DRAFT--Oct. 15, 1997
For Conference Distribution Only

IN THE GENOME, SYMMETRY SEEMS TO CODE FOR SYMMETRY

by

Guido Pincheira
Professor of Genetics
University of Chile
Santiago, Chile

The Twenty-first International Conference on the Unity of the Sciences
Washington, D.C. November 24-30, 1997

Committee 2

Symmetry in its various aspects: search for order in the Universe.

DRAFT 9/30/97

For Conference Distribution only

In the genome, symmetry seems to code for symmetry

Prof. Guido Pincheira V.
Universidad de Chile

Symmetry is a very general phenomenon in nature. It exists in many different ways at the different levels of organization of matter and may be defined as the "harmony or balance in the proportions of parts to the whole" (2).

Biology provides countless situations in which structural and physiological symmetries are present in the different levels of organization of living matter. Moreover, biological symmetries are very often broken giving rise to assymmetries or to new structural and operational symmetries. Symmetry, as such, appears as an interesting characteristic of life associated with the periodicity, regularity and harmonious arrangement of structures that are the essence of profound developmental and evolutionary connotations for the organisms.

The very rich variability of biological symmetry may be considered related to the paths imposed by evolution. Symmetry in micromolecules, such as H_2O; NH_3. may have

appeared in prebiotic phenomena as the result or consequence of physicochemical events. A better performance of this molecules may have been related to the forces and energy fields present in the common matrix in which they were embedded. These forces, specially electromagnetism must have been responsible for a rich variety of structural symmetries that later, gradually, influenced the organization of larger biomolecules, structures, cells and organisms.

In fact, the symmetry of molecular structures stem from the symmetries present in the forces that produced them. As such, the symmetry elements of small molecules must have influenced the processes of formation of aminoacids, nitrogen bases, sugars etc. and, subsequently, the processes of polimerization to form proteins, nucleic acids of certain characteristics. The structural and dynamical symmetries of these new molecules are, of course, inherent to the dynamical properties of the structures they form, such as de genome, which contains the coded information for the transient or permanent characteristics of the organisms.

The genome.

The genome is made of nucleic acids (DNA; RNA) which in many cases are associated with proteins of different types.

When studying the structural and physiological characteristics of the genome one cannot be more impressed by its structural regularity and the precise dynamics exhibited. Such characteristics generate a sense of beauty to the observations and interpretations of genetic phenomena.

DNA (desoxiribonucleic acid)

a) Structural characteristics.

A double helix is the usual representation of the structure of the DNA molecule (Fig 1). For quite a long time, the DNA double helix was looked upon as a very static structure, with very uniform characteristics and a considerable degree of rigidity. This was in accordance with its role of serving as a true blueprint to make new DNA or RNA for the different needs of the organism. Today, the new technologies available to study the molecular aspects of genetics have demonstrated that this is not the case. Instead, the DNA double helix is characterized by a considerable conformational flexibility and a very dynamic structural heterogeneity .

The DNA double helix is formed by two antiparallel polinucleotide helical chains, coiled with a clockwise (dextral) axial rotation based on the tilt imposed by the staggering of many nucleotide pairs. These helical chains (excluding the nitrogen bases) are related by a dyad perpendicular to the helix axis. The dyad operation is restricted to the glycosyl-C1-N-linkages and it is not applicable to the bases. It is therefore called a pseudo-dyad.

The structural studies of natural DNA generally means to deal with heterogeneous populations of molecules, being difficult to obtain them as single crystals, suitable for x-ray analysis. Nevertheless, synthetic oligonucleotides of a defined sequence may be crystallized alone or forming complexes with metal ions, drugs or proteins bound with them. X-ray crystallography can yield precise structure for molecules of weight up to 2000 daltons. This correspond to hexanucleotides (350 daltons = average MW/nucleotide). A few DNA oligomers have been atomically resolved, and they exhibit a perfectly regular structure with asymmetric units consisting of one nucleotide (3). In addition to the high conformational

flexibility of one nucleotide, it has been possible to detect the same characteristic at the junction of two neighbor nucleotides in a polymer.

The crystal structure of a B-DNA dodecamer has been studied at high resolution. Nevertheless, it is too short and irregular to precisely estimate the pitch, diameter and offset of a large DNA molecule, made of many thousands of nucleotides.

