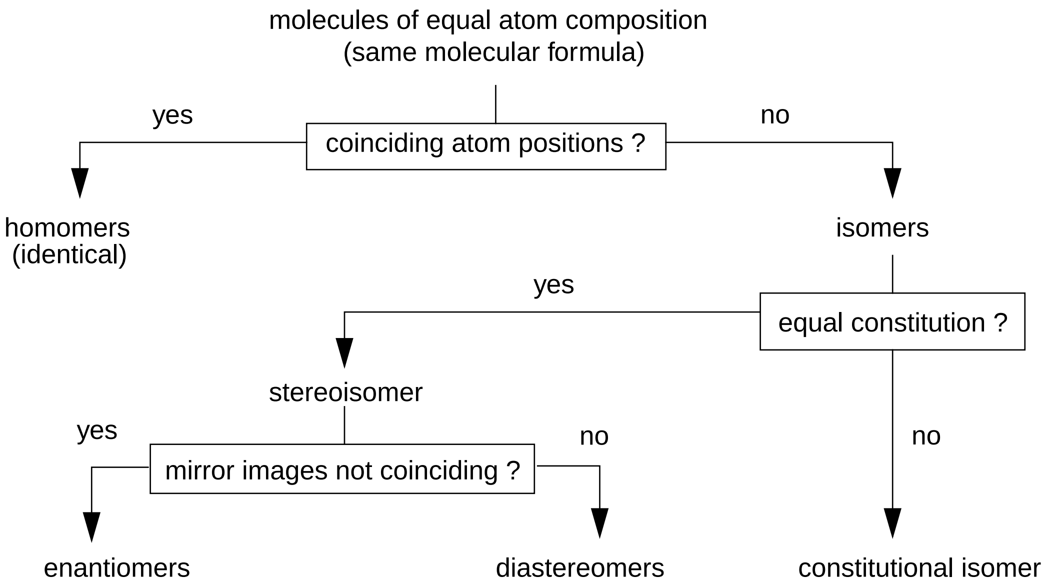
3. Computational methods in NMR-spectroscopy

3.1 Nomenclature of proteins and nucleic acids

3.1.1. Classification of structural relationships between molecules

For the **description of** small organic **molecules** several notations were introduced to classify the chemical and structural relations. These are: *isomer, constitutional isomer, stereoisomer, enantiomer and diastereomer*. The following flow-chart has been designed to find the stereochemical relationship between two molecules.

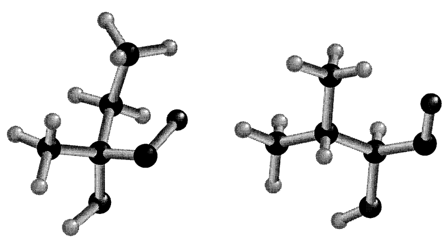


There are a lot of computer programs performing various kinds of structure calculations for biopolymers, which differ partially or substantially in the force fields or algorithms used. All programs, however, rely on structure relevant geometric data (covalent bound atoms, bond angles, dihedral angles, improper dihedral angles etc.) which build the **topology** of the biomacromolecule building blocks often collected in a **library** file. Some stereochemical relationships are also reflected in the topology.

The following terms should be distinguished when discussing structures:

- **constitutional isomers** = molecules having an identical chemical formula but different binding pattern. For instance, both valine and isovaline have the same chemical net formula but differ substantially in the covalent bonding pattern;

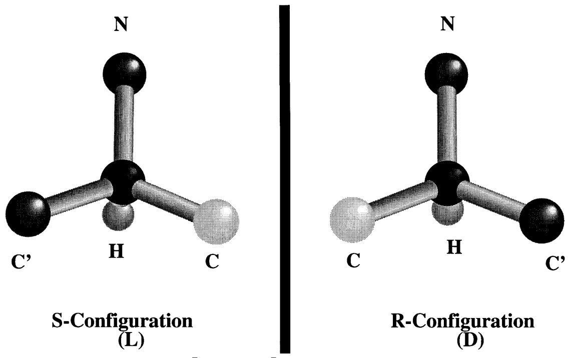
C5H9NO



D-iso-valine **valine**

Therefore, constitutional isomers will be found in the library as fragments with separate residue names.

- **configuration** = spatial arrangements of atoms around a center of chirality [see tetrahedron], not to made identical without breaking a bond. In nature this conversion can be done by chemical reactions or enzymes:



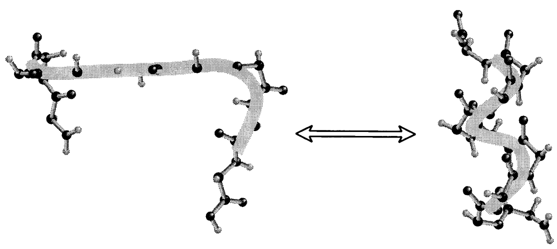
The α–carbon in amino acids is the most prominent atom for which the configuration is important. In topological libraries the configurations are defined by the specification of either the covalent bound atoms together with a coordinate set, or by improper dihedral angles.

