Geometrical Aspects of Molecular Structures

by Devapriya Choudhury

School of Biotechnology, JNU, New Delhi- 110067, India

1 Why study molecular structures?

The analysis of molecular structures is perhaps one of the most important and productive areas in chemistry as well as biology. The material world consists of an astonishing diversity of molecules which interact among themselves and in myriad different ways. A thorough knowledge of the structure and properties of these molecules is an important step in understanding their functions. Often a detailed knowledge of the structure and function of molecules help us to manipluate them in ways beneficial to mankind. The functions of biomolecules are most often intricately linked up with their structures which in turn makes the study of molecular structures an interesting and important undertaking.

To take a specific example, studying the enzymes that catalyse various reactions in the living system constitute one of the most important activities of the science of biochemistry. The functional properties of the enzymes including their substrate specificity and catalytic mechanism is intricately linked with the precise three dimensional arrangement of atoms in the enzyme active site. Hence a detailed understanding of the functional properties of enzymes depend on the precise understanding of the principle that govern structure formation in such molecules. Structural principles lie at the root of amost every property in the material world, and the analysis of molecular structures can contribute to almost every field, be it understanding enzyme catalysis or the tensile strength of various materials.

2 Molecular Modeling

Nowadays the term 'Molecular Modeling' has become closely associated with the analysis and manipulation of molecular structures particularly when such studies are carried theoretically or using a computer. Modeling implies building a simplified and often idealised description that is used to mimic the essential properties of some real system. When applied to molecules modeling implies the construction of ways and means that can mimic some of their essential properties. Often such models are constructed with the help of a computer although one can perform molecular modeling studies using simple mechanical models of molecules or even with purely abstract representations.

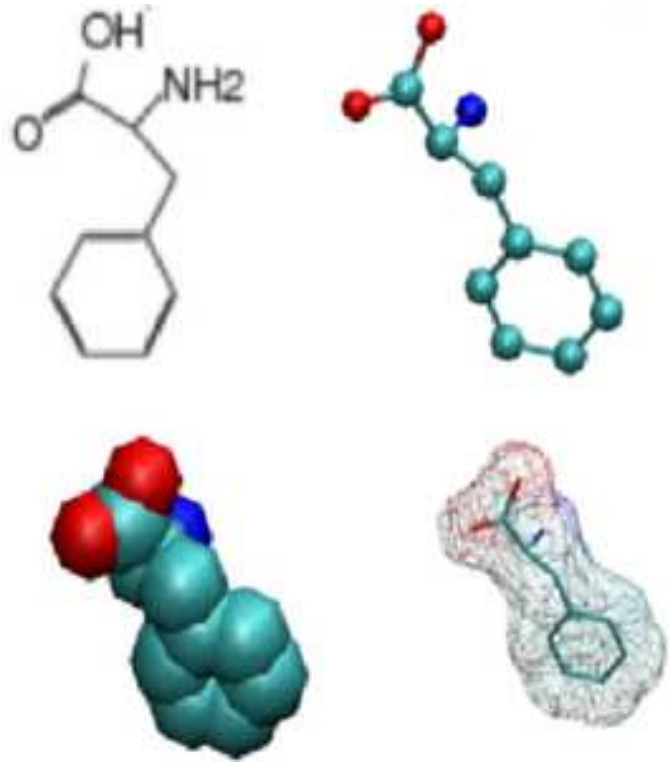


Fig. 1. Pictorial representations of molecules. Different ways to represent the amino acid phenylalanine. (Top Left) A simple two dimensional representation depicting the chemical structure. This type of representations provide only minimal conformational information. (Top Right) A ball and stick representation. The atoms are depicted as small spheres and the bonds are represented by sticks. (Bottom Left) A space filling model, where the molecule is depicted by a system of overlapping spheres representing atoms. (Bottom Right) The solvent accessible surface of the molecule is depicted in the form of a wire-mesh and a stick representation of the molecule is superimposed.

An important aspect of molecular modeling is the construction of pictorial models of molecules. Pictures are used not only to beautify a molecular modeling text but also to highlight specific molecular properties. A large number of pictorial representations have been developed in order to highlight different aspects of molecular structure (Figures 1 & 2). Representations that are suitable for small molecules may not always be suitable for large molecules. In case of large molecules often highly simplified "cartoon" representations are

used that can highlight specific aspects of their structure without burdening the viewer with an unnecessary overload of information.

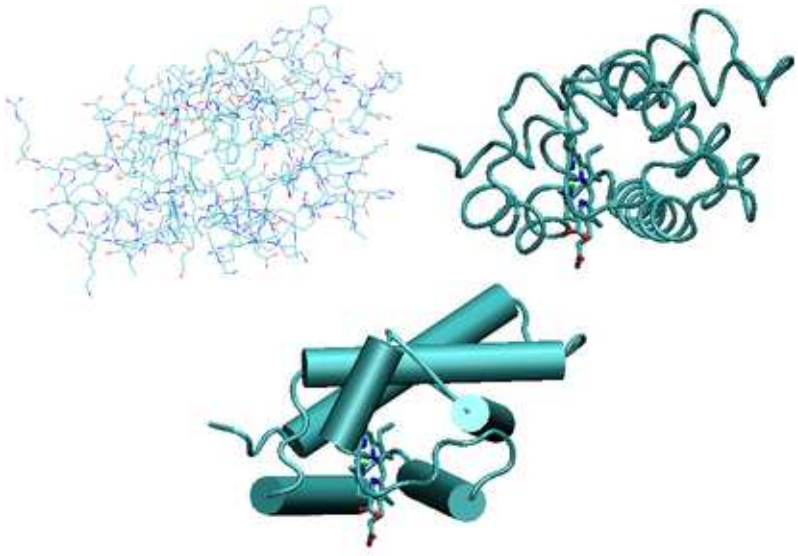


Fig. 2. Pictorial representations of large molecules. Different views of the α-chain of Hemoglobin. (Top Left) An all-atom view where every bond in the molecule is depicted by a thin line. (Top Right) A tube view where the polypeptide chain backbone is represented by a smooth curve. The heme group is shown using a stick representation. (Bottom) A cartoon representation, where the helices are drawn as cylinders, the coil regions are shown as tubes, and the heme group is depicted using a stick diagram.

3 Configuration and Conformation

We can define molecular structure as the precise three dimensional arrangement of the constituent atoms in a molecule. Variations in the structure of a molecule can be described in terms either configurational or conformational changes. Configuration denotes the spatial arrangement of atoms that is conferred by the presence either double bonds (around which there is no freedom of rotation) or chiral centers around which the substituent groups are arranged in a specific sequence. It is not possible for a change in configuration to happen without a concomitant breaking and rejoining of bonds in the molecule. Molecules with the same chemical arrangement of atoms but in different configurations are called *isomers*. There may be *geometric* or *cis-*

trans isomers, when they differ in their arrangement of substituent groups around a rigid double bond, or they may be *optical isomers* or *stereoisomers* when they differ in the specific order of substituents around one or more chiral centers. If there are more than one chiral centre in a molecule then some of the stereoisomeric pairs are mirror images of each other while others are not. Those pairs of stereoisomers that are also mirror images of each other are called *enantiomers* and those pairs that are not are called *diastereoisomers*

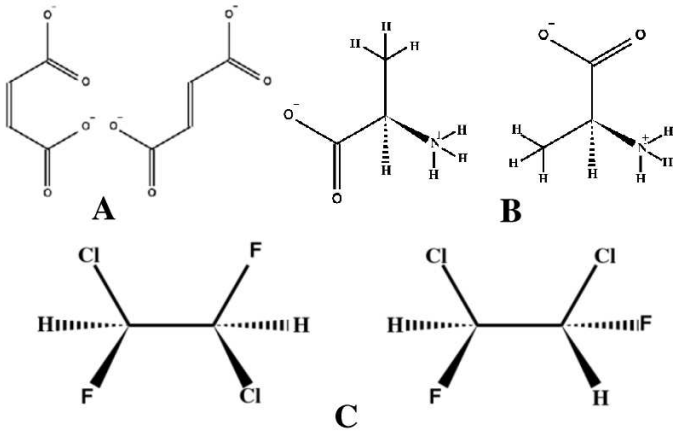


Fig. 3. Configurational and conformational isomers. Panels A and B show two types of configurational isomers while panel C shows a pair of conformational isomers. (A)Geometrical (or *cis-trans*) isomeric pairs (maleate and fumarate) (B)stereoisomeric pairs (D-alanine and L-alanine) (C) Conformational isomers (or *conformers*) of Dichloro difluro ethane. The two conformers differ only in their rotation about the central C-C bond. Conformational isomers that differ only in the the rotational isomeric state about one or few bonds are also called *rotamers*

4 External and Internal Coordinates

While the pictorial molecular models shown in figures 1 and 2 are informative and highly useful, many situations arise when an abstract mathematical representation of a molecule is needed. Such situations may arise, for example, when one needs to accurately compare the structures of a pair of molecules. Abstract representations are most useful when any kind of structure manipulation is to be done in the computer. Most such representations are also highly compact and therefore very useful for storage purposes.

4.1 External Coordinates

This is perhaps one of the simplest of all abstract representations of a molecule. Essentially all that it requires is to define a suitable three-dimensional coordinate system. Each atom in the molecule is now represented by point in this coordinate system and its position is marked with a triple of numbers (e.g., x,y,z) representing the projections drawn from the point to the coordinate axes. The coordinate system usually chosen is perpendicular and rectilinear. Such systems are called Cartesian coordinate systems.[1] However there may be situations (for example in crystallography) when the three axes are not necessarily orthogonal to each other. Finally in special cases various forms of curvilinear coordinate systems can also be used.

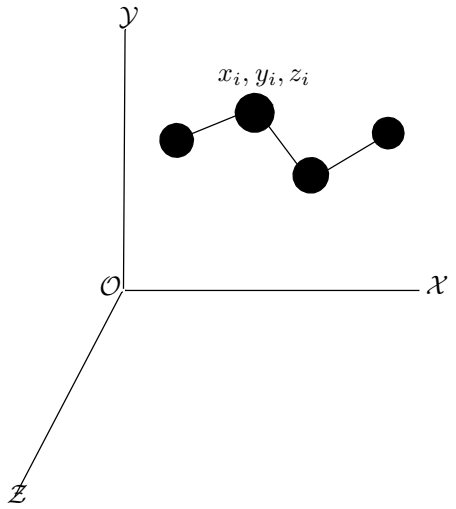


Fig. 4. Representation of a molecule using external coordinates. The external coordinate system is completely specified by the origin ($\mathcal{O}$) and the mutually perpendicular axes (x, y and z). Each atom has, as its coordinates the triple (x_i, y_i, z_i) that specify the shortest distance from the point to the respective coordinate axes.

A molecule with N atoms requires $3N$ coordinates to be described (each atom requires three coordinates - x,y and z). External coordinates are sensitive to both the position and orientation of the molecule *i.e.*, the external coordinates of a molecule will change if the molecule is moved or rotated even without any change in conformation.

1. So named after the French philosopher René Descartes (1596-1650). Who in the year 1637 introduced this type of coordinate system for describing plane curves.

4.2 Internal Coordinates

There are several different internal coordinate representations of a molecule. Out of these, the one based on bond lengths, bond angles and torsion angles is the most important. In this representation, molecules are described in terms of certain parameters called bond lengths, bond angles and torsion angles (defined below) that are internal to the molecule and do not depend upon an external reference system. Internal coordinates of a molecule are therefore independent of any external reference system and do not change if the molecule is moved or rotated. In other words they are invariant to translations and rotations carried out on the entire molecule. Thus changes in internal coordinates are solely due to conformational changes, a fact that makes them a very useful descriptor of conformational properties of a molecule.

While there are $3N$ external coordinates for a molecule, there are only $3N - 6$ internal coordinates. This is because the 6 parameters that are required to describe the position and orientation with respect to an external coordinate system, are missing from an internal coordinate description of a molecule.

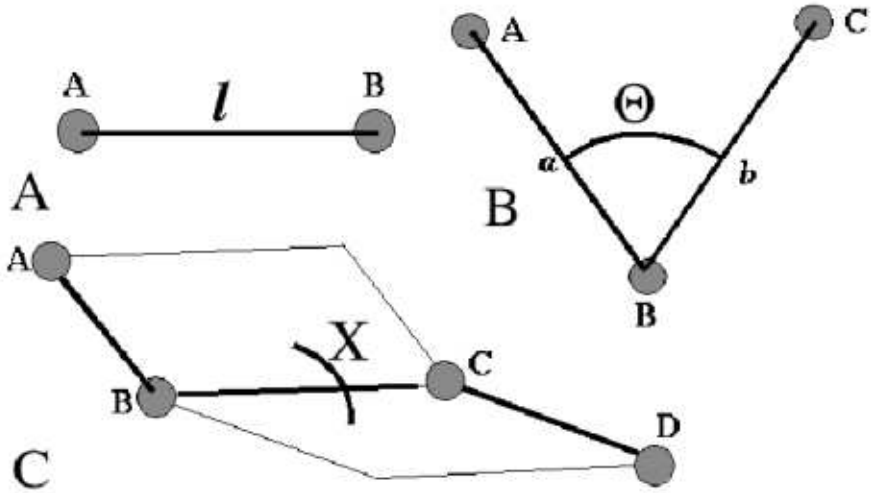


Fig. 5. Internal coordinates. The circles shown in the figure represent atoms, while bonds are shown with thick lines. The atoms are also labeled with capital letters. (A) Bond length (B) Bond angle (C) Torsion angle. The two planes that define the torsion angle are also shown.

Bond lengths, bond angles and torsion angles are defined and formulas for obtaining them from the external coordinates of atoms are given in the following sub-sections.

4.2.1 Bond Length

The bond length is the distance between a pair of bonded atoms. If x_i, y_i, z_i and x_j, y_j, z_j are the coordinates of the two atoms, then the bond length (l) is given by the formula:

$$l_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{1}$$

4.2.2 Bond Angle

Let $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ be the position vectors[2] of the three atoms, A, B and C shown in figure 5B. We define vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that:

$$\boldsymbol{a} = (\boldsymbol{A} - \boldsymbol{B})/|\boldsymbol{A} - \boldsymbol{B}|$$
$$\boldsymbol{b} = (\boldsymbol{C} - \boldsymbol{B})/|\boldsymbol{C} - \boldsymbol{B}|$$

The bond angle is now given by the formula:

$$\theta = \cos^{-1}(\boldsymbol{a} \cdot \boldsymbol{b}) \tag{2}$$

where $\boldsymbol{a} \cdot \boldsymbol{b}$ is the dot product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

4.2.3 Torsion Angle

Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ be the position vectors specifying atoms A,B,C and D in the linked four-atom system shown in figure 6A. We further define vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ such that:

$$\boldsymbol{a} = (\boldsymbol{A} - \boldsymbol{B})/|\boldsymbol{A} - \boldsymbol{B}|$$
$$\boldsymbol{b} = (\boldsymbol{C} - \boldsymbol{B})/|\boldsymbol{C} - \boldsymbol{B}|$$
$$\boldsymbol{c} = (\boldsymbol{D} - \boldsymbol{B})/|\boldsymbol{D} - \boldsymbol{B}|$$

2. See Note 1 for an introduction to essential vector algebra.

We define perpendiculars $\boldsymbol{p}$ and $\boldsymbol{q}$ such that:

$$\boldsymbol{p} = \boldsymbol{b} \times \boldsymbol{a}$$
$$\boldsymbol{q} = \boldsymbol{b} \times \boldsymbol{c}$$

The torsion angle χ is then given by:

$$\chi = [(\boldsymbol{p} \times \boldsymbol{q}) \cdot \boldsymbol{b}] \cos^{-1}(\boldsymbol{p} \cdot \boldsymbol{q}) \tag{3}$$

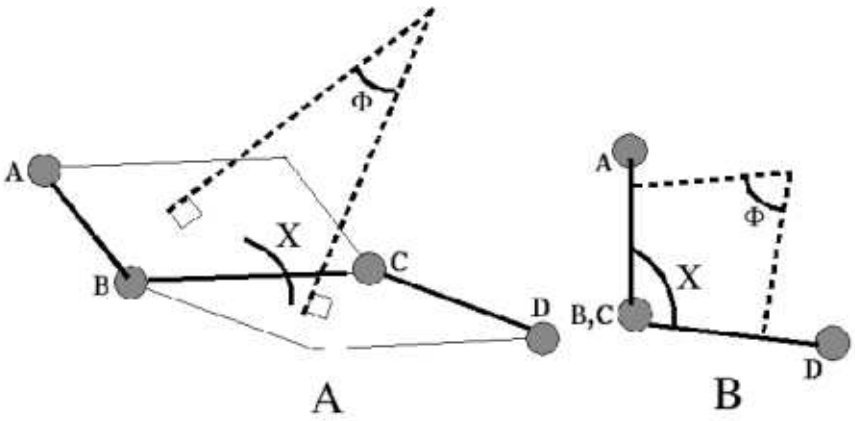


Fig. 6. Definition of torsion and dihedral angles.(A) Torsion angle χ(A-B-C-D) gives the angle between the planes A-B-C and B-C-D. The dihedral angle ϕ is also shown which is the angle between the normals to the two planes. (B) The torsion angle χ(A-B-C-D) can also be thought to be describing the orientation of the bonds A-B and C-D with respect to the central bond. If atoms A and D are on the same side of the B-C bond then χ is defined to be 0° conversely if all the atoms are again co-planar and A and D are on opposite sides then χ is defined to be 180°. The convention for the sign of the torsion angle can be understood with reference to figure 6B. Looking down the B-C bond if the far bond C-D rotates clockwise with respect to the near bond A-B the χ is considered to be positive. Conversely if the far bond C-D is rotated anticlockwise with respect to the near bond A-B then it is negative. The sign of the torsion angle does not change regardless of which direction (B-C or C-B) one looks at the central bond.

Rather than always describe a torsion angle with its numerical value it is customary to use certain names for ranges of torsion angle values. The nomenclature preferred by organic chemists are *syn* ($\backsim 0°$), *anti* ($\backsim 180°$), $\pm$ *synclinal* ($\backsim \pm 60°$) and $\pm$ *anticlinal* ($\backsim \pm 120°$). Spectroscopists and cystallographers, on the other hand, prefer the following notation *cis* ($\backsim 0°$),

trans ($\backsim 180°$) and $\pm$ *gauche* ($\backsim \pm 60°$). A detailed description of the names of torsion angle ranges is given in figure 7.

