BIOPHYSICS, FACULTY OF SCIENCE UNIVERSITY OF SPLIT

Protein - DNA interaction in chromatin

Nucleic acids and proteins

Professor: Rudolf Podgornik

Student: Tomislav Donđivić

February 2013

Physics of DNA, chromatin and viruses

Biopolymers

Biopolymers are polymers produced by living organisms and it contain monomeric units that are covalently bonded to form larger structures. Three main classes of biopolymers are present classified according to the monomeric units and structure of the biopolymer formed: nucleic acids, proteins and polysaccharides. Polynucleotides (RNA and DNA) are long polymers composed of 13 or more nucleotide monomers; polypeptides are short polymers of amino acids; and polysaccharides are often linear bonded polymeric carbohydrate structures.

Biopolymers are strings or sequences of monomeric units or monomers for short and in many cases these strings are linear. Sometimes they are closed and circular, branched or even cross-linked. It's structure is determined by the nature of the building blocks (monomeric units) in combination with environmental conditions such as the temperature, the solvent (water) and the presence of salts and/or other molecular components.

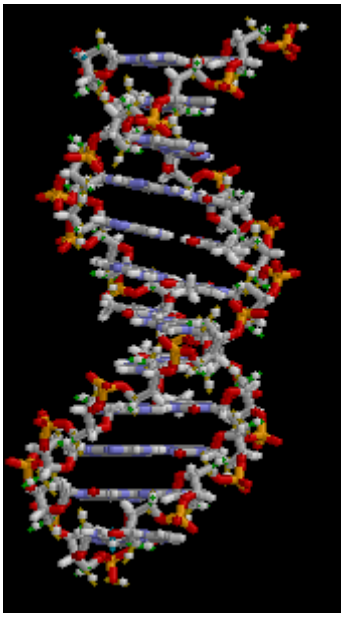


Figure 1. In this microstructure of DNA is a pair of biopolymers, polynucleotides, forming the double helix found in DNA

Most of biopolymers are essentially heteropolymers, because they may contain a variety in monomeric units. The biological relevance of a biopolymer is ultimately based on the sequence of the monomers, the primary tructure. In the case of DNA, the primary structure is the sequence of bases attached to the sugar rings, which determines the genetic code.

For proteins, it is the amino acid sequence, which eventually determines, together with environmental conditions, their 3D shapes and biological functions.

In contrast to synthetic polymers which have a simpler and more random structure, biopolymers are complex molecular assemblies that adopt precise and defined 3D shapes and structures. This feature is essential because this is what makes biopolymers active molecules in vivo. Their defined shape and structure are indeed keys to their function. For example, hemoglobin would not be able to carry oxygen in the blood if it was not folded in a quaternary structure.

Cellulose is the most common organic compound and biopolymer on Earth. About 33 percent of all plant matter is cellulose. The cellulose content of cotton is 90 percent, while wood's is 50 percent.

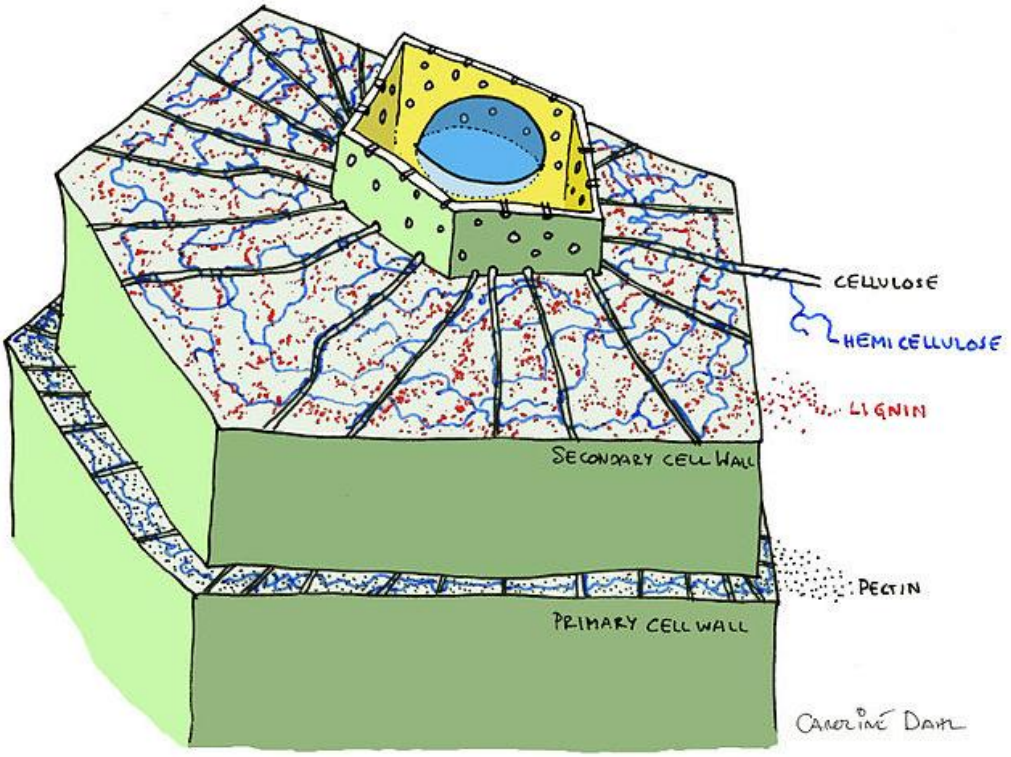


Figure 2. Cellulose in a plant cell

The characteristic of biopolymers are the formation of hierarchical structures at successive length scales. Starting from the primary structure, the monomeric units are organized in a certain local molecular conformation. This local conformation is commonly referred to as the secondary structure. Examples of secondary structures are the famous doublehelical arrangement of the two opposing strands in the DNA molecule (the duplex) and α − helixes and β − sheets formed by the polypeptide chains in proteins. At a larger distance scale, a biopolymer can adopt a defined 3D conformation: the so-called tertiary structure.

Nucleic acids

In nature, there are two type of nucleic acids, RNA (ribonucleic acid) and DNA (deoxyribonucleic).

