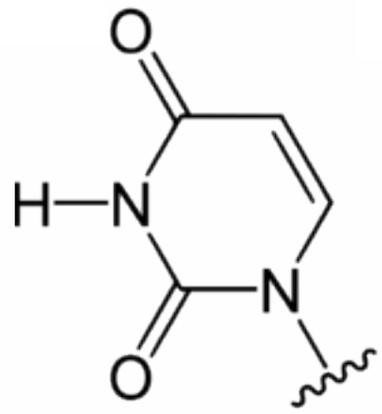
Adenine (**A**)

DNA

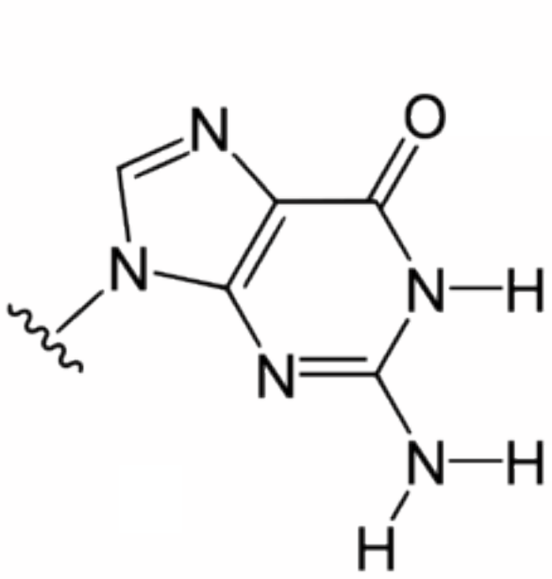


Thymine (**T**)

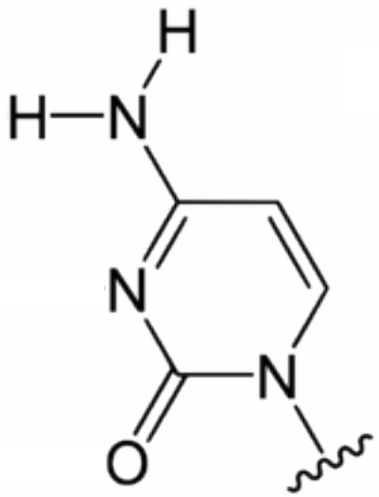
RNA



Uracil (**U**)

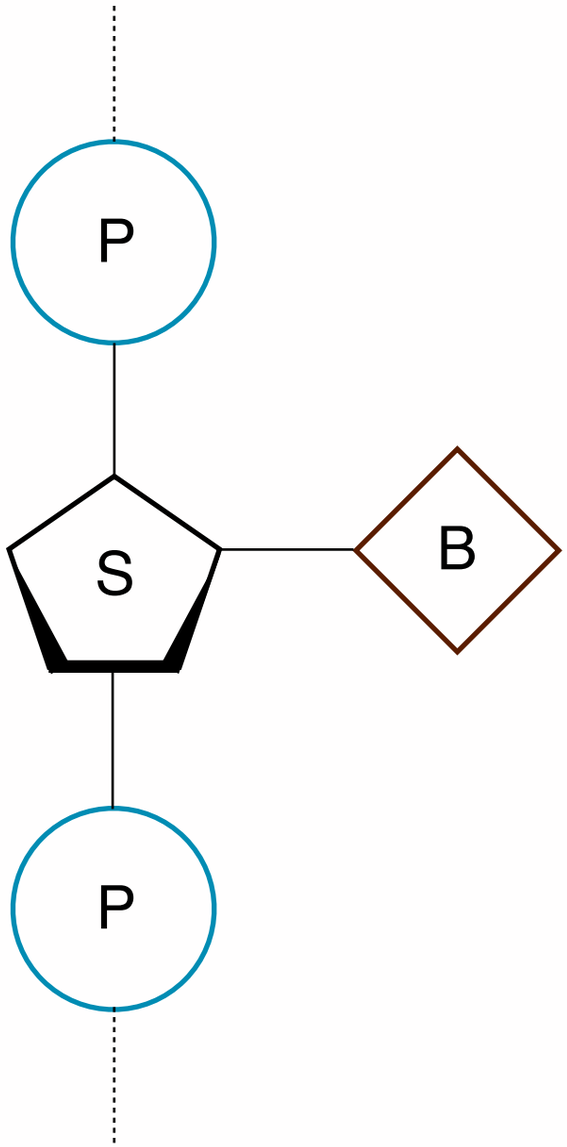


Guanine (**G**)

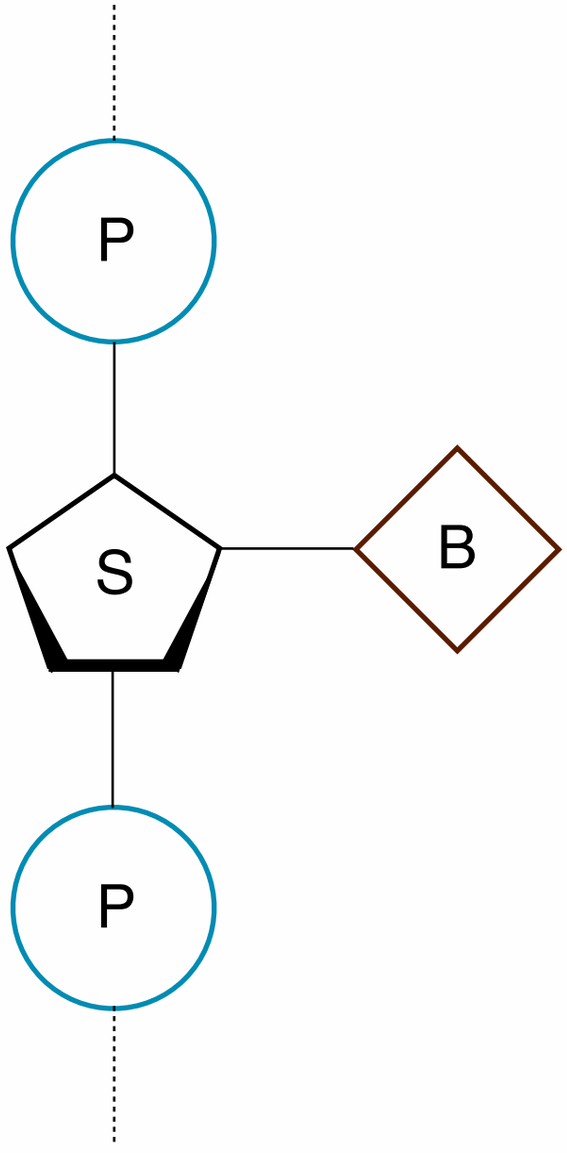


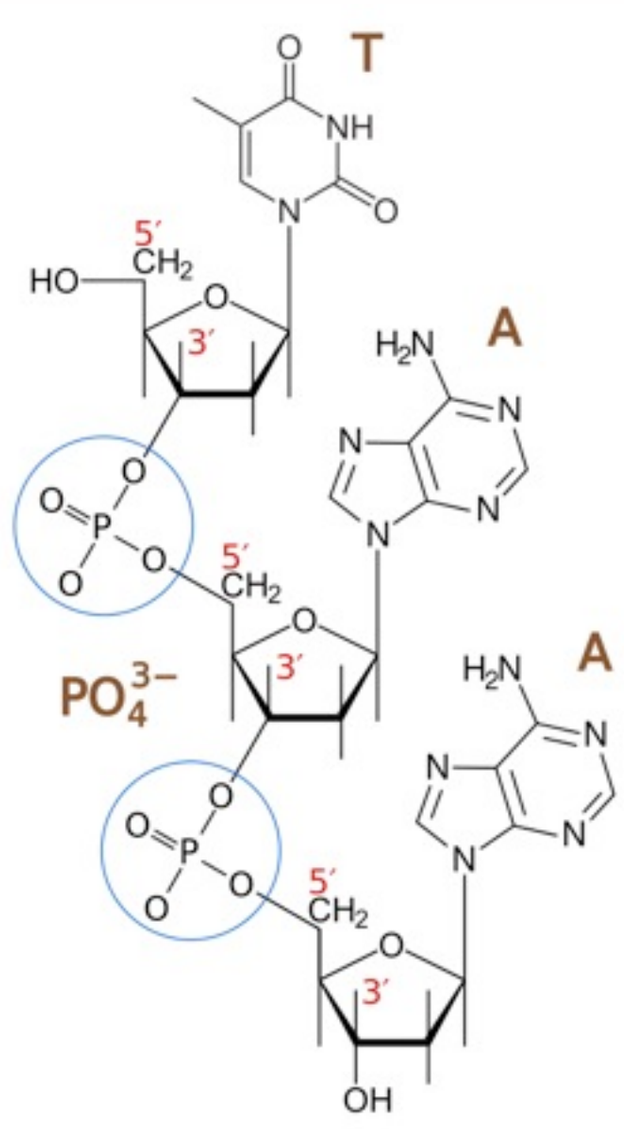
Cytosine (**C**)

The phosphodiester bond



The phosphodiester bond





Helix stability

Hydrogen bonds and base-stacking interactions

The two types of base pairs form different numbers of hydrogen bonds (**2 for AT, 3 for GC**).

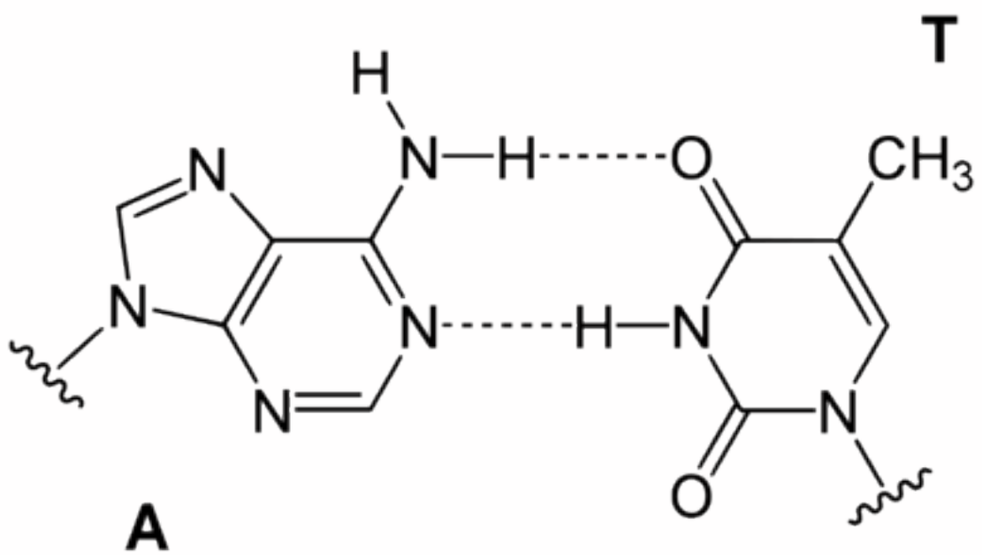
The DNA double helix is maintained largely by the intra-strand **base stacking interactions** (GC > AT).

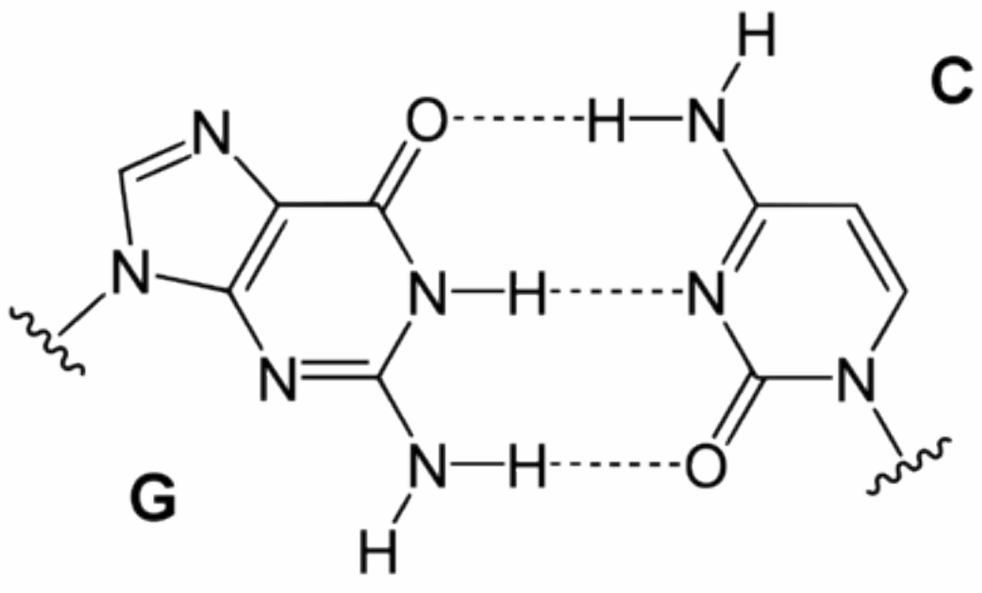
The stability of the dsDNA form depends also on **sequence** and **length**.