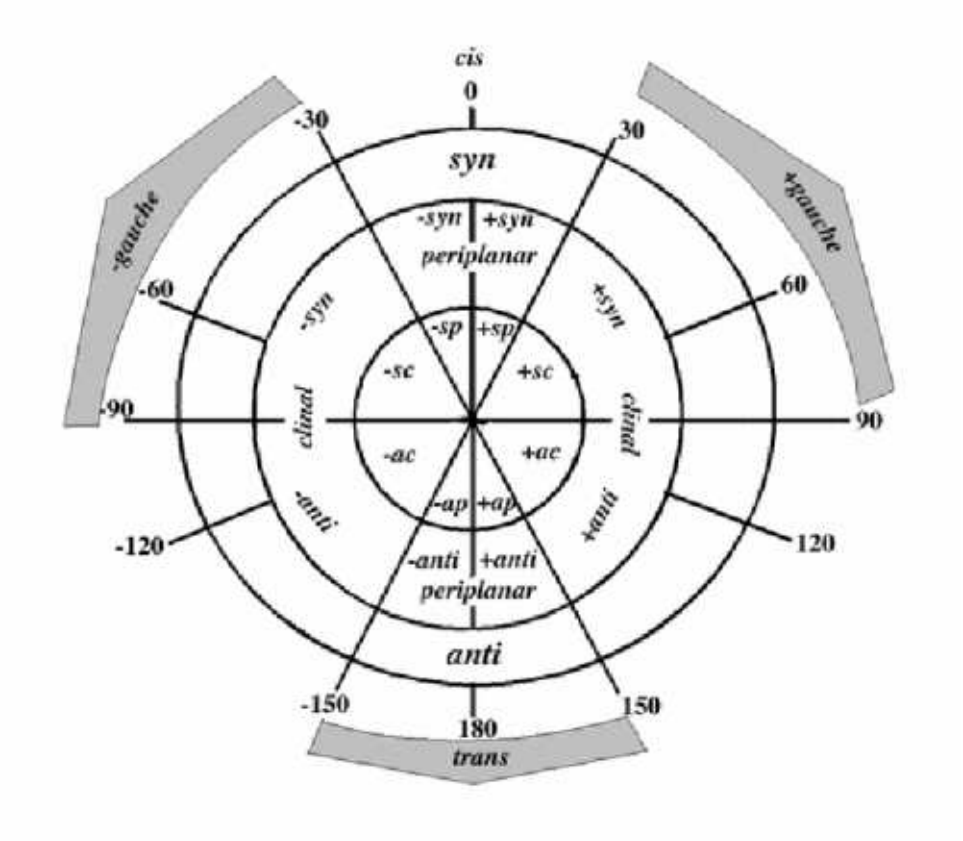


Fig. 7. A schematic description of the ranges of torsion angles and their names. The numerals indicate the value of the torsion angle in degrees. The names of the various torsion angle ranges used by organic chemists as well as spectroscopists/crystallographers are indicated. The figure is not to scale.

4.3 Comparison between external and internal coordinate representations of molecules

The major points of contrast between external and internal coordinates are as follows. For a molecule with N atoms there are $3N$ external coordinates, however only $3N-6$ internal coordinates are sufficient for a complete description[3].This is because external coordinates also specify the overall position and

orientation of the molecule. The position of the molecule can be described by a translation of any one atom of the molecule (assuming that the molecule is rigid) from the origin of the coordinate reference frame. The orientation can be described by a series of three rotations about the coordinate axes from some standard orientation. It can be shown[4] that three parameters are required for describing the translation.[5] The rotations are also described using three parameters. In case of internal coordinates, the overall position and orientation of the molecule cannot be described. Hence the number of internal coordinates that are required for a full description of the molecule is six short of the corresponding number of external coordinates. The above discussion implies that internal coordinates are invariant to translation and rotations of the molecule. In most cases this is not of much importance. However, in specific situations like in case of simulations of intermolecular interactions, it may be useful to specify the overall position and orientation of the molecule. In such situations the use of external coordinates to describe the state of the molecule might be necessary.

In most cases the bond lengths and bond angles of molecules undergo little or no perturbation during conformational changes. Often the bond lengths and bond angles depend solely upon the chemical nature of the constituent atoms and can be predicted with confidence from chemical knowledge. Hence in many studies involving molecular conformation, it is assumed that bond lengths and bond angles remain constant and only the torsion angles variations are important. Restricting attention to only the torsion angles tremendously reduces the conformational space available to a molecule and thus greatly simplifies many problems involving conformational search methods. Usually this assumption makes otherwise impossible problems tractable with current methodology. However there are a few issues that may need to be addressed when conformational studies are carried out in torsion angle space. One of them is that the amount of conformational change in a molecule depends not only on the amount of change in torsion angle values but also on the location of the particular torsion angle with respect to the rest of the molecule. This can be understood from figure 8 where it can be seen that for a simple linear chain, torsion angles that are located centrally in the molecule have a greater effect on the conformation of the molecule than those torsion angles that are located peripherally.

---

3. For a diatomic molecule only the bond length is required. Hence the number of internal coordinates in this case is $3N-5$. In general, if there are N atoms in a perfectly linear chain, then the molecule can be described by just $N-1$ bond lengths.

4. See for example, Goldstein H., Poole C. and Safko J. *Classical Mechanics* (Third edition) Pearson Education Asia (2002)

5. One each for translations along the three axial directions.

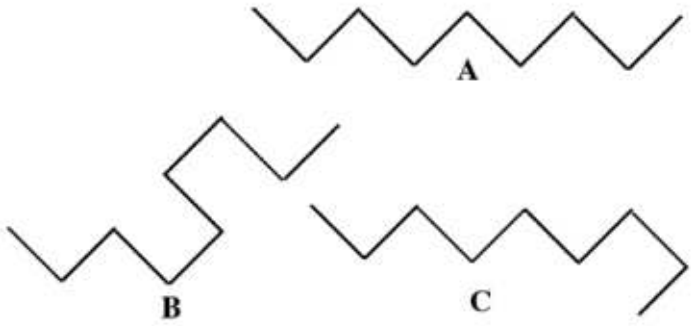


Fig. 8. Effect of torsion angle position on molecular conformation. (A) A simple linear chain with all torsion angles in the *trans* position. (B) One torsion angle from the middle of the molecule is changed to *cis*. (C) One torsion angle from the end of the molecule is changed to *cis*.

4.4 Specifying a molecule in internal coordinates. The Z-matrix

The Z-matrix is way to specify the positions of every atom in the molecule using internal coordinates. The following table shows an example Z-matrix for ethane (C_2H_6).

1	C						
2	C	1.54	1				
3	H	1.0	1	109.5	2		
4	H	1.0	2	109.5	1	180.0	3
5	H	1.0	1	109.5	2	60.0	4
6	H	1.0	2	109.5	1	-60.0	5
7	H	1.0	1	109.5	2	180.0	6
8	H	1.0	2	109.5	1	60.0	7

Table 1. Example Z-matrix for Ethane. The first column is the atom number (unique for each atom). The second column gives the corresponding atom name. The 3^{rd}, 5^{th} and 7^{th} columns give the bond length, bond angle and torsion angle respectively. The 4^{th}, 6^{th} and 8^{th} are atom numbers which identify respectively atoms shifted from the current atom by one, two and three bonds. These atoms must be predecessors of the current atom in the sense that they must have been described before the current atom can be described. It is important to note that the first, second and third atoms have missing values for one bond length, two bond angles and three torsion angles. Together these six missing parameters signify the six rigid body parameters that are not specified in internal coordinates.

The main utility of specifying the conformation of a molecule in internal coordinates is that it uses bond lengths and bond angles as its parameters, both of which have chemical meaning. Thus errors in specification of the conformation can often be quickly diagnosed by looking at the bond lengths and bond angles and corrected accordingly.[6]

4.5 Converting internal to external coordinates: The Fourth Atom Fixing Algorithm

In the previous sections we described the formulas for transforming a set of external coordinates into internal coordinates. In this section we discuss the reverse problem i.e., the conversion of internal to external coordinates. One problem in converting internal coordinates to external coordinates is that since in the former, there does not exist any information about the overall position and orientation of the molecule, it is necessary to either obtain this information from some other source or generate the molecule in some suitable standard position and orientation. A very elegant and general solution of the problem goes by the name of 'Fourth Atom Fixing Algorithm'.[7] In this algorithm one supplies the information of the overall position/orientation of the atom by specifying the external coordinates of the first three atoms.[8] Given the external coordinates of the first three atoms, and the bond length, bond angle and torsion angle leading to the fourth atom (referring to figure 9 the bond length l, the bond angle θ and the torsion angle χ) one can generate the external coordinates of the fourth atom in following steps.

6. Given the external coordinates of a molecule one can easily generate the corresponding internal coordinates using the formulas given in the text. For doing this with the help of a computer, one often needs to maintain lists of atoms which are connected by one, two or three bonds respectively. Data structures and algorithms relating to making of such lists are discussed in Note 2.

7. Ramachandran G.N. and Sasisekharan V. Conformation of polypeptides and proteins. *Adv. Prot. Chem.* **23** (1968) 283-438. The algorithm is described in section III pp 308-312 of the aforementioned review.

8. If such information is not provided, one can then generate the external coordinates in some standard orientation. One way to do this is shown in Note 3.

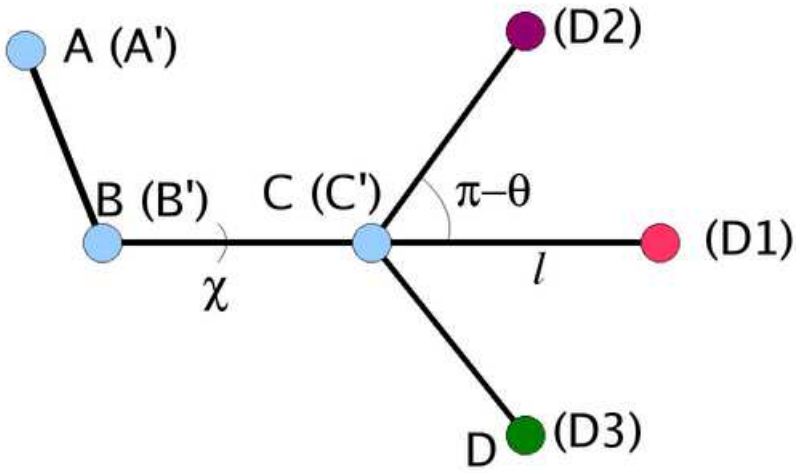


Fig. 9. Fourth Atom Fixing Algorithm. The atoms in blue represent the first three atoms whose positions must be known *a priori*. The primed labels represent the same atoms after the origin of the coordinates have been shifted. Atoms in red and magenta represent the first and the second intermediate positions of the fourth atom whose coordinates are to be determined. The atom in green represents the final position of the fourth atom at the end of the procedure. l, θ and χ respectively denotes the bond length, bond angle and the torsion angle values that are essential inputs to the algorithm.

Input:

Coordinates of the first three atoms (labelled as A, B, and C in figure 9) bond length l, bond angle θ and torsion angle χ

Step 1:

Translate the origin of the external coordinate system to coincide with the position of atom C. Thus if $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ represents the position vectors of atoms A, B and C generate:

$$\boldsymbol{A}' = \boldsymbol{A} - \boldsymbol{C}$$
$$\boldsymbol{B}' = \boldsymbol{B} - \boldsymbol{C}$$
$$\boldsymbol{C}' = [0, 0, 0]^T$$

Step 2:

Consider unit vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$\boldsymbol{a} = (\boldsymbol{B}' - \boldsymbol{A}')/|\boldsymbol{B}' - \boldsymbol{A}'|$$
$$\boldsymbol{b} = (\boldsymbol{C}' - \boldsymbol{B}')/|\boldsymbol{C}' - \boldsymbol{B}'|$$

and the perpendicular

$$\boldsymbol{q} = \boldsymbol{a} \times \boldsymbol{b}$$

Generate atom D1 along $\boldsymbol{b}$ and at a distance l from the origin (atom C). Thus if $\boldsymbol{b}$ has components $(a,\ b,\ c)$ and the coordinates of D1 are given by (α, β, γ) then we have:

$$\alpha = al$$
$$\beta = bl$$
$$\gamma = cl$$

Step 3:
Rotate atom D1 about the axis $\boldsymbol{q}$ and angle $(\pi - \theta)$.
Let the new position be designated D2 (See Note 4 for the theory of rotations).
The position vector of atom D2 is given by:

$$\boldsymbol{D}2 = \mathcal{R}^{\boldsymbol{q}}_{\pi-\theta}\boldsymbol{D}1$$

where $\mathcal{R}^{\boldsymbol{q}}_{\pi-\theta}$ is the rotation matrix (in polar form) about the axis $\boldsymbol{q}$ and angle $\pi - \theta$

Step 4:
Rotate atom D2 about the vector $\boldsymbol{b}$ (parallel to bond B--C) and angle χ. Let the new atom have position vector $\boldsymbol{D}3$. Thus:

$$\boldsymbol{D}3 = \mathcal{R}^{\boldsymbol{b}}_{\chi}\boldsymbol{D}2$$

Step 5: Shift the origin of the coordinate system back to its original position. Thus:

$$\boldsymbol{D} = \boldsymbol{D}3 + \boldsymbol{C}$$

The coordinates of the atom D, thus obtained, are the required external coordinates.

5 Torsion angle designations in proteins and nucleic acids

5.1 Proteins

Since the protein backbone consists of a repeating series of N,C^{α}, and C atoms, only three types of torsions angles are required to define the protein backbone. These are:

Symbol	Definition
ϕ_i	C_{i-1}-N_i-C_i^α-C_i
ψ_i	N_i-C_i^α-C_i-N_{i+1}
ω_i	C_{i-1}^α-C_{i-1}-N_i-C_i^α

Table 2. Definition and symbols of the backbone torsion angles in polypeptides. The subscript i refers to the i^{th} residue in the polypeptide.

It is reasonable to assume that bond lengths and bond angles of any molecule always remain within very narrow limits around their equilibrium values. Thus, without serious loss of accuracy, one may say that the conformation of the protein backbone is solely defined by the three torsion angles ϕ, ψ and ω. Moreover it was shown by Pauling that the peptide bond is always planar and hence ω is a constant around 180°.[9] This leaves only the torsion angles ϕ and ψ as the variable parameters. G.N. Ramachandran and colleagues showed for the first time that free variation is not permited even for ϕ and ψ and only certain allowed combinations of these angles are allowed for a polypeptide. The allowed combinations of the ϕ and ψ angles are described in the celebrated Ramachandran map, which has made a tremendous contribution to our understanding of protein structure.[10] Like the main chain conformation, the conformation of the side chains can also be specified by a series of torsion angle values. The side chain torsion angle are labelled χ_i where i is an index varying with each single bond in the amino acid side chain. The torsion angle closest to the main chain is labelled χ_1.

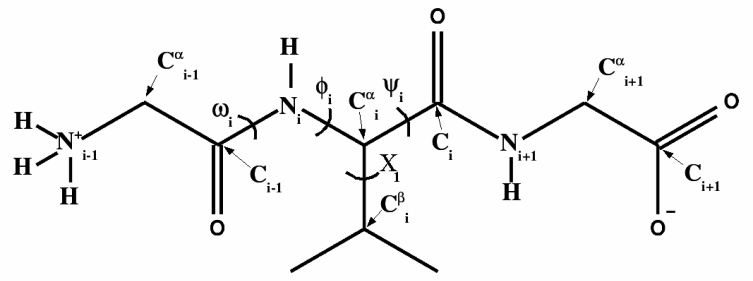


Fig. 10. Pictorial representation of a tripeptide (Glyclyl valyl glycine) with the main chain torsion angles ω, ϕ, ψ and the sidechain torsion angle χ_1 shown for the central residue.

9. Analysis of a large number of polypeptide crystal structures show that a certain amount of non-planarity is permitted for the peptide bond. However the allowed variation in ω is still restricted to $\sim 180 \pm 20°$.

10. We will discuss Ramachandran maps in more detail later in this article.

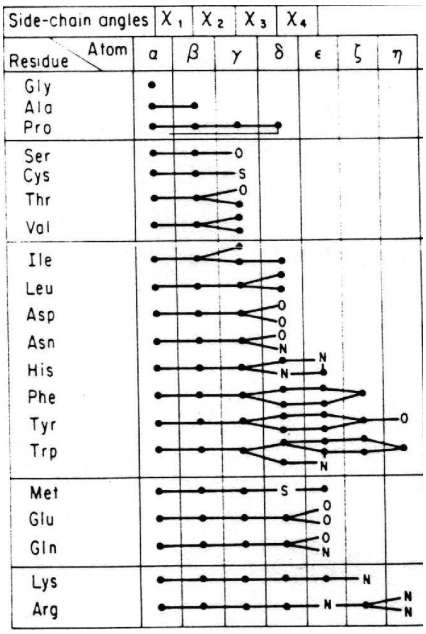


Fig. 11. Schematic view of the definition of χ torsion angles of amino acid side-chains.

Detailed analysis of the amino acid sidechain conformations in protein crystal structures have led to the realisation that the conformations of particular amino acid sidechains cluster around particular sets of χ values. Such clusters have been dubbed *rotamers*. A number of groups[11] have identified these clusters and carried out detailed statistical analysis. Such *rotamer libraries* have proved to be highly useful during predictions of protein sidechain conformation.

5.2 Nucleic Acids

The backbone of a polynucleotide chain consists of a repeating series of atoms

11. See for example, Lovell S.C., Word J.M., Richardson J.S. and Richardson D.C. The penultimate rotamer library *Proteins Struct. Func. Genet.* (2000) **40** 389-408 and references cited therein.

viz., -P-O5'-C5'-C4'-C3'-O3'-. Torsion angles specifying the backbone conformation are defined as rotations around every single bond in this series. Each torsion angle is symbolically identified with a greek letter starting from α though ζ. The definitions and symbols of the backbone torsion angles are shown in table 3. Beside the backbone atoms, the five endocyclic torsion angles of the (deoxy)ribose sugar is also important for nucleic acid conformations and need to be specified. Finally one more torsion angle is required to specify the orientation of the glycosidic bond. Unlike in the case of polypeptides, the larger number of variable torsion angles that are required to specify a polynucleotide results in a greater potential for conformational flexibility in case of polynucleotides. However in case of polynucleotide double helices, enough conformational constraints are generated which allows specification of the backbone structure of polynucleotides with much fewer parameters.[12]

Symbol	Definition
α	${}_{(n-1)}O_{3'}$-P-$O_{5'}$-$C_{5'}$
β	P-$O_{5'}$-$C_{5'}$-$C_{5'}$
γ	$O_{5'}$-$C_{5'}$-$C_{4'}$-$C_{3'}$
δ	$C_{5'}$-$C_{4'}$-$C_{3'}$-$O_{3'}$
ε	$C_{4'}$-$C_{3'}$-$O_{3'}$-P
ζ	$C_{3'}$-$O_{3'}$-P-${O_{5'}}_{(n+1)}$
χ	$O_{4'}$-$C_{1'}$-N_1-C_2 (pyrimidines) $O_{4'}$-$C_{1'}$-N_9-C_4 (purines)
υ_o	$C_{4'}$-$O_{4'}$-$C_{1'}$-$C_{2'}$
υ_1	$O_{4'}$-$C_{1'}$-$C_{2'}$-$C_{3'}$
υ_2	$C_{1'}$-$C_{2'}$-$C_{3'}$-$C_{4'}$
υ_3	$C_{2'}$-$C_{3'}$-$C_{4'}$-$O_{4'}$
υ_4	$C_{3'}$-$C_{4'}$-$O_{4'}$-$C_{1'}$

Table 3. Definition of poynucleotide torsion angles. The subscripts $(n-1)$ and $(n+1)$ refer to the preceding and succeeding nucleotide repeat unit respectively.