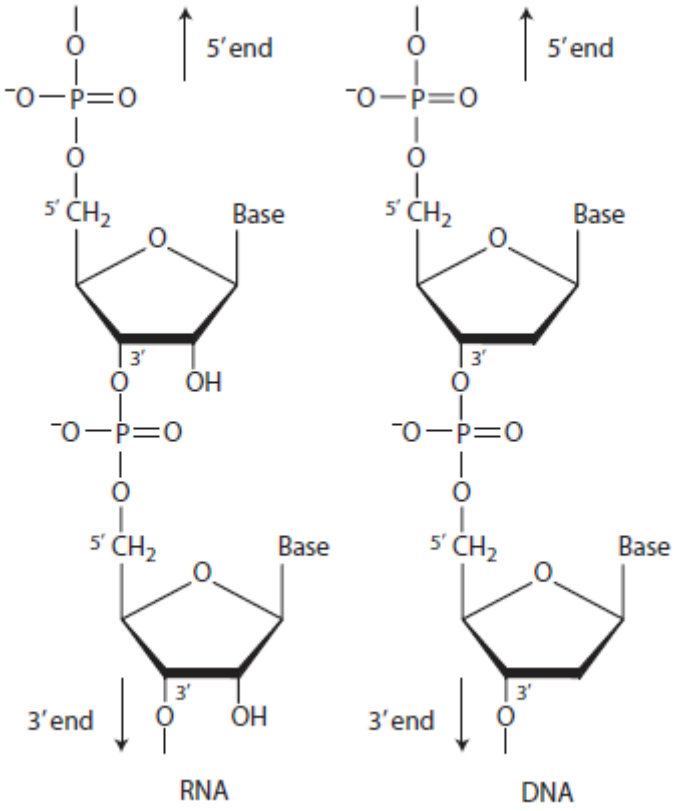


Figure 3. Chemical structures of ribonucleic acid (RNA, left) and deoxyribonucleic acid (DNA, right)

Nucleic acids are large biological molecules essential for all known forms of life. They include DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). Nucleic acids were discovered by Friedrich Miescher in 1869. Nucleic acids are linear polymers or chains of nucleotides, where each nucleotide consists of three components: a purine or pyrimidine base, a pentose sugar, and a phosphate group.

Since there is two type of nucleic acids in nature, they differ in the structure of the sugar in their nucleotides - DNA contains 2'-deoxyribose while RNA contains ribose. The bases found in the two nucleic acid types are different: adenine, cytosine, and guanine are found in both RNA and DNA, while thymine occurs in DNA and uracil occurs in RNA.

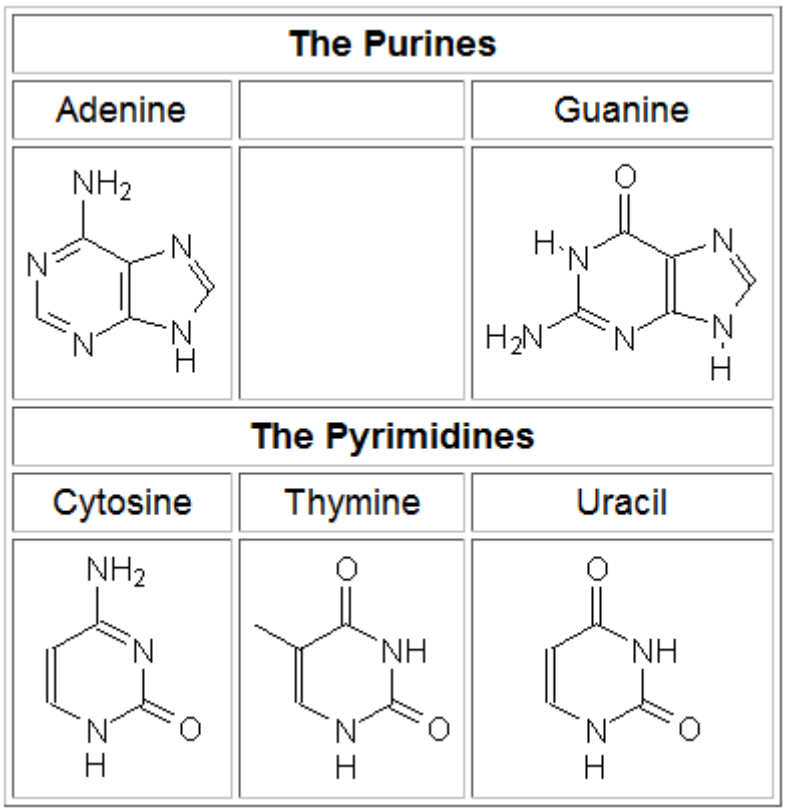


Figure 4. Chemical structures of bases, The Purines and The Pyrimidines

DNA and RNA are synthesized in cells by DNA polymerases and RNA polymerases. Short fragments of nucleic acids also are commonly produced without enzymes by oligonucleotide synthesizers. In all cases, the process involves forming phosphodiester bonds between the 3' carbon of one nucleotide and the 5' carbon of another nucleotide.

Most DNA exists in the famous form of a double helix, in which two linear strands of DNA are wound around one another. The major force promoting formation of this helix is complementary base pairing: A's form hydrogen bonds with T's (or U's in RNA), and G's form hydrogen bonds with C's.

The two strands of DNA are arranged antiparallel to one another: viewed from left to right the top strand is aligned 5' to 3', while the bottom strand is aligned 3' to 5'. This is always the case for duplex nucleic acids.

G-C base pairs have 3 hydrogen bonds, but A-T base pairs have 2 hydrogen bonds: one consequence of this disparity is that it takes more energy (a higher temperature) to disrupt GC-rich DNA than AT-rich DNA.