DNA with high GC-content is more stable than DNA with low GC-content.

Base pairing

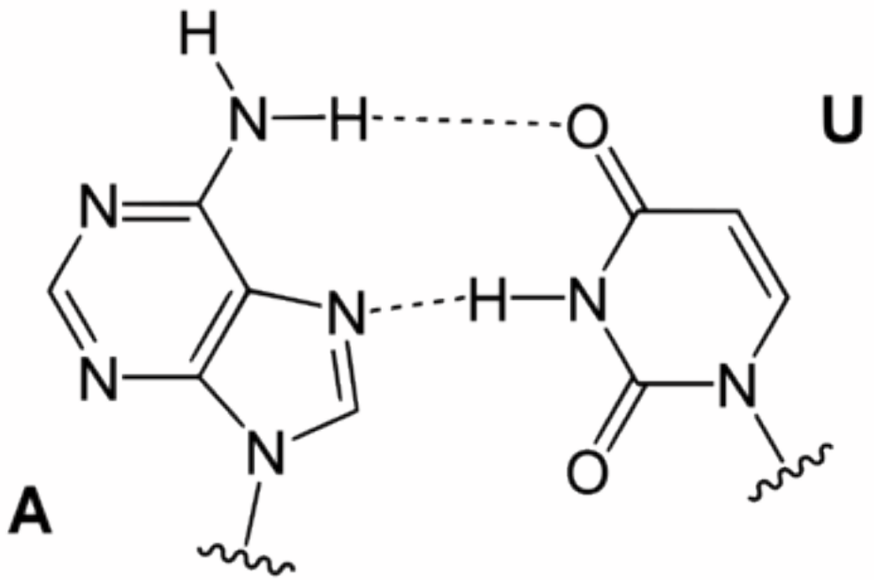
DNA

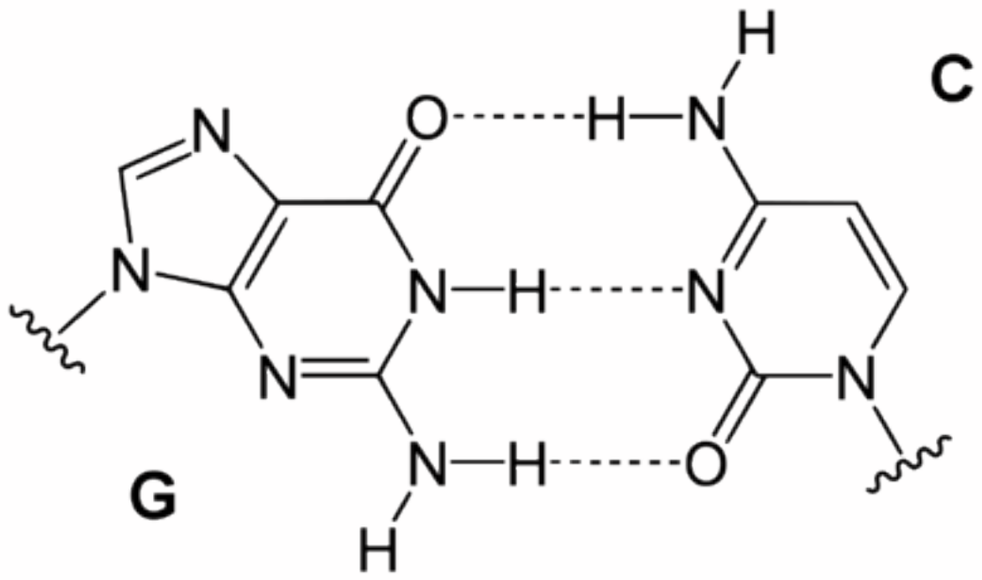




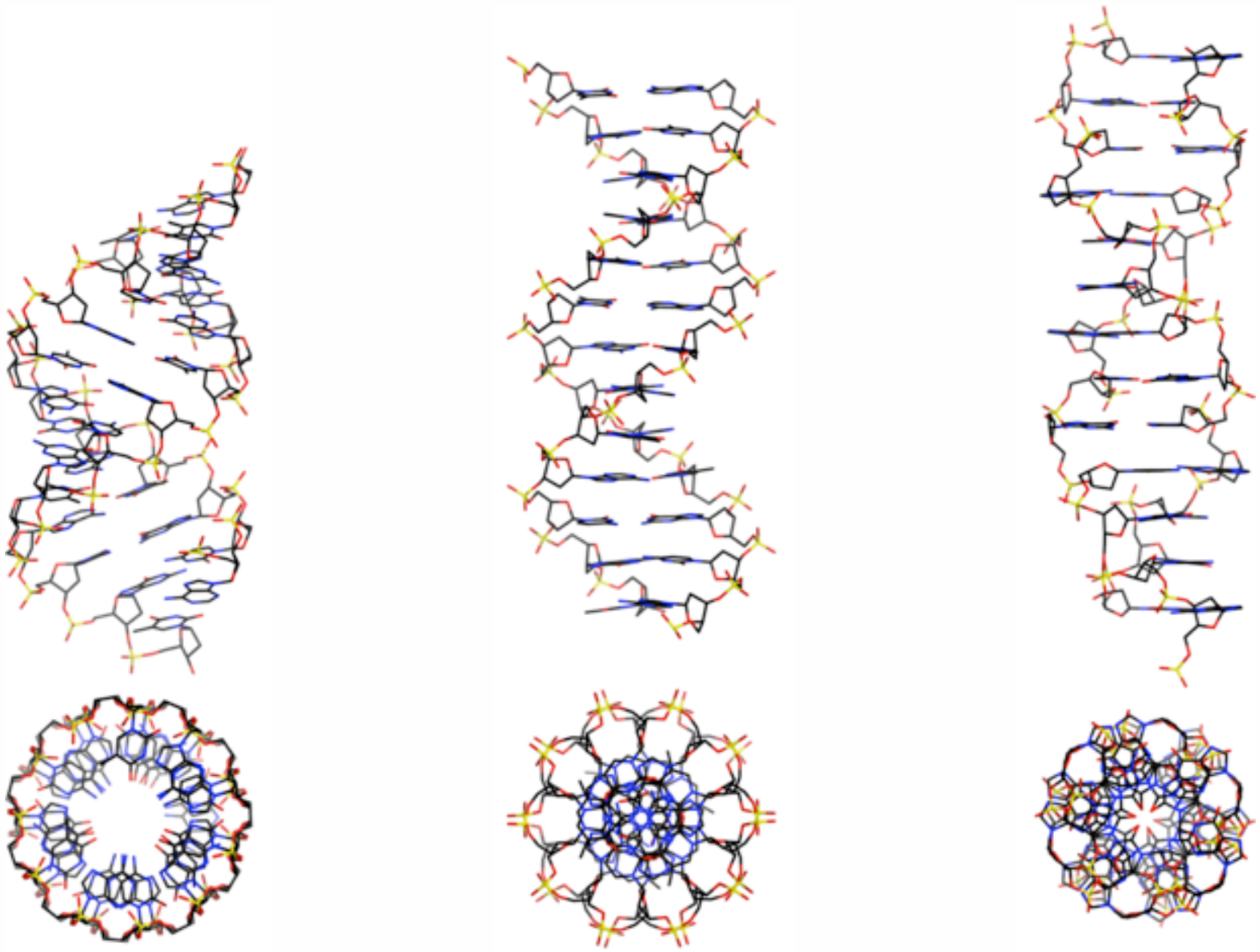
Base pairing

RNA





Nucleic acids helical structures

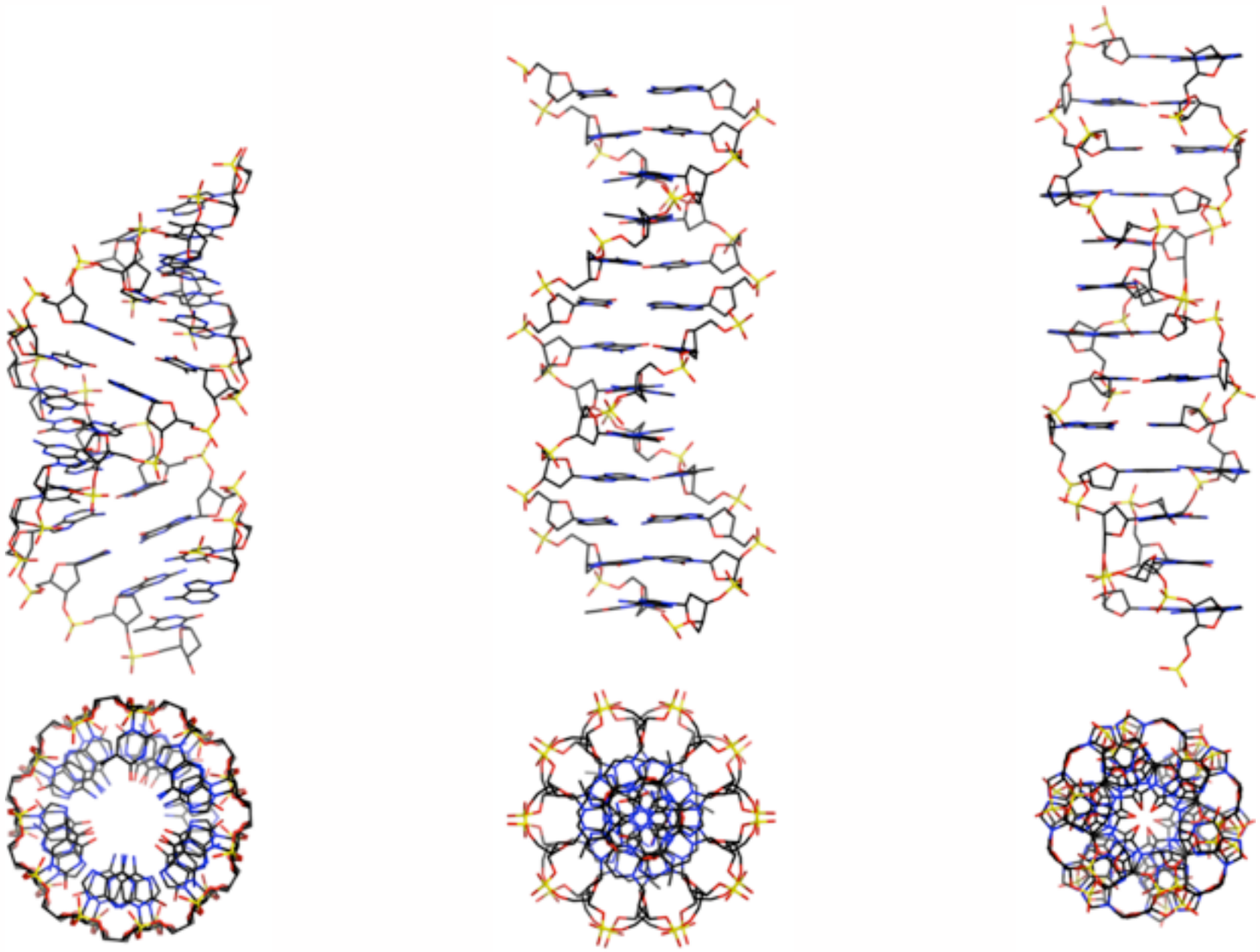


A-DNA B-DNA Z-DNA

Nucleic acids helical structures

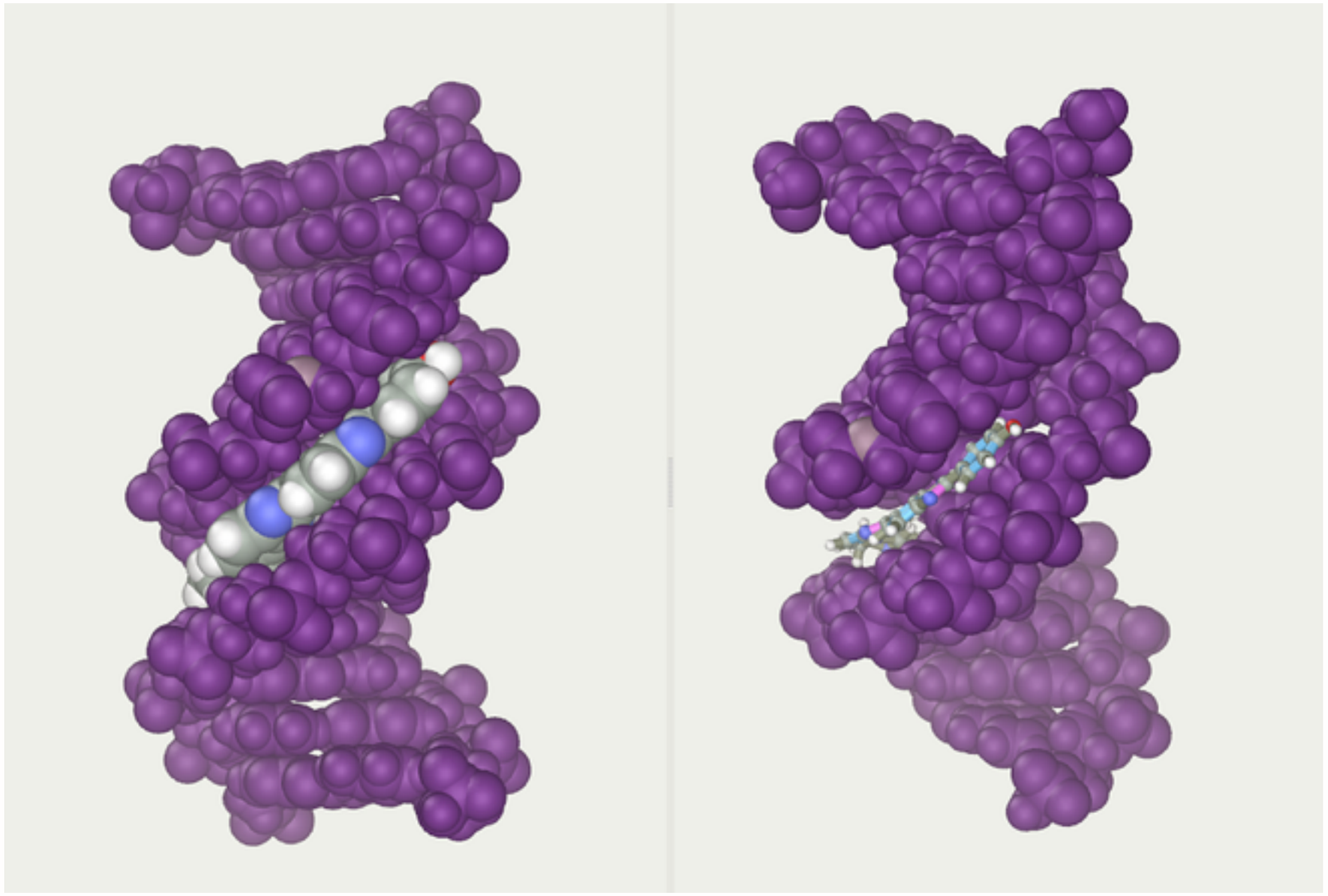
	A	B	Z
Helix sense	R	R	L
bp per turn	11	10	12
Vertical rise per bp (Å)	2.56	3.4	3.7