12. See for example, Yathindra N. and Sundaralingam M. Analysis of possible helical structures of nucleic acids and polynucleotides. Application of (n-h) plots. *Nucleic Acids Res.* (1976) **3**,729-747

Sasisekharan V. and Pattabiraman N. Structure of DNA predicted from stereochemistry of nucleoside derivatives. *Nature* (1978) **275**,159-162

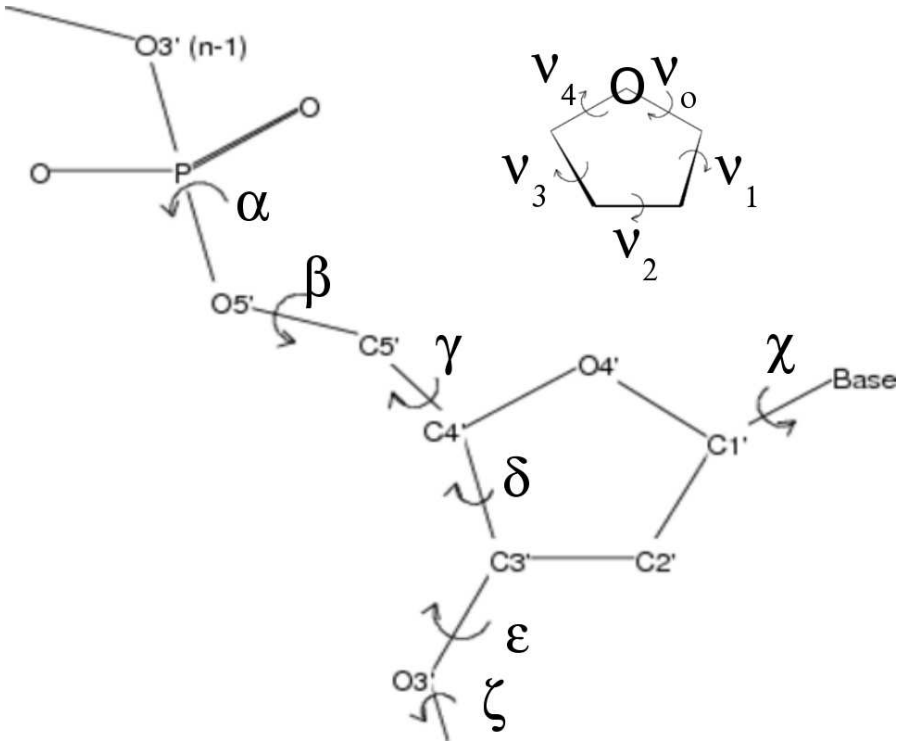


Fig. 12. Pictorial view of torsion angle definitions in a nucleotide unit. The endocyclic torsion angles of the sugar ring is also shown

6 Conformation of Cyclic Systems

The conformation of a molecule being fully specified by its internal coordinates *viz*., its bond lengths, bond angles and torsion angles, the entire conformational space accessible to a molecule can be visited by variations in its bond lengths, bond angles and torsion angles. Ordinarily the bond lengths and bond angles remain within narrow ranges specified by the type of their constituent atoms, the torsion angles however, are relatively free to vary with the exception of certain special cases like the ω torsion in peptides. In cyclic systems however, the additional ring closure interaction exerts peculiar constraints on the other internal parameters often drastically reducing the conformational variability of the molecule.

6.1 Conformation of Five-membered rings

Consider the five membered cyclic molecule in figure 13. Assuming that the bond lengths and bond angles are fixed to ideal values, the conformation of the molecule becomes only dependent on the torsion angles. However, not all the five torsion angles are required to uniquely determine the conformation. This is illustrated with the following simple argument.

The positions of any three successive atoms (for example O4', C1' and C2') can be fixed without reference to any of the torsion angles.[13] The position of atom C3' requires knowledge of the torsion angle ν_2 and fixing the position of atom C4' requires torsion angle ν_3. Thus only two out of the five endocyclic torsion angles are necessary as well as sufficient to specify the conformation of the five-membered ring in torsion angle space.

Five-membered rings like the ribose sugar and the amino acid proline play important roles in biochemistry. The conformation of these rings have a crucial role to play in many biochemical systems. Because of steric reasons five-membered rings of the furanose sugar or the amino acid proline cannot be completely planar, one or more atoms are always pushed out of the plane defined by the other atoms, a phenomenon known as *puckering*.

If four atoms in a five-membered ring are approximately in the same plane and only one atom is out of the plane then the conformation is described as an *envelope*. A furanose ring has various envelope puckering modes depending on the out of plane atom and also the direction of its displacement. For example if the C2$'$ atom in the furnaose sugar goes out of plane in the direction of the C5$'$ atom then it is known as C2$'$-*endo* pucker abbreviated ${}^2E$. DNA in the B-conformation has most of its deoxyribose sugars in the ${}^2E$ conformation. RNA and DNA in the A-conformation has the C3$'$ atom of the ribose sugar moving out of plane and in the direction of the C5$'$ atom. This conformation of the sugar ring is known as C3$'$-*endo* or ${}^3E$.

Besides the envelope forms, one may also have the so called *Twist* form, where more two atoms move out of plane[14] usually in opposite directions. Thus C3$'$-*endo*/C2$'$-*exo* or ${}^3T_2$ is a type of twist pucker where the C3$'$ atom moves out of plane in the direction *away* from the C5$'$ atom and the C2$'$ atom also moves out of plane in the direction *towards* the C5$'$ atom. The different puckered conformations found in (deoxy)ribose sugars together with their pucker designations are shown pictorially in figure 13.

13. See Note 3

14. In this case the reference plane is defined with respect to the three atoms that are closest to the mean plane of the five-membered ring.

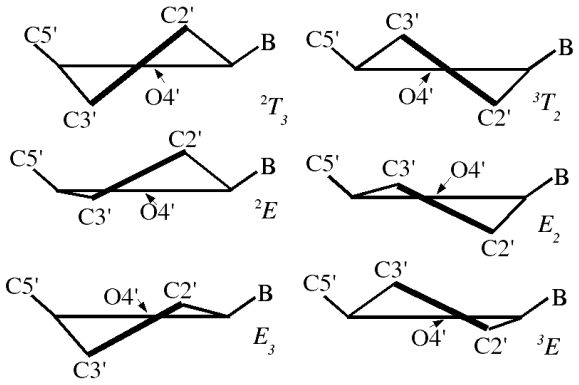


Fig. 13. Diagrammatic representation of β-D nucleoside conformers and their puckering designations

6.2 Pseudorotation

Pseudorotation is defined as form of *stereoisomerisation*[15] resulting in a structure that appears to have been produced by rotation of the entire molecule and is superimposable on the initial structure unless different positions in the molecule are distinguished by substitution including isotopic substitution. Conformational isomerism of furanose rings among the different envelope or twist puckered forms is just one example of pseudorotation. Another example, the so called *Berry pseudorotation* is a polytopal resaarangement that provides a pathway for the isomerisation of trigonal bipyramidal compounds e.g., λ^5_1-phosphanes. The five bonds incident on the central atom E are designated as e_1, e_2, e_3, a_1 and a_2 respectively (figure 14). On the basis of their orientation about the central atom, the first three of them are called equatorial bonds and the last two are known as apical bonds. In Berry pseudorotation two of the equitorial bonds move apart to become apical bonds and at the same time the two apical bonds come together to become equitorial bonds.

15. This term can include both configurational as well as as conformational isomerism.

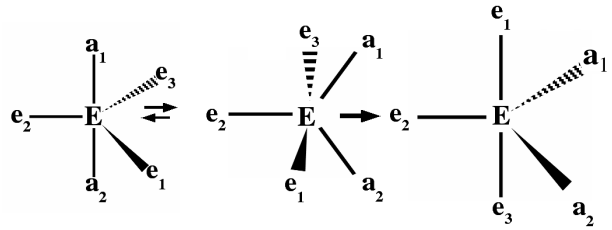


Fig. 14. Schematic view of Berry pseudorotation.

Altona and Sundaralingam[16] carried out an elegant analysis of pseudorotation in five-membered rings and defined a set of new parameters for its description that has found a lot of use in the analysis of sugar ring conformational dynamics. The parameters defined by Altona and Sundaralingam were ν_m the maximum angle of torsion (also called the puckering amplitude) and P the pseudorotation phase angle.[17] The phase angle locates a ring conformation on the pseudorotational pathway relative to some standard conformation and dependent on any pair of endocyclic torsion angles.

Since only two of the five endocyclic torsion angles in a furanose sugar ring are really independent one can express the interdependence between them using the following set of equations:

$$\nu_2 = \nu_m \cos P \tag{4}$$

$$\nu_3 = \nu_m \cos (P + \delta) \tag{5}$$

$$\nu_4 = \nu_m \cos (P + 2\delta) \tag{6}$$

$$\nu_o = \nu_m \cos (P + 3\delta) \tag{7}$$

$$\nu_1 = \nu_m \cos (P + 4\delta) \tag{8}$$

where $\delta = 4\pi/5$ and other symbols are as defined previously.

Using equations 4 to 8 one can derive a very useful relation.[18]

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_o)}{2\nu_2[\sin(\pi/5) + \sin(2\pi/5)]} \tag{9}$$

16. Altona C. and Sundaralingam M. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Amer. Chem. Soc* (1972) **94** 8205-8212

17. The notation used here differes trivially from that used by Altona and Sundaralingam.

18. See Note 5 and the following article for a derivation.

Altona C., Geise H.J., and Romers C. *Tetrahedron* (1968) **24** 13

Once the value of the phase angle (P) is known from equation 9, the maximum angle of torsion (ν_m) can be calculated using equation 4. it is important to realise that the maximum torsion travels twice along the ring in one pseudorotational cycle (i.e., P going from 0° to 360°). When P changes by 180°, a mirror image conformation results with the sign of all the endocyclic torsion angles being reversed.

In this analysis the standard conformation ($P = 0°$) is chosen to be that conformation where the value of the C1′-C2′-C3′-C4′ torsion angle (ν_2) is maximally positive. This corresponds to the symmetrical twist C2-*exo*-C3-*endo* or ^{3}T$_2$ conformation. By systematically varying P from 0° to 360° one can generate all the puckered conformations of a five-membered ring. Such a plot known as the *pseudorotational wheel* is shown in figure 15. It is known that in most ring systems the energy of the diffrent puckered conformations is relatively independent of ν_m but depend strongly upon P. Thus the pseudorotational wheel depicted in figure 15 can be thought of as a conformational map of the five-membered ribose sugar that also denote a pathway (known as the pseudorotational pathway) for conformational transitions in the sugar ring. Sugar ring conformations with P ranging from 0° to 36° are very common in RNA and in DNA with the A-conformation. These are commonly known as the N (for north) family of conformations. DNA molecules in the B-conformation have their sugar rings with P ranging from 136° to 216°. Such sugar conformations care commonly said to belong to the S (for south) family.

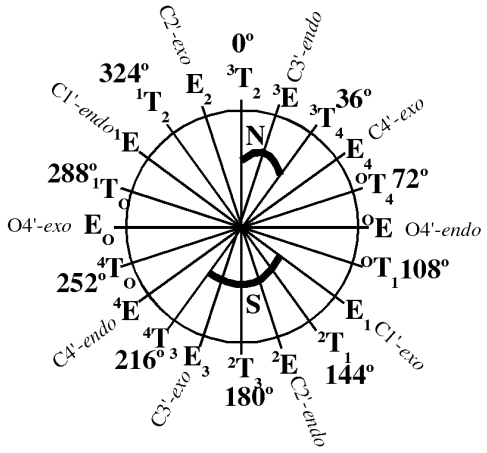


Fig. 15. Pseudorotational wheel. Circular plot of the pseudorotational phase angle P values with the corresponding sugar ring conformations. The range of P values for the N and S families of sugar ring conformations are also shown.

6.3 Cremer-Pople Parameters

The Altona-Sundaralingam parameters P and ν_m provide an extremely powerful tool for pseudorotational analysis of five-membered rings on account of their elegant simplicity, however they suffer from some slight problems. For example, all the five endocyclic torsions are not treated identically. There exist a slight dependence of the parameters on the choice of the reference torsion angle (ν_2 in this case). Cremer and Pople[19] suggested a different formulation of the pseudorotational problem that overcomes these problems. Their treatment also has the added attraction that it is completely general and can be applied to any cyclic system and not just five-membered rings.

Instead of defining the ring pucker in terms of torsion angle values, Cremer and Pople define them in terms of distances of the atoms from a ring reference plane. The reference plane is defined as follows:

Let the positions of the nuclei of N atoms in a puckered ring be characterised by position vectors $\boldsymbol{R}_j$ $j = (1, 2, 3...N)$. The origin is shifted to the geometrical centre (or the centre of mass if the N nuclei have the same mass). Then we have:

$$\sum \boldsymbol{R}_j = \mathbf{0} \tag{10}$$

The coordinate system is now chosen that the origin is the geometrical centre and the z-axis is perpendicular to the reference plane. Let z_j be the displacement of the individual atoms from this reference plane. It follows that:

$$\sum_{j=1}^{N} z_j = 0 \tag{11}$$

Cremer and Pople now impose the additional conditions:

$$\sum_{j=1}^{N} z_j \cos[2\pi(j-1)/N] = 0 \tag{12}$$

$$\sum_{j=1}^{N} z_j \sin[2\pi(j-1)/N] = 0 \tag{13}$$

Equations 12 and 13 uniquely fixes the reference plane. In case of small puckering displacements they lead to the condition of no overall angular momentum. However they are generally applicable to finite displacements, to differences in bond lengths and bond angles and to non-equivalent atoms. Most importantly these conditions are invariant to differences in numbering of the atoms.

Given the position vectors of the atoms, the orientation of the reference plane is derived as follows:

19. Cremer D. and Pople J.A. A general definition of ring puckering coordinates. *J. Amer. Chem. Soc.* (1975) **97** 1354-1358

Define a pair of vectors

$$\boldsymbol{R}' = \sum_{j=1}^{N} \boldsymbol{R}_j \sin\left[2\pi(j-1)/N\right] \tag{14}$$

$$\boldsymbol{R}'' = \sum_{j=1}^{N} \boldsymbol{R}_j \cos[2\pi(j-1)/N] \tag{15}$$

The z-axis is chosen perpendicular to $\boldsymbol{R}'$ and $\boldsymbol{R}''$ and is given by:

$$\boldsymbol{n} = \boldsymbol{R}' \times \boldsymbol{R}'' / |\boldsymbol{R}' \times \boldsymbol{R}''| \tag{16}$$

The full set of displacements from there reference plane are then given by:

$$z_j = \boldsymbol{R}_j \cdot \boldsymbol{n} \tag{17}$$

Generalised ring puckering parameters are now defined in the following way:

If N is odd and $N > 3$ define q_m and ϕ_m by:

$$q_m \cos\phi_m = \sqrt{2/N} \sum_{j=1}^{N} z_j \cos[2\pi m(j-1)/N] \tag{18}$$

$$q_m \sin\phi_m = -\sqrt{2/N} \sum_{j=1}^{N} z_j \sin[2\pi m(j-1)/N] \tag{19}$$

where $m = (2, 3..., (N-1)/2)$. $q_m, (q_m > 0)$ and $\phi_m, (0 \leqslant \phi_m < 2\pi)$ represent respectively, the amplitudes and phase angles of puckering. When N is even, values of m can reach upto $N/2 - 1$ and there is an additional puckering coordinate:

$$q_{N/2} = (1/\sqrt{N}) \sum_{j=1}^{N} z_j \cos[(j-1)\pi] = (1/\sqrt{N}) \sum_{j=1}^{N} (-1)^{j-1} z_j \tag{20}$$

Unlike other q values, $q_{N/2}$ can have either sign. Thus the total number of puckering parameters for a ring with N atoms is $N - 3$. Equations 11 to 13 and 18 to 20 sets up N independent linear equations for the N displacements z_j and may be solved in terms of the puckering parameters q_m and ϕ_m. The results are:

$$z_j = \sqrt{(2/N} \sum_{m=2}^{\frac{N-1}{2}} q_m \cos[\phi_m + 2\pi m(j-1)/N] \tag{21}$$

(for odd N)

$$z_j = \sqrt{2/N} \sum_{m=2}^{\frac{N}{2}-1} q_m \cos[\phi_m + 2\pi m(j-1)/N] + \frac{1}{\sqrt{N}} q_{N/2} (-1)^{j-1} \tag{22}$$

(for even N)

The quantity

$$\sum_{j=1}^{N} z_j^2 = \sum_{m} q_m^2 = Q^2 \tag{23}$$

is called the *total puckering amplitude.*

For a five-membered ring, the value of m in equations 18 and 19 is restricted to 2. Solving the equations results in a single amplitude-phase pair (q,ϕ). The displacements of individual atoms from the reference plane (z_j) can be calculated from (q, ϕ) using equation 21 and is given by:

$$z_j = \sqrt{(2/5)}\, q \cos[\phi + 4\pi(j-1)/5] \tag{24}$$

For a six-membered ring again $m = 2$ only. Thus there are three parameters defining the puckering viz., the amplitude-phase pair (q_2, ϕ_2) and a single puckering coordinate q_3. Alternately one can express the parameters in terms of a sperical polar set (Q, θ, ϕ) where the total puckering amplitude Q is given by:

$$Q = \sqrt{q_2^2 + q_3^2} \tag{25}$$

and θ, $(0 < \theta < \pi)$, is a parameter given by:

$$q_2 = Q \sin\theta \tag{26}$$

$$q_3 = Q \cos\theta \tag{27}$$

Every conformation of a six membered ring can be plotted on the surface of a sphere in terms of these three parameters.

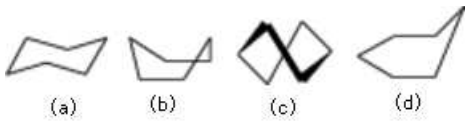


Fig. 16. Typical conformations of the cyclohexane ring. The chair form (a) is characterized by $\theta = 0$ or 180^o, $q_2 = 0$ and $q_3 = \pm Q$. Varying the phase angle ϕ (ϕ = 0, 60, 120, 180, 240, 300^o) leads to six different boat conformations exemplified by (b). Six twist boat conformations exemplified by (c) appear when ϕ = 30, 90, 150, 210, 270, 300^o. These can be interconverted along a pseudorotational path. The half-chair form (d) is intermediate between (a) and (c) and is characyterized by $\tan\theta = \sqrt{3/2}$ and $\phi = 90^o$.

Comparing the Cremer-Pople parameters with the Altona-Sundaralingam parameters, we find that although they serve the same function to characterise a cyclic molecule, the former set is more general, being applicable to rings with any number of atoms and not just five like the Altona-Sundaralingam parameters. The Cremer-Pople parameters are also more general than the Altona-Sundaralingam parameters in the sense that they are applicable also to conformationally deformed rings where bond lengths and bond angles cannot be assumed to remain essentially constant. Both parameter sets work equally well in the special case when there are only five atoms in the ring (like in furanose sugars) and when it is safe to assume that bond lengths and bond angles remain constant during conformational transitions.

7 Conformation of helices

A regular structure in a chain molecule which can be described by a repeated set of torsion angles $(\phi_1, \phi_2, \ldots \phi_m)_n$ is a helix.[20] A *helix*[21] is defined as the locus of a point that undergoes simultaneous rotation by an angle Δ about an axis $\boldsymbol{A}$ (called the *helix axis*) and is also being translated along $\boldsymbol{A}$ by a fixed amount t. Any sequence of atoms that can be joined by a space curve with these properties is defined as a helix. A vector valued function

$$\boldsymbol{r}(n) = a\cos(n\Delta)\boldsymbol{i} + a\sin(n\Delta)\boldsymbol{j} + nt\boldsymbol{k} \tag{28}$$

mathematically defines a helix. The curve traced by $\boldsymbol{r}$ is a helix lying on the surface of a cylinder whose $\boldsymbol{i}$ and $\boldsymbol{j}$ components satisfy the equation:

$$x^2 + y^2 = a^2[\cos(n\Delta)]^2 + a^2[\sin(n\Delta)]^2 = a^2 \tag{29}$$

The radius of the cylinder, and hence the radius of the helix is given by a. The curve rises as the $\boldsymbol{k}$-component $z = nt$ increases. One turn of the helix is completed when $n\Delta$ just crosses 2π. The equation

$$x = a\cos(n\Delta),\ y = a\sin(n\Delta),\ z = nt \tag{30}$$

describes the helix.