R-/S-nomenclature rules [by Cahn, Ingold & Prelog]:
I.) Find the priority of the four substituents on the central atom according to:
a.) Priority decreases with lower ordinal number/atom weight.
(J > Br > Cl > S > P > F > O > N > C > H)
b.) If one level of comparison shows no difference, go to the next level.
c.) Double or triple bonds are substituted by duplicates of the bound atoms.
d.) If two atoms are of equal ordinal number (like C and C^{β} in a protein), their substituents have to be arranged and compared pairwise.
(Thus, O, O & N at C and $H^{\beta 2}$, $H^{\beta 3}$ & C^{γ} at C^{β} will be compared to find the priorities on center $C\alpha$ resulting in a higher priority for C.)
e.) Isotopes are ordered according to decreasing mass (T > D > H).
II.) Arrange the substituents for decreasing priority (a > b > c > d).
III.) Put the substituent with lowest priority (d) in the background direction.
IV.) Configuration is R, if substituent a is on the top and substituent b left of it.

Notice that there is a difference between the historical grown D-/L-nomenclature for amino acids based on a Fischer-projection of the molecule (which always leads to a L-configuration for natural amino acids) and the systematic R-/S-nomenclature (which leads to a R-configuration for the amino acids cysteine, only)!

- **conformation** = spatial arrangements of atoms which can be interconverted by changing one or several torsions [see model of extended peptide chain and α-helix]. In nature this can be realized by temperature changes since these transitions need only a few kJ per mole.

Bradykinin (9 amino acids)



Extended chain *α-helix*

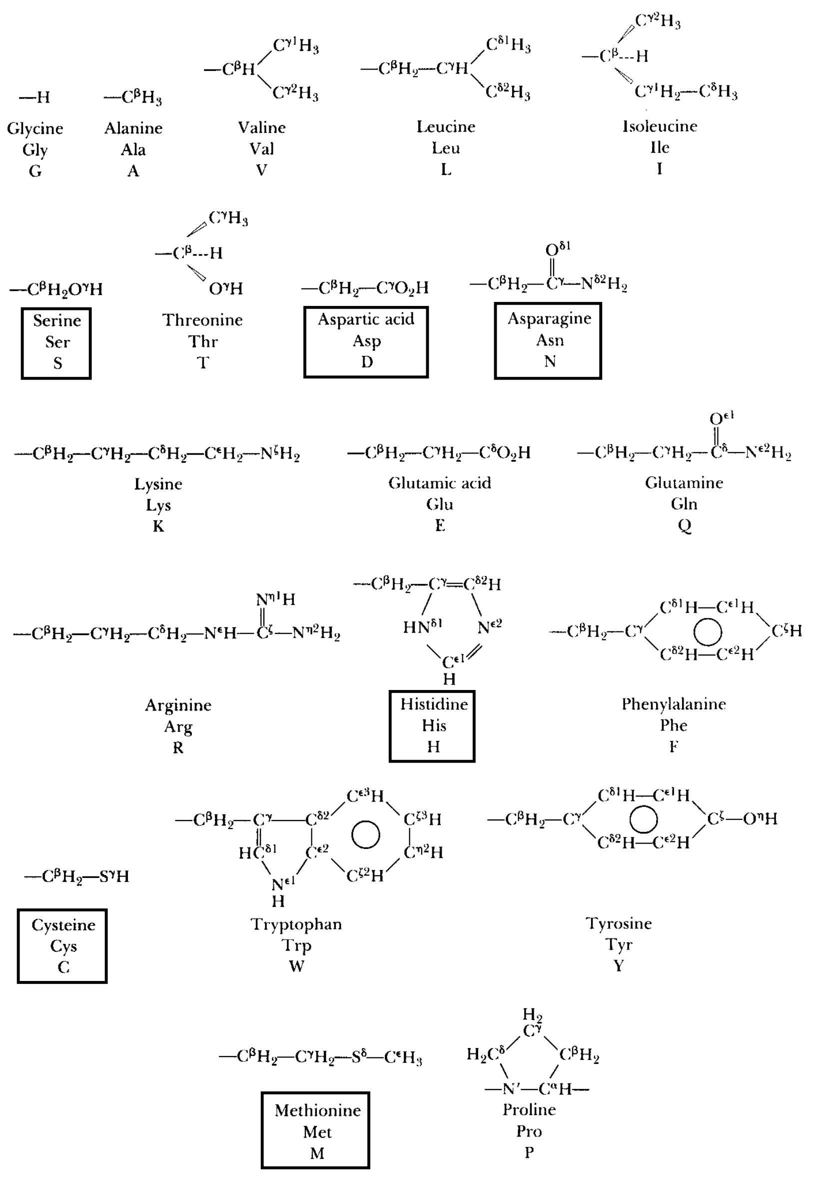
These conformational changes are calculated by distance geometry or molecular dynamics programs.