Rotation per bp (degrees)	+33	+36	-30
Helical diameter (Å)	23	19	18

Nucleic acids helical structures



A-DNA B-DNA Z-DNA

Major and minor groove

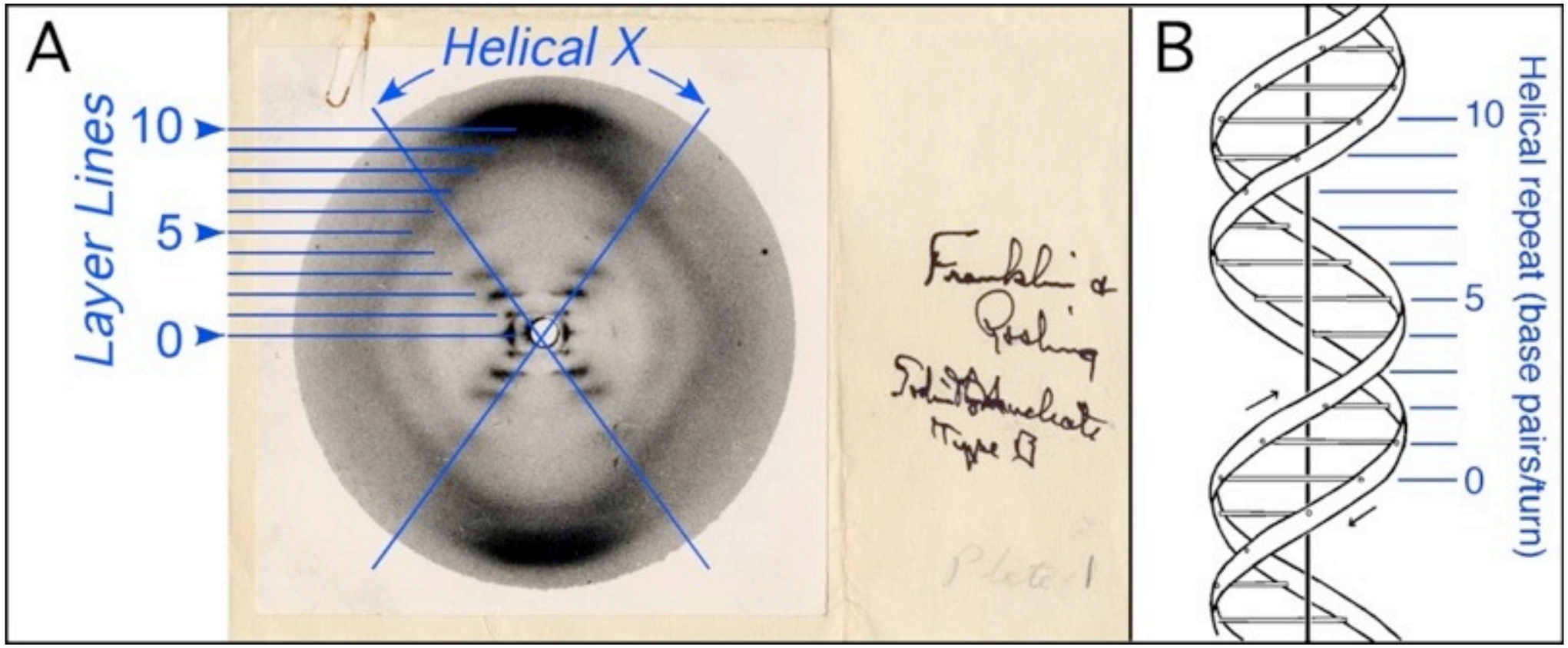


Major groove

Minor groove

The helical structure and DNA

Rosalind Franklin



Take home message

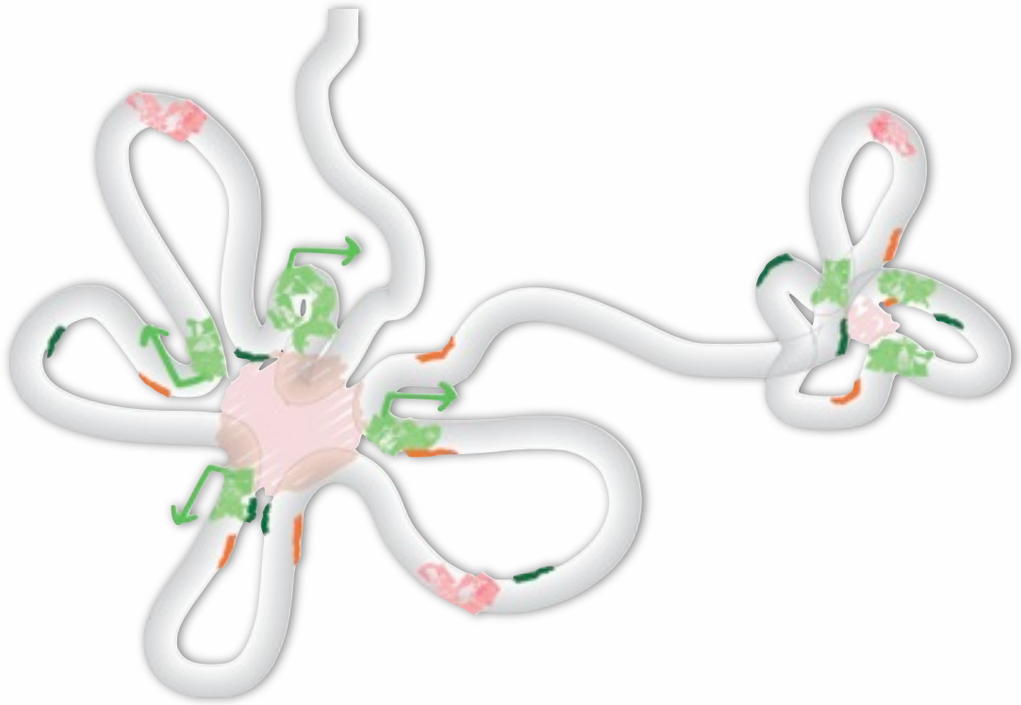
DNA and RNA
Polymers of nucleotide units

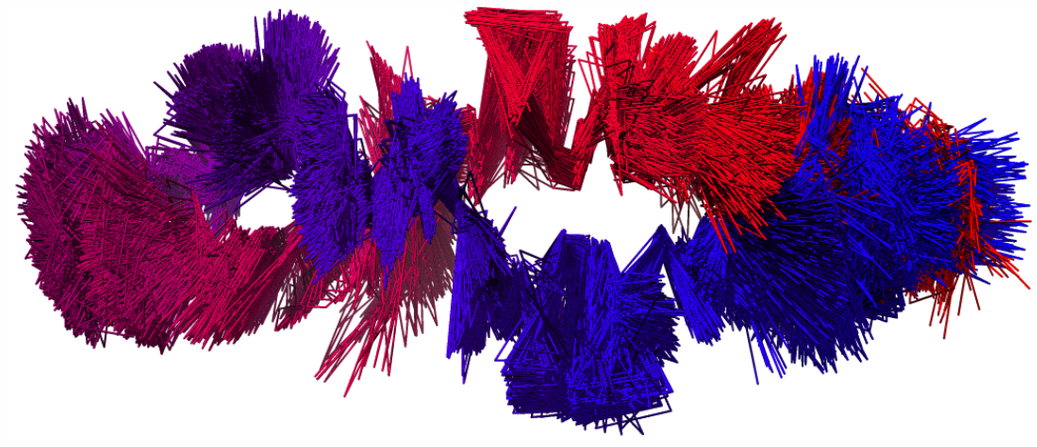
Nucleotides
Nucleobase (G,C,A,T - U)
+ sugar +phosphate