If $n\Delta = 2\pi$ for some integer n. Then it means that the n^{th} point (or repeating unit) is exactly above the first repeating unit.[22] Helices with this property are called *integral helices*. In many cases $n\Delta \neq 2\pi$ for any integer n. But one may have $n\Delta = m\pi$ where m is an integer multiple of 2. In such cases the number of atoms per turn of the helix is a real number ($n = 2\pi/\Delta$) but not an integer. Although such statements do not necessarily have any physical meaning, they are in common use and must be viewed as an approximation only. Helices with non-integral values of n are obviously known as *non-integral helices*.[23] Depending on the direction of the rotation as one goes from one repeating unit to another, a helix can be either *right-handed* or *left handed*. When viewed perpendicular to the helix axis, a right-handed helix is one when the nearer part of the helix (to the viewer) travels from the left to the right.[24]

---

20. For peptides, this set comprises of three torsion angles viz., ϕ, ψ and ω. For nucleic acids the set consists of the six backbone torsion angles $\alpha, \beta, \gamma, \delta, \varepsilon$ and ζ. When the set of backbone torsion angles are exactly repeated the conformation of the so called *repeating unit* (i.e., a peptide or a nucleotide unit) remains identical throughout the length of the structural segment.

21. The term helix is derived from an old Greek word for "spiral".

22. Alternatively stated, there are n repeating units per turn of the helix.

23. A correct way to describe non-integral helices would be to say that they have n' atoms per m turns of the helix, where $n' = mn$ and is an integer.

24. A right-handed helix is the path traced by a screw-driver that moves forward when turned clockwise, i.e., with the right hand.

Right-handed helices have a positive angle of rotation. When the rotation angle is 2π, a so called 2-fold helix (i.e., with $n=2$) results. Such a helix does not have any handedness. A helix, therefore, is characterised by the following set of parameters:

* A helix axis $\boldsymbol{A}$, which is a vector with components (A_x, A_y, A_z) about which rotations of the repeating unit take place.
* The unit *twist* (Δ), which is the angle of rotation about the helix axis. A right-handed helix has a positive twist.
* The number of repeating units (n) per turn of the helix. $n=2\pi/\Delta$. n may or may not be an integer.
* The unit *rise* along the helix axis (t). The *pitch* (p) of a helix is defined as the rise along the helix axis per turn of the helix. Hence $p=nt$.
* The *radius* (a) of the helix.

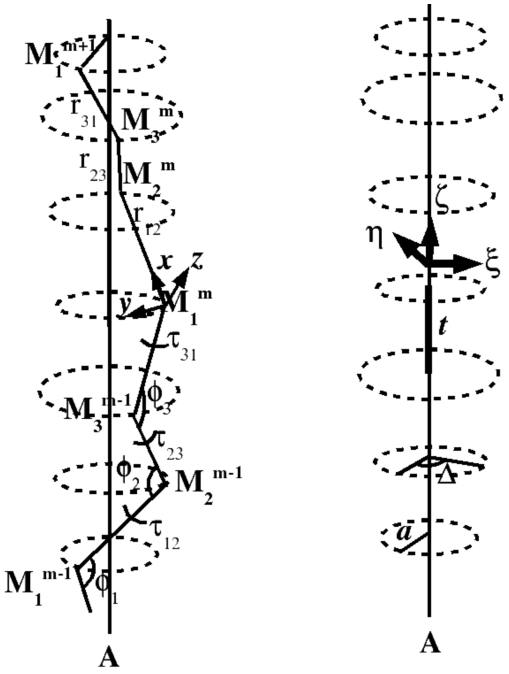


Fig. 17. The left panel shows a schematic representation of a helix. The internal coordinates (bond lengths, bond angles and torsion angles), and right-handed Cartesian coordinate system (x, y, z) that defines the repeating unit are also depicted. The right panel depicts the helical parameters (a, Δ, t) that can be used to characterise the helix. A right handed Cartesian coordinate system (ξ, η, ζ) that can be used to characterise the helix is also shown.The helix axis is also indicated on both figures.

7.1 Calculation of the Helical Parameters

The helical parameters (a, Δ, t) together with the helix axis are extremely useful parameters for the characterisation of a helix. Hence it is important to be able to derive these parameters from either internal or external coordinates of a repeating unit. A general method for calculation of helical parameters due to Sugeta and Miyazawa[25] is given here. The central principle of their method is that since a regular helix has a screw symmetry, one can bring repeating unit into coincidence with the next repeating unit by a rotation of unit twist (Δ) about the helical axis and a translation (t) along the same axis. Alternatively, the same operation can be done by a sequence of translations along directions defined by the bond vectors connecting the repeating units and rotations by the appropriate bond angles and torsion angles about (or perpendicular to) the corresponding bonds. Since these two aproaches are independent but completely equivalent, they can be used to set up a series of equations in terms of the helical parameters $(a, \Delta, t,)$ and the components of the helix axis (l, m, n) and solved. The mathematical details of the above procedure is now given below:

We assume the helical main chain to be made up of p atomic repeat units, with bond lengths r_{ij}, bond angles ϕ_i and torsion angles τ_{ij} (shown in figure 17 on the left). Set up right-handed Cartesian coordinates $\boldsymbol{X}(x, y, z)$ with the origin on the atom M_1^m. The x-axis is parallel to the vector $M_1^m - M_2^m$ while the y-axis lies in the plane $M_p^{m-1} - M_1^m - M_2^m$ and makes an acute angle with the bond $M_p^{m-1} - M_1^m$. Figure 17 (right panel) also shows the helical parameters (a, Δ, t), right-handed Cartesian coordinates $\Xi(\xi, \eta, \zeta)$ are also set up with the origin on the helix axis. The ξ-axis being perpendicular to the helix axis and points towards atom M_1^m, while the η-axis is parallel to the helix axis.

The position vector ($\boldsymbol{X}_i^m$) of the i^{th} atom of the m^{th} unit is now given by[26]:

$$\boldsymbol{X}_i^m = \boldsymbol{B}_{12} + \mathcal{R}_{12}\boldsymbol{B}_{23} + \mathcal{R}_{12}\mathcal{R}_{23}\boldsymbol{B}_{34} + \cdots + \mathcal{R}_{12}\mathcal{R}_{23}\cdots\mathcal{R}_{i-2,i-1}\boldsymbol{B}_{i-1,i} \quad (i > 2) \tag{31}$$

25. Sugeta H. and Miyazawa T. General Method for Calculating Helical Parameters of Polymer Chains from Bond Lengths, Bond Angles, and Internal-Rotation Angles. *Biopolymers* (1967) **5** 673-679.

26. The rotation matrices $\mathcal{R}^\tau$ and $\mathcal{R}^\phi$ correspond to rotations about the x, and z-axes respectively. One recognises that the matrix product $\mathcal{R}^\tau\mathcal{R}^\phi$ would come in the 'Fourth-atom fixing algorithm' if one chooses to have the origin on the 3$^{\text{rd}}$ atom (C) and the x-axis along the CB (where B denotes the 2$^{\text{nd}}$ atom) bond.

where

$$\mathcal{R}_{ij} = \mathcal{R}^{\tau}_{ij}\mathcal{R}^{\phi}_{j}$$

$$\mathcal{R}^{\tau}_{ij}\mathcal{R}^{\phi}_{j} = \begin{pmatrix} \cos\phi_j & -\sin\phi_j & 0 \\ \sin\phi_j & \cos\phi_j & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\tau_{ij} & -\sin\tau_{ij} \\ 0 & \sin\tau_{ij} & \cos\tau_{ij} \end{pmatrix}$$

$$\Rightarrow \mathcal{R}^{\tau}_{ij}\mathcal{R}^{\phi}_{j} = \begin{pmatrix} -\cos\phi_j & -\sin\phi_j & 0 \\ \sin\phi_j\cos\tau_{ij} & -\cos\phi_j\cos\tau_{ij} & -\sin\tau_{ij} \\ \sin\phi_j\sin\tau_{ij} & -\cos\phi_j\sin\tau_{ij} & \cos\tau_{ij} \end{pmatrix} \tag{32}$$

and

$$\boldsymbol{B}_{ij} = \begin{pmatrix} r_{ij} \\ 0 \\ 0 \end{pmatrix} \tag{33}$$

Defining

$$\mathcal{R}^0 = \boldsymbol{I} \tag{34}$$

$$\mathcal{R}^i = \mathcal{R}_{12}\mathcal{R}_{23}\cdots\mathcal{R}_{i,i+1} \qquad (0 < i < p) \tag{35}$$

$$\mathcal{R} = \mathcal{R}_{12}\mathcal{R}_{23}\cdots\mathcal{R}_{p-1,p}\mathcal{R}_{p,1} \tag{36}$$

Thus the position vector $\boldsymbol{X}^m_i$ can be written as:

$$\boldsymbol{X}^m_i = \sum_{n=1}^{i-1} \mathcal{R}^{(n-1)}\boldsymbol{B}_{n,n+1} \tag{37}$$

The vector $\boldsymbol{B}$ from atom M^m_1 to atom M^{m+1}_1 is given by:

$$\boldsymbol{B} = \boldsymbol{X}^{m+1}_1 = \sum_{n=1}^{p-1} \mathcal{R}^{n-1}\boldsymbol{B}_{n,n+1} + \mathcal{R}^{p-1}\boldsymbol{B}_{p,1} \tag{38}$$

Accordingly the position vector of the i^{th} atom of the $(m+1)^{\text{th}}$ unit i given by:

$$\boldsymbol{X}^{m+1}_i = \mathcal{R}\boldsymbol{X}^m_i + \boldsymbol{B} \tag{39}$$

$$\Rightarrow \boldsymbol{X}^m_i = \mathcal{R}^T(\boldsymbol{X}^{m+1}_i - \boldsymbol{B}) \tag{40}$$

Where $\mathcal{R}^T$ is the transpose of the $\mathcal{R}$ matrix.

Using equations 39 and 40 one can define the vector $\boldsymbol{B}'$ going from atom M^{m-1}_1 to atom M^m_1 and also the vector

$\boldsymbol{B''}$ going from atom M_1^{m+1} to atom M_1^{m+2} as follows:

$$\boldsymbol{B'} = \boldsymbol{X}_1^m - \boldsymbol{X}_1^{m-1} = \boldsymbol{X}_1^m - \mathcal{R}^T(\boldsymbol{X}_1^m - \boldsymbol{B}) = \mathcal{R}^T\boldsymbol{B} \tag{41}$$

$$\boldsymbol{B''} = \boldsymbol{X}_1^{m+2} - \boldsymbol{X}_1^{m+1} = \mathcal{R}\boldsymbol{X}_1^{m+1} + \boldsymbol{B} - \boldsymbol{X}_1^{m+1} = \mathcal{R}\boldsymbol{B} \tag{42}$$

and $\boldsymbol{X}_1^m$ is a null vector.

Two vectors $\boldsymbol{C}$ and $\boldsymbol{C'}$ are further constructed as follows:

$$\boldsymbol{C} = \boldsymbol{B'} - \boldsymbol{B} \tag{43}$$

$$\boldsymbol{C'} = \boldsymbol{B} - \boldsymbol{B''} \tag{44}$$

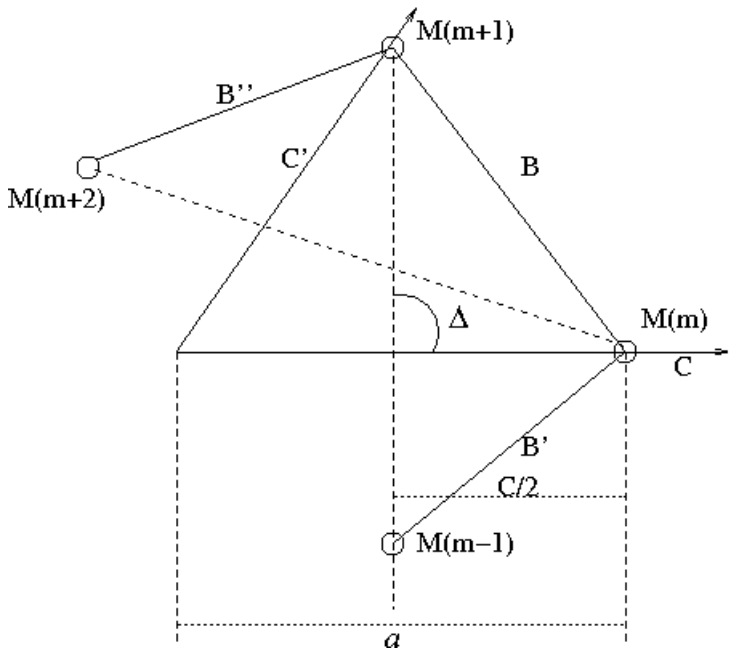


Fig. 18. Helical Parameters Δ and a together with the vectors $\boldsymbol{B}, \boldsymbol{B'}, \boldsymbol{B''}$, $\boldsymbol{C}$ and $\boldsymbol{C'}$.

The helical parameters viz., twist (θ) , radius (ρ) and rise (d) are now given by the relations:

$$\cos\Delta = (\boldsymbol{C}\cdot\boldsymbol{C'})/(\boldsymbol{C}\cdot\boldsymbol{C}) \tag{45}$$

$$a(1-\cos\Delta) = \sqrt{(\boldsymbol{C}\cdot\boldsymbol{C})}/2 \tag{46}$$

$$t^2 + 2a^2(1-\cos\Delta) = (\boldsymbol{B}\cdot\boldsymbol{B}) \tag{47}$$

$$t\sin\Delta = \boldsymbol{B}\cdot(\boldsymbol{C}\times\boldsymbol{C'})/(\boldsymbol{C}\cdot\boldsymbol{C}) \tag{48}$$

The unit vectors $\boldsymbol{e}_\xi, \boldsymbol{e}_\eta, \boldsymbol{e}_\zeta$, along the ξ, η and ζ axes are given by:

$$\boldsymbol{e}_\xi = \boldsymbol{C}/\sqrt{(\boldsymbol{C}\cdot\boldsymbol{C})} \tag{49}$$

$$\boldsymbol{e}_\zeta = (\boldsymbol{C}\cdot\boldsymbol{C}')/[(\boldsymbol{C}\cdot\boldsymbol{C})\sin\Delta] \tag{50}$$

$$\boldsymbol{e}_\eta = \boldsymbol{e}_\zeta \times \boldsymbol{e}_\xi \tag{51}$$

Thus the transformation between the Cartesian coordinate systems $\boldsymbol{X}(x, y, z)$ and $\boldsymbol{\Xi}(\xi, \eta, \zeta)$ is given by:

$$\boldsymbol{\Xi} = \mathcal{T}\boldsymbol{X} + \boldsymbol{L} \tag{52}$$

Where:

$$\mathcal{T} = \begin{pmatrix} (\boldsymbol{e}_\xi)_x & (\boldsymbol{e}_\xi)_y & (\boldsymbol{e}_\xi)_z \\ (\boldsymbol{e}_\eta)_x & (\boldsymbol{e}_\eta)_y & (\boldsymbol{e}_\eta)_z \\ (\boldsymbol{e}_\zeta)_x & (\boldsymbol{e}_\zeta)_y & (\boldsymbol{e}_\zeta)_z \end{pmatrix} \tag{53}$$

and

$$\boldsymbol{L} = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix} \tag{54}$$

Given the coordinates of the i^{th} atom in the m^{th} unit in the $\boldsymbol{\Xi}(\xi, \eta, \zeta)$ system the coordinates of the corresponding atom in the $m+s^{\text{th}}$ is given by:

$$\boldsymbol{\Xi}_i^{m+s} = \mathcal{N}^s\boldsymbol{\Xi}_i^m + s\boldsymbol{T} \tag{55}$$

where:

$$\mathcal{N} = \begin{pmatrix} \cos\Delta & -\sin\Delta & 0 \\ \sin\boldsymbol{\Delta} & \cos\Delta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{56}$$

$$\boldsymbol{T} = \begin{pmatrix} 0 \\ 0 \\ t \end{pmatrix} \tag{57}$$

From the $\boldsymbol{\Xi}$ vectors the helical parameters a_i, t_{ij} and Δ_{ij} are given by:

$$a_i = \sqrt{(\xi_i^m)^2 + (\eta_i^m)^2} \tag{58}$$

$$t_{ij} = \zeta_j^m - \zeta_i^m \tag{59}$$

$$\Delta_{ij} = \cos^{-1}[(\xi_i^m \xi_j^m + \eta_i^m \eta_j^m)/a_i a_j] = \sin^{-1}[(\xi_i^m \eta_j^m - \eta_i^m \xi_j^m)/a_i a_j] \qquad (60)$$

Thus the protocol for the calculation of helical parameters requires the Cartesian coordinates $\boldsymbol{X}(x, y, z)$ of three successive repeating units. These are then transformed to a new set of Cartesian coordinates $\boldsymbol{\Xi}(\xi, \eta, \xi)$ and thereafter the helical parameters are obtained using equations 58-60.

8 Comparison between Molecular Structures

There can be a large number of situations when an accurate comparison of a pair of molecular structres is required. During the course of evolution protein sequences and structures undergo small but significant changes. One may perhaps like to have a quantitative measure of the effect of changes in sequence of a protein to changes in backbone structure. This would undoubtedly require an accurate comparison of the backbone structures of different proteins with slightly different sequences. Similarly an enzymologist light like to accurately compare the active site structures of enzymes with different substrate specificities. Finally a drug designer would like to use structure comparison methods in order to identify compounds that are structurally similar to active lead compounds. Careful structural comparison is a very important part of major drug discovery algorithms like CoMFA.

8.1 Measures for Structure Comparison

The first requirement for structure comparison at atomic resolution is a quantitative measure that accurately reflects the differences in atomic structures. Unfortunately there seems to be no ideal measure of molecular structure difference (or similarity) at this time. However, among the various measures proposed, two of them *viz*., cRMSD and dRMSD are far ahead of others in terms of popularity. cRMSD and dRMSD stands for the Root Mean Squared Deviation of structures defined in terms of coordinates or distances respectively. These can be described mathematically as follows:

$$\text{cRMSD} = \sqrt{\frac{\sum_{i=1}^{N} [(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2]}{N}} \qquad (61)$$

where x_i, y_i, z_i ($i = 1 \cdots N$) and x'_i, y'_i, z'_i ($i = 1 \cdots N$) refers to the atomic coordinates of pairs of equivalent atoms in the two molecular structures being compared, N being the total number of equivalent atoms.

$$\mathrm{dRMSD} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij} - d'_{ij})^2} \tag{62}$$

where d_{ij} and d'_{ij} are the equivalent interatomic distances between pairs of atoms in the two structures being compared. N being the total number of such distances.

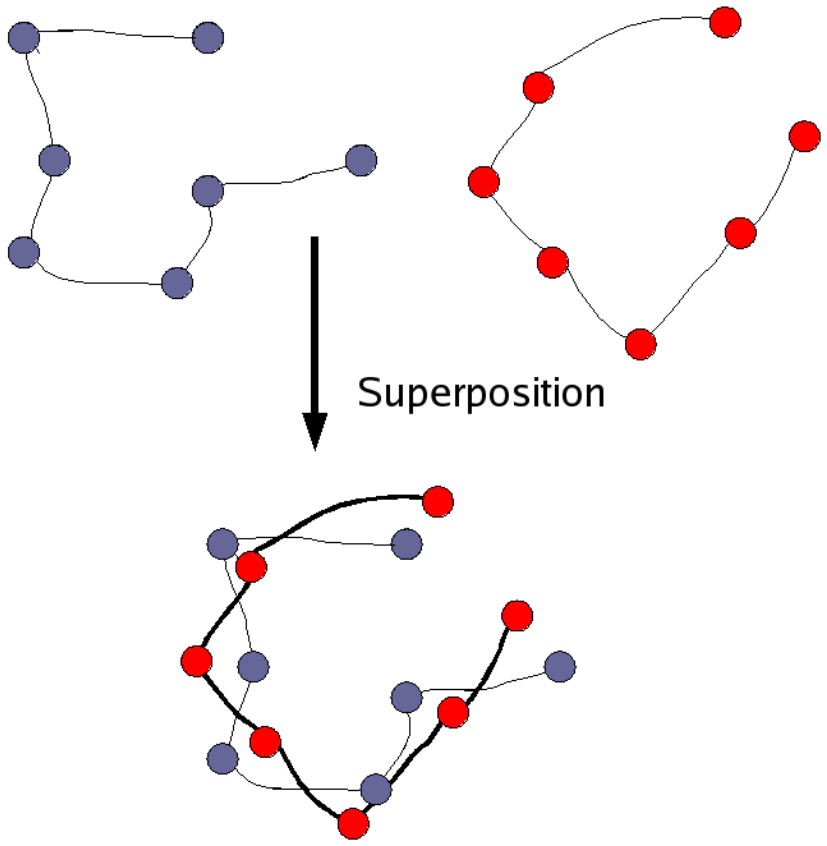


Fig. 19. Schematic view of structure superposition. Two different molecules before superposition (top). The same molecules after superpostion (bottom). Note the change in orientation for one of the molecules.