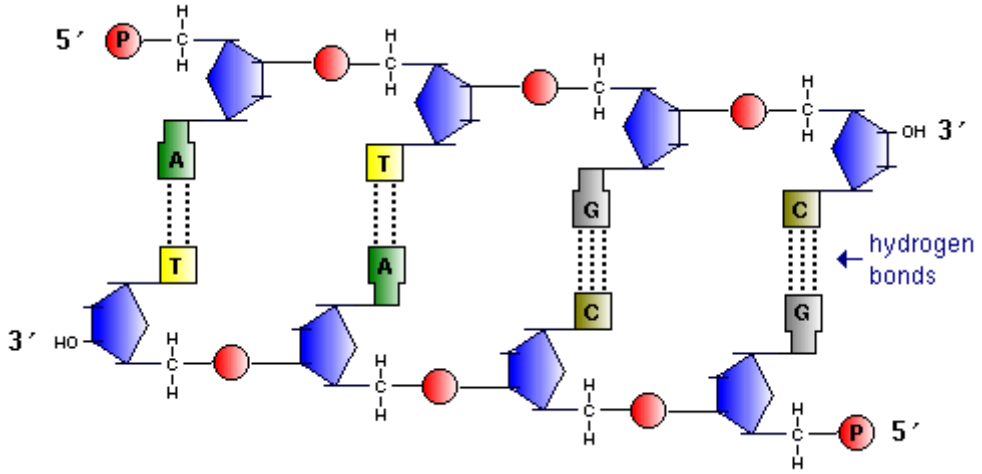


Figure 5. If we mix two ATGC's together, this will form following duplex

A nucleic acid sequence is a succession of letters that indicate the order of nucleotides within a DNA (using GACT) or RNA (GACU) molecule. By convention, sequences are usually presented from the 5' end to the 3' end. Because nucleic acids are normally linear (unbranched) polymers, specifying the sequence is equivalent to defining the covalent structure of the entire molecule. For this reason, the nucleic acid sequence is also termed the primary structure.

Secondary structure is the general three-dimensional form of proteins and nucleic acids (DNA/RNA). In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases.

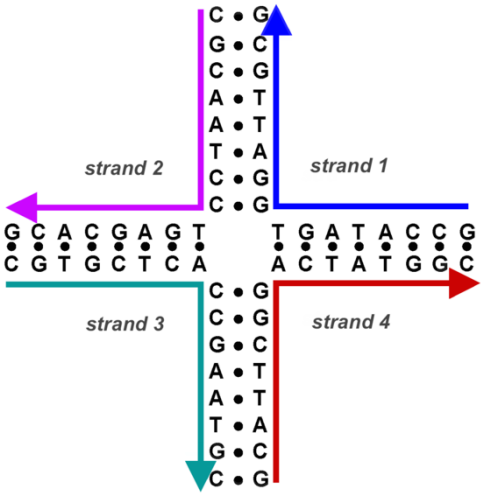


Figure 6. Nucleic acid design can be used to create nucleic acid complexes with complicated secondary structures such as this four-arm junction. These four strands associate into this

structure because it maximizes the number of correct base pairs, with A's matched to T's and C's matched to G's.

The precise three-dimensional shape od the molecule can vary, however, depending on the conditions in which the DNA is placed. There is A, B and Z form. The main distinguishing features of these different secondary structures of DNA are:

The A- and B-forms are right-handed and can be found in any sequence. B is the dominant form under physiological conditions. The A-form is found at low hydration levels, such as in spun fibres. The Z-form is lefthanded and occurs in alternating purine-pyrimidine sequences, particularly guanine-cytosine (GC).

The double-stranded duplex in the A–form is thick and compressed along the helix; in the Z–form it is elongated and thin whereas in the B–form it is intermediate.

Table 10.2 Characteristics of DNA secondary structures

Characteristic	A-DNA	B-DNA	Z-DNA
Conditions required to produce structure	75% H_2O	92% H_2O	Alternating purine and pyrimidine bases
Helix direction	Right-handed	Right-handed	Left-handed
Average base pairs per turn	11	10	12
Rotation per base pair	32.7°	36°	−30°
Distance between adjacent bases	0.26 nm	0.34 nm	0.37 nm
Diameter	2.3 nm	1.9 nm	1.8 nm
Overall shape	Short and wide	Long and narrow	Elongated and narrow

Figure 7. Characteristics of DNA secondary structures

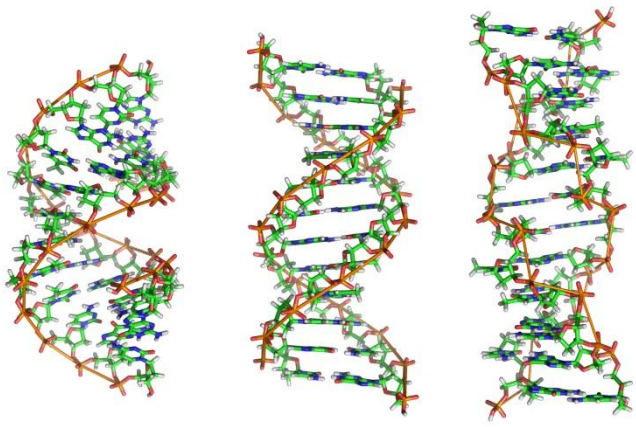


Figure 8. A, B and Z forms of DNA

Type of RNA

There are three types of RNA: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

mRNA is formed by transcription of complementary sections of DNA. After its formation, it is transported selectively from the nucleus to the ribosomes.

tRNA act as 'carriers' of amino acids during protein synthesis. They have a characteristic 'clover leaf' structure and they are relatively small, 73-93 nucleotides long.

rRNa is a structual component of ribosomes. The molecules consist of single strands of RNA. The specific function of rRNA is not fully established, but it binds proteins and mRNA to provide the site of proteine synthesis.

Molecular composition and size

Nucleic acids can vary in size, but are generally very large molecules. Indeed, DNA molecules are probably the largest individual molecules known. Well-studied biological nucleic acid molecules range in size from 21 nucleotides (small interfering RNA) to large chromosomes (human chromosome 1 is a single molecule that contains 247 million base pairs). Nucleic acids are linear polymers (chains) of nucleotides. Each nucleotide consists of three components: a purine or pyrimidine nucleobase (sometimes termed nitrogenous base or simply base), a pentose sugar, and a phosphate group.

The sugars and phosphates in nucleic acids are connected to each other in an alternating chain (sugar-phosphate backbone) through phosphodiester linkages. In conventional nomenclature, the carbons to which the phosphate groups attach are the 3'-end and the 5'-end carbons of the sugar. This gives nucleic acids directionality, and the ends of nucleic acid molecules are referred to as 5'-end and 3'-end. The nucleobases are joined to the sugars via an N-glycosidic linkage involving a nucleobase ring nitrogen (N-1 for pyrimidines and N-9 for purines) and the 1' carbon of the pentose sugar ring.