3.1.2. Atom names

Amino acids

Proteins are composed of amino acids linked together by peptide bonds. The 20 naturally occurring amino acids consist of a dipolar ionic group $H_3^+N\text{-}RC^{\alpha}H^{\alpha}\text{-}COO^-$ with different **sidechains** R attached. All atoms without the one corresponding to R are designated as the **backbone** or **mainchain** atoms. Amino acids can be linked together in forming biopolymer chains since by elimination of one water molecule. Conventionally, the heavy atoms of amino acids are designated by Greek letters with the mainchain carbon being called α, the first sidechain atom β, and so on. The α-carbon (except in glycine, where R = H) carries four different substituents mentioned above. Of the two possible enantiomorphs of the amino acids, only the L-form occurs in proteins.

The figure shows the configuration of the 20 natural amino acids and their 3- and 1-letter abbreviation.



The torsion angles in an amino acid backbone are called φ (C-N-C$^{\alpha}$-C), ψ (N-C$^{\alpha}$-C-N) and ω

(C^{α}-C-N-C^{α}). Sidechain torsions are enumerated χ_1, χ_2 etc.

The table presents an example for a selected amino acid topology stored in the distance geometry program library of DIANA (Güntert et al., 1991).

Residue name / Number of dihedral angles / Number of atoms in the residue / Start atom for current residue / End atom of current residue

```
RESIDUE   ARG+      7   32    3   31
    1 OMEGA   -1    2   10.0000    2    1    3    4    0
    2 PHI      0    0    0.0000    1    3    5   30    0
    3 CHI1     1    3    1.3500    3    5    7   11   29
    4 CHI2     1    3    1.3500    5    7   11   15   29
    5 CHI3     1    3    1.3500    7   11   15   19   29
    6 CHI4     0    0    0.0000   11   15   19   21   29
    7 PSI      0    0    0.0000    3    5   30   32    0
    1 C     C_BYL   2    8.2024   -0.6824   -1.1357    0.0000    2    3    0    0    0
    2 O     O_BYL   2   -7.0004   -0.1723   -2.2550    0.0000    1    0    0    0    0
    3 N     N_AMI   2   -6.4912    0.0000    0.0000    0.0000    1    4    5    0    0
    4 HN    H_AMI   0    3.2103   -0.4226    0.9063    0.0000    3    0    0    0    0
    5 CA    C_ALI   3    1.1597    1.4530    0.0000    0.0000    3    6    7   30    0
    6 HA    H_ALI   0    0.3677    1.7317   -0.5213    0.9158    5    0    0    0    0
    7 CB    C_ALI   3   -0.5515    2.0038   -0.7402   -1.2205    5    8    9   11    0
    8 HB2   H_ALI   0    0.2687    1.6375   -1.7668   -1.2235    7    0    0    0    9
    9 HB3   H_ALI   0    0.2687    1.6375   -0.2685   -2.1323    7    0    0    0    8
   10 QB    PSEUD   0    0.0000    1.6375   -1.0177   -1.6779    0    0    0    0    0
   11 CG    C_ALI   3   -0.5515    3.5338   -0.7388   -1.2182    7   12   13   15    0
   12 HG2   H_ALI   0    0.5515    3.9001    0.2878   -1.2152   11    0    0    0   13
   13 HG3   H_ALI   0    0.5515    3.9001   -1.2106   -0.3064   11    0    0    0   12
   14 QG    PSEUD   0    0.0000    3.9001   -0.4614   -0.7608    0    0    0    0    0
   15 CD    C_ALI   3    2.2062    4.0846   -1.4791   -2.4387   11   16   17   19    0
   16 HD2   H_ALI   0    0.1980    3.7230   -2.5073   -2.4444   15    0    0    0   17
   17 HD3   H_ALI   0    0.1980    3.7230   -1.0090   -3.3532   15    0    0    0   16
   18 QD    PSEUD   0    0.0000    3.7230   -1.7581   -2.8988    0    0    0    0    0
   19 NE    N_AMI   2   -5.4589    5.5643   -1.4643   -2.4144   15   20   21    0    0
   20 HE    H_AMI   0    4.1861    6.0160   -1.0017   -1.6515   19    0    0    0    0
   21 CZ    C_VIN   2   10.5783    6.3367   -2.0321   -3.3506   19   22   26    0    0
   22 NH1   N_AMO   2   -7.0852    7.6712   -1.9691   -3.2468   21   23   24    0    0
   23 HH11  H_AMI   0    5.1195    8.2477   -2.3929   -3.9454   22    0    0    0   24
   24 HH12  H_AMI   0    5.1195    8.0907   -1.4983   -2.4706   22    0    0    0   23
   25 QH1   PSEUD   0    0.0000    8.1692   -1.9456   -3.2080    0    0    0    0    0
   26 NH2   N_AMO   2   -7.0852    5.7747   -2.6630   -4.3906   21   27   28    0    0
   27 HH21  H_AMI   0    5.1195    6.3512   -3.0868   -5.0893   26    0    0    0   28
   28 HH22  H_AMI   0    5.1195    4.7788   -2.7101   -4.4681   26    0    0    0   27
   29 QH2   PSEUD   0    0.0000    5.5650   -2.8985   -4.7787    0    0    0    0    0
   30 C     C_BYL   2    8.2024    1.9838    1.4350    0.0000    5   31   32    0    0
   31 O     O_BYL   2   -7.0004    1.2064    2.3881    0.0000   30    0    0    0    0
   32 N     N_AMI   2   -6.4912    3.3043    1.5436    0.0000   30    0    0    0    0
```