Several important points need to kept in mind when using both the above measures. Some of these points are given below:

- Both the RMSD measures are suitable for the comparison of only two structures. Generalization of the concept to three or more structures are neither unique, nor straightforward.
- Both cRMSD and dRMSD are pairwise measures defined over coordinates or distances repectively. Hence a strict equivalence relationship between pairs of atoms (or distances) must be known *a priori*. In the event such a relationship is unknown or when such a relationship is disturbed[27] in some way, both these measures become meaningless.
- Since cRMSD is defined over external coordinates it is susceptible to a change in the position or orientation of at least one of the molecules even without any conformational change. Usually cRMSD is used to compare different conformations, hence it becomes meaningful only when one of the two structures is overlaid on top of the other. This is done by an algorithm that rigidly translates and rotates one of the molecules for best structural superposition (*i.e* with minimum cRMSD) with the other. dRMSD, being defined on distances (which is a type of internal coordinate representation) is invariant to rigid rotations or translations and thus does not require any overlay algorithm.
- In many cases it is difficult to assign a statistical significance to an RMSD (cRMSD or dRMSD) value. For example it is often difficult to determine whether two molecules that have a cRMSD of 3 $\mathring{A}$ over 1000 equivalent pairs of atoms is more (or less) similar than another pair of molecules that have a cRMSD of 2.5 $\mathring{A}$ over only 10 equivalent pairs of atoms.
- Both cRMSD and dRMSD are weak in considering inherent molecular flexibility and/or imprecision in the structures of the molecules being compared.
- RMSD measures also appear to be sensitive to outliers. Thus if it happens that most of the equivalent atom-pairs after a superposition are very close to each other, but a few atoms are very far away, then these few atoms tend to influence the RMSD measure resulting in an artifactually large value.

27. By a change in atom-ordering in one of the molecules for example.

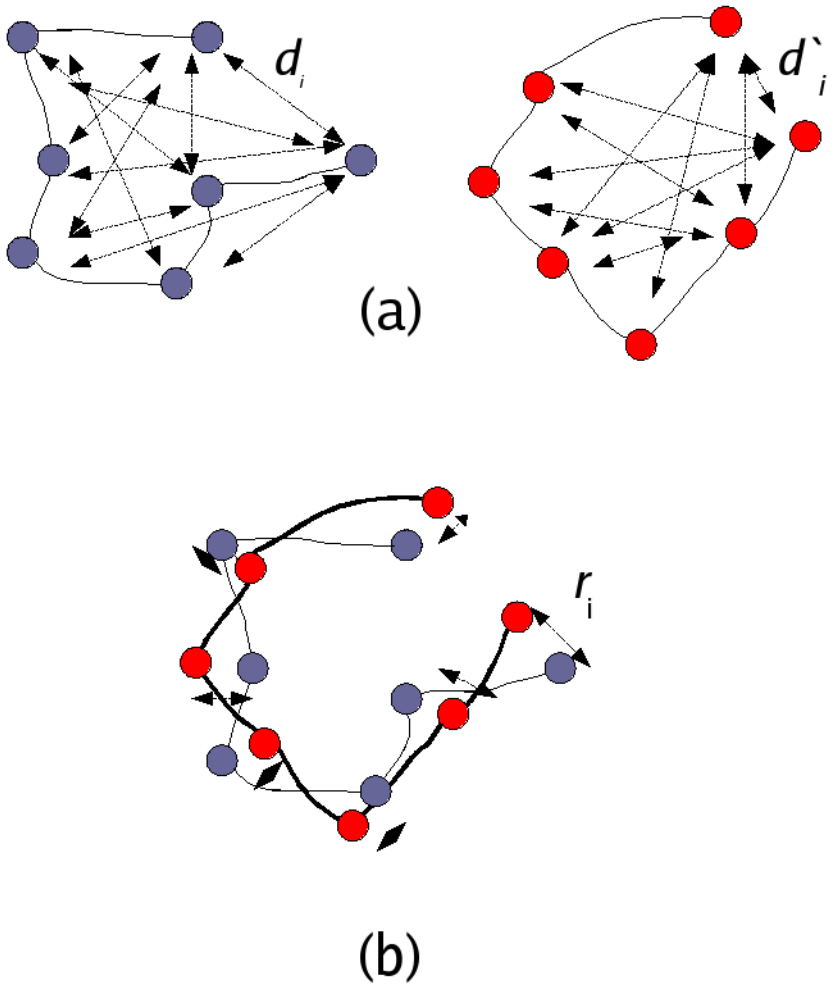


Fig. 20. Schematic illustration of the difference between cRMSD and dRMSD. (a) Two molecules being compared on the basis of internal distances to yield dRMSD. The molecules need not be superposed on each other. (b) The intermolecular distances between equivalent atoms are being considered for cRMSD. The molecules must be optimally superposed.

8.2 Structure superposition or overlay algorithms

As has been mentioned in the previous section, structure comparison measure like cRMSD are only meaningful if the molecules under comparison are overlaid on top of each other in such a way that the average sum of squared distances between equivalent atoms in the two molecules comes to a minimum. For this we can assume that one of our molecules is static and the other

one is moving. Our objective is to bring the bring the moving molecule in best superposition with the static molecule using only rigid transformations. Mathematically the problem can be stated as follows[28]:

Let matrices $\mathcal{Y}_{N\times 3}$ and $\mathcal{X}_{N\times 3}$ denote the coordinates of the static and moving molecules. The k^{th} row of the matrices specifies the position vectors of the k^{th} atom ($\boldsymbol{y}_k^T$ and $\boldsymbol{x}_k^T$ $1\leqslant k\leqslant N$) of the static and moving molecules respectively.

Find an orthogonal transformation $\mathcal{R}$ and a translation T such that the residual E (weighted by w_k) is minimized.

$$E=\frac{1}{N}\sum_{k=1}^{N} w_k\big(\left|\boldsymbol{y}_k-\mathcal{R}^T\boldsymbol{x}_k+\boldsymbol{T}_k\right|\big)^2=\frac{1}{N}(|\mathcal{Y}-\mathcal{X}\mathcal{R}+\boldsymbol{T}|)^2 \tag{63}$$

There a number of approaches to determine the optimum rotation matrix $\mathcal{R}$ and the translation $\boldsymbol{T}$. Taking the translation first, we consider the case where the optimum rotation matrix $\mathcal{R}$ is such that there is no residual. In this case we have (assuming that weights w_k are all unity):

$$\sum_{k=1}^{N}(\boldsymbol{y}_k-\mathcal{R}^T\boldsymbol{x}_k+\boldsymbol{T}_k)=0$$

so that:

$$\boldsymbol{T}_k=\mathcal{R}^T\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k-\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{y}_k=\mathcal{R}^T\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}} \tag{64}$$

i.e the optimum translation is the difference between the position vectors of the center of masses of the target ($\bar{\boldsymbol{y}}$) and the model after the optimum rotation ($\mathcal{R}^T\bar{\boldsymbol{x}}$). Shifting both sets of coordinates such that their geometric centers coincide with the origin is thus the most effective way of getting rid of the translation. The residual is now given by (assuming the weights are all unity):

$$E=\frac{1}{N}\sum_{k=1}^{N}\left(\left|\mathcal{R}^T\boldsymbol{x}_k-\boldsymbol{y}_k\right|\right)^2=\frac{1}{N}(|\mathcal{Y}-\mathcal{X}\mathcal{R}|)^2 \tag{65}$$

There are two major approaches to obtain the optimum rotation matrix $\mathcal{R}$. The first one proposed by Kabsch[29] is a special case of the *orthogonal Procrustes problem* commonly studied in multivariate statistics. An alternate method formulated independently by others[30] involves quaternion algebra. Both methods are mathematically equivalent and produce identical results.

The Kabsch method can be stated as:

Minimize $\left|\mathcal{Y}-\mathcal{X}\mathcal{R}\right|$ with respect to $\mathcal{R}$ subject to the condition $\mathcal{R}^T\mathcal{R}=\boldsymbol{I}$.

---

28. Coutsias E.A., Seok C. and Dill K.A. *J. Comput. Chem* (2004) **25** 1849-1857

29. Kabsch W *Acta Crystallogr.* (1976) **A32** 922 and Kabsch W. *Acta Crystallogr.* (1978) **A34** 827

30. Horn B.K.P. *J.Opt.Soc.Am. A.* (1987) **4** 629; Diamond R. *Acta Crystallogr.* (1989) **A44** 211; Kearsley S.K. *Acta Crystallogr.* (1989) **A45** 208

This is equivalent to the following statement:

Maximize $trace(\mathcal{R}^T\mathcal{X}^T\mathcal{Y})$ subject to the condition that $\mathcal{R}^T\mathcal{R} = \boldsymbol{I}$[31].

This problem can be approached by means of a technique common in linear algebra called *Singular Value Decomposition*(SVD)[32]. Computing the SVD of the correlation matrix $\mathcal{X}^T\mathcal{Y}$ we have:

$$\mathcal{X}^T\mathcal{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{66}$$

Thus:

$$trace(\mathcal{R}^T\mathcal{X}^T\mathcal{Y}) = trace(\mathcal{R}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T) = trace(\boldsymbol{V}^T\mathcal{R}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}) = trace(\boldsymbol{Z}\boldsymbol{\Sigma})$$

where: $\boldsymbol{Z} = \boldsymbol{V}^T\mathcal{R}^T\boldsymbol{U}$

Thus $\max trace(\mathcal{R}^T\mathcal{X}^T\mathcal{Y}) = \max trace(\boldsymbol{Z}\boldsymbol{\Sigma})$. As $\boldsymbol{Z}$ is an orthogonal matrix, the maximum is achieved when $\boldsymbol{Z} = \boldsymbol{I}$.

Hence we have for the maximum value of the trace:

$$\boldsymbol{V}^T\mathcal{R}^T\boldsymbol{U} = \boldsymbol{I}$$

This leads to the required optimum transformation matrix and is given by:

$$\mathcal{R} = \boldsymbol{U}\boldsymbol{V}^T \tag{67}$$

However this matrix may represent an improper rotation[33] if the left and right singular vector matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ have opposite chirality. To correct for this situation the optimum rotation matrix is written as:

$$\mathcal{R} = \boldsymbol{U}\begin{pmatrix} 1 & & \\ & 1 & \\ & & \chi \end{pmatrix}\boldsymbol{V}^T \tag{68}$$

where $\chi = sign(\det \mathcal{R})$.

8.3 Determining equivalent atom-pairs for superposition

A major limitation of the RMSD based superposition algorithms is the strict

31. This is due to a result first formulated by Schöneman that states:

$(|\mathcal{Y} - \mathcal{X}\mathcal{U}|)^2 = trace(\mathcal{Y} - \mathcal{X}\mathcal{U})(\mathcal{Y} - \mathcal{X}\mathcal{U})^T = trace(\mathcal{Y}^T\mathcal{Y}) + trace(\mathcal{X}^T\mathcal{X}) - 2trace(\mathcal{U}^T\mathcal{X}^T\mathcal{Y})$

Schöneman P *Psychometrika* (1966) **31** 1

32. Using SVD one can decompose a matrix $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are called left and right singular vectors respectively. The diagonal matrix $\boldsymbol{\Sigma}$ gives the singular values $(\sigma_1, \sigma_2, \sigma_3 ... \sigma_N)$ that usually sorted in non-increasing order. Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ should be orthogonal *ie.*, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$.

33. A rotation accompanied by a reflection is called an *improper rotation*. Otherwise they are called *proper rotations*. If an improper rotation is applied to a chiral molecule, then there will be a change in handedness of the result. In most cases in molecular biophysics this is not the desired result. (*See Note 4 for details regarding the theory of rotations*).

requirement of equivalent atom-pairs. In many real life cases[34] the molecules to be compared may have different numbers of atoms, or may be structurally different enough so that an equivalence relationship is not immediately apparent. A number of different algorithms have been developed that analyze the molecules to be compared and suggest equivalence relationships that can then be used for superposition. Most such algorithms exploit specific chemical information about the molecules to be compared, hence they are applicable only to specific classes of molecules like proteins or drug-like small molecules. In this section some of the common approaches for finding equivalence relationships will be discussed. It should be emphasized that there may not always exist a unique best equivalence relationship. Algorithms that exploit chemical information for determining suitable equivalences may consider different sets of chemical informations. Finally the methods of acquiring and processing of the information may also vary between algorithms. As a result of all this, there may be a number of algorithms that differ from each other only in small ways. Rather than provide an exhaustive list of algorithms, the following sections will try to group together similar methods into classes and only the general concepts common to each class will be discussed leaving out the operational differences between them.

8.3.1 Iterative Alignment and Superposition

Algorithms belonging to this class are usually used for protein structure comparisons. They usually start with an initial equivalence set E_o that is used to find an optimum transformation operator ($\boldsymbol{T}_o$) that minimizes an RMSD measure over the set of atom-pairs E_o. The entire structure is then superposed using $\boldsymbol{T}_o$, distances between all pairs of atoms (or residues) are then calculated and a scoring matrix created. This matrix is then used to align the two molecules usually by a dynamic programming method like the Smith-Waterman algorithm. From the aligned atom (or residue) pairs a new equivalence relation E_1 is created and the whole process repeated until the RMSD value between successive superpositions converges to a suitably low value. The initial equivalence set E_o may be obtained by sequence alignment methods or even by large scale superposition of small fragments (3-4 residues) taken from both the structures and looking for pairs of fragments whose structures are highly similar with the constraints that different fragment pairs must respect collinearity of sequence. Different programs within this class also differ with each other in the way the scoring matrix is created for the intermediate dynamic programming phase. One approach is to base the scoring on a probability function that depends upon the distances between equivalent atom-pairs obtained in the previous superposition step. Other approaches make use of local structural similarities or even chemical similarities to suitably convert

34. Say for example, when two proteins having similar (not identical) sequences or structures but having slightly different lengths.

the distances to scores. The critical aspect of algorithms in this class is the determination of the initial equivalence set E_o. It is known that choice of different equivalence sets can give rise to substantially different final results, and it is also not trivial to rank the final results in terms of quality.

8.3.2 Double Dynamic Programming

Rather than relying on an alternating series of alignment and superposition the double dynamic programming (DDP) approach[35]. Traditional local or global dynamic programming based methods[36] are not suitable for structure-structure alignments because because the scores of matched pairs of elements are not independent of each other. In the DDP method, this limitation is bypassed by an ingenious two level method in which first a low-level dynamic programming alignments are carried out, the highest scoring dynamic programming paths are propagated to a high-level dynamic programming matrix that is then used to find the overall best structural alignment.

The scoring matrices for the low-level dynamic programming steps are constructed as follows:

Suppose residue a_i and b_j are the i^{th} and j^{th} residues of structures A and B respectively that are to be compared. Local coordinate frames are derived from the local geometry around the two residues and the structures are rotated and translated so that the two local coordinate frames become coincident. The score ${}^{ij}S_{kl}$ between all pairs of residues a_k and b_l ($k \neq i$ and $l \neq j$) when residues a_i and b_j are aligned is a function of the corresponding distance ${}^{ij}d_{kl}$. To ensure that the low-level alignment goes through the pair (a_i, b_j) the score for (a_i, b_j) is given so high a value that the optimum path is forced to pass through it.

In practice, only a small subset of the residue-pairs (a_i, b_j) are used for the low-level dynamic programming step. The initial pair-list may be chosen randomly or derived from a comparison of the secondary structure elements of the two proteins.

The results from the low-level computations are summed up to the high-level scoring matrix. This is done letting the contribution from the low-level matrix ${}^{ij}S_{pq}$ such that (a_p, b_q) lies on the optimal low-level path when ${}^{ij}S$ is used as a scoring matrix. Since the elements high-level scoring matrix may contain the sum of many entries from low-level scoring matrices, their value may be quite large at times. Hence it is usual to normalize the scoring matrix values or use their logarithmic values at the start of the high-level dynamic programming alignment.

The optimal paths generated at the end of the high-level dynamic programming alignment now gives an equivalence set. This is then used to create

35. Tawlor W. and Orengo C. *J.Mol.Biol.* (1989) **208** 1-22

36. Like the Smith-Waterman and Needleman-Wunsch methods used in sequence analysis.

residue-pairs of another round of low-level dynamic programming. The process continues until the equivalence set does not change between iterations following which the set is used for final structure-structure superposition.

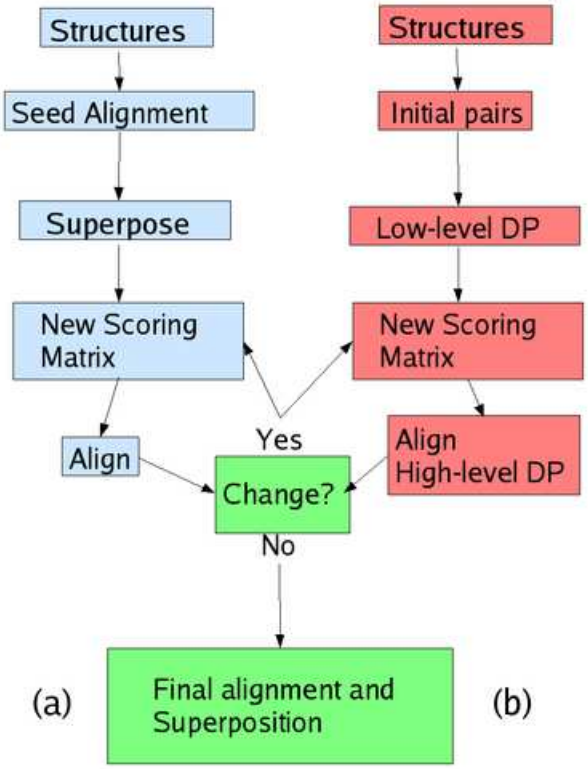


Fig. 21. Schematic comparison of the Alternating Alignment and Superposition (a) with the Double Dynamic Programming algorithm

8.3.3 Geometric Hashing

Originally developed for problems in computer vision, geometric hashing is a technique that has been extended to the task of matching 3D structures of biomolecules. The idea behind geometric hashing is to find all substructures (in the query) that are similar to substructures in a database of target structures (or models). In a preprocessing step, local reference frames are created for each non-collinear triplets of points in the model (these may taken from atoms/residues or secondary structure elements). Information about the reference frames are then stored in a specially designed highly redundant has table that facilates very quick lookups. In the search phase, the query structure is also broken down to substructures and reference frames are created for each

substructure. These frames are then looked up in the has table and similar reference frames together with their associated points are then merged taking into account of sequence collinearity. This set now constitutes the equivalence set for actual superposition.