Nucleic acid hybridization

Hybridization is the process of complementary base pairs binding to form a double helix. Melting is the process by which the interactions between the strands of the double helix are broken, separating the two nucleic acid strands. These bonds are weak, easily separated by gentle heating, enzymes, or physical force. Melting occurs preferentially at certain points in the nucleic acid. T and A rich sequences are more easily melted than C and G rich regions. Particular base steps are also susceptible to DNA melting, particularly T A and T G base steps. These mechanical features are reflected by the use of sequences such as TATAA at the start of many genes to assist RNA polymerase in melting the DNA for transcription.

Protein

Proteins are polymers of amino acids covalently linked through peptide bonds into a chain. Within and outside of cells, proteins serve a myriad of functions, including structural roles (cytoskeleton), as catalysts (enzymes), transporter to ferry ions and molecules across membranes, and hormones to name just a few.

Proteins are polymers of amino acids joined together by peptide bonds. There are 20 different amino acids that make up essentially all proteins on earth. Each of these amino acids has a fundamental design composed of a central carbon (also called the alpha carbon) bonded to:

- a hydrogen

- a carboxyl group

- an amino group

- a unique side chain or R-group

Thus, the characteristic that distinguishes one amino acid from another is its unique side chain, and it is the side chain that dictates an amino acids chemical properties.

Amino acids

The 20 amino acids that are found within proteins convey a vast array of chemical versatility. The precise amino acid content, and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein. The chemical properties of the amino acids of proteins determine the biological activity of the protein.

The unique side chains confer unique chemical properties on amino acids, and dictate how each amino acid interacts with the others in a protein. Amino acids can thus be classified as being hydrophobic versus hydrophilic, and uncharged versus positively-charged versus negatively-charged. Ultimately, the three dimensional conformation of a protein - and its activity - is determined by complex interactions among side chains. Some aspects of protein structure can be deduced by examining the properties of clusters of amino acids.

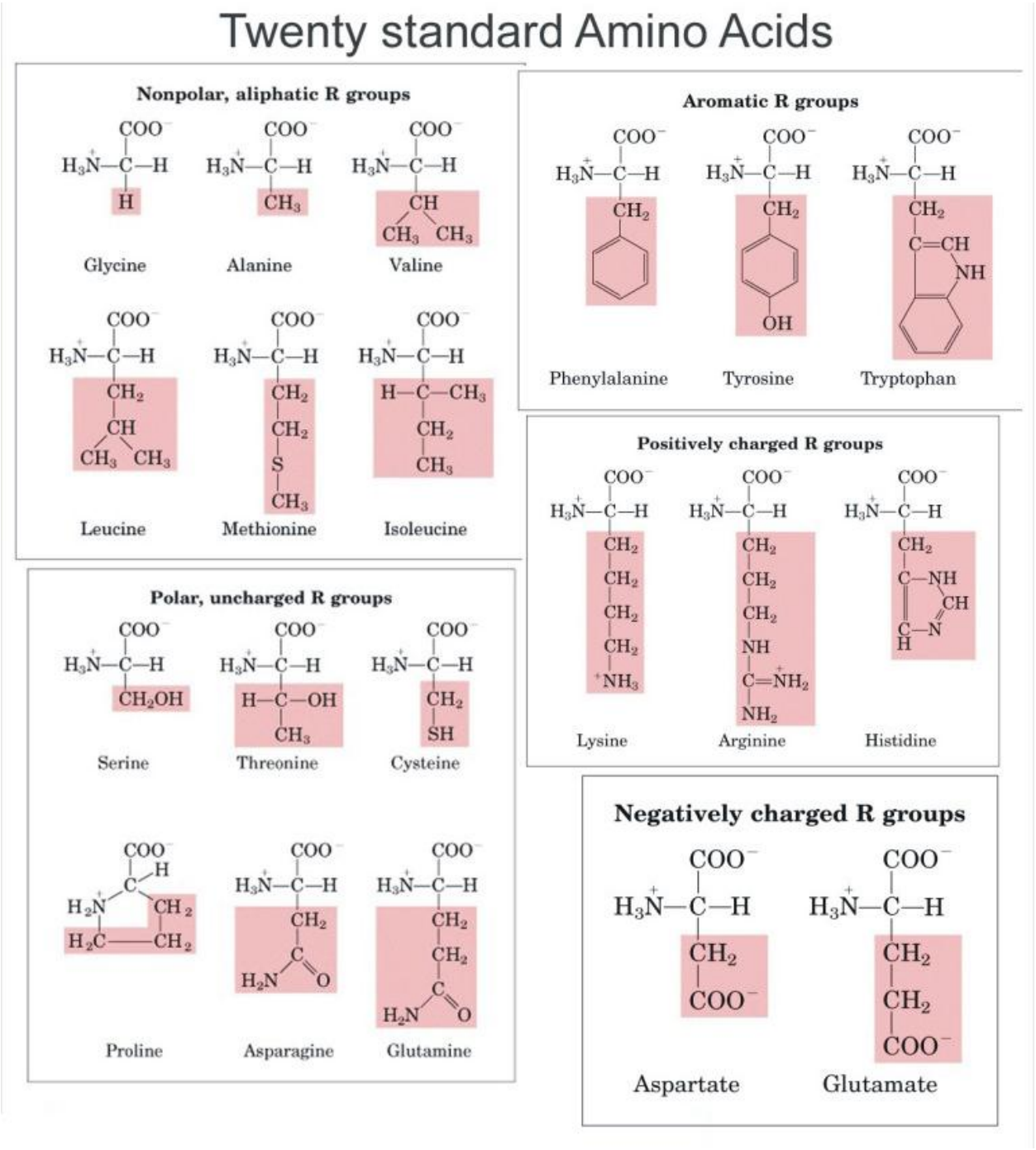


Figure 9. Table of amino acids

Peptides

If the chain length is short (say less than 30 amino acids) it is called a peptide; longer chains are called polypeptides or proteins. Peptide bonds are formed between the carboxyl group of one amino acid and the amino group of the next amino acid. Peptide bond formation occurs in a condensation reaction involving loss of a molecule of water.