DNA
Store the genetic information
RNA
Implicated in various
biological processes

Genomes

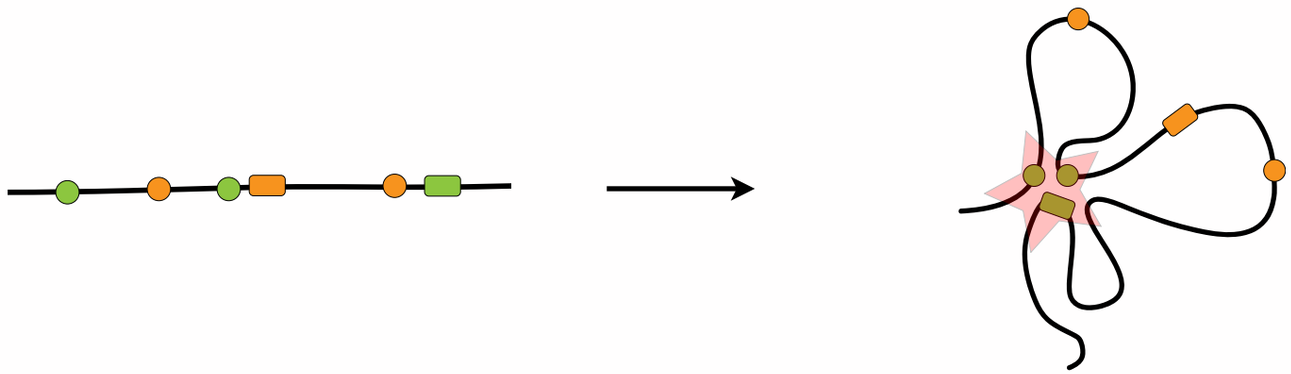
Limited data types





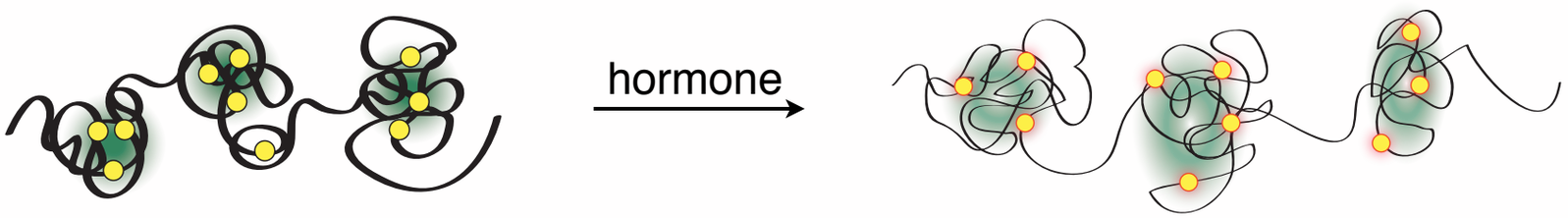
The role of chromatin structure

Activity



Organization

Processes



Chromatin definition

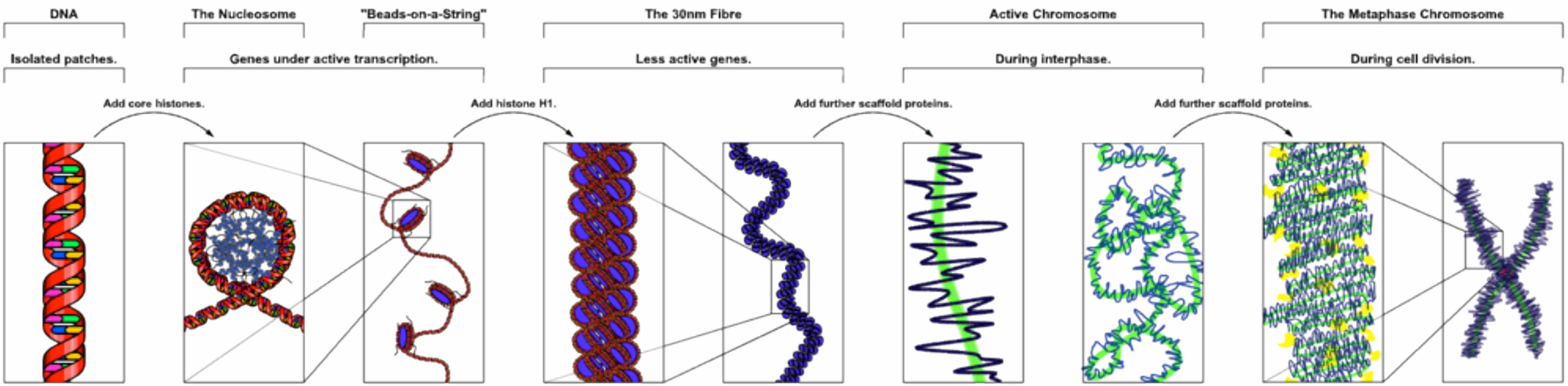
Chromatin is composed of **DNA** complexed with **histone proteins** and other **bio-molecules**.

Chromatin formation enables the genome to be hierarchically **packaged** or **condensed** so that it can fit inside the nuclear space.

The compaction allows to modulate gene **transcription**, **DNA repair**, **recombination**, and **replication**.

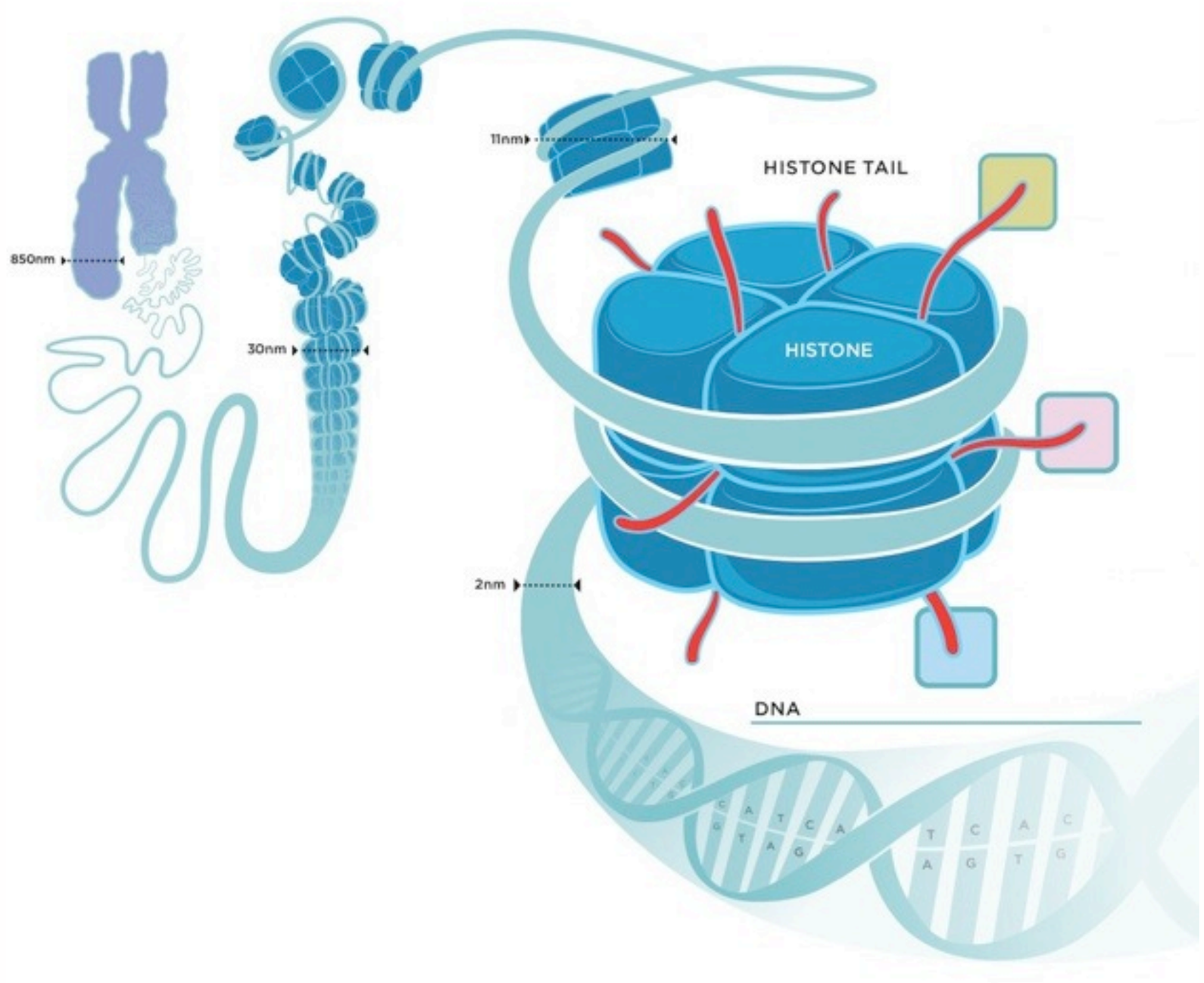
Chromatin structure is considered **highly dynamic**.

Chromatin structures



The nuclear organization of DNA

Chromosome **Chromatin fibre** **Nucleosome**

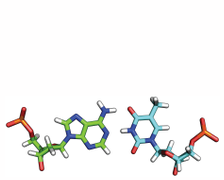


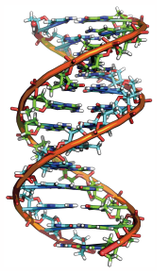
Adapted from Richard E. Ballermann, 2012

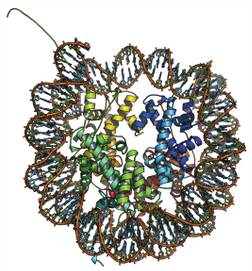
The resolution gap

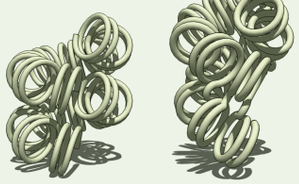
What do we “really” know?