Symmetry elements are found in these synthetic polynucleotides. They are related to the glycosil bonds linking the nitrogen bases with sugars and to the attached sugar phosphate backbones. At least, two cyclic hydrogen bonds must form to produce a stable nitrogen base pair, which may display two kinds of two-fold symmetry, depending on the orientations of the -glycosyl C-N bonds. Two-fold rotation axis can be arranged, either perpendicular to or within base-pair planes. This has an influence upon the orientation of the attached sugar-phosphate groups or backbones of DNA (5).

The outer DNA double helix is not cylindrically smooth, but displays two grooves of different width and depth (Fig. 1). They are called the major (12 A° wide, 8.5 A° deep). and the minor (6 A° wide, 7.5 A° deep) grooves. As we shall see later, they play an important function in the interaction with other molecules, specially proteins.

Opposite rotational torque (left handed helix) may occur in DNA under physiological conditions, forming what is called Z-DNA. Its coexistence with B-DNA has been demonstrated, suggesting the occurrence of variations of DNA structure in particular regions of the same molecule. In fact, the structure of Z-DNA is formed in nucleotide sequences in which there is an alternation of purines and pyrimidines, such as CpGpCpGpCpGp, indicating that DNA codes its own conformational flexibility.

Other forms of DNA structure are know as A-DNA, C-DNA and D-DNA. They have been obtained in experimental conditions: A-DNA and C-DNA have also a dextral

rotation and are found in any sequence of nucleotides. These findings indicate that DNA has the possibility to adopt different types of conformations, depending on physiological factors (humidity, temperature, ionic strength) that may have an influence upon it.

A linear DNA molecule in solution, usually adopts the B configuration, with ten nucleotides pairs per turn of the double helix. This is the minimal energy configuration. If the molecule experiences additional coiling, its energy increases. Such effect may be minimized by the adoption of new structural configurations, such as a closed circularization, supercoiling, etc. Many of these configurational changes are made possible by the action of enzymes called topoisomerases and girases, which are synthesized under the control of the genome. This means that DNA is able to adjust its torsional pressure by adjusting the structure of the double helix itself.

An extreme situation of this type of molecular structural adjustment would be the local or regional conversion of a right-handed helix into a left-handed helix. Structural fluctuations in the DNA double helix are due to intrinsic conformational preferences derived from its chemical constitution and the competing influence of the components of the immediate aqueous ionic environment.

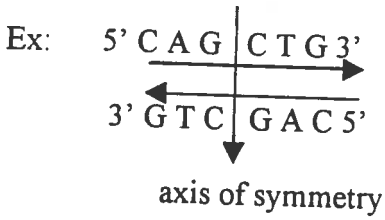
Additional structural complexity of the genome is derived from the twisting of DNA molecules around its axis, a characteristic designated as "supercoiling" because it generates other helical structures very important to accommodate the genome in the chromosomes, specially in higher organisms.

These new structures have, of course, important consequences for DNA physiology. In fact, different types and degrees of DNA supercoiling affect the physical (hydrodynamic behavior), chemical (sensitivity to enzymes and mutagens, etc.) and biological properties (replication, transcription, repair, recombination) of the genome.

The DNA molecule may have many regions with elasticities varying by a 3-4 factor, different energies, etc. All of these may affect processes the kinetic and the binding to other molecules.

b) Organization.

Specially important for certain aspects of DNA physiology is the existence of inverted sequences of nucleotides, called palindromes, in which a sequence of nucleotides read in the 5' 3' direction of one DNA strand is identical to a sequence of nucleotides read from the other end of the palindrome on the other strand in the 5'→3' direction.



A DNA palindrome is a region of dyad symmetry, in which the axis of symmetry identifies the central point about which the sequence of nucleotides is the same on either side.

As we shall see later, palindromes play a very important role in the process of recognition of DNA by enzymes (nucleases) that are able to cleave this molecule in the processes of replication, transcription or repair of the genome.

RNA (ribonucleic acid)

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

In the case of RNA, symmetry is also derived of palindromic nucleotide sequences that make possible that single RNA strands may fold back on itself promoting intrastrand H bonding, forming loops or hairpins, giving rise to new symmetry elements which have very important biological implications.

Perhaps, the most thoroughly studied RNA is transfer RNA (tRNA). These molecules are polinucleotide chains made of 75-90 units, with a molecular weight around 26.000 daltons. They can be crystallized from aqueous medium by addition of alcohol or salts. These tRNA crystals contain 30-80% solvent and may be considered a concentrated solution.