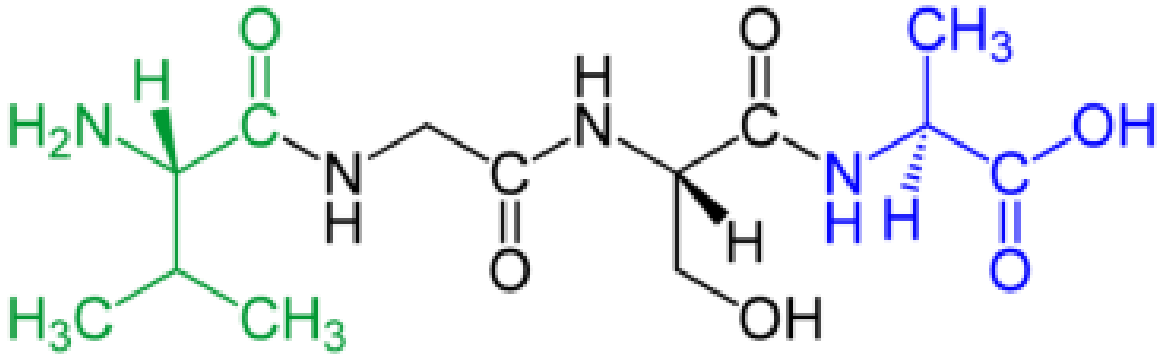


Figure 10. A tetrapeptide (example Val-Gly-Ser-Ala) with green marked amino end (L-Valine) and blue marked carboxyl end (L-Alanine).

The shortest peptides are dipeptides, consisting of 2 amino acids joined by a single peptide bond, followed by tripeptides, tetrapeptides, etc. A polypeptide is a long, continuous, and unbranched peptide chain. Hence, peptides fall under the broad chemical classes of biological oligomers and polymers, alongside nucleic acids, oligo- and polysaccharides, etc.

Peptides have recently received prominence in molecular biology for several reasons. The first is that peptides allow the creation of peptide antibodies in animals without the need to purify the protein of interest. This involves synthesizing antigenic peptides of sections of the protein of interest. These will then be used to make antibodies in a rabbit or mouse against the protein.

Levels of Protein Structure

Primary structure: the linear arrangment of amino acids in a protein and the location of covalent linkages such as disulfide bonds between amino acids.

Secondary structure: areas of folding or coiling within a protein; examples include alpha helices and pleated sheets, which are stabilized by hydrogen bonding.

Tertiary structure: the final three-dimensional structure of a protein, which results from a large number of non-covalent interactions between amino acids.

Quaternary structure: non-covalent interactions that bind multiple polypeptides into a single, larger protein. Hemoglobin has quaternary structure due to association of two alpha globin and two beta globin polyproteins.

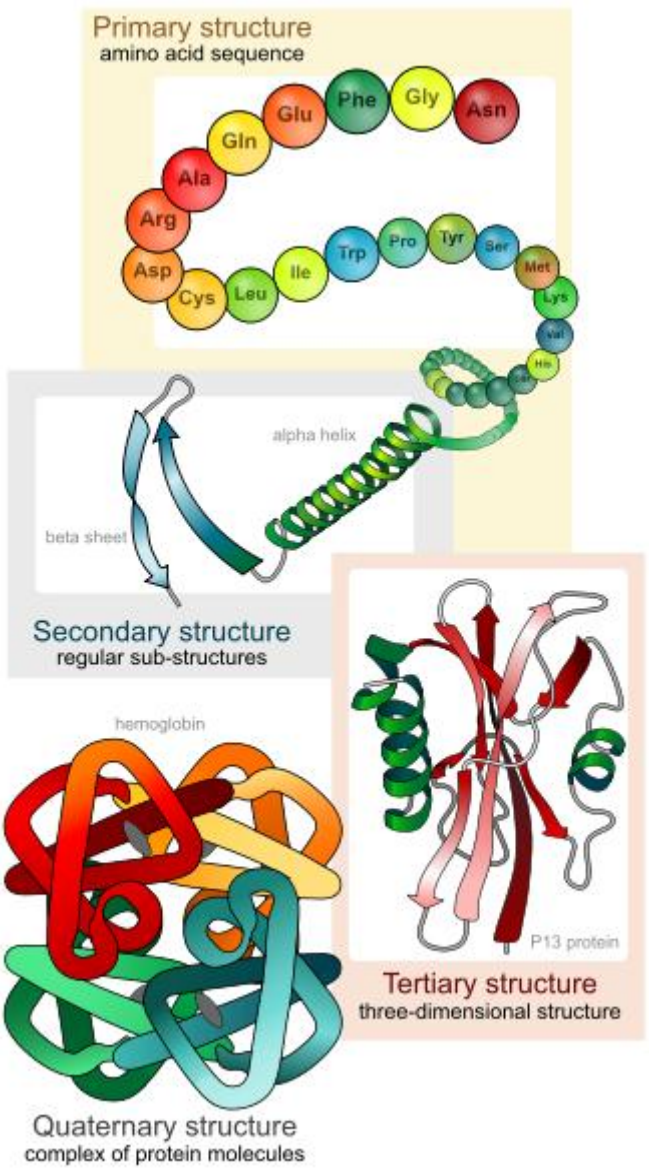


Figure 11. Protein structure, from primary to quaternary structure.

Domains, motifs, and folds in protein structure

Protein are frequently described as consisting from several structural units.

1. A structural domain is an element of the protein's overall structure that is self-stabilizing and often folds independently of the rest of the protein chain. Many domains are not unique

to the protein products of one gene or one gene family but instead appear in a variety of proteins. Domains often are named and singled out because they figure prominently in the biological function of the protein they belong to; for example, the "calcium-binding domain of calmodulin". Because they are independently stable, domains can be "swapped" by genetic engineering between one protein and another to make chimeras.

2. The structural and sequence motifs refer to short segments of protein three-dimensional structure or amino acid sequence that were found in a large number of different proteins.

3. The supersecondary structure refers to a specific combination of secondary structure elements, such as beta-alpha-beta units or helix-turn-helix motif. Some of them may be also referred to as structural motifs.

4. Protein fold refers to the general protein architecture, like helix bundle, beta-barrel, Rossman fold or different "folds" provided in the Structural Classification of Proteins database.

Despite the fact that there are about 100,000 different proteins expressed in eukaryotic systems, there are many fewer different domains, structural motifs and folds.