9 Geometric Features of Biomolecular Structures

9.1 Proteins

The structure of proteins can best described at two-levels, (a) the structure of the protein backbone and (b) the structure of the side-chains[37]. The backbone structure of the protein essentially defines the overall shape of the protein. Essentially it consists of three bonds repeating in sequence. The sequential linearity of the the protein chain is broken by the relatively rare occurrence of disulfide bonds that can link different parts of the protein chain through cysteine residue side-chains. A key feature of protein structure is the peptide bond that links adjacent amino acids. This bond, although notionally a single bond has substantial double bond character due to a resonance effect. As a result free rotation about this bond is severely hindered and the torsion angle w defined by the peptide bond is generally restricted to the trans-conformation[38].

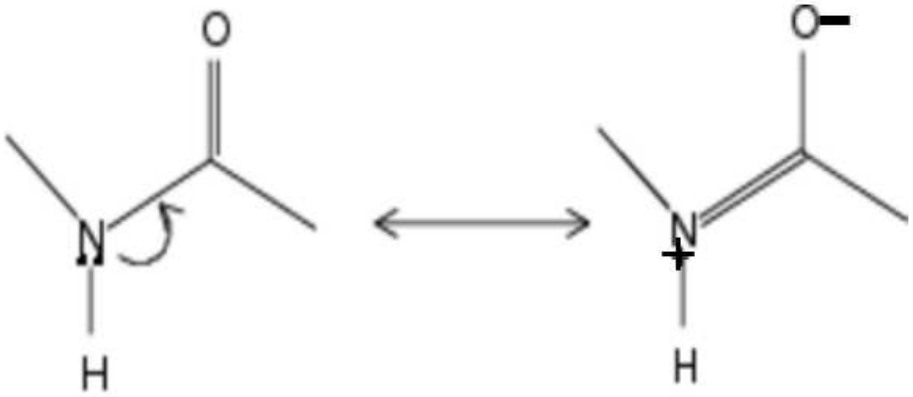


Fig. 22. The peptide bond can be considered to be a resonance hybrid of two forms, that results in a significant double bond character.

37. Refer to Figures 10, 11 and Table 2 for definition of the backbone and side-chain bonds and torsions.

38. Exceptions can occur for prolyl peptide bonds that can often be in the cis-conformations. Very rarely asparginyl peptide bonds are also observed to be in the cis- conformation.

On account of the restricted rotation aboyt the peptide bond, variations of only two torsion angles *viz.,* ϕ and ψ account for almost all of the conformational variation in the protein backbone. It was recognized by G.N.Ramachandran and co-workers[39] variations about the ϕ, ψ torsions are restricted considerably by the fact that certain combinations of these torsion angles brings different atoms impossibly close together. For example for the $\phi = 0$, $\psi = 0$ pair, would bring the carbonyl oxygen of the peptide at a distance of $0.35\mathring{A}$ from an amide hydrogen, which is obviously not possible. Based on a detailed analysis of high-resolution crystal structures available at that time, Ramachandran and co-workers prepared a list of distances that represented the lower limit of the distance that a pair of non-bonded atoms can approach. Further analysis lead them to define a pair of limiting distances for each non-bonded atom pair, the first limit which they called the "normal" limit is the shortest distance of approach between a pair of non-bonded atoms found in a majority of the structures. The second limit, which they called the "extreme" limit was an even shorter distance of approach that could be observed only in a few cases. No non-bonded atom pair could approach each other closer than the extreme limit defined by them. The values of the normal and extreme limit distances obtained by Ramachandran and co-workers are given in the following table[40].

Type of Contact	Normal Limit	Extreme Limit
H···H	2.0	1.9
H···O	2.4	2.2
H···N	2.4	2.2
H···C	2.4	2.2
O···O	2.7	2.6
O···N	2.7	2.6
O···C	2.8	2.7
N···N	2.7	2.6
N···C	2.9	2.8
C···C	3.0	2.9
C···C(H)	3.2	3.0
C(H)···C(H)	3.2	3.0

Table 4. Values of limiting distances ($\mathring{A}$) for various interatomic contacts. C(H) stands for united atom carbons *ie.,* CH_2 or CH_3 groups where the hydrogen atoms have not been definitely located.

39. Ramachandran G.N., Ramakrishnan C. and Sasisekharan V. *J.Mol.Biol.* (1963) **7** 95-99

40. Ramachandran G.N. and Sasisekharan V. *Adv. Prot. Chem* (1968) **23** 283-438

Ramachandran and co-workers then took a system of two adjacent peptide units around a C^{α} carbon, such that a pair of ϕ, ψ torsion angles could be defined. They then varied both the torsion angles and determined the conformational region where all the non-bonded interatomic distances are higher than the normal or extreme limits. The result was what is now famous as the Ramachandran plot (figure 24). A look at the plot in figure 24 shows

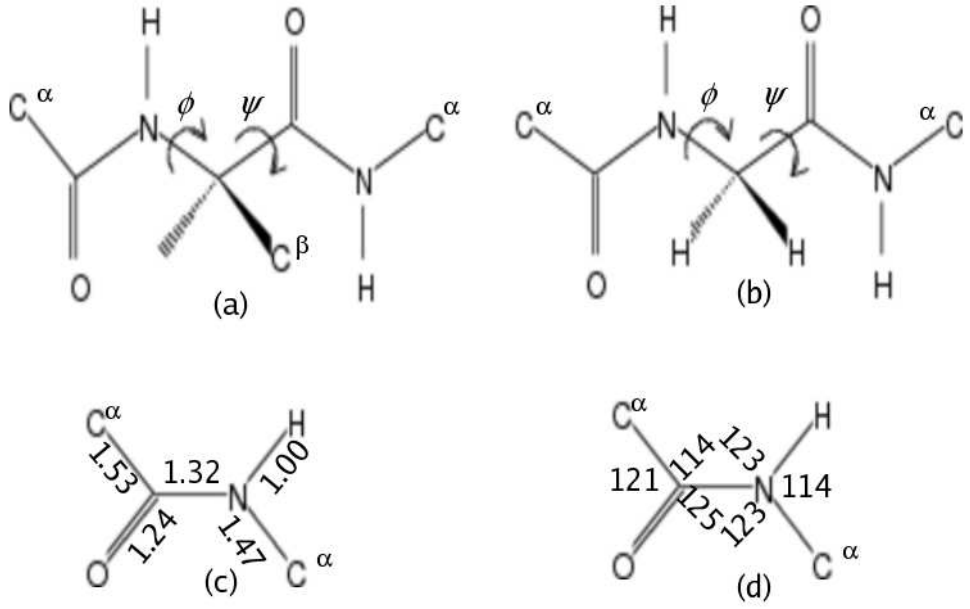


Fig. 23. Two linked peptide units used to calculate the Ramachandran plot. Note depending on the presence or absence of the C^{β} atom [panels (a) and (b) respectively] different results are obtained. The peptide bond geometries used for the calculation are also given [(c) bond lengths ($\mathring{A}$) and (d) bond angles (^{o}).

that almost three-quarters of the conformational space of a dipeptide unit becomes disallowed when the limits of interatomic distances are taken into account. The remaining quarter of the conformational space is now home to almost all possible secondary and tertiary structures of proteins with a few exceptions. The exceptions arise primarily due to glycyl and prolyl residues in proteins. Glycyl residues lack a C^{β} carbon (figure 23b) and hence have a greater conformational freedom. This is observed in the Ramachandran map of glycine that has a much greater extent of fully allowed conformational

space. Prolyl residues on account of their ring structure have their ϕ angle

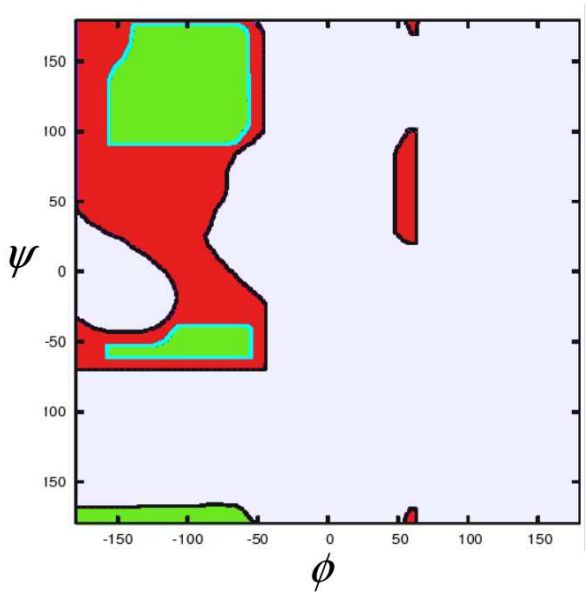


Fig. 24. Ramachandran plot for amino acids other than glycine and proline. The green regions are fully allowed with interatomic distances greater than the normal limits. The red regions are partially allowed which have at least one interatomic distance less than the normal limit but greater than the extreme limit. The pale blue regions are sterically disallowed and have at least one interatomic distance that is less than the extreme limit.

restricted to around 60^o. The ψ angle however, varies more or less in its normal range of -180^o to +180^o. The residue preceding proline is also affected to some extent by the prolyl residue and has a smaller allowed region in the map. Analysis of the ϕ, ψ angles from a large number of protein structures bears out the prediction that almost all the conformational space allowed to a protein lies within the bounds of the Ramachandran map (figure 25). Most of the exceptions that are observed can be traced to errors in experimental determination of the structures. In a few cases however, the exceptions are genuine and point towards specific structural peculiarities that often have important structural or functional consequences.

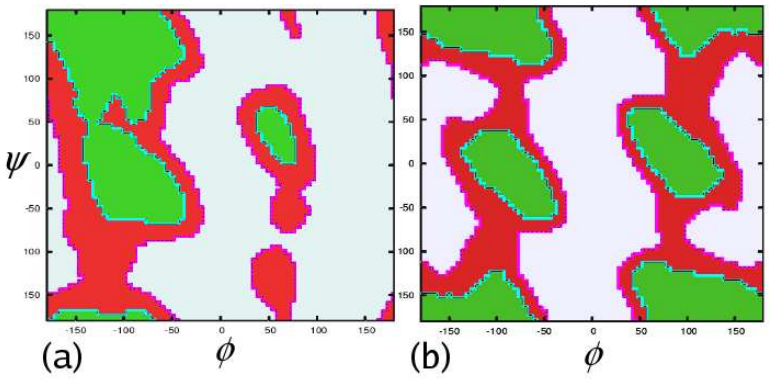


Fig. 25. ϕ, ψ distributions gerenated from a large number of experimentally determined protein structures. Green regions are those where approximately 98% of the ϕ, ψ apirs lie, while the red region include 99.95% of all ϕ, ψ pairs. Panel (a) shows the distributions for all residues except gly, pro and the residue preceding pro. Panel (b) shows the distribution for gly residues only. Data used for the figure were taken from Lovell *et. al.* [*Prot. Struct. Func. Genet.* (2003) **50** 437-450].

Genuine Ramachandran violations *i.e.*, occurrence of ϕ, ψ pairs in disallowed regions of the map are often a pointer to interesting structural peculiarities. In most cases the deviation is due an alteration in the peptide bond geometry. Cis-peptides, for example account for a significant proportion of Ramachandran plot violations. Other reasons can be pyramidalization of the nitrogen atom or distortion of some of the bond lengths and bond angles of the peptide bond. It is important to try and locate the source of the strain that causes the geometric distortions in peptide bond structure and consequent violations of the Ramachandran sterechemical criteria because very often such strained residues have important roles to play in the stability or functional activity of the protein.

9.2 Nucleic Acids

The basic repeating unit in a nucleic acid polymer is the nucleotide. There are six rotatable bonds in the nucleotide backbone (figure 12), which allows for a much greater degree of conformational freedom than protein chains. Indeed, for single-stranded nucleic acids, the conformational freedom can be considered total. Nucleic acids strands, both DNA and RNA however have the tendency to to specifically pair-up with each other forming double helices. Specific hydrogen bonding between Guanine (G) and Cytosine (C) bases and also Adenine (A) and Thymine (T) bases (in DNA) and Adenine and Uracil (U) bases (in RNA) leads to the formation of double helices. Depending on

environmental conditions nucleic acid double helices acquire different polymorphic forms. The structures of the different polymorphic forms were first deciphered at low resolution using the technique of x-ray fiber diffraction. The shape and dimensions of AT and GC base-pairs are quite similar hence both of them could be easily accomodated in the various polymorphic forms. Furthermore the inherently low resolution views obtainable from the early fiber diffraction studies did not allow the observation of sequence dependent structural variation in nucleic acid double helices. Hence early models made for nucleic acids are idealized forms that showed only the general features of the double helices.

It was observed that DNA fibers under low humidity conditions as well as most RNA double helices adopted a structure now known as the A-form. In this form there are 11 base pairs/turn of the double helix and the helical rise between successive base pairs is about 2.6 $\mathring{A}$. The base pairs are somewhat destacked and moved away from the helix axis, a stricking feature was that the base pairs are not perpendicular to helix axis, rather the base pair normals are tilted about 20^o away from the helix axis. The isosteric base pairs as defined by Watson and Crick give rise to two distinct edges that lead to two distinct grooves or clefts in the double helix known as the major and the minor groove respectively. Much of protein-DNA recognition events through the major groove by a hydrogen bond based sequence readout process. In the A-form double helix, however, the major groove is very narrow and deep that does allow anything to pnetrate deep into the groove and make sequence specific hydrogen bonds with the bases. The ribofuranose sugar in the A-form structure normally exists in the C3'-endo conformation.

Under conditions of high humidity DNA, but not double helical RNA is known to exist in the B-form structure. In the B-form there are 10 base pairs/turn of the double helix and the rise along the helix axis is 3.4 $\mathring{A}$/base pair. The base pairs are very well stacked on top of each other and are perpendicular to the helix axis. The major groove is somewhat wider than the minor groove but both of them are sufficiently deep that allows hydrogen bond formation between DNA interacting proteins molecules. The sugar ring in the B-form structures exist in the C2'-endo conformation.

While the A and B forms of nucleic acids are generally sequence neutral and only environment dependent,[41] the double helices made out of the alternating co-polymer of C and G residues can in presence of about 4-5M NaCl or 60% ethanol produce a third polymeric form known as the Z-form. Z-form DNA is unique because it is a left-handed double helix.In Z-DNA alternating C and G nucleotides are in distinctly different environments, hence the repeat unit in this double helix is considered to be a dinucleotide rather than a mononucleotide as in the other two forms. Thus the glycosidic torsion for the

41. although homopurine-homopyrimidine sequences of G or C prefer the A-conformation even under physiological conditions of salt and humidity.

guanosine residues are in the *syn* region ($\sim 70^o$) while it is *anti* ($\sim -160^o$) for the cytidine residues. The sugar puckering for the guanosine residues varies in the range from C2'-endo to C2'-exo while the sugar puckering for the cytidine residues is locked in the C2'-endo region. The phosphates are located on two different sequentially alternating radii and neighbouring sugar units point in opposite directions. A curve joining adjacent phosphorus atoms is not smooth as in A or B-DNA but traces out a zig-zag path, hence the name Z-DNA.

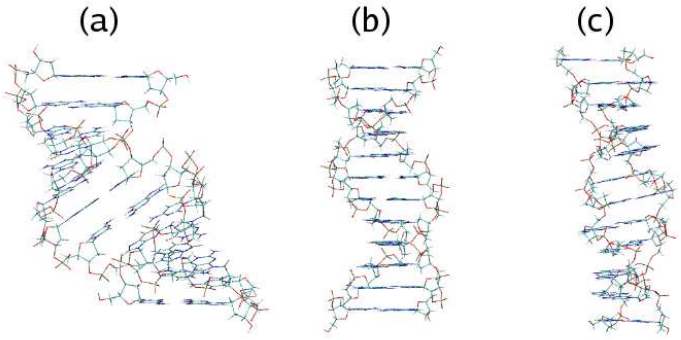


Fig. 26. Canonical polymorphs of double helical DNA. The structures are shown along the major groove. (a) A-DNA (b) B-DNA (c) Z-DNA

The structural polymorphs of nucleic acid double helices just described are idealized forms with hardly any sequence dependent structural variations. It turns out however that there are significant sequence dependent variations in nucleic acid structure that may include sequence dependent curvature and other subtle but functionally important changes. The availability of high resolution crystal structures of oligonucleotides and their complexes has opened a window to study sequence dependent structural variability in nucleic acids. It increasingly becoming apparent that in case of nucleic acids, it is not just the the most stable structure alone, but the entire conformational ensemble accessible to the nucleic acid molecule under functionally active situations that is important. In an effort to clarify subtle sequence dependent variability in nucleic acid structures a common definition for structural parameters at the base pair and dinucleotide step level has been defined. The definition of these parameters are pictorially shown in figure 27. Statistical analysis of the base pair and dinucleotide step parameters from a large number of oligonucleotide crystal structures, as well as from molecular dynamics simulations point towards significant sequence dependent conformational preferences. How these variations of conformational preferences by diufferent bases and dinucleotides are exploited by sequence dependent nucleic acid binding molecules is now a subject of intense research.

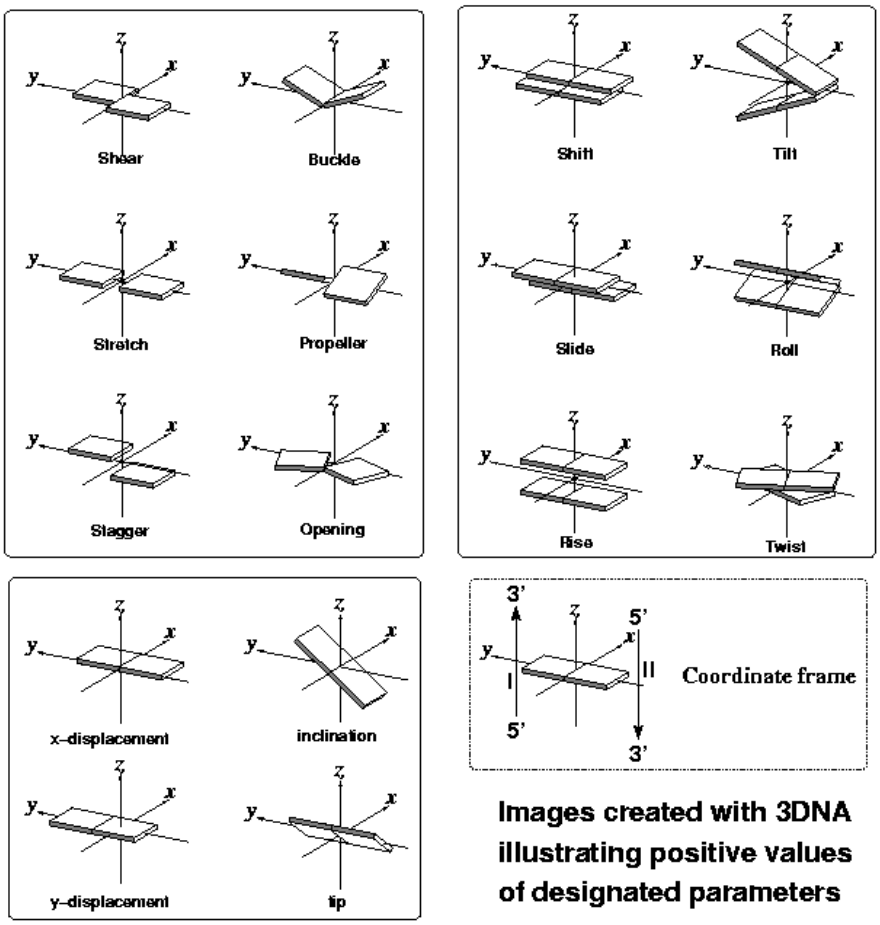


Fig. 27. Definitions of base base pair and dinucleotide step parameters in nucleic acids. [Figure downloaded from http://ndbserver.rutgers.edu/archives/report/tsukuba_sup/bp_step_hel.html]

10 Epilogoue

Geometrical methods have played a fundamental role in the understanding of biomolecular structures. The brief description given here is neither complete in terms of coverage nor sufficient in depth. Nevertheless, it is hoped that the article could provide a glimpse of the tremendous importance of geometry in the fascinating world of biomolecular conformation analysis.