Dihedral angle definitions

Atom number / Atom name / Atom type

x y z
Cartesian coordinates

Covalent bonding partners & stereo partner atom

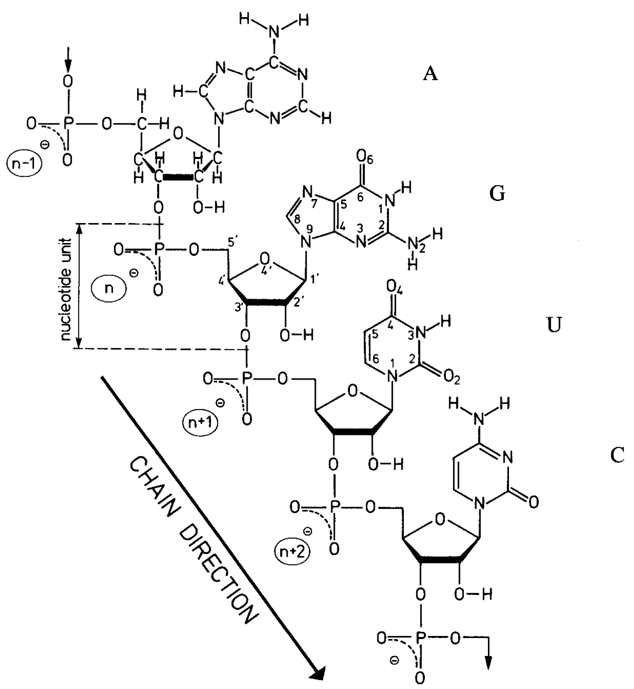
Since computers are unable to handle these characters, Greek letters used in the IUPAC nomenclature are substituted by their arabic equivalents.

Sometimes backbone atoms N, H, C and O are also denoted N’, H’, C’ and O’.

The atom numbering for geminal sidechain protons changes between different amino acids according to the rules of the R-/S-nomenclature and different substituents. The highest priority number (1) is given to the following heteroatom in the sidechain. The two prochiral hydrogens of equal priority are numbered in clockwise direction. In the previous figure, those carbons are marked by boxes at which $H^{\beta2}$ and $H^{\beta3}$ are reversed.
Since in NMR the first step is an assignment of resonances to a specific nucleus whose prochirality is *a priori* not known, the upfield shifted proton is often designated by >“< and the downfield shifted proton by a >‘< (e.g. $H^{\beta'}$ 2.87 ppm and $H^{\beta''}$ 2.72 ppm).

Nucleosides/Nucleotides
Nucleic acids consist of repeating units, the nucleotides or nucleosides. The backbone of these is formed by a phosphate, linking the 3’-hydroxyl group of one pentose ring to the 5’-hydroxyl group of the next one and the pentose itself. A different nitrogenous base is linked to the C1’ of the pentose depending on the nucleotide/nucleoside. In all nucleic acids the two purine bases are adenine (1-letter-code: A) and guanine (G). In DNA the two pyrimidine bases are thymine (T) and cytosine (C), whereas in RNA uracil (U) and cytosine (C) form the pyrimidines. RNA and DNA ribose units differ only in the substituents at C2’. In a DNA nucleoside two hydrogens are attached to this carbon. In a RNA nucleotide one position is substituted by a hydroxyl group. The four bases are the only variable constituents along the nucleic acid chain and their specific sequence constitutes the genetic information. In order to distinguish base atoms from the pentose, the sugar atom names are designated by a >‘<.

The figure shows a polynucleotide consisting of the 4 different nucleotides and the atom numbering scheme (Saenger, 1983).