Knowledge



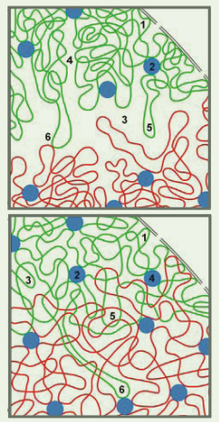


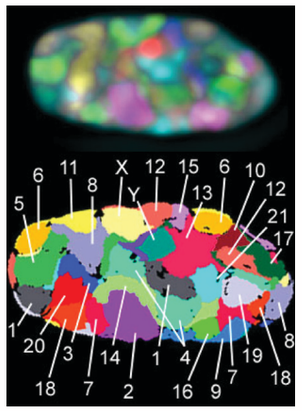




IDM

INM





DNA length

10^0 10^3 10^6 10^9 nt

Volume

10^{-9} 10^{-6} 10^{-3} 10^0 10^3 μm^3

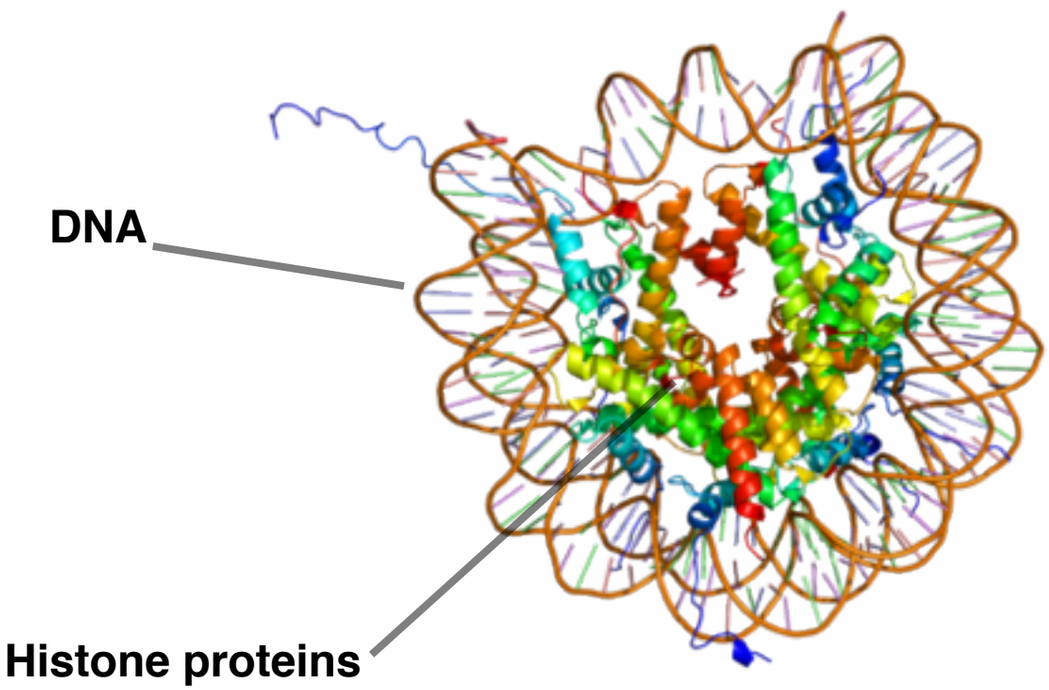
Time

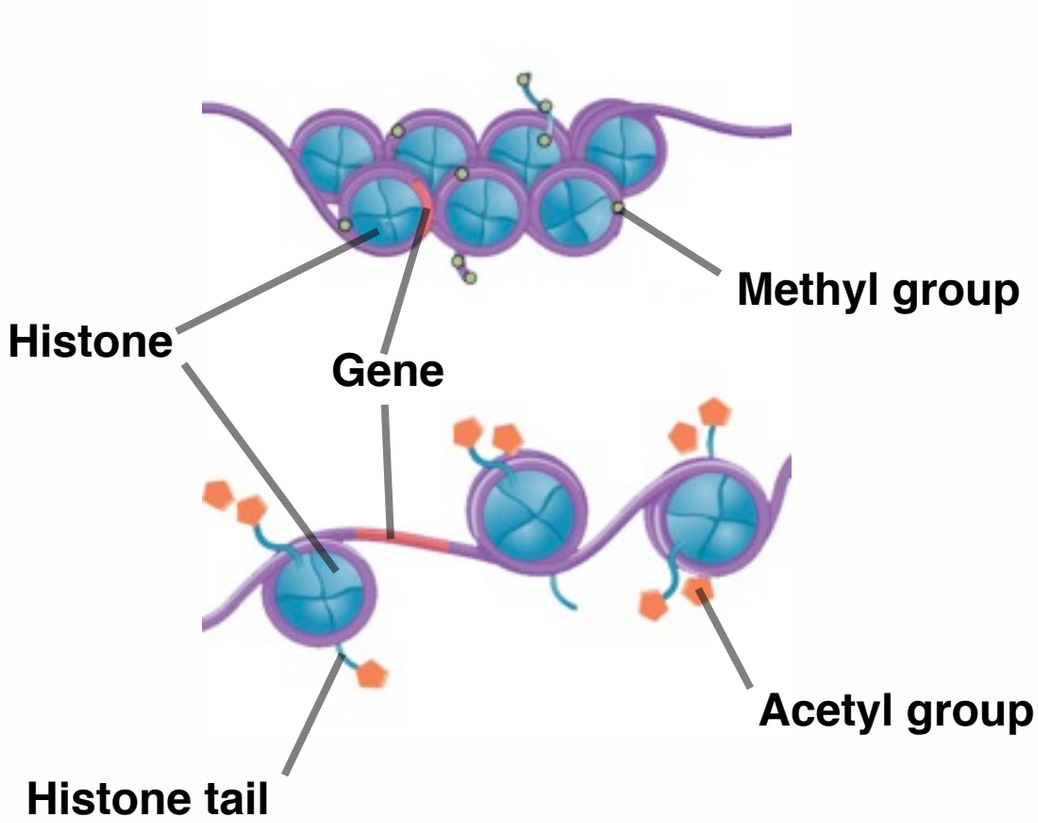
10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^0 10^2 10^3 s

Resolution

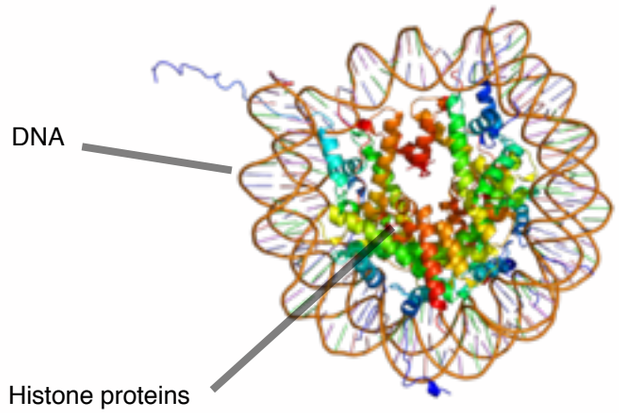
10^{-3} 10^{-2} 10^{-1} μ

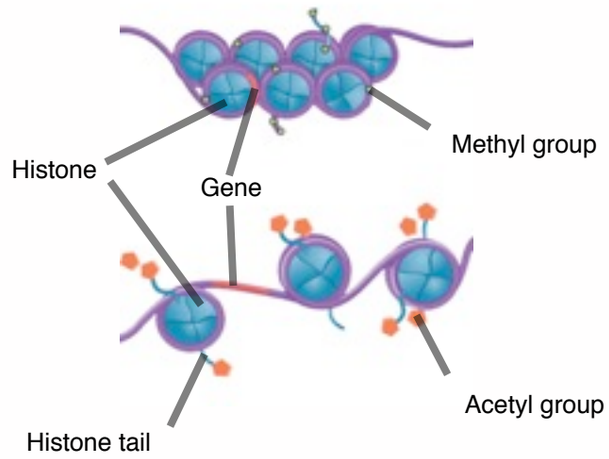
The nucleosome





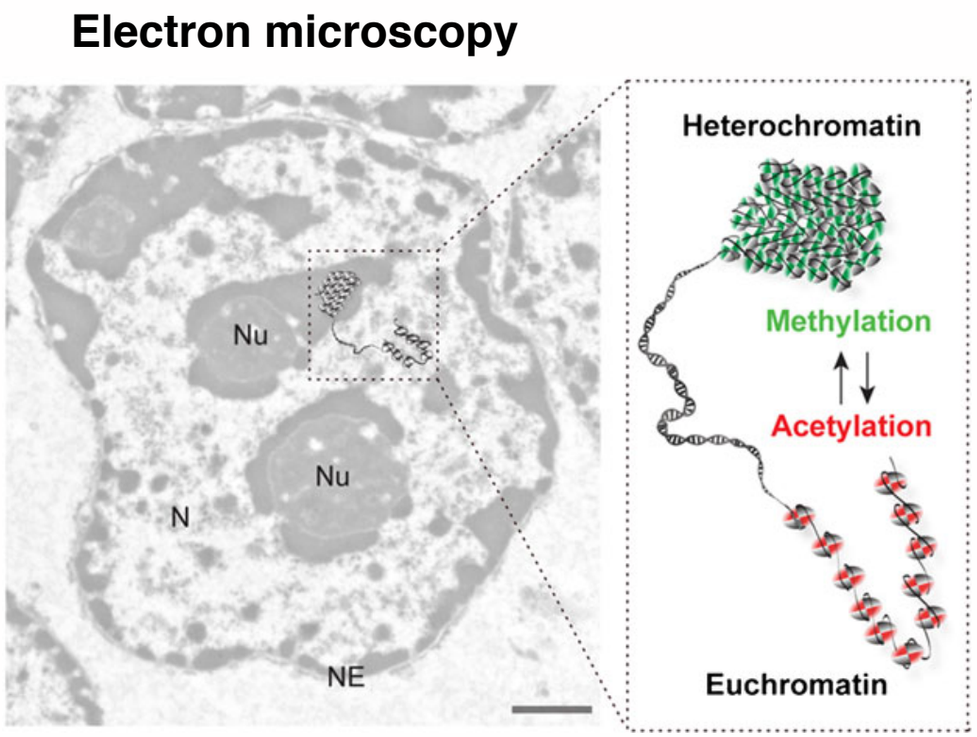
The nucleosome & chromatin marks





Modification	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation	activation		activation	activation	activation	activation
di-methylation	activation	repression		repression	activation		
tri-methylation	activation	repression		repression	activation, repression		repression
acetylation		activation	activation				

Euchromatin and heterochromatin

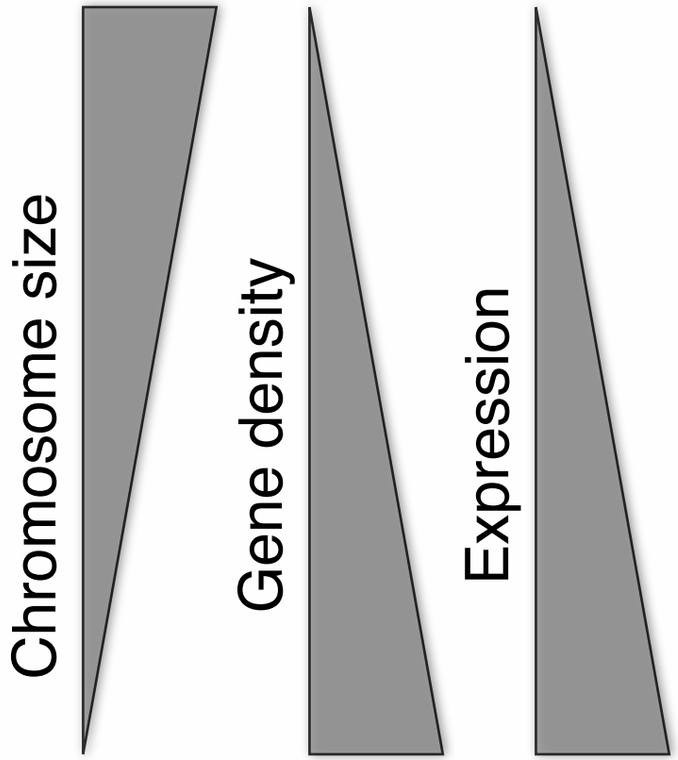


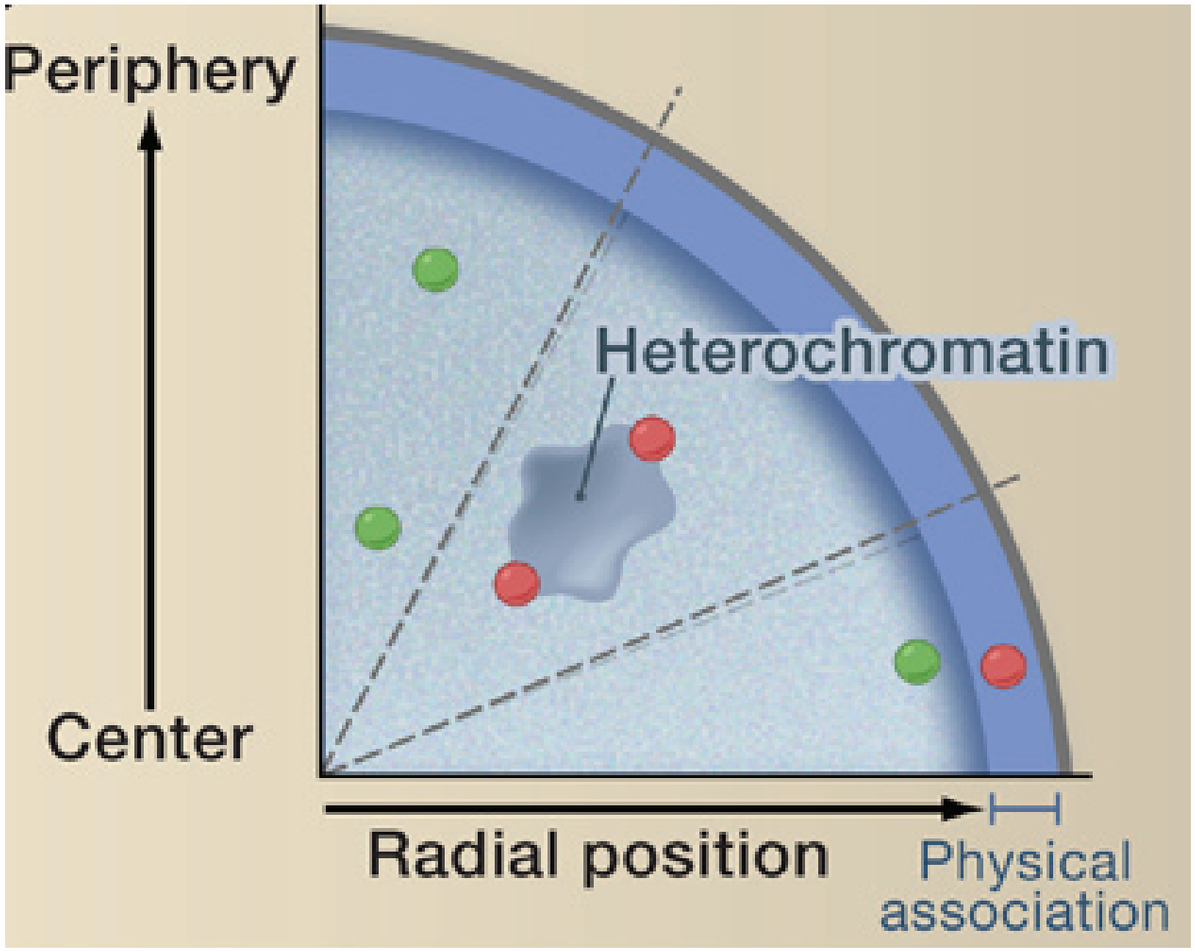
Euchromatin:
chromatin that is located away from the nuclear lamina, is generally less densely packed, and contains actively transcribed genes