Note 1. **Essential Mathematics**[42]

Vectors and Vector Algebra

A *vector* is a quantity that has both magnitude and direction. The velocity of an object or the force acting on it are vectorial quantities because they have both magnitude and direction. One can contrast these with quantities like mass and temperature which have a magnitude only. Such quantities are called *scalars*. One can represent a vector with an arrow as shown in figure 1.9A. The length of the arrow represents the magnitude of the vector while the arrowhead points to the direction. Another way to represent vectors is by a set of numbers representing its *components* along a set of well defined directions i.e., the coordinate axes. It is customary to refer to a vector using a boldface letter such as $\boldsymbol{A}$ (or a character with some sort of embellishment such as $\mathbb{A}$, $\tilde{A}$, $\hat{A}$ etc.) and its components (which are scalar quantities) by ordinary characters e.g., (A_x, A_y).

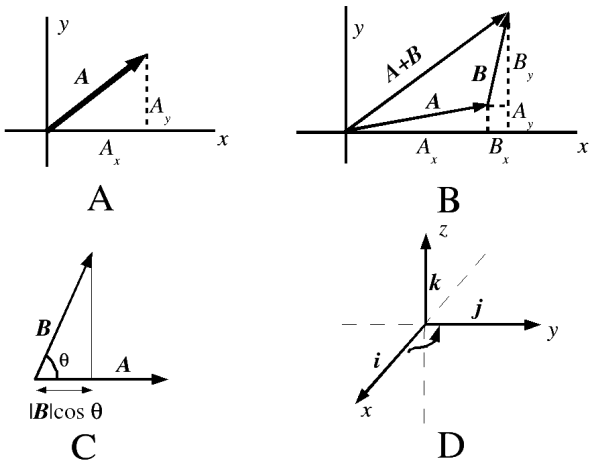


Fig. 28. Vectors and their properties. (A) A single vector shown along with its components. (B) Summation of a pair of vectors (C) A projection of a vector onto another (D) A right handed Cartesian coordinate system shown along with the unit vectors defining the axial directions

42. This note is intended merely to provide a few pointers to the necessary mathematical knowledge for studying molecular conformation. The discussion given here will neither be comprehensive nor sufficiently rigourous. The interested student is referred to standard mathematical texts for further reading. A good and very readable book is the following:

Boas M.L. *Mathematical Methods in the Physical Sciences* (Second Edition)Wiley (2003).

The length of vector $\boldsymbol{A}$ also called the *norm* of $\boldsymbol{A}$ (written $\|\boldsymbol{A}\|$) or the magnitude of $\boldsymbol{A}$ (wriiten $|\boldsymbol{A}|$). In terms of its components one can write the magnitude of a vector as follows[43]:

$$\left|\boldsymbol{A}\right| = \sqrt{A_x^2 + A_y^2} \qquad \text{(in two dimensions)} \tag{69}$$

$$\left|\boldsymbol{A}\right| = \sqrt{A_x^2 + A_y^2 + A_z^2} \qquad \text{(in three dimensions)} \tag{70}$$

Vector quantities may be added or subtracted from one another. Figure 1.8B demonstrates the addition of two vectors. To add vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ we need to place the tail of vector $\boldsymbol{B}$ at the head of vector $\boldsymbol{A}$. The tail of vector $\boldsymbol{C}$(the sum) coincides with the tail of vector $\boldsymbol{A}$and its head coincides with the head of vector $\boldsymbol{B}$. The components of the sum of a pair of vectors can be obtained by summing the components of the vectors themselves. Thus the components of vector $\boldsymbol{A} + \boldsymbol{B}$ is given by $(A_x + B_x,\ A_y + B_y,\ A_z + B_z)$. from an examination of figure 1.9B the following laws on vector addition become apparent.

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A} \quad \text{(Commutative law of addition)} \tag{71}$$

$$(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) \quad \text{(Associative law of addition)} \tag{72}$$

Vectors may be multiplied by a scalar. Thus $\boldsymbol{A} + \boldsymbol{A}$ can be written as $2\boldsymbol{A}$. Multiplying a vector $\boldsymbol{A}$ by a scalar n gives a vector that is n times as long as $|\boldsymbol{A}|$ but having the same direction. The the negative of a vector $\boldsymbol{A}$ (written $-\boldsymbol{A}$) has the same length as $|\boldsymbol{A}|$ but points to the opposite direction. The negative of a vector has the sign of each of its components reversed with respect to the corresponding component of $\boldsymbol{A}$. The subtraction of a vector from another can be defined as addition of a vector with the negative of another vector. Like in case of vector addition, the components of the vector $\boldsymbol{A} - \boldsymbol{B}$ can be obtained from the components of the vevtors $\boldsymbol{A}$ and $-\boldsymbol{B}$ i.e., $(A_x - B_x,\ A_y - B_y,\ A_z - B_z)$. If two vectors are equal in magnitude but opposite in sign than their summation will lead to a zero vector i.e., $\boldsymbol{A} + (-\boldsymbol{A}) = \boldsymbol{A} - \boldsymbol{A} = \boldsymbol{0}$. Zero vectors have all its components equal to 0 and it does not have a direction. A vector with length 1 is called a *unit vector*. For any $\boldsymbol{A} \neq \boldsymbol{0}$, the vector $\boldsymbol{A}/|\boldsymbol{A}|$ is a *unit vector*[44]. Unit vectors are very useful when a particular direction has to be specified. For example a Cartesian coordinate system can be denoted by the unit vectors $\boldsymbol{i}, \boldsymbol{j}$ and $\boldsymbol{k}$ representing the x, y and z directions respectively. A vector $\boldsymbol{A}$ with components $(A_x, A_y,\ A_z)$ can be written as:

$$\boldsymbol{A} = \boldsymbol{i}A_x + \boldsymbol{j}A_y + \boldsymbol{k}A_z \tag{73}$$

---

43. It is very easy to prove the expressions by means of the Pythagorean theorem.

44. The components of a unit vector are given by $A_x/|\boldsymbol{A}|, A_y/|\boldsymbol{A}|, A_z/|\boldsymbol{A}|$. These are also known as *direction cosines* because they represent the cosine of the angle between the vector $\boldsymbol{A}$ and the x, y and z axes respectively.If l, m, n are the direction cosines of a vector, then $l^2 + m^2 + n^2 = 1$.

Multiplication of a vector with another vector can be defined in two different ways. The first type called the *Dot product* (also called the *scalar product* or the *inner product*) and the second type is called the *Cross product* (also called the *vector prduct* or the *outer product*). The dot product of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ (written $\boldsymbol{A}\cdot\boldsymbol{B}$ or sometimes $\langle\boldsymbol{AB}\rangle$) produces a result that is a scalar quantity.The cross product of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ (written $\boldsymbol{A}\times\boldsymbol{B}$) produces a vector result. The different physical meanings of these two types of vector multiplication are explained below.

Dot Product The dot product of a pair of vectors is defined to be the product of the magnitudes of the vectors times the cosine of the angle between them. Thus:

$$\boldsymbol{A}\cdot\boldsymbol{B}=|\boldsymbol{A}||\boldsymbol{B}|\cos\theta \tag{74}$$

Where θ denotes the angle[45] between the two vectors.

Since $|\boldsymbol{B}|\cos\theta$ gives the length of the projection of vector $\boldsymbol{B}$ on $\boldsymbol{A}$ (figure 1.9C) one can interpret the dot product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as the magnitude of $\boldsymbol{A}$ times the projection of $\boldsymbol{B}$ on $\boldsymbol{A}$. The following properties of the dot product can be easily derived:

$$\boldsymbol{A}\cdot\boldsymbol{B}=\boldsymbol{B}\cdot\boldsymbol{A}\quad(\text{Commutative Law}) \tag{75}$$

$$\boldsymbol{A}\cdot(\boldsymbol{B}+\boldsymbol{C})=\boldsymbol{A}\cdot\boldsymbol{B}+\boldsymbol{A}\cdot\boldsymbol{C}\quad(\text{Distributive Law}) \tag{76}$$

From the above it is apparent that:

$$\boldsymbol{A}\cdot\boldsymbol{A}=|\boldsymbol{A}|^2\cos 0=|\boldsymbol{A}|^2 \tag{77}$$

Similarly if two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ are perpendicular (or *orthogonal*) to each other we have:

$$\boldsymbol{A}\cdot\boldsymbol{B}=|\boldsymbol{A}||\boldsymbol{B}|\cos(\pi/2)=0 \tag{78}$$

One can easily express the dot product in terms of the components of the vectors. Thus:

$$\boldsymbol{A}\cdot\boldsymbol{B}=(\boldsymbol{i}A_x+\boldsymbol{j}A_y+\boldsymbol{k}A_z)\cdot(\boldsymbol{i}B_x+\boldsymbol{j}B_y+\boldsymbol{k}B_z)$$

Using the distributive law we can write:

$$\begin{aligned}\boldsymbol{A}\cdot\boldsymbol{B}=&\boldsymbol{i}\cdot\boldsymbol{i}A_xB_x+\boldsymbol{i}\cdot\boldsymbol{j}A_xB_y+\boldsymbol{i}\cdot\boldsymbol{k}A_xB_z+\\&\boldsymbol{j}\cdot\boldsymbol{i}A_yB_x+\boldsymbol{j}\cdot\boldsymbol{j}A_yB_y+\boldsymbol{j}\cdot\boldsymbol{k}A_yB_z+\\&\boldsymbol{k}\cdot\boldsymbol{i}A_zB_x+\boldsymbol{k}\cdot\boldsymbol{j}A_zB_y+\boldsymbol{k}\cdot\boldsymbol{k}A_zB_z\end{aligned}$$

Because $\boldsymbol{i},\boldsymbol{j}$ and $\boldsymbol{k}$ have unit lengths and orthogonal to each other we have:

$\boldsymbol{i}\cdot\boldsymbol{i}=\boldsymbol{j}\cdot\boldsymbol{j}=\boldsymbol{k}\cdot\boldsymbol{k}=1$ and $\boldsymbol{i}\cdot\boldsymbol{j}=\boldsymbol{i}\cdot\boldsymbol{k}=\boldsymbol{j}\cdot\boldsymbol{k}=0$ Thus:

$$\boldsymbol{A}\cdot\boldsymbol{B}=A_xB_x+A_yB_y+A_zB_z \tag{79}$$

45. For all calculations the angle must be represented in radians.

Cross Product The cross product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ give a new vector $\boldsymbol{C} = \boldsymbol{A} \times \boldsymbol{B}$ that is orthogonal to both $\boldsymbol{A}$ and $\boldsymbol{B}$.The direction of the new vector $\boldsymbol{C}$ is given by the right hand rule.[46] The magnitude of the cross product is given by the formula:

$$\boldsymbol{A} \times \boldsymbol{B} = |\boldsymbol{A}||\boldsymbol{B}| \sin\theta \tag{80}$$

where θ is the positive angle ($\leq 180°$) between $\boldsymbol{A}$ and $\boldsymbol{B}$. An important characteristic of the cross product is that it is not commutative i.e., $\boldsymbol{A} \times \boldsymbol{B} \neq \boldsymbol{B} \times \boldsymbol{A}$. In fact, $\boldsymbol{A} \times \boldsymbol{B} = -\boldsymbol{B} \times \boldsymbol{A}$. The cross product of a pair of parallel (or antiparallel) vectors has 0 magnitude ($\boldsymbol{A} \times \boldsymbol{B} = |\boldsymbol{A}||\boldsymbol{B}| \sin 0 = 0$) and the magnitude of the cross product of a pair of orthogonal vectors is the product of the magnitude of the vectors themselves i.e., ($\boldsymbol{A} \times \boldsymbol{B} = |\boldsymbol{A}||\boldsymbol{B}| \sin(\pi/2) = |\boldsymbol{A}||\boldsymbol{B}|$). Applying these results to the unit vectors along the axial directions of a cartesian coordinate system we obtain:

$$\begin{aligned} \boldsymbol{i} \times \boldsymbol{i} = \boldsymbol{j} \times \boldsymbol{j} = \boldsymbol{k} \times \boldsymbol{k} = \boldsymbol{0} \\ |\boldsymbol{i} \times \boldsymbol{j}| = |\boldsymbol{i} \times \boldsymbol{k}| = |\boldsymbol{j} \times \boldsymbol{k}| = 1 \end{aligned} \tag{81}$$

Using the right hand rule and referring to figure 1.9C we have the following additional expressions:

$$\begin{array}{lll} \boldsymbol{i} \times \boldsymbol{j} = \boldsymbol{k} & \boldsymbol{j} \times \boldsymbol{k} = \boldsymbol{i} & \boldsymbol{k} \times \boldsymbol{i} = \boldsymbol{j} \\ \boldsymbol{j} \times \boldsymbol{i} = -\boldsymbol{k} & \boldsymbol{k} \times \boldsymbol{j} = -\boldsymbol{i} & \boldsymbol{i} \times \boldsymbol{k} = -\boldsymbol{j} \end{array} \tag{82}$$

The above expressions depend on the way the coordinate axes are labelled in figure 1.9D. Such coordinate systems are called right handed where rotation of the x-axis onto the y-axis corresponds to the rotation of a right handed screw advancing in the z-direction. One can also have left handed coordinate systems,[47] which are obtained by interchanging any two of the coordinate axes in figure 1.9C. The cross product expressions in (1.17))will have their sign reversed in left handed corrdinate systems. Unless explicitly mentioned, a right handed coordinate system is always assumed. The student should therefore be careful while drawing diagrams of coordinate systems.

---

46. To understand the right hand rule imagine that you are grasping the vector C with your right hand. The fingers then curl in the direction of rotation $\boldsymbol{A}$ onto $\boldsymbol{B}$ and the thumb points towards the direction of $\boldsymbol{C}$. An alternative way to think about this is to imagine a screw driver driving a right handed screw. The rotation of the screw driver gives the direction of rotation of $\boldsymbol{A}$ onto $\boldsymbol{B}$ and the direction of its forward movement gives the direction of the cross product $\boldsymbol{C}$.

47. Left handed coordinate systems are quite common in the Computer graphics literature.

It is possible to calculate the cross product of a pair of vectors from their components.To do this we state without proof the *distributive law* of cross products[48] viz.,

$$\boldsymbol{A}\times(\boldsymbol{B}+\boldsymbol{C})=\boldsymbol{A}\times\boldsymbol{B}+\boldsymbol{A}\times\boldsymbol{C}$$

$$\text{Thus } \boldsymbol{A}\times\boldsymbol{B}=(\boldsymbol{i}A_x+\boldsymbol{j}A_y+\boldsymbol{k}A_z)\times(\boldsymbol{i}B_x+\boldsymbol{j}B_y+\boldsymbol{k}B_z)$$

$$=\boldsymbol{i}(A_yB_z-A_zB_y)+\boldsymbol{j}(A_zB_x-A_xB_z)+\boldsymbol{k}(A_xB_y-A_yB_x)$$

$$=\begin{vmatrix}\boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k}\\ A_x & A_y & A_z\\ B_x & B_y & B_z\end{vmatrix} \tag{83}$$

Points, Lines and Planes

A point can be specified by a set of three coordinates (x,y,z). One can also think of a point as the head of a vector $\boldsymbol{r}=\boldsymbol{i}x+\boldsymbol{j}y+\boldsymbol{k}z$ with its tail at the origin (0,0,0). Vector $\boldsymbol{r}$ is called a *position vector* or *bound vector*. If there are two points with position vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ respectively. Then the distance between them is equal to the magnitude of the difference vector $\boldsymbol{s}-\boldsymbol{r}$.

Thus we have (using equations 1.10 and 1.6):

$$\left|\boldsymbol{s}-\boldsymbol{r}\right|=\left|\boldsymbol{i}(s_x-r_x)+\boldsymbol{j}(s_y-r_y)+\boldsymbol{k}(s_z-r_z)\right|=$$

$$\sqrt{(s_x-r_x)^2+(s_y-r_y)^2+(s_z-r_z)^2} \tag{84}$$

Consider the line a fixed point $A(x_o,y_o,z_o)$ to any point $B(x,y,z)$. It can be written as the difference of the position vectors of points B and A respectively. Thus we can describe the line as:

$$\boldsymbol{B}-\boldsymbol{A}=\boldsymbol{i}(x-x_o)+\boldsymbol{j}(y-y_o)+\boldsymbol{k}(z-z_o)$$

Consider another vector $C=\boldsymbol{i}a+\boldsymbol{j}b+\boldsymbol{k}c$ that is parallel to $\boldsymbol{B}-\boldsymbol{A}$. Then we can write (for $a,b,c\neq 0$)

$$\frac{x-x_o}{a}=\frac{y-y_o}{b}=\frac{z-z_o}{c} \tag{85}$$

These are known as the *Symmetric equations* of a straight line passing though the point (x_o,y_o,z_o). Alternatively we can write:

$$\boldsymbol{B}-\boldsymbol{A}=\boldsymbol{C}t \tag{86}$$

Where t is some scalar multiple. Thus we have:

$$\boldsymbol{B}=\boldsymbol{A}+\boldsymbol{C}t \tag{87}$$

or in terms of the components we have:

48. The proof is not difficult but tedious. The student may check the following textbook:

Thomas G.B. and Finney R.L. *Calculus and Analytic Geometry* 9th Edition. Pearson Education Asia (2002) pp 817-818

$$\begin{aligned} x &= x_o + at \\ y &= y_o + bt \\ z &= z_o + ct \end{aligned} \tag{88}$$

These are the equations of a straight line in *parametric form*. The parameter t can have various possible interpretations. For example, if a particle is moving along a straight line with uniform velocity then t can be interpreted as the time.

The equation of a plane may be derived using a similar argument. If (x_o, y_o, z_o) is some point in a plane and (x, y, z) is another point in the same plane, then the vector connecting the two points is given by:

$$\boldsymbol{r} - \boldsymbol{r}_o = (x - x_o)\boldsymbol{i} + (y - y_o)\boldsymbol{j} + (z - z_o)\boldsymbol{k}$$

If $\boldsymbol{N} = a\boldsymbol{i} + b\boldsymbol{j} + c\boldsymbol{k}$ be a vector perpendicular to $\boldsymbol{r} - \boldsymbol{r}_o$ then we have:

$$\begin{aligned} \boldsymbol{N} \cdot (\boldsymbol{r} - \boldsymbol{r}_o) &= 0 \\ \Rightarrow a(x - x_o) + b(y - y_o) + c(z - z_o) &= 0 \qquad (89) \\ \Rightarrow ax + by + cz &= d \qquad (90) \\ \text{where } d &= ax_o + by_o + cz_o \end{aligned}$$

which gives the equation of a plane.

Linear and Orthogonal Transformations

Whenever we transform a set of variables to another in such a way that each new variable is a linear combination of the old variables we have carried out a *linear tranformation*. For example we can use linear transformation to generate a set of new variables x', y', z' from a set of old variables x, y, z using the following way:

$$\begin{aligned} x' &= ax + by + cz \\ y' &= dx + ey + fz \\ z' &= gx + hy + iz \\ \text{where } a, b, c, d, e, f, g, h &\text{ and } i \text{ are all constants.} \end{aligned} \tag{91}$$

Linear transformations can be interpreted geometrically in the following way. Let $\boldsymbol{r}$ be some vector. We can write:

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$$

Consider the linear transformation in equations 1.28. Since each x, y, z is related to each x', y', z'. We can consider the vector $\boldsymbol{r}$ to be expressed in some new coordinate system x', y', z'. Thus we have *two* sets of coordinate systems (x, y, z) and (x', y', z') and *one* vector $\boldsymbol{r} = \boldsymbol{r}'$ with components relative to each set of axes given by:

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} = \boldsymbol{r}' = x'\boldsymbol{i}' + y'\boldsymbol{j}' + z'\boldsymbol{k}' \tag{92}$$

where $\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}'$ are unit vectors along the x', y', z' axes respectively.