DNA, RNA and proteins

Comparison of DNA, RNA and proteins

	DNA	RNA	Proteins
Encodes genetic information	Yes	Yes	No
Catalyzes biological reactions	No	Yes	Yes
Building blocks (type)	Nucleotides	Nucleotides	Amino acids
Building blocks (number)	4	4	20
Strandedness	Double	Single	Single
Structure	Double helix	Highly complex	Highly complex
Stability to degradation	Extremely high	Variable	Variable
Repair systems	Yes	No	No

Figure 12. Table of properties of DNA, RNA and proteins

Why DNA is best for encoding genetic information

DNA and RNA are both capable of encoding genetic information, because there are biochemical mechanisms which read the information coded within a DNA or RNA sequence and use it to generate a specified protein. On the other had, the sequence information of a protein molecule is not used by cells to functionally encode genetic information.

DNA has three primary attributes that allow it to be far better than RNA at encoding genetic information. First, it is normally double-stranded, so that there are a minimum of two copies of the information encoding each gene in every cell. Second, DNA has a much greater stability against breakdown than does RNA, an attribute primarily associated with the absence of the 2'-hydroxyl group within every nucleotide of DNA. Third, highly sophisticated DNA surveillance and repair systems are present which monitor damage to the DNA and repair the sequence when necessary. Analogous systems have not evolved for repairing damaged RNA molecules.

Why proteins are best for catalyzing biological reactions

The single-stranded nature of protein molecules, together with their composition of 20 or more different amino acid building blocks, allows them to fold in to a vast number of different three-dimensional shapes, while providing binding pockets through which they can specifically interact with all manner of molecules. In addition, the chemical diversity of the different amino acids, together with different chemical environments afforded by local 3D structure, enables many proteins to act as enzymes, catalyzing a wide range of specific biochemical transformations within cells. In addition, proteins have evolved the ability to bind a wide range of cofactors and coenzymes, smaller molecules that can endow the protein with specific activities beyond those associated with the polypeptide chain alone.

Why RNA is multifunctional

RNA encodes genetic information that can be translated into the amino acid sequence of proteins, as evidenced by the messenger RNA molecules present within every cell, and the RNA genomes of a large number of viruses. The single-stranded nature of RNA, together with tendency for rapid breakdown and a lack of repair systems means that RNA is not so well suited for the long-term storage of genetic information as is DNA.

In addition, RNA is a single-stranded polymer that can, like proteins, fold into a nearly infinite number of three-dimensional structures. Some of these structures provide binding sites for other molecules and chemically-active centers that can catalyze specific chemical reactions on those bound molecules. The limited number of different building blocks of RNA (4 nucleotides vs >20 amino acids in proteins), together with their lack of chemical diversity, results in catalytic RNA (ribozymes) being generally less-effective catalysts than proteins for most biological reactions.

Histones

Histones are proteins, family of basic proteins, whic are in interaction with DNA in the nucelus and help condense it into chromatin. Nuclear DNA does not appear in free linear strands, it is highly condensed and wrapped around histones in order to fit inside of the nucleus and take part in the formation of chromosomes. Positive charges of histones allow them to interact with DNA, which is negatively charged. Some histones function as spools for the thread-like DNA to wrap around. Under the microscope in its extended form, chromatin looks like beads on a string. The beads are called nucleosomes. Each nucleosome is made of DNA wrapped around eight histone proteins that function like a spool and are called a histone octamer. Each histone octamer is composed of two copies each of the histone proteins H2A, H2B, H3, and H4. The chain of nucleosomes is then wrapped into a 30 nm spiral called a solenoid, where additional H1 histone proteins are associated with each nucleosome to maintain the chromosome structure.

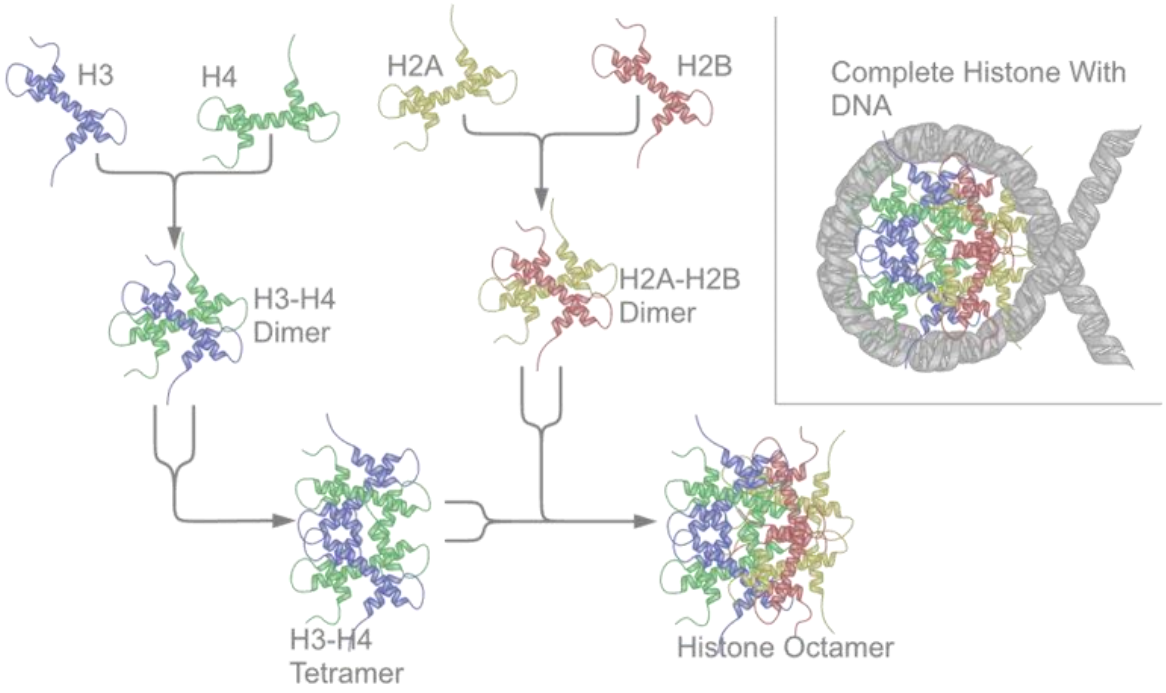
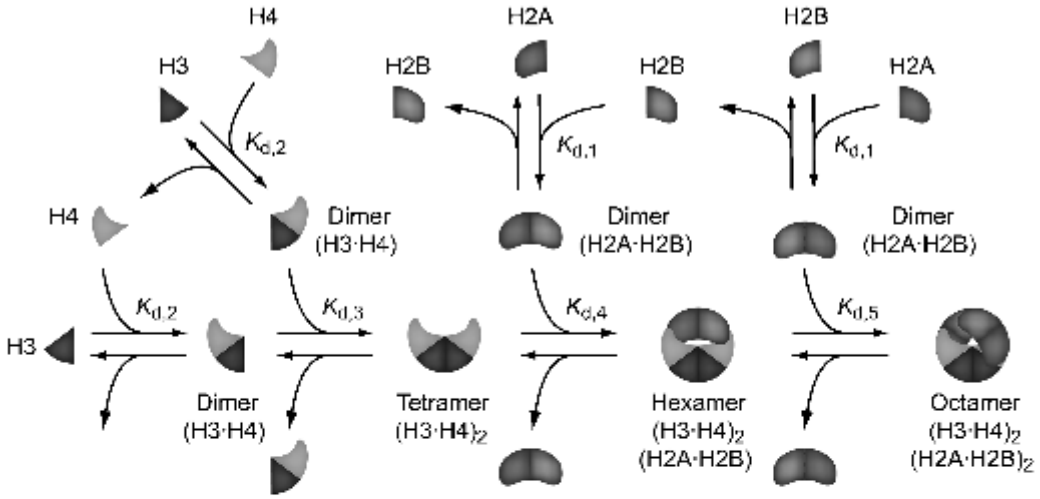