The primary sequences of nucleotides of tRNA give rise to secondary structures that ressemble a cloverleaf form, with four double-helical stems and three loop regions (see Fig. 2).

Each one of the four stem regions of this cloverleaf contains 4-7 base pairs organized as double helices. They are called the acceptor (it accepts the aminoacid that will be transported to the site of protein synthesis), the anticodon (it matches the codon in mRNA), the D (for having dihydrouridine) and the T (for having ribothymidine) stems.

The subsequent folding of the cloverleaf into a tertiary structure confers an L shape to the tRNA molecule (Fig. 3). The L shape is composed of two nearly perpendicular RNA helices, each one having a length of 70 A° and a thickness of 20 A°. The two ends of the L

configuration are: the acceptor CCA triplet that will bind the aminoacid, and the anticodon loop that will recognize the messenger RNA codon in the process of protein synthesis.

The two arms of the L tRNA tertiary configuration may be separated by a pseudo-dyad axis of symmetry that bisects the L angle and relates the anticodon/D helix to the acceptor/T helix, a very important condition for the process of protein synthesis.

In fact, this aspect of tRNA symmetry seems to be very important for the interaction with the enzymes responsible to carry the aminoacids to the acceptor loop. The substrate recognition is facilitated and, perhaps, it may give rise to an enzyme-tRNA complex with a two fold symmetrical structure.

Similar situations with symmetry characteristics occur in the other types of RNA, specially rRNA.

Since the characteristics of these molecules are coded in the genome, it is possible to indicate that the most important break in their symmetry properties is derived from the effect that mutations (changes of nucleotides in DNA) may have upon them. These changes are the primary source of variations feeding the process of evolution of the organisms, in which new symmetries may arise. From the huge variety of possible new symmetries, nature selects only a few and tries to preserv them as much as they proove to be advantageous for living processes.

PROTEINS

Long polypeptide chains have also an helical configuration with an helical rotation derived from the tilt of peptide bonds between L aminoacid residues.

The α-helix is the most important regular configuration in the polypeptidic chain. Nevertheless, the presence of irregular lateral groups of aminoacids connected to the regular central skeleton determines that the protein structure is, usually, irregular. Practically there are no proteins with a simple helical structure. In fact, proteins may have helical regions and non-helical regions. One reason is that one of the aminoacids, proline, has no amino group, and when present, H bonding is interrupted. Another reason is the formation of -S-S linkages between cysteine residues, producing distortions of the helical structure. A third factor that determines irregularities of proteins is the diverse chemical nature of aminoacid lateral groups, which will tend to form the most favorable secondary interactions from the energy point of view.

Some proteins may have no helical structure at all. Proteins, such as silk, and specially those rich in glycine and alanine, exhibit a secondary laminar structure, called β-sheet, formed by fully extended polypeptidic chains, held together by H bonds between groups of the different chains. A protein β-sheet may exhibit a parallel or an antiparallel disposition. The last possibility happens when an extended polypeptide folds back and forth upon itself, satisfying the requirements of a dyad symmetry.

In addition, the possibility of proteins to fold up gives rise to tertiary structures and increases the number of possible shapes in proteins, even in those of small size. In biological conditions the protein usually adopts one of these-conformations, generally the most energetically convenient.

Although it is possible to figure out or to get pictures of three dimensional structures of proteins illustrating the positions of all atoms, we still do not understand clearly how proteins fold up.

Besides, the idea that these structures have a high degree of rigidity is not correct. Instead they should be considered as highly dynamic structures, involved in cooperative phenomena between similar or different polypeptide units in which symmetry may be broken or may emerge playing an important physiological role.

Similar to RNAs, proteins are also coded in the genomic DNA. The information is transmitted via mRNA and processed by tRNAs and rRNAs.

Isomorphism between nucleotide sequences in DNA, RNA and aminoacid sequences in polypeptide chains is a symmetrical phenomenon. The transformation take place thanks to the genetic code that makes possible the formation of new and identical DNA, complementary RNAs and proteins through mechanisms in which symmetries code for symmetries.

PROTEIN-NUCLEIC ACID INTERACTIONS

It is clear that the properties or functions of the genome can not be understood in isolation or without consideration of its interdependence with many other molecules of the organism.