The table lists an AMBER topology entry for one nucleotide (Pearlman et al., 1995).

```
R-ADENOSINE - with 5' - phosphate group and 3' - O(minus) group

 RA   INT     1
 CORRECT OMIT DU   BEG
   0.0
   1   DUMM  DU    M    0  -1  -2    0.00      0.00      0.00       0.0000
   2   DUMM  DU    M    1   0  -1    1.00      0.00      0.00       0.0000
   3   DUMM  DU    M    2   1   0    1.00     90.00      0.00       0.0000
   4   P     P     M    3   2   1    1.60    119.04    200.00       1.1662
   5   O1P   O2    E    4   3   2    1.48    109.61    150.00      -0.7760
   6   O2P   O2    E    4   3   2    1.48    109.58     20.00      -0.7760
   7   O5'   OS    M    4   3   2    1.60    101.43    -98.89      -0.4989
   8   C5'   CT    M    7   4   3    1.44    119.00    -39.22       0.0558
   9   H5'1  H1    E    8   7   4    1.09    109.50    -60.00       0.0679
  10   H5'2  H1    E    8   7   4    1.09    109.50     60.00       0.0679
  11   C4'   CT    M    8   7   4    1.52    110.00    180.00       0.1065
  12   H4'   H1    E   11   8   7    1.09    109.50   -200.00       0.1174
  13   O4'   OS    S   11   8   7    1.46    108.86    -86.31      -0.3548
  14   C1'   CT    B   13  11   8    1.42    110.04    105.60       0.0394
  15   H1'   H2    E   14  13  11    1.09    109.50   -240.00       0.2007
  16   N9    N*    S   14  13  11    1.52    109.59   -127.70      -0.0251
  17   C8    CK    B   16  14  13    1.37    131.20     81.59       0.2006
  18   H8    H5    E   17  16  14    1.08    120.00      0.00       0.1553
  19   N7    NB    S   17  16  14    1.30    113.93    177.00      -0.6073
  20   C5    CB    S   19  17  16    1.39    104.00      0.00       0.0515
  21   C6    CA    B   20  19  17    1.40    132.42    180.00       0.7009
  22   N6    N2    B   21  20  19    1.34    123.50      0.00      -0.9019
  23   H61   H     E   22  21  20    1.01    120.00      0.00       0.4115
  24   H62   H     E   22  21  20    1.01    120.00    180.00       0.4115
  25   N1    NC    S   21  20  19    1.34    117.43    180.00      -0.7615
  26   C2    CQ    B   25  21  20    1.33    118.80      0.00       0.5875
  27   H2    H5    E   26  25  21    1.08    120.00    180.00       0.0473
  28   N3    NC    S   26  25  21    1.32    129.17      0.00      -0.6997
  29   C4    CB    E   28  26  25    1.35    110.80      0.00       0.3053
  30   C3'   CT    M   11   8   7    1.53    115.78   -329.11       0.2022
  31   H3'   H1    E   30  11   8    1.09    109.50     30.00       0.0615
  32   C2'   CT    B   30  11   8    1.53    102.80    -86.30       0.0670
  33   H2'1  H1    E   32  30  11    1.09    109.50    120.00       0.0972
  34   O2'   OH    S   32  30  11    1.43    109.50    240.00      -0.6139
  35   HO'2  HO    E   34  32  30    0.96    107.00    180.00       0.4186
  36   O3'   OS    M   30  11   8    1.42    116.52   -203.47      -0.5246

CHARGE RA
  1.1662 -0.7760 -0.7760 -0.4989  0.0558
  0.0679  0.0679  0.1065  0.1174 -0.3548
  0.0394  0.2007 -0.0251  0.2006  0.1553
 -0.6073  0.0515  0.7009 -0.9019  0.4115
  0.4115 -0.7615  0.5875  0.0473 -0.6997
  0.3053  0.2022  0.0615  0.0670  0.0972
 -0.6139  0.4186 -0.5246

IMPROPER
 C8   C4   N9   C1'
 C6   H61  N6   H62
 N7   N9   C8   H8
 N1   N3   C2   H2
 C5   N1   C6   N6

LOOP CLOSING EXPLICIT
 C1'  C2'
 C4   C5
 C4   N9
```

Pseudo atoms are included in the libraries mostly for NMR spectroscopic reasons. There is a possibility of magnetic equivalence of two or more nuclei. This results in one resonance line for the magnetically equivalent protons. In other words, separate resonance lines cannot be resolved. Distance measurement for cross peaks including such resonances can be performed but the distances have to be corrected to reflect to the position of a mean atom rather than to the contributing individual atoms, since any member of the group may be responsible for the NOE cross peak. Thus, topology libraries have to provide virtual atoms at the geometric center of atom pairs which may be magnetically equivalent.

Typical NMR examples are the magnetic equivalence of α-hydrogens in glycine, aromatic ring hydrogens or sidechain-methylene protons.

A list of pseudoatom geometries is given below (Wüthrich, 1986).

Amino acid	Pseudo-structure[a]	intra-residue correction[b]	inter-residue correction[b]
Gly	NH–LA–CO ⋮ PA		m
Ala	CH | MB	i(N → β)	m
Ile (Thr)	CH | CH PG ··· LG MG	m(N → Mγ, Pγ) i(α → Mγ, Pγ)	m
	| MD	m(α → Mδ)	m
Val	CH | CH MG1 QG MG2	m(N → Mγ) s(N → Qγ) i(α → Mγ) i(α → Qγ)	q, m
Leu	CH | LB ··· PB	i(N → Pβ)	m
	| CH MD1 QD MD2	m(α → Mδ) s(α → Qδ)	q, m
Lys (Ser, Asp, Asn, Cys, His, Trp, Glu, Gln, Met, Arg, Pro)	CH | LB ··· PB	i(N → Pβ)	m
	| LG ··· PG	m(N → Pγ) i(α → Pγ)	m
	| LD ··· PD	m(α → Pδ)	m
	| LE ··· PE		m
	| NZ		
Phe (Tyr)	CH | LB ··· PB	i(N → Pβ)	m
	| C1	r(N → C1) m(α → C1)	r
	K6 K2 QR K5 K3	q(N → QR) r(α → QR)	q
	C4 | H4	m(N → C4) i(α → C4)	r