Heterochromatin:
chromatin that is near the nuclear lamina, tightly condensed, and transcriptionally silent

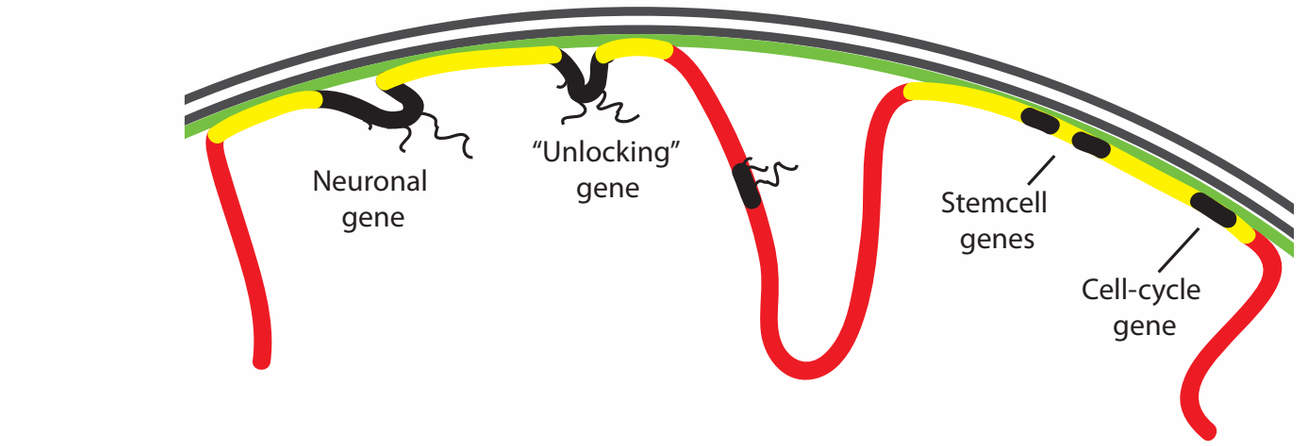
Complex genome organization

Takizawa, T., Meaburn, K. J. & Misteli, Cell 135, 9–13 (2008)





Lamina-genome interactions

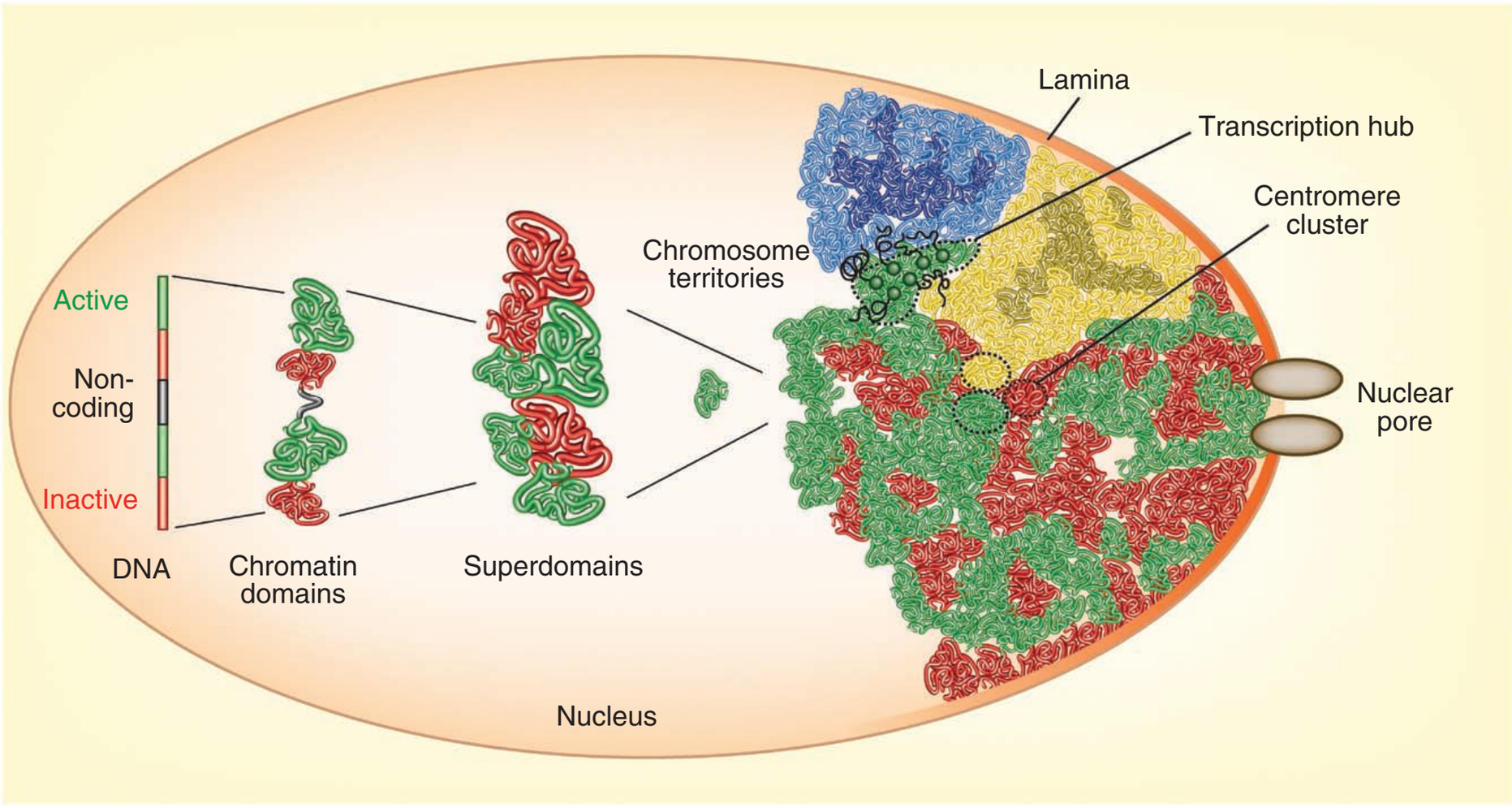


Most genes in Lamina Associated Domains are transcriptionally silent, suggesting that **lamina-genome interactions** are widely involved in the control of **gene expression**

Adapted from Molecular Cell 38, 603-613, 2010

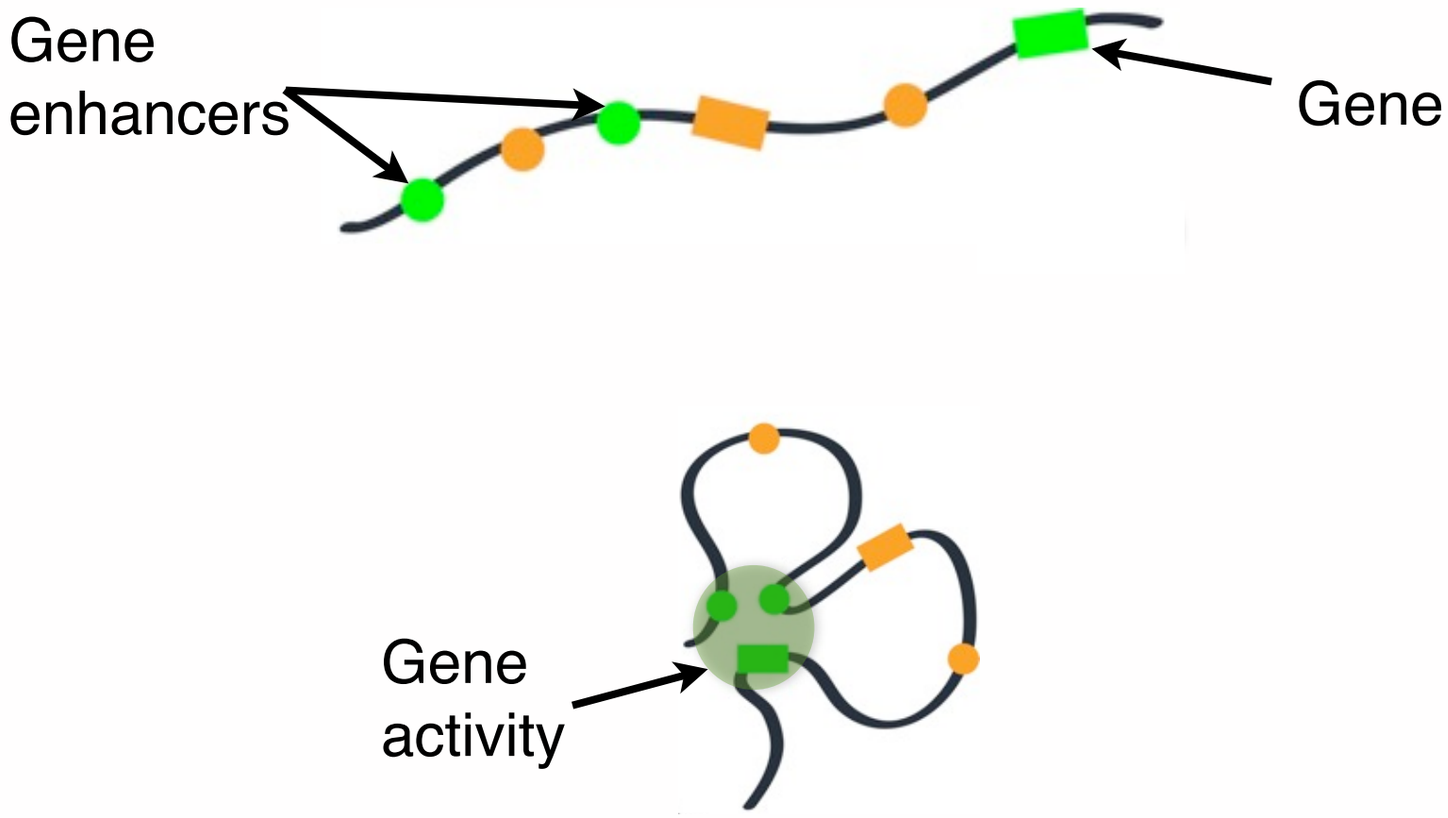
Complex genome organization

Cavalli, G. & Misteli, Nat Struct Mol Biol 20, 290–299 (2013)



Marina Corral

Chromatin loops



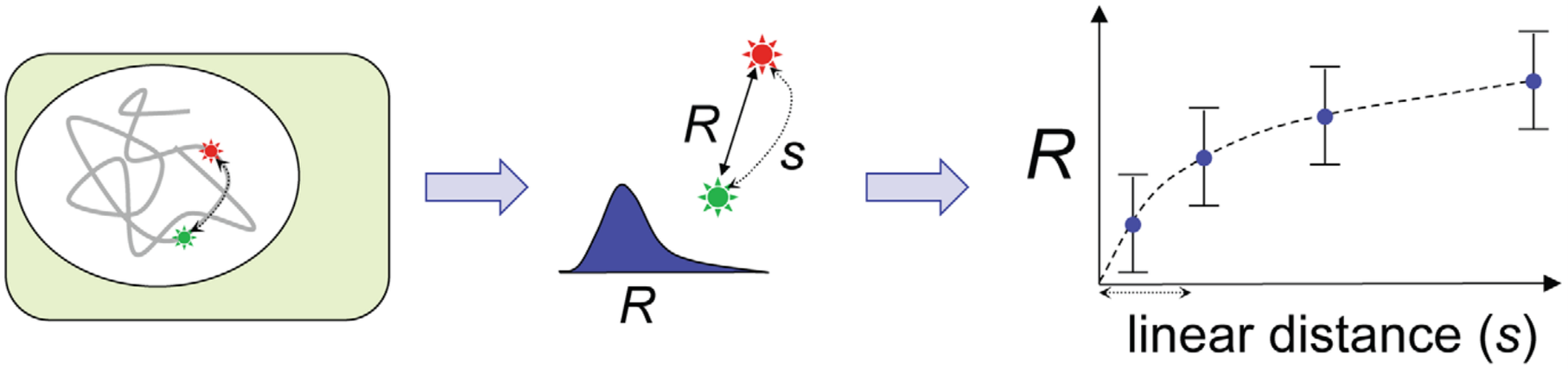
Loops bring **distal** genomic regions in **close** proximity to one another.

This in turn can have profound effects on **gene transcription**.