The interaction between nucleic acids and proteins is one of the central processes in all living cells. The interactions occur at all levels of nucleic acid physiology (replication, transcription, translation, repair, etc.).

Proteins may interact with single or double stranded nucleic acids. They also may interact with segments of nucleic acids that adopt a globular structure.

It is also possible to distinguish interactions that are based in the specific recognition of a protein for particular sequences of nucleotides from interactions that do not discriminate special sequences of nucleotides. In the last case, the protein must primarily

recognize structural details of the sugar-phosphate backbone of the nucleic acid. One type of compatibility is the distance of 7 A° between adjacent phosphate groups of the double strand DNA backbone, when measured on the outer perimeter of the phosphate groups, and an equivalent distance between two positively charged groups of arginine or lysine in a protein α-helix, when separated by three other aminoacids.

Many DNA binding proteins contain two α helical recognition units related by a two fold axis of symmetry and separated by 34 A°, equivalent to the pitch of the B-DNA double helix. A similar dimensional correspondence exists between double stranded DNA and protein antiparallel β-sheets of fully extended polypeptides.

This intermolecular compatibility may be related to elements of rotational symmetry in the DNA and the protein. As already indicated, A and B DNA have two types of pseudo twofold axis per base pair, one on the plane of each base pair and one in between two adjacent base pairs. A protein antiparallel β-sheet also exhibits the two types of pseudo two-fold axis (6).

So, we have structural elements in double stranded nucleic acids and proteins exhibiting similar dimensions and symmetries.

These compatibilities facilitate the formation of hydrogen bonding between the backbone of a peptide and the backbone of the nucleic acid. Several models have been proposed to explain a general type of double stranded DNA-protein recognition.

Portions of each base pair as a constituent of a DNA molecule are exposed in two separate grooves. Each of the four possible base pair arrangements (A-T; T-A; G-C or C-G) may he recognized by the different arrangements of the protruding atoms, allowing a unique hydrogen bonding pattern for each of the four base pairs.

When one examines the contact region of a major groove of DNA with the α-helix of a protein, it is possible to find structural alignments in which certain C atoms of the α-helix are roughly coplanar with a nitrogen base pair plane. This situation may facilitate the formation of H bonds according a sort of "recognition code" between aminoacids and nucleotides in which a nucleotide may be recognized by different aminoacids, and each aminoacid may bind to different nitrogen bases. This interaction may be crucial for the initiation or termination of many genetic activities

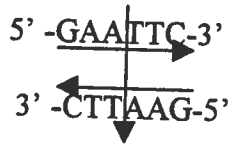
In addition, steric recognition seems to play a very important role. In this respect, the most important group would be the CH3 group of thymidine, because it has the largest accessible area of any group on the major groove.

A similar situation takes place in the interaction of proteins with nucleotides in the minor groove of B-DNA.

Specific DNA-protein interactions.

a) Interaction with enzymes. The action of nucleases.

The nucleases are enzymes that cleave nucleic acids. Some endonucleases recognize and cleave DNA regions characterized by the presence of symmetrical sequences of nucleotides, particularly tetra or hexanucleotides. As indicated previously, these sequences are called palindromes and exhibit a two-fold axis of symmetry and a two-fold rotational symmetry. An example is the enzyme Eco RI, which recognizes the following DNA palindromic sequence:



a) Interaction of B-DNA which cro repressor.

The cro protein is the product of a gene in the bacteriophage lambda. It is a polypeptide made of 66 aminoacids. This molecule is able to recognize, at least, 6 near-palindromic sequences in double stranded DNA of phage lambda. After binding the DNA, it prevents the expression of several genes.

The structure of the peptide cro consists of three α-helices and three antiparallel β-sheets. When binding to DNA, cro protein may be a dimer or a tetramer. The dyad axis of symmetry of a cro dimer coincides with the dyad symmetry of the near-palindromic region of phage lambda B-DNA to which it binds, facilitating the formation of hydrogen bonds between certain aminoacids and the nucleotides.

b) Interaction of B-DNA with the catabolite gene activator protein (CAP).

The catabolite activator protein (CAP) activates several genes in a very specific way. In order to do that, it requires to be activated by cyclic adenosine monophosphate (cAMP), which acts as an allosteric effector.

The activation of gene expression in bacteria by CAP must include the binding of cyclic AMP to CAP and the interaction of this complex with segments, called promoters, in the target genes.