Figure 13. Schematic representation of the assembly of the core histones into the nucleosome.



$$H2A + H2B \overset{K_{d,1}}{\rightleftharpoons} (H2A{\cdot}H2B) \qquad \text{(Eq. 1)}$$

$$H3 + H4 \overset{K_{d,2}}{\rightleftharpoons} (H3{\cdot}H4) \qquad \text{(Eq. 2)}$$

$$2\,(H3{\cdot}H4) \overset{K_{d,3}}{\rightleftharpoons} (H3{\cdot}H4)_2 \qquad \text{(Eq. 3)}$$

$$(H3{\cdot}H4)_2 + (H2A{\cdot}H2B) \overset{K_{d,4}}{\rightleftharpoons} (H3{\cdot}H4)_2(H2A{\cdot}H2B) \qquad \text{(Eq. 4)}$$

$$(H3{\cdot}H4)_2(H2A{\cdot}H2B) + (H2A{\cdot}H2B) \overset{K_{d,5}}{\rightleftharpoons} (H3{\cdot}H4)_2(H2A{\cdot}H2B)_2 \qquad \text{(Eq. 5)}$$

Figure 14. Reaction scheme for assembly of the histones octamer. The equiibrium for each step is described by the dissociation constant Kd. The designation (H2AxH2B) represents the heterodimer between H2 and H2B, (H3xH4) the heterodimer between H3 and H4, $(H3xH4)_2$ the tetramer, $(H3xH4)_2$(H2AxH2B) the hexamer and $(H3xH4)_2$(H2AxH2B)2 the octamer complex.

Histones interaction with DNA

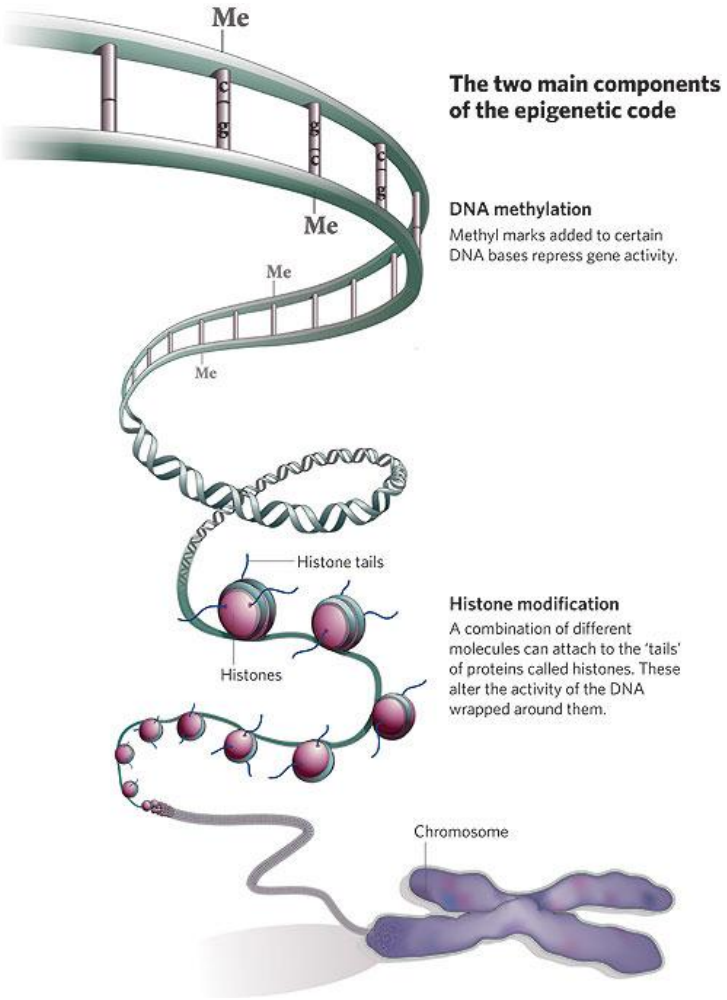


Figure 15. Interaction histone with DNA

Histones contain a large proportion of the positively charged (basic) amino acids, lyseine and arginine in their structure and DNA is negatively charged due to the phosphate groups on its backbone. These result of these opposite charges is strong attraction and therefore high binding affinity between histones and DNA. Hydrogen bonding involving hydroxyl amino acids in the histone peptide and the phosphodiester backbone of DNA and are also important in further stabilizing the structure. One of the advantages of histones interacting mainly with the backbone of DNA is it means the interaction is not sequence dependent. This means that despite an apparent preference of histones to some sequences of DNA, they are able to bind anywhere.

The packaging of DNA into nucleosomes shortens the fiber length about sevenfold. In other words, a piece of DNA that is 1 meter long will become a 'string-of-beads' chromatin fiber

just 14 centimeters long. Despite this shortening, a half-foot of chromatin is still much too long to fit into the nucleus, which is typically only 10 to 20 microns in diameter. Therefore, chromatin is further coiled into an even shorter, thicker fiber, termed the '30-nanometer fiber', because it is approximately 30 nanometers in diameter. Histone H1 is very important in stabilizing chromatin higher-order structures, and 30-nanometer fibers form most readily when H1 is present.