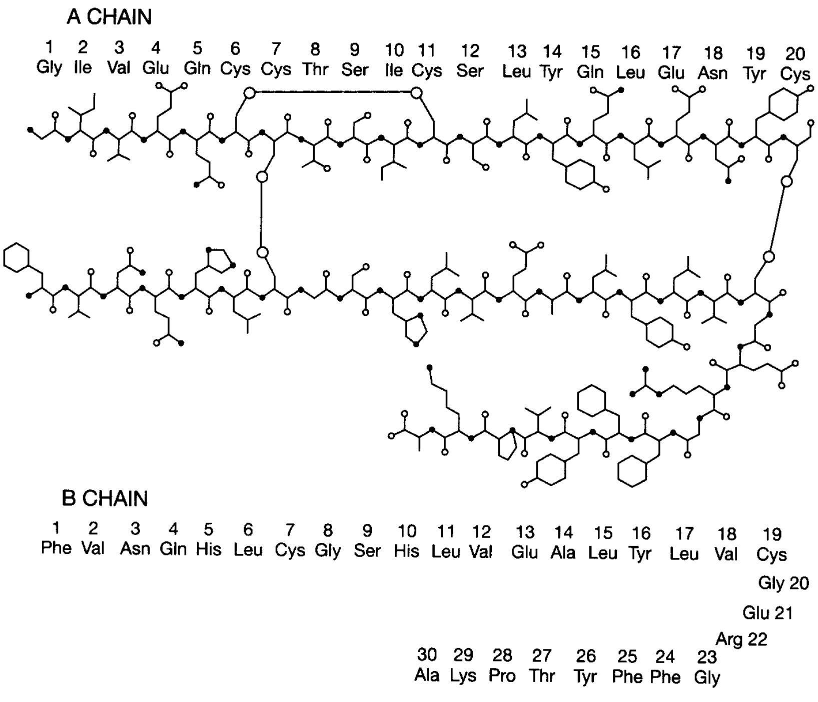
[a] P, Q, and QR are dimensionless pseudoatoms used as reference points for NOE distance constraints. P and Q are centrally located relative to the two protons of CH_2 groups and the two methyl groups in Val and Leu, respectively. QR is in the ring center in Phe and Tyr. K, L, and M are spheres of radii 1.5, 1.6, and 1.8 Å representing, respectively, the volume occupied by ring CH groups in Phe and Tyr, by CH_2 groups, and by CH_3 groups. K and L are at the carbon positions. M is in the center of the three methyl protons at a distance of 0.36 Å from the methyl carbon atom, and it serves also as a reference point for NOE distance constraints to the methyl protons. The indices A, B, G, D, E, and Z, are used in place of the more common Greek letters to identify the side chain atom positions.

[b] i = 0.6 Å, m = 1.0 Å, s = 1.7 Å, r = 2.0 Å, q = 2.4 Å. Intraresidue corrections are only indicated when they are different from the corresponding interresidue corrections. (From Wüthrich et al., 1983.)

3.1.3.1. Primary structure

Proteins consist of one or more polypeptide chains that contain between hundred and several thousand amino acid residues arranged in a definite, genetically determined sequence. The number of chains and the sequence of residues within them form the primary structure of the protein.
By convention, the residues are numbered in sequence along each chain, beginning at the N-terminus or amino end. Several chains within one molecule are numbered with Roman or Greek letters, which are added to the residue numbers.

The figure depicts two ways to illustrate a protein primary structure, sequence or chemical constitution for the example pig-insuline (Perutz, 1992).



3.1.3.2. Secondary structure

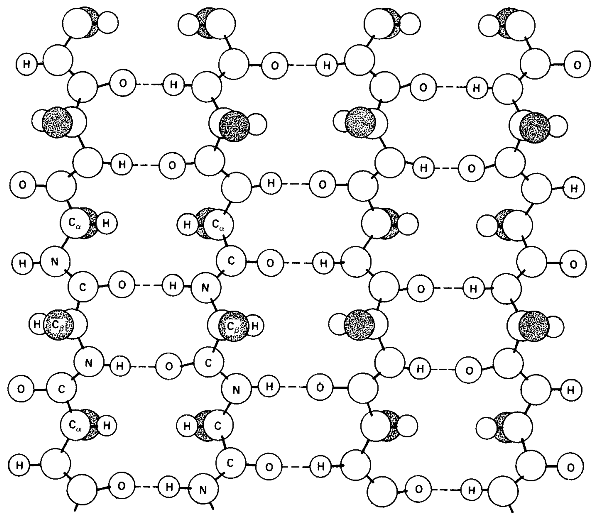
Several residues within a biopolymer chain can build up local elements of high structural order and symmetry. These elements are called secondary structure.
Biopolymers in which a similar atomic pattern (main-chain hydrogen bonding combined with specific ϕ,ψ-values) repeats at regular intervals tend to have screw symmetry. This means that each unit of pattern is brought into congruence with its neighbours assumed along the chain by a rotation about a common axis and translation along it. This holds true also for polypeptide chains when we neglect the different sidechains.
The multitude of mainchain torsion angle rotations is reduced and only a few secondary structure elements are adopted because of steric and electrostatic reasons. The main elements show a right-handed twist that alleviates close contacts between β-carbons of one residue and carbonyl oxygens of the next residue.