When the intracellular levels of cAMP increases, CAP activates the expression of bacterial genes coding for enzymes necessary for the catabolism of lactose, arabinose, maltose and other sugars.

The cAMP-CAP complex makes its major contacts with a DNA region that includes two successive major grooves, the minor groove situated between them and the minor grooves situated on each side of the DNA region (4).

The molecular structure of CAP.

CAP is composed of two identical polypeptide subunits, each made of 210 aminoacids (1). The two subunits are related by a dyad axis of symmetry (see Fig. 4).

CAP has two domains: a carboxyl-terminal, or DNA binding domain and a larger amino-terminal domain, responsible for: a) the interaction with cAMP; b) the interaction between the subunits to form the dimmer and, c) probably the regulation of the whole binding process with DNA.

The carboxyl-terminal domain of each CAP subunit consist of three α-helices, designated as D, E and F. connected by short β-sheet structures. In each subunit, the F α-helices are thought to provide the major interaction sites with the target DNA.

DNA structure for binding with CAP.

The segment of DNA for binding with CAP has an extension of 28 nucleotide pairs. In this stretch of DNA we have:

A nucleotide sequence 5' TGTGA 3', which is critical for CAP binding,

A block of 6 nucleotides that shows little sequence preferences,

A second sequence that contains an inverted repeat of the TGTGA motif.

In some cases, the sequence of 6 nucleotides generates a two-fold symmetry in the DNA sequences. This symmetry is clearly functional in the gene lac, because mutations of this gene prevent the binding of CAP (1).

Experimental results indicate that, the binding of CAP to the lac gene, requires that the symmetrically arranged sequences of nucleotides must recognize the symmetrically protein configuration. The affinity of CAP for DNA appears to be higher when these sequences of nucleotides are symmetrically arranged than when they are not symmetrical. The two peptides in the CAP dimer are in near-parallel arrangement and related by a pseudo-dyad axis running between them. An adequate interaction for a CAP dimer with DNA may depend on the coincidence of the dyad axis of the protein with the dyad axis of the DNA palindromic sequence.

The binding of cAMP-CAP complex to a fragment containing a lac promoter is favored by 4 kcal/mole over the binding to fragments that do not contain a binding site.

A hierarchy of CAP binding affinities to different DNA regions may explain why different cAMP sensitive points are activated by different concentrations of cyclic AMP *in vivo*.

EVOLUTIONARY ASPECTS

x-ray diffraction studies of the molecular structures of cro and of an other repressor, called cI, also in phage lambda, and CAP in the bacterium *Escherichia coli* reveal a striking structural homology in the relative position and orientation of two consecutive α-helices (E and F) in the carboxyl terminal domain of CAP and two α-helices (II and III) in the amino terminal part of cro protein.

The path of 24 α-carbon atoms in the structural unit that contains these two helices in CAP can be superimposed on the path of the α-carbons of the two homologous helices in

cro. A similar structural homology is found between 2 helices of cro and two similar helices of lambda cI repressor.

Thus, a conserved helix-turn-helix motif, essential for interactions with DNA, is found in each of the three proteins providing an example of evolutionary stability of similar functional molecular structures.

x-ray diffraction data of cro and cI proteins allow the possibility to suggest models for the interaction with DNA. Lambda cI repressor recognizes the same sites that cro proteins binds, but with a difference in affinity.

Cro, cI repressor and CAP interact with DNA in a basically similar manner. Each of these proteins use symmetrically related subunits to interact with two-fold related nucleotide sequences in the regions that operate the corresponding genes. These DNA operator sequences have approximate two-fold symmetries and the complexes that they may form with the regulatory proteins may also be symmetric.

Although the basic design of DNA interaction unit appears to be conserved, most of these different proteins recognize different sites in DNA. This situation suggests that the specificity in DNA recognition most also be provided by the aminoacid side chains in the motif, and possibly, by other adjacent residues in the proteins.

Molecular recognition provides a sort of inertia to evolutionary innovation because it demands the conservation of genes or DNA segments and their products or functions (3).

In the case of cro protein, a model proposes that one α-helix interacts with a major groove in DNA, and one β-sheet contacts a minor groove.