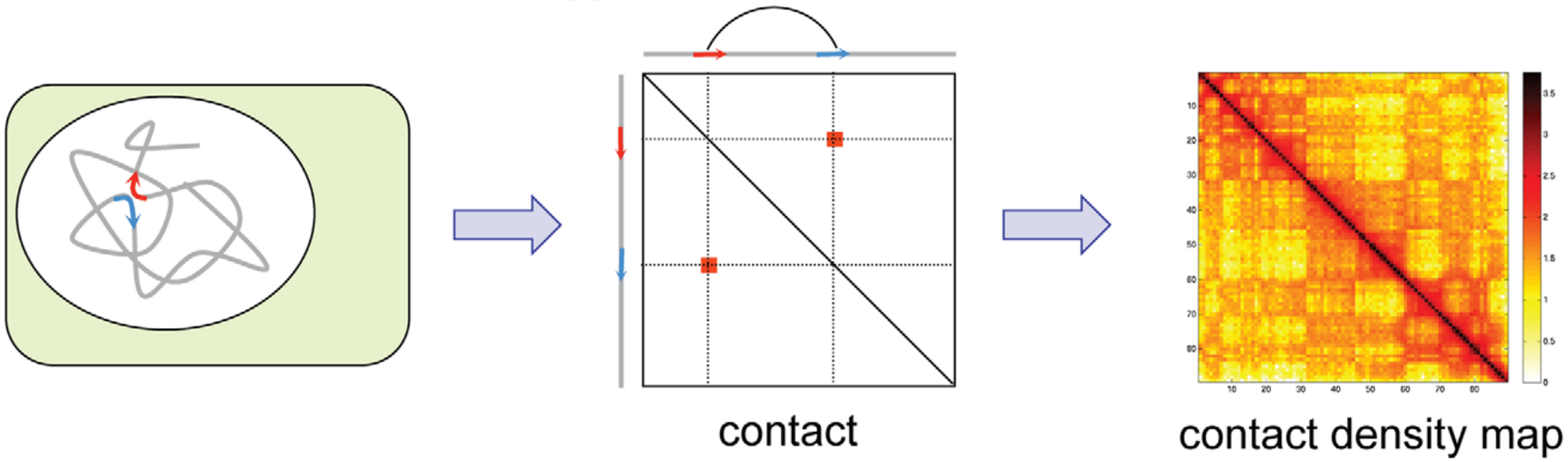
Enhancers can be thousands of kilobases away from their **target genes** in any direction (or even on a separate chromosome).

Main approaches

Light microscopy (FISH)



Cell/molecular biology (3C-based methods)

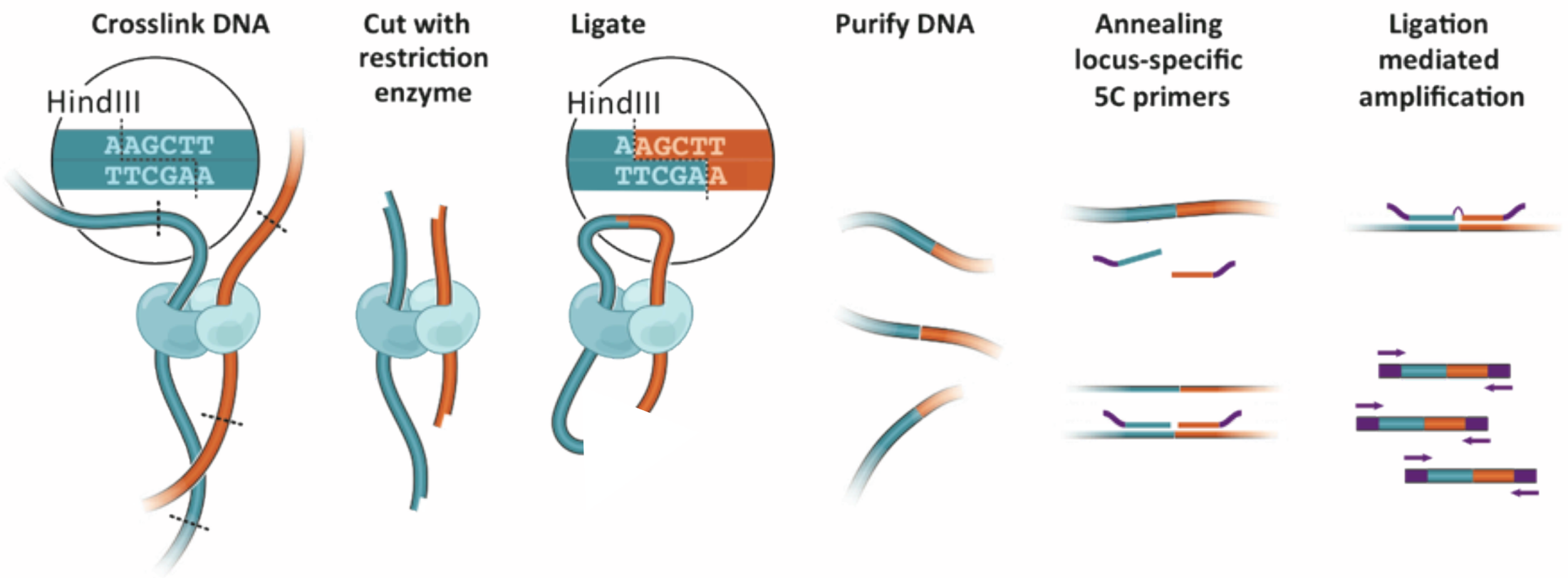


Job Dekker

5C technology

http://my5C.umassmed.edu

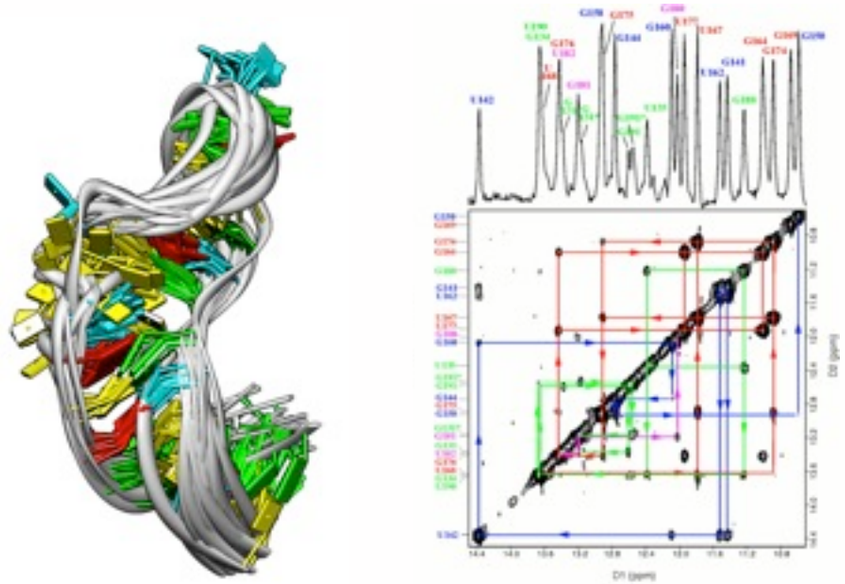
UMASS



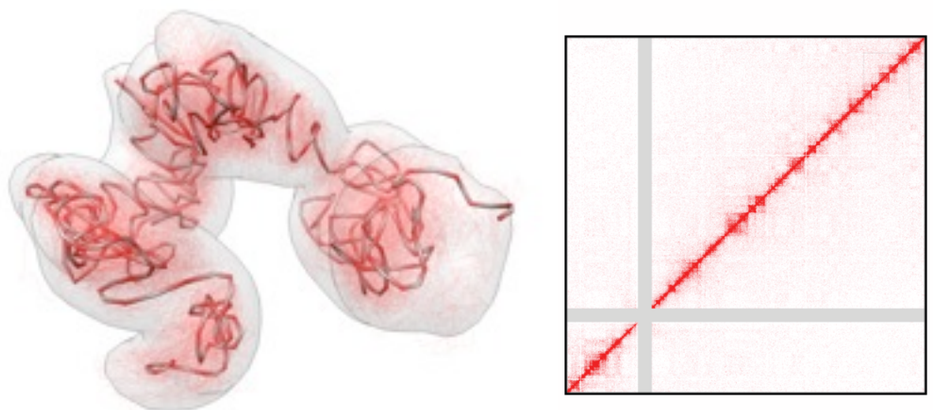
Dostie et al. Genome Res (2006) vol. 16 (10) pp. 1299-309