Core particles is made of 2 molecules of each histone protein: H2A, H2B, H3, H4, and linker histones H1 and H5. Core proteins have similar central domain composed of three alpha-helicase connected by two loops. There are 14 sticking points on the octamer surface where wrapped DNA contacts the octamer core. Energy of adsorpion is in order of 1.5 - 2 k_bT. That is the energy which is left after DNA has been around the octamer to make contact with the sticking point. The elastic energy needed to bind 127 bp of DNA around octamer is given by:

$$\frac{E_{elastic}}{k_BT} = \frac{l_P l}{2R_0^2}$$

l is bent part of wrapped DNA, R0 is the radius of curvanture of the wrapped DNA. When equation is solved with adding all known qantities, we get bending energy per sticking point to be in order of 58 k_bT. This is aproximatelly 4 k_bT per sticking point. Together with the observation then the net gain per sticking point is around 2 k_bT. It is concluded that the pure adsorption energy is on average 6 kbT per binding site.

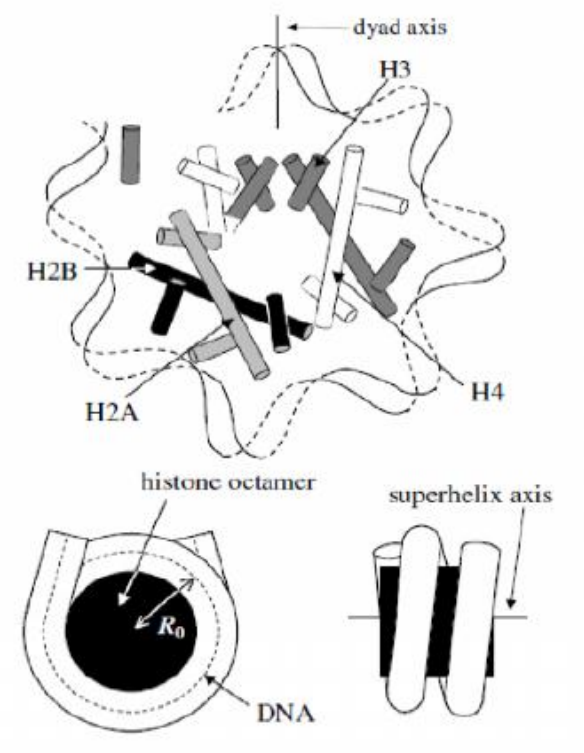


Figure 16. Schematic view of the nucleosome core particle. Top: upper half of the 8 core histones and the nuceosomal DNA. Bottom: a simplified model where the octamer is replaced by the cylinder and DNA by the worm-like chain (WLC). Also indicated are dyad axis and DNA superhelix axis.

The nucleosome is a complex in which 147 bp of the DNA are wrapped in 1.67 turns around the histone octamer. Most nucleosomes are augmented by linker histones to a complex referred to as a chromatosome. Many experiments have been carried in order to characterize the DNA binding of the H1 group of proteins. It was demonstrated that linker histones bind cooperatively to linear double-stranded DNA and also linker histones can form large complexes with relatively long DNA fragments hinting at more than one DNA binding site for the histone. The other putative DNA binding domain is a loop in the globular domain which is less conserved and comprises a stretch of basic amino acids at the opposite surface of the globular domain. A few models have been proposed for the integration of H1 and/or its globular domain in the nucleosomal structure, and three of these are shown for the globular domain in figure 17. Binding of linker histone directly affects the geometry of the DNA entering/exiting the nucleosome. This will translate into changes of the higher order chromatin structure upon binding of linker histone. It is conceivable that multiple positions can be adopted by linker histones, which could explain the divergent findings of several groups.

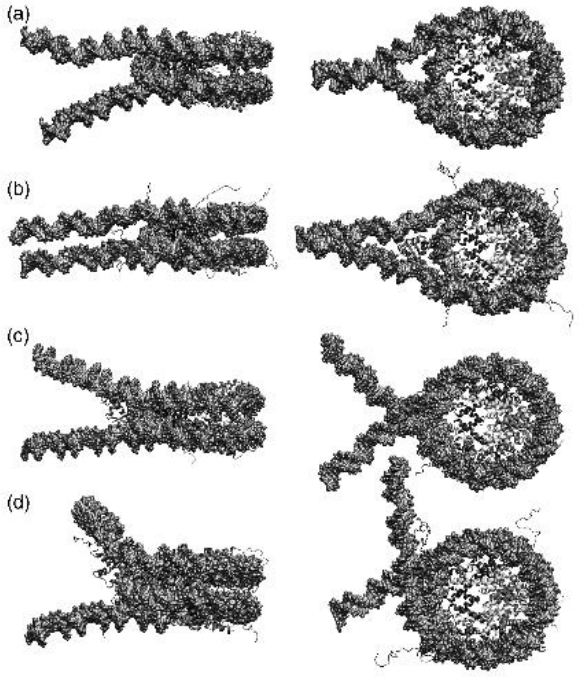


Figure 17. Model structure of the linker histone H1/H5 globular domain bound to the nucleosome. (a) For comparison a nucleosome with 199bp of DNA was extracted from the tetranucleosome crystal structure. The interactions between the nucleosomes led to some bending of the linker DNA. (b) Model for the globular domain interacting with the nucleosome derived from a chromatosome structure with full-length H1.

Literature

Introduction to biopolymer physics, Johan R.C. van der Maarel

Lehninger Principles of Biochemistry, 5th Edition, Nelson, Cox

The physics of chromatin, Helmut Schiessel, J. Phys.: Condens Matter 15 (2003) R699-R774

http://biology.about.com

http://www.vivo.colostate.edu/hbooks/genetics/biotech/basics/chem.html

http://en.wikipedia.org/wiki/DNA,_RNA_and_proteins:_The_three_essential_macromolecules_of_life

http://en.wikipedia.org/wiki/Protein

http://www.mun.ca/biology/scarr/A_B_Z_DNA.html

http://www.sbs.utexas.edu/genetics/Fall05/Handouts/ABZ-DNA.pdf

http://www.biology.arizona.edu/biochemistry/problem_sets/aa/aa.html

http://strength-health-alliance.com/eating-for-strength-and-health-part-ii-protein-biochemistry-public-health-and-athletic-performance-part-a/

http://en.wikipedia.org/wiki/Histone