Since the active form of cro-repressor is a dimer, it must interact with DNA symmetrical sequences in two successive major grooves along one face of the DNA double

helix. Adjacent α-helices of cro-repressor make contact with the DNA backbone and may help to orient the recognition process. Nevertheless these models and suggestions need to be tested with additional crystallographic, biochemical and genetic studies.

The process of evolution from small biomolecules to macrobiomolecules, to cells and to organisms is associated with a continuous symmetry break. The emergence of new symmetries may be traduced in many new properties, for the organisms, affecting structural, physiological, and behavioral traits. Among these we may have subtle changes in the organization of neurons in the brain that could affect the perception of symmetry in nature. These changes may influence the application or creation of symmetry in the art, in music, in architecture, in paintings as pathways to enjoy and dream.

CONCLUSIONS

Biomolecules display a wide variety of symmetries at the different levels of their organization; an aspect that is projected in the process of morphogenesis of unicelular and multicellular structures of the organisms.

The symmetrical characteristics of the genomic biomolecules may be interpreted in terms of the relationships between their structure, forces and energy fields displayed in basic living phenomena. Symmetries may be primarily related to the better performance or organization of molecules to fullfil requirements for more efficient living processes in the context of an evolutionary pattern.

The symmetrical characteristics of these biomolecules are associated with a periodicity, regularity and harmonious arrangement of the constituents, which in some way provide a sense of beauty to the models proposed to illustrate their different functions.

The different products or aspects of the genome functions, such as the formation of new DNA, RNA, proteins or regulatory systems of gene activity may be interpreted as the possibility that symmetry is, in part, responsible to code for itself and evolves in response to the numerous actors participating in biological phenomena.

REFERENCES

1) Crombrugghe, Benoit de; Stephen Busby & Henri Buc. 1984. Cyclic AMP receptor protein: Role in transcription activation. Science **224**: 831-837.

2) Ferraro, J.R. and J.S. Ziomek. Introductory group theory and its application to molecular structure. Plenum Press. New York. 1969.

3) Garcia-Bellido Antonio. 1996. Symmetries through organic evolution. PNAS USA **93**: 14229-14232.

4) Mizobuchi, Kiyoshi; Itaru Watanabe and James D. Watson (eds.) Nucleic Acid Research Future Development. Academic Press, 1983.

5) Saenger, Wolfram. Principles of nucleic acid structure. Springer-Verlag. New York, Berlin, Heidelberg, Tokyo. 1984.

6) Sung-Hou Kim. Structural modes for DNA-Protein recognition. In Nucleic Acid Research Future Development. Ed. Mizobuchi, K. *et al.* Academic Press. 1983

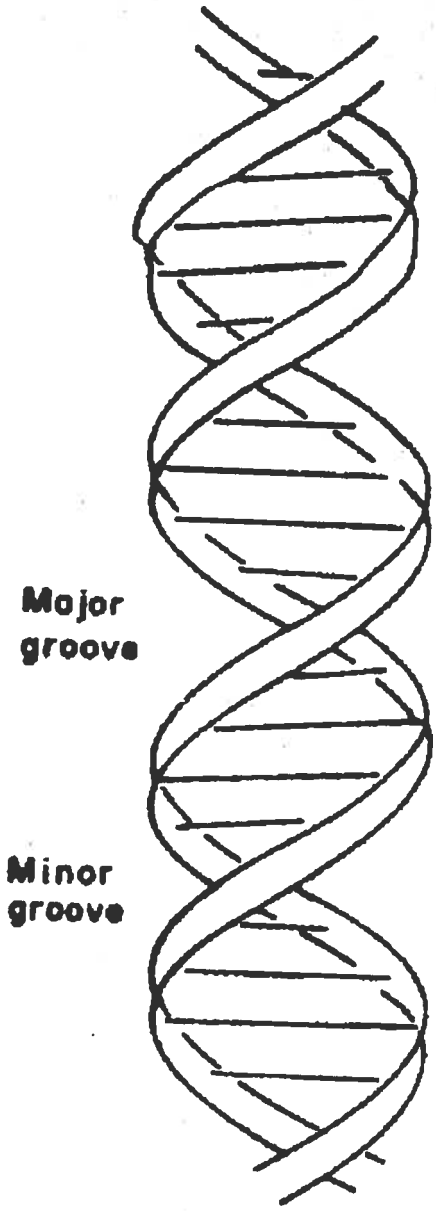


Fig. 1. Scheme of B-DNA

Double helix indicating the positions of major and minor grooves.