Two different ***structures with hydrogen bonds between the mainchains*** are known:

- In a **pleated β-sheet** each residue is related to its neighbours by a rotation of 180^{o} and a translation of 3.25-3.5 Å. A sheet involves hydrogen bonding between distant parts of the backbone.

- A **left-handed** α**-helix** (collagen helix) is obtained by a rotation of 120^o and translation of 3.1 Å. It is stabilized against the unfavourable sidechain repulsions by coiling of several helices.

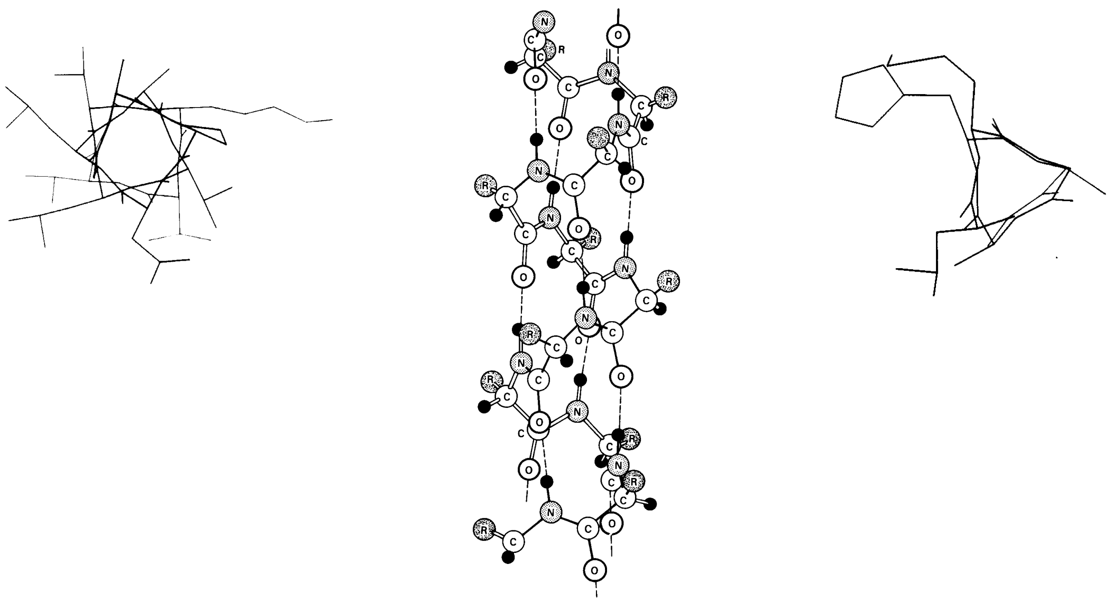
Figure of a pleated β*-sheet (Perutz, 1992).*



A variety of ***structures with hydrogen bonds within the same chain*** are found.
repetitive elements

- A **right-handed** α**-helix** (or 3.6_{13}-helix) is defined by a regular pattern of local H-bonding between O’ of residue n and H’ of residue n+4. The backbone torsion ϕ adopts a value around -60^o and ψ around -40^o. The pitch or repeat of an ideal α-helix is 3.6 residues per turn. This leads to a rise per turn along the helix axis of 1.5 Å or 5.4 Å per turn. The H-bond in an α-helix closes a loop that contains 13 backbone atoms starting at an O’ at one and ending at H’ on the other end of the segment resulting in parallel H-bonds.

Figure of a α*-helix [(a) end view, (b) side view] and a 3_{10}-helix [(c) end view] (middle: Perutz, 1992).*



- The **3_{10}-helix** rarely occurs and is defined by a H-bond between O'_n and H'_{n+3} with ϕ- and ψ-values of -70^o and -5^o. This helix is more tightly wound and shows a distinct triangular appearance in an end view where the α-carbons on successive turns are in line resulting in tilted H-bonds. It is frequently found that a 3_{10}-helix forms the last turn at the C-terminus of an α-helix.

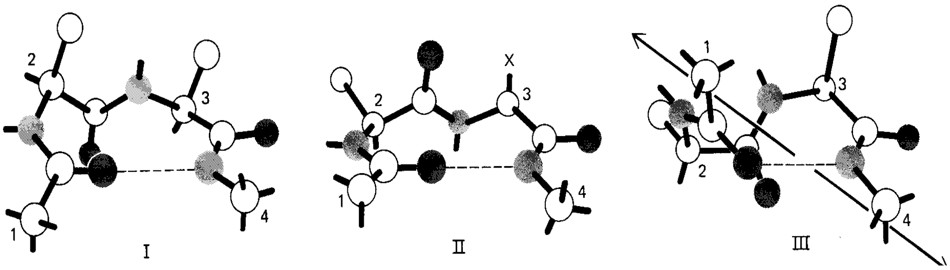
non-repetitive elements

- Various types of **reverse turns** are observed in protein structures. These are non-repetitive but well-ordered regions exhibiting a distict torsion angle set and hydrogen-bonding pattern. Turns are often connecting elements between repetitive extended secondary structure elements (e.g. helix-turn-helix motif), thereby reversing the chain direction and forming the globularity of a protein.
  Formally four consecutive residues are involved in a turn. Between the O'_n and H'_{n+3} a hydrogen bond normally exists. For the classification the torsion angles of residue n+1 and N+2 are critical.
  Mirror images of the ideal standard β-turns are designated by a >'<.
  Turn type I can be constructed by any amino acid except proline in position 3, whereas the I' β-turn can only be built if positions 2 and 3 are glycines.
  Type III turns are a portion of a 3_{10}-helix.