cnag CRG

Structure determination using Hi-C data

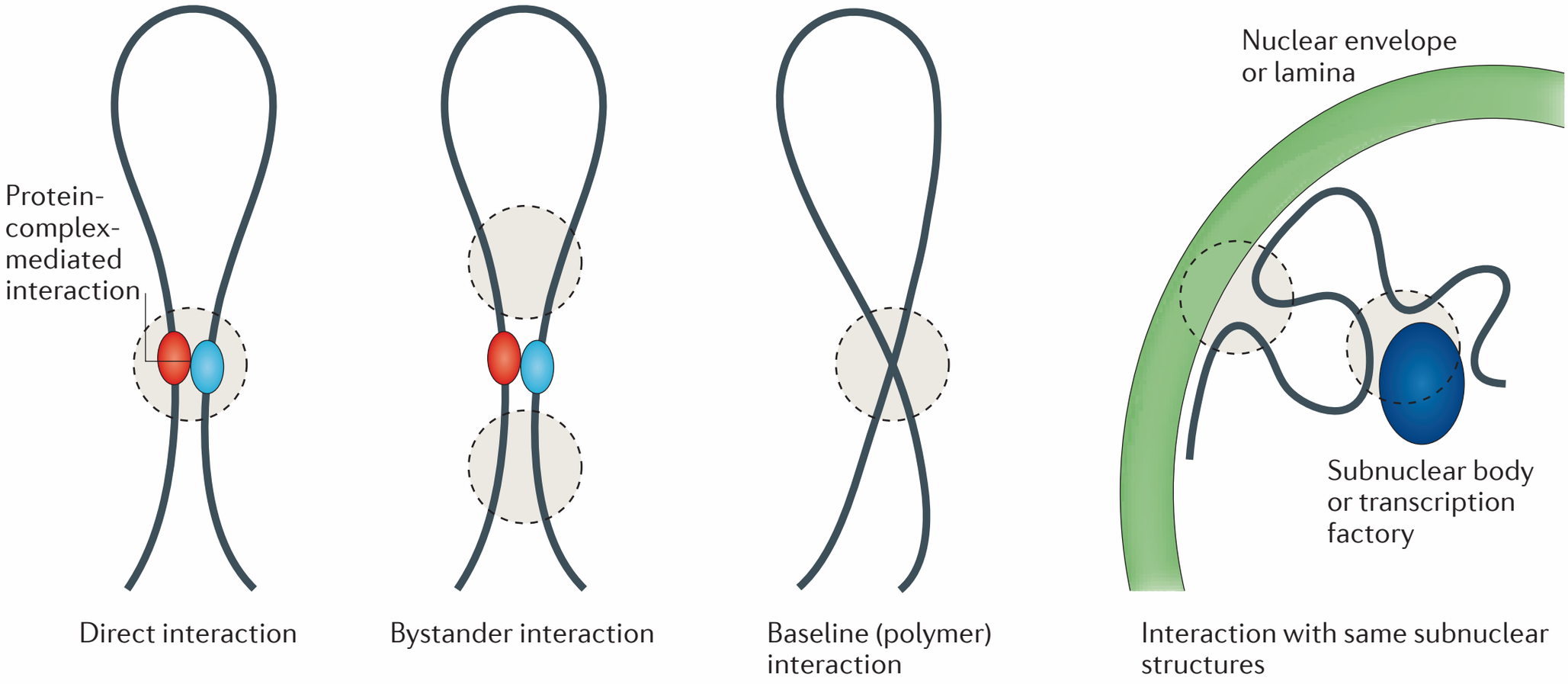


Biomolecular structure determination
2D-NOESY data



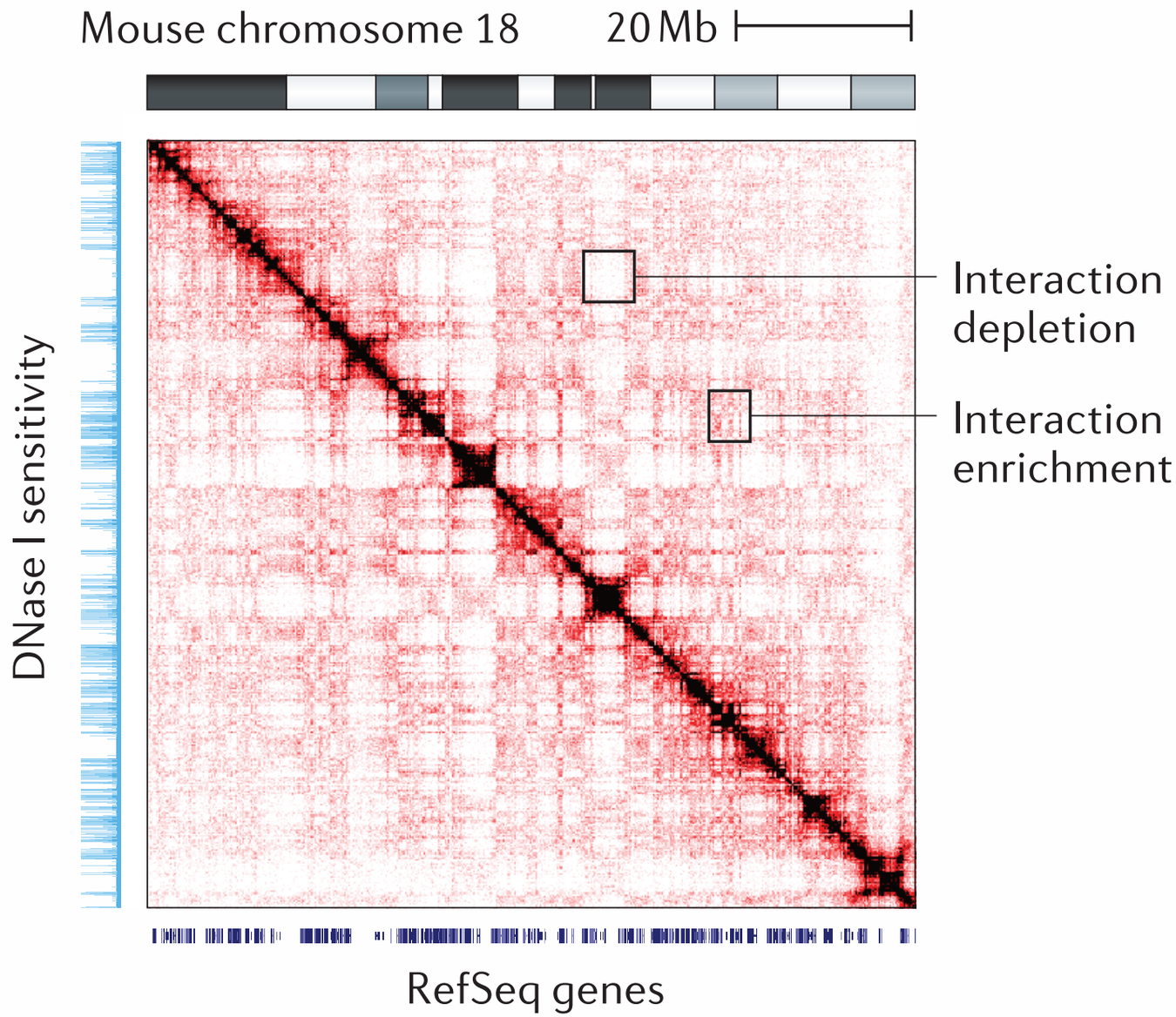
Chromosome structure determination
3C-based data

Interpreting chromatin interaction data



Adapted from Dekker et all, (2013) Nat Rev Genetics

Hi-C data and genomic tracks data

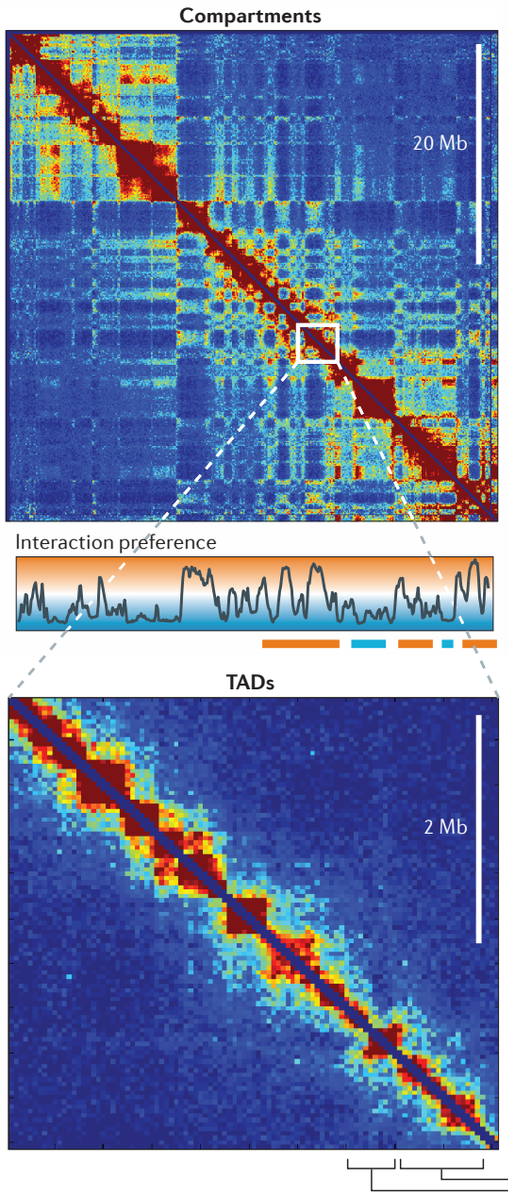


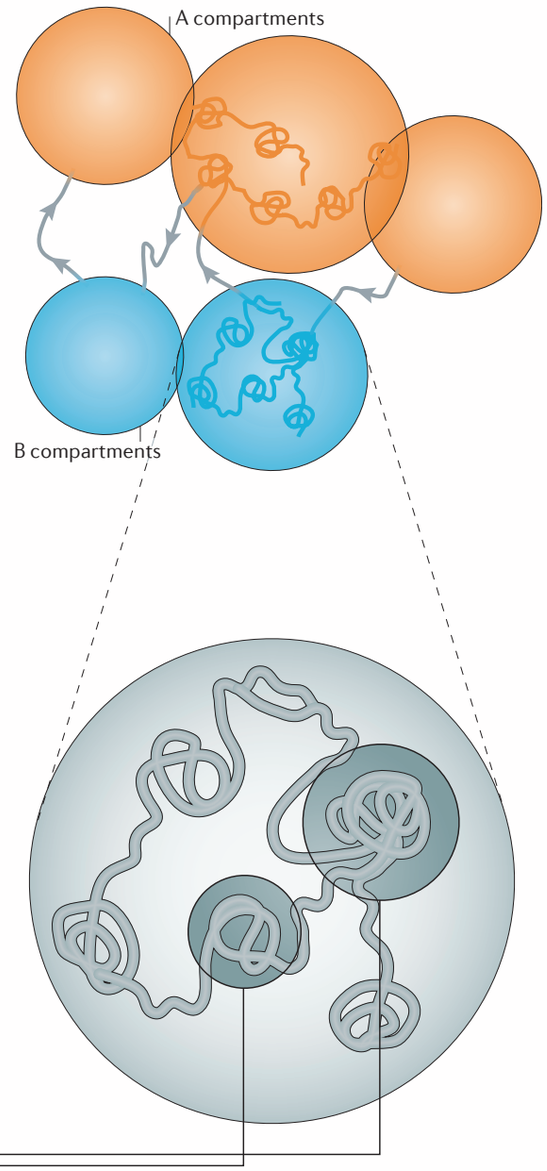
Adapted from Dekker et all, (2013) Nat Rev Genetics

Genome Organization

Dekker, J., Marti-Renom, M. A. & Mirny, L. A.Nat Rev Genet (2013)

Human chromosome 14





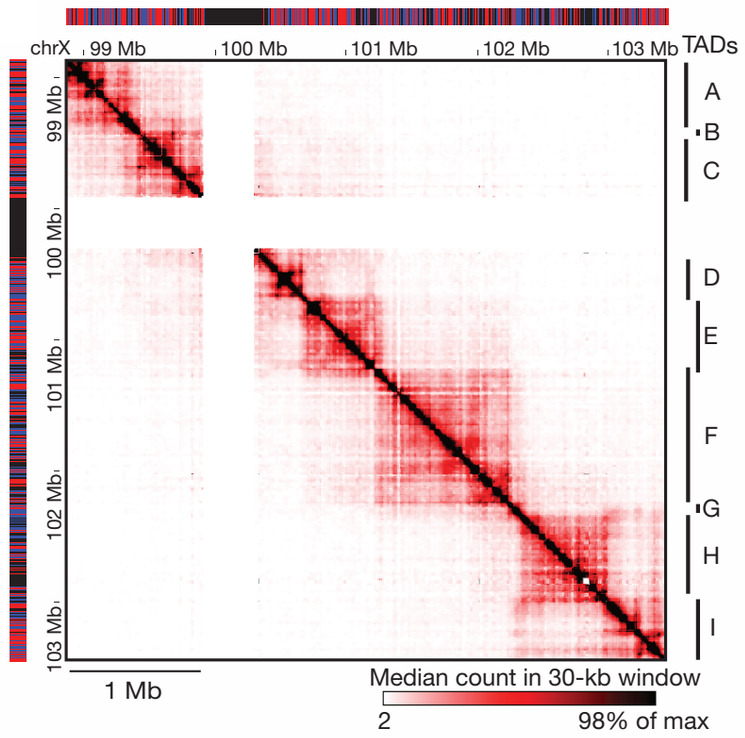
A-B compartments

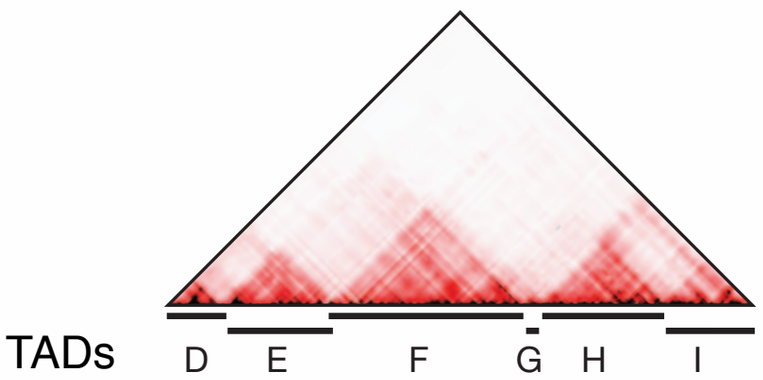
TADs

Adapted from Dekker et all, (2013) Nat Rev Genetics

cnag CRG

Topologically Associating Domains (TADs)





Topologically associating domains (TADs) can be made of up to hundreds of kb in size

Loci located within TADs tend to interact more frequently with each other than with loci located outside their domain

The human and mouse genomes are each composed of over 2,000 TADs, covering over 90% of the genome

Take home message

Chromatin = DNA + (histone) proteins + other biomolecules

The genome is well organized and hierarchically packaged

Histone modifications affect chromatin structure and activity

3C-like data measure the frequency of interaction between distant loci

[1] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. Proc Natl Acad Sci U S A, 27(11):499–506, 1941.

[2] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431(7011):931–45, 2004.

[3] F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. Nature,192:1227–32, 1961.

[4] M. Grunberg-Manago, P. J. Oritz, and S. Ochoa. Enzymatic synthesis of nucleic acidlike polynucleotides. Science, 122(3176):907–10, 1955.

[5] H. G. Khorana. Polynucleotide synthesis and the genetic code. Fed Proc, 24(6):1473–87, 1965.

[6] P. Leder and M. W. Nirenberg. Rna codewords and protein synthesis, 3. on the nucleotide sequence of a cysteine and a leucine rna codeword. Proc Natl Acad Sci U S A, 52:1521–9, 1964.

[7] J. H. Matthaei, O. W. Jones, R. G. Martin, and M. W. Nirenberg. Characteristics and composition of rna coding units. Proc Natl Acad Sci U S A, 48:666–77, 1962.

[8] F. Sanger and A. R. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. J Mol Biol, 94(3):441–8, 1975.

[9] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. i. the identification of lower peptides from partial hydrolysates. Biochem J, 49(4):463–81, 1951.

[10] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171(4356):737–8, 1953