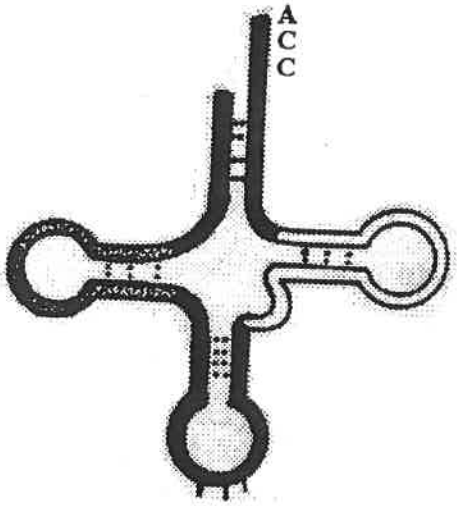


Fig. 2. tRNA secondary structure

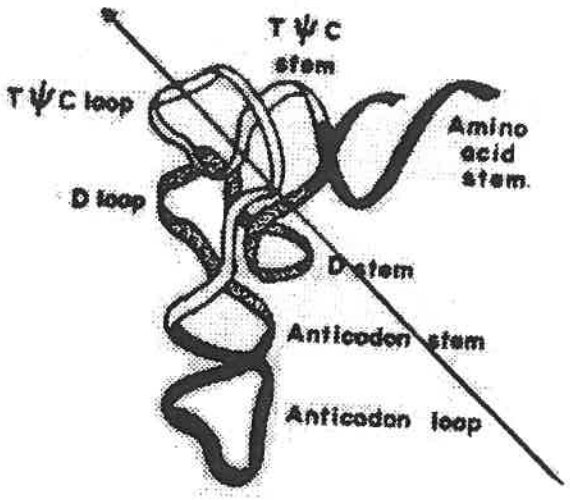


Fig.3. tRNA tertiary structure. Arrow indicates the pseudo-dyad axis of symmetry.

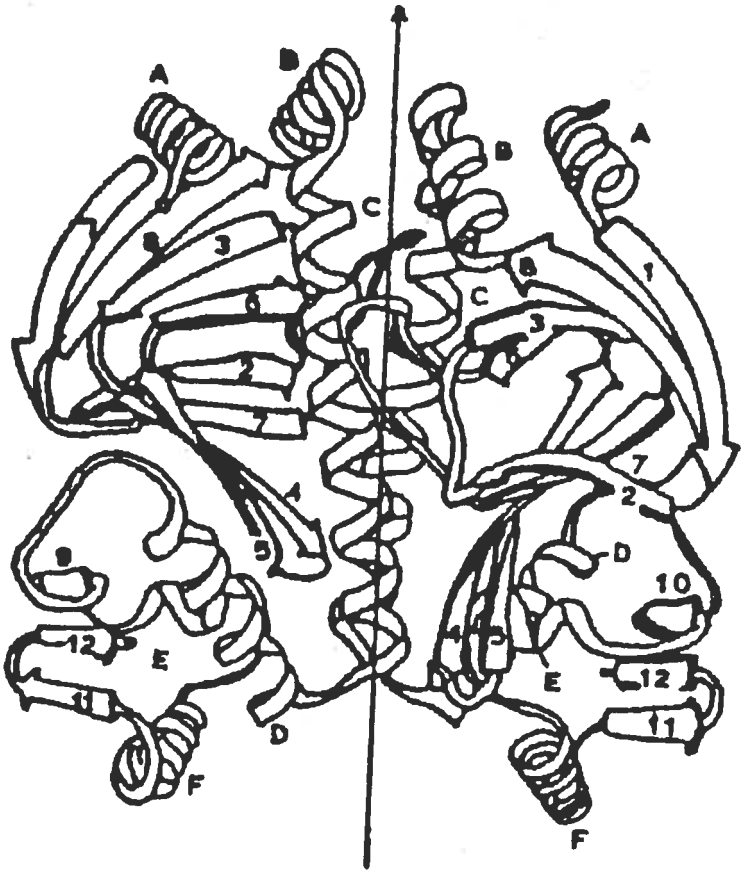


Fig. 4. Scheme of CAP as a dimer.

The drawing shows the existence of α-helices (designated by letters A-F) and β-sheets (designated by numbers 1-12). The arrow represents a dyad axis of symmetry relating the two identical subunits of the dimer.