Ideal values for the dihedral angles and classification of bend types (Ghelis&Yon, 1982).

DIHEDRAL ANGLES OF TWO CENTRAL RESIDUES (°)[a]				NUMBER OF OBSERVED BENDS			
ϕ_2	ψ_2	ϕ_3	ψ_3	Ideal bends[b]	Nonideal bends	Total bends	H-bonded bends[c]
−60	−30	−90	0	130	46	176	99
60	30	90	0	8	5	13	10
−60	120	80	0	41	23	64	43
60	−120	−80	0	15	5	20	16
−60	−30	−60	−30	66	11	77	45
60	30	60	30	11	2	13	7
A bend with two or more angles differing by at least 40° from those given above				0	35	35	5
−80	80	80	−80	1	2	3	0
80	−80	−80	80	0	4	4	2
A *cis* Pro at position 3				8	0	8	6
A kink in the protein chain created by $\psi_2 \approx 180°$ and $\lvert\phi_3\rvert < 60°$ or $\lvert\psi_2\rvert < 60°$ and $\phi_3 \approx 180°$				8	0	8	1
				288	133	421	234

Classical types of reverse turns (Perutz, 1992).



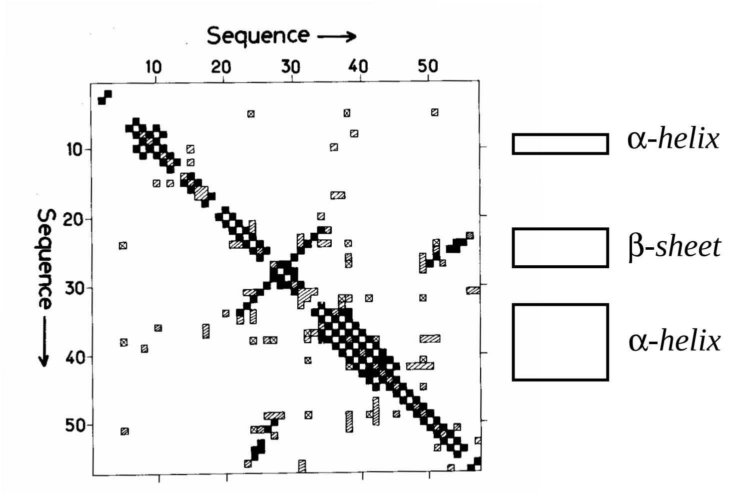
3.1.3.3. Tertiary structure

The overall three-dimensional architecture of a biomolecule arises from the intrinsic rotational ability of single covalent bonds. Rotations of several torsion angles result in a different, non-superimposable, **three-dimensional folding** which we already introduced as conformation. Molecules in which the different conformations are not seperated by high energy barriers will populate many possible rotational states and therefore lack a defined tertiary structure. A small polypeptide chain of 100 amino acids may thus adopt up to 10^{100} conformations. The absence of a stable tertiary structure is called **random coil**. In other words, a tertiary structure is a rather fixed arrangement of several stable secondary structure elements.

Tertiary structure representation of lactate dehydrogenase domain 1. β*-sheets are indicated by arrows,* α*-helices by a ribbon.*

The arrangement of secondary structure elements in a tertiary structure can be visualized by **contact** or **distance maps**. Here, short C^{α}-C^{α} distances (X-ray) or H'-H'/H^{α} NOE's (NMR) between all pair residues are marked. α-helices are evident by a number of contacts parallel along the main diagonal since the atoms H' and H^{α} in the relation i to i+3 are in close contact. An antiparallel β-sheet is illustrated as a diagonal perpendicular to the main diagonal since H^{α}_{i} to H^{α}_{j}, H^{α}_{i+2} to H^{α}_{j-2} and other contacts repeat in a regular pattern.

Diagonal plot of BUSI (bull seminal inhibitor, 57 amino acids) (Wüthrich, 1992).



3.1.3.4. Quarternary structure

Polypeptides are able to associate to bigger complexes by specific interactions. As seen for pig-insuline two amino acid chains form the vital biomolecule. Each subunit has an own tertiary structure which exhibits interfaces for an interaction of hydrophobic or polar kind. Shape and hydrophobicity of the subunits have to be highly complementary in order to allow a close packing of the protein. If the subunits associate isologously, i.e. build a dimeric structure with a twofold axis of symmetry, it will be possible to detect the same resonance signal set for both subunits in an NMR experiment.

Two modes of association (Monod et al., 1965).