In general $\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}'$ need not be orthogonal to each other. When they are the transformation corresponds to a rotation[49] and the length of the vector is left unchanged. Linear transformations of this type are called *orthogonal transformation*. By definition an orthogonal transformation is a linear transformation that transforms varables x, y, z to x', y', z' such that:

$$x^2 + y^2 + z^2 = x'^2 + y'^2 + z'^2 \tag{93}$$

Matrices

Matrices[50] (*sing.* Matrix) provide a way compact way to describe linear transformations. For example, equation 1.28 may be written in matrix notation as follows:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{94}$$

One can easily see from the above equation that vectors can be represented as a matrix with a single column (or row). The transformation itself is written in the form of a 3 x 3[51] matrix. Like vectors matrices are also denoted with bold (or with other forms of embellishment) letters. Thus a matrix might be denoted as $\boldsymbol{A}$ and its elements (like the components of a vector) are denoted $A_{r,c}$ where the subscripts r and c denote the respective row and column indices of the element. Several mathematical operations can be carried out on matrices. Some of these are described below:

---

49. The rotation is of the vector $\boldsymbol{r}$ by some angle θ or its equivalent, the coordinate system $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ by angle $-\theta$. We will discuss rotations in detail in note 1.4.

50. The term matrix seems to have been first used by Sylvester in 1851 who stated, "From the rectangular matrix consisting of n rows and $(n+1)$ columns ... Then all the $(n+1)$ determinants can be formed by rejecting any one column at pleasure out of this matrix are identically zero". However it was Cayley who first used this term in its modern form in a series of papers between 1855 to 1858. Incidentally C.L.Dodgson (also known as Lewis Carroll, the author of *Alice in Wonderland*), objected to the word 'Matrix' and instead preferred the word 'block'. However his objections remained unheeded. (*source:* http://mathworld.wolfram.com/Matrix.html)

51. Matrices with m rows and n columns are often described as $m \times n$ matrices. Sometimes an $m \times n$ matrix $\boldsymbol{A}$ is written in the form $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Here $\mathbb{R}$ is the set of real numbers and $\mathbb{R}^{m \times n}$ is the *vector space* of all $m \times n$ real matrices. The symbol $\in$ stands for *is an element of*. In simple words, the statement means that $\boldsymbol{A}$ is a matrix of real numbers having m rows and n columns.

Transpose of a matrix To obtain the transpose of a matrix $\boldsymbol{A}$ (usually written as $\boldsymbol{A}^T$) one simply needs to interchange the rows and columns of $\boldsymbol{A}$. For example:

$$\text{If } \boldsymbol{A} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix} \text{ then } \boldsymbol{A}^T = \begin{pmatrix} a & e & i \\ b & f & j \\ c & g & k \\ d & h & l \end{pmatrix} \tag{95}$$

Thus if $\boldsymbol{A}$ is an $m \times n$ matrix,$\boldsymbol{A}^T$ is an $n \times m$ matrix.

Multiplication of a matrix by a number To multiply a matrix with a number one needs to multiply every element of the matrix with the number. The result is of course a matrix. Thus for example:

$$n\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} = \begin{pmatrix} na & nb & nc \\ nd & ne & nf \end{pmatrix} \tag{96}$$

Addition (or subtraction) of matrices In order to add (or subtract) a pair of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Every element of $\boldsymbol{A}$ must be added to (or subtracted from) the corresponding element of $\boldsymbol{B}$. Obviously a pair of matrices can be added or subtracted only if they have the same number of rows and columns. For example:

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} + \begin{pmatrix} g & h & i \\ j & k & l \end{pmatrix} = \begin{pmatrix} a+g & b+h & c+i \\ d+j & e+k & f+l \end{pmatrix} \tag{97}$$

Matrix Multiplication To multiply a pair of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, consider each row of $\boldsymbol{A}$ and each column of $\boldsymbol{B}$ as a vector. The product $\boldsymbol{AB}$ is a matrix whose elements are the dot products of every row vector of $\boldsymbol{A}$ with every column vector of $\boldsymbol{B}$. One can describe matrix multiplication with the following formula:

$$c_{i,k} = \sum_j a_{i,j} b_{j,k} \tag{98}$$

Where $a_{i,j}$ is the element at i^{th} row,j^{th} column of matrix $\boldsymbol{A}$, $b_{j,k}$ is the element at j^{th} row and k^{th} column of matrix $\boldsymbol{B}$ and $c_{i,k}$ is the element at i^{th} row and k^{th} column of the product matrix. The summation is carried out over the index j. An algorithm for matrix multiplication is given below in pseudocode.

```
Let A ∈ ℝ^{m×p}, B ∈ ℝ^{p×n} be a pair of matrices. The product
AB = C ∈ ℝ^{m×n} can be obtained as follows:
  for i = 1 to m
     for k = 1 to n
       for j = 1 to p
          C(i,k)= A(i,j)*B(j,k) + C(i,k)
       end
     end
```

```
end
```

It is obvious that in order to multiply a pair of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ it necessary that the number of columns of $\boldsymbol{A}$ be equal to the number of rows of $\boldsymbol{B}$. Another very important characteristic of matrix multiplication is that it is not commutative i.e., $\boldsymbol{AB} \neq \boldsymbol{BA}$.

Inverse of a matrix An *identity matrix* or *unit matrix* is matrix whose diagonal elements are all 1 and the off-diagonal elements are all 0. Identity matrices are often called $\boldsymbol{I}, \boldsymbol{E}$ or $\boldsymbol{U}$. The following is a 3×3 identity matrix.

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The inverse of a matrix $\boldsymbol{A}$ (written $\boldsymbol{A}^{-1}$) is a matrix which satisfies the following equation.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \tag{99}$$

Only *square* matrices[52] can have inverses[53]. Actually some square matrices also do not have inverses. The determinant of such matrices have a value of 0. Matrices that do not have inverses are called *singular*.

Note 2. **Generating Atomic Neighbour Lists**

While the calculation of individual bond lengths, bond angles and torsion angles can be easily carried out using the formulas given the text, complications arise when all the bond lengths, bond angles and torsion angles in a molecule has to be calculated. The main problem is the enumeration all pairs, triples or quartets of atoms which define the bond lengths, bond angles and torsion angles repectively. A branch of mathematics known as *graph theory* has long been seized with these kind of problems (and much else besides). We therefore solve our particular problem by borrowing some of the ideas and concepts from graph theory.

A graph is defined as a collection of points (also called *nodes* or *vertices* (sing. *vertex*)) connected by lines (also called *links* or *edges*). A molecule can be thought of as a graph where the atoms are the vertices and the bonds are the edges.[54] In the simplest form of graph representation of a molecule double or triple bonds are not taken into account and only the mere presence of a bond is shown. A graph is completely specified by the connectivity information among its constituent vertices. There are many ways to specify this connectivity information some of these are the following.

---

52. Matrices of the type $m \times m$ i.e., those that have the same number of rows and columns are called *square matrices*.

53. Otherwise we could not multiply $\boldsymbol{A}\boldsymbol{A}^{-1}$

54. This way of looking at molecules does not take any account of molecular conformation. A graph will remain the same regardless how one arranges the vertexes in space, as long as the edges connecting them are intact.

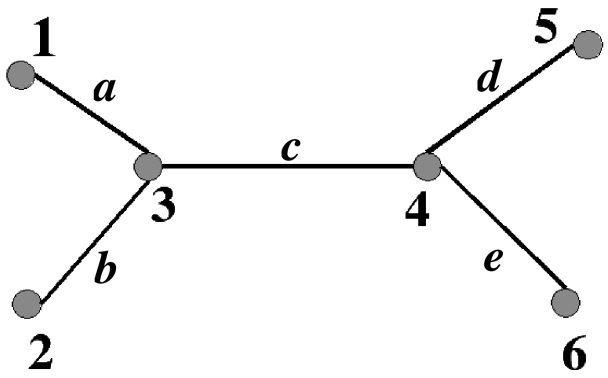


Fig. 29. ictorial view of a molecular graph. Each node represents an atom and each edge represents a bond. The nodes have been numbered with arabic numerals and the edges labelled with arabic letters.

Adjacency Matrix. The adjacency matrix is a $n \times n$ matrix where n is the number of vertices. If nodes i and j have an edge between them (i.e., atoms i and j are linked by a bond), then the element of the adjacency matrix $a_{ij} = 1$ else $a_{ij} = 0$. Adjacency matrices of molecular graphs are always symmetric and its elements have only 0 or 1 values. Since there is no edge between an atom and itself,[55] the diagonal elements of the adjacency matrix are always zero. The sum of the elements in each node (or column) of an adjacency matrix gives the *degree* of connectivity of the corresponding node in the graph.[56] The adjacency matrix of the graph shown in figure 1.10 is shown below:

0	0	1	0	0	0
0	0	1	0	0	0
1	1	0	1	0	0
0	0	1	0	1	1
0	0	0	1	0	0
0	0	0	1	0	0

Fig. 30. The adjacency matrix of the graph shown in figure 29

Incidence Matrix The incidence matrix is an $m \times n$ matrix for a graph with m nodes and n edges. An element a_{ij} in the incidence matrix is 1 if edge j has node i in one of its termini. Figure 1.12 shows the incidence matrix of

55. Known as a *self loop* in graph theory terminology.

56. Which in case of molecular graphs will be the valency of the corresponding atom (considering all multiple bonds as single bonds).

the graph in figure 1.10. Since every edge has exactly two termini the number of 1s in each column of the incidence matrix is two. Moreover the number of sum of the elements in each row of an incidence matrix gives the degree of connectivity of the corresponding node.

	a	b	c	d	e
1	1	0	0	0	0
2	0	1	0	0	0
3	1	1	1	0	0
4	0	0	1	1	1
5	0	0	0	1	0
6	0	0	0	0	1

Fig. 31. The incidence matrix of the graph shown in figure 29. The node and edge labels of the graph are also included for clarity.

Edge List The edge list is simply an array of node pairs that are connected by an edge. It can be written as an $m \times 2$ matrix, where m is the number of edges in the graph and the columns contain the two node indices corresponding to an edge. In order to avoid double counting of every edge one may insist that the node index in the first column of the edge matrix is always less (or more) than the node index in the second column.

1	3
2	3
3	4
4	5
4	6

Fig. 32. Edge list of the graph shown in figure 29.

VertexListArray The vertex list array is a simple array of n lists where n is the number of nodes in the graph. In each list corresponding to a node in the graph the list elements are all other node indices that share an edge with the given node.

1	3		
2	3		
3	1	2	4
4	3	5	6
5	4		
6	4		

Fig. 33. Graph shown in figure 29 represented in VertexListArray format.

It is clear that lists of atoms in a molecule that are bonded to each other can be provided in any of the four graph representations discussed above. Often however, when a molecule is described in external coordinates the connectivity information (i.e., lists of atoms that are bonded) is not provided explicitly. However taking advantage of the fact that the properties of chemical bonds in molecules are often determined solely by the type of atoms that constitute the bond, one can make safely extract the required connectivity information. Thus, given the external coordinates of a molecule, one first calculates the distance between all pairs of the constituent atoms. Bonds are assigned to individual pairs of atoms if the distance between them is less than the following set of empirically established cutoffs.

Atom Type	Bond length cutoff ($\mathring{A}$)
X-H	1.1
X-X	1.6
S-S	1.85

Table 5. Empirical cutoff distances for assigning bonds. X indicates any atom type not specified in the table.

One should excercise utmost caution while using the cutoffs in table 5 for assigning bonds. There may be true bonds which are missed out by injudicious use of these cutoffs. Furthermore, even small errors in the external coordinates are likely to generate spurious bonds. In case one assigns bonds using this method, one should also check the valencies of all atoms and particularly the values of all bond angles. In most cases, errors in bond assignment can be detected from these checks which can be corrected manually.

The bonds assigned can be stored in any of the four formats described above. It is simple to convert any of the formats to another. Once all the bonds in a molecule are assigned it is simple to generate lists of bond and torsion angles. For generating bond angle lists, one can start with a vertexlistarray, since for each atom (node) the elements in its corresponding vertex list are other atoms (nodes) it is bonded to, one can simply consider the particular atom to be the central atom in a bond angle, the terminal atoms can be generated by taking pairs of non-identical atoms from the corresponding vertex list. In order to avoid double counting of bond angles one can further enforce that the index of the first atom in a bond angle triple is always less (or more) than the index of the third atom. The following pseudocode gives the algorithm for bond angle enumeration described here.

Let n be the number of atoms in the molecule (and hence the size of the VertexListArray V). Further let p be the number of elements in a particular vertex list.

```
for i = 1 to n
  v2 = V(i,1)
   for j = 2 to p -1
      v1 = V(i,j)
      for k = j+1 to p
         v3 = V(i,k)
          if(v1 < v3)
                write v1,v2,v3
          else
            continue
       end
    end
  end
```

For generating the list of torsion angles one can use an edgelist and a vertexlistarray. The central bond defining a torsion angle can be obtained from the edge list, i.e., the second and third atoms in the torsion angle quartet can come from the bonded atom pairs in the edge list. For each atom in the edge list, consider the atoms connected to it from the corresponding vertexlists. Pick up pairs of such atoms and assign them to the first and fourth positions in the torsion angle quartet taking care that no atom is used twice in a quartet. Again double counting can be avoided by enforcing that the index for the first atom is less (or more) than the index of the fourth atom and the index of the second atom is less (or more) than the third atom.

It should be noted that if the two atoms involved in a bond are linked to m and n other atoms respectively, then there are mn independent torsion angles sharing the same central bond. The values of all these torsion angles are related to each other and can be obtained if the value of just one torsion angle is given. Often therefore just one torsion angle (the so called *principal torsion angle*) among a related set is reported. The principal torsion angle can be determined by setting a priority order[57] to the substituents on second and third atoms (i.e., those that define the central bond). That quartet of atoms is picked for the principal torsion angle whose termini have the highest priority. the following pseudocode describes the algorithm for generating neighbour quartets using the ideas just described.

```
Let V be the VertexListArray and E be the edge list of a molecule with n atoms
for i = 1 to n
  v2 = E(i,1)
  v3 = E(i,2)
```

57. The priority order can be set by rules like the Cahn-ingold-Prelog rules that are also used for assigning absolute configuration for chiral atoms.

```
if(v2 ≮ v3) continue
Let p = size of list V(v2,:)
Let q = size of list V(v3,:)
for j = 2 to p
   v1 = V(v2,j)
   if(v1 = v3) continue
   for k = 2 to q
      v4 = V(v3,k)
      if(v2 = v4) continue
      if(v1 ≮ v4) continue
      write v1,v2,v3,v4
   end
end
end
```

Note 3. **External coordinates for a three atom system**

The fourth atom fixing algorithm requires the external coordinates of the first three atoms. In many cases these corrdinates may be available from a previous iteration of the algorithm or may be supplied in some way be the problem at hand. For those cases where this information is not available, the coordinates of the first three atoms can be generated in some standard orientation. There are many ways this may be done depending on the type of orientation one chooses. The following is an example of one such method.

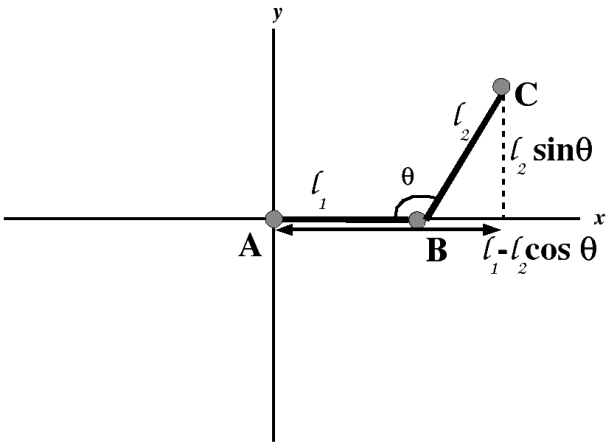


Figure 1.11 A standard orientation for a three atom molecule. The three atoms are labelled A,B and C. The bond length AB and BC are designated l_1 and l_2 respectively and the bond angle A-B-C is designated θ.

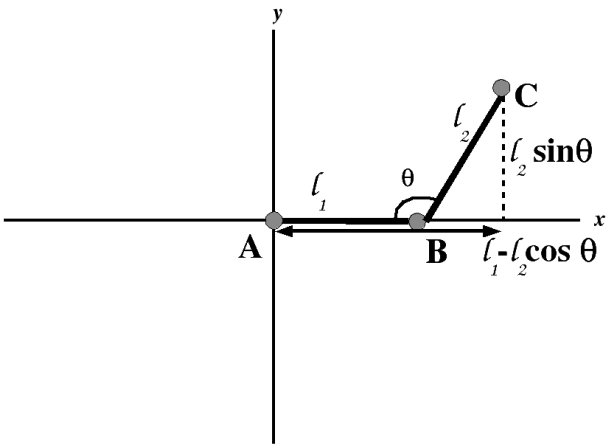


Fig. 34. A standard orientation for a three atom molecule. The three atoms are labelled A,B and C. The bond length AB and BC are designated l_1 and l_2 resopectively and the bond angle A-B-C is designated θ.

Set the first atom A in the origin of the coordinate system. The coordinates of atom A are therefore (0, 0, 0). Assume that atom B lies on the $x-$ axis and the bond length AB is l_1. The coordinates of atom B are therefore (l_1, $0, 0$). Finally assume that atom C lies on the xy-plane with the bond length BC being equal to l_2 and the bond angle A-B-C being equal to θ. Clearly the z-coordinate of C is 0. The y-coordinate is given by $l_2 \sin(\pi - \theta) = l_2 \sin\theta$. The x-coordinate of C is given by the length of the projection of the vector $\overline{\mathrm{AC}}$ on the x-axis which is equal to $l_1 + l_2 \cos(\pi - \theta) = l_1 - l_2 \cos\theta$. Thus the coordinates of the three atoms are given by:

$$A = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, B = \begin{pmatrix} l_1 \\ 0 \\ 0 \end{pmatrix} \text{and } C = \begin{pmatrix} l_1 - l_2 \cos\theta \\ l_2 \sin\theta \\ 0 \end{pmatrix}$$

Note 4. **Theory of Rotations**

If an object (or a molecule) has to be moved without any reflection or change in its shape (i.e., retaining the configuration and conformation of a molecule) then this may done only in the form of a translation or rotation. Thus if a point has a position vector $\boldsymbol{r}$ which after movement becomes $\boldsymbol{r}'$ we may write:

$$\boldsymbol{r}' = \mathcal{R}\boldsymbol{r} + \boldsymbol{T} \tag{100}$$

where $\mathcal{R}$ is the so called *rotation matrix*[58] and $\boldsymbol{T}$ is a translation vector.[59]

58. Rotations always preserve the length of the vectors they transform. Hence they may be classified under orthogonal transformations which by definition can be written in matrix form.

There are several important properties of rotation matrices, some which are listed below:

1. *The columns of a rotation matrix are orthogonal unit vectors.*

   Since the rotation matrix can transform any set of vectors, one can use them to transform the three unit vectors along the x, y and z axes which are orthogonal to each other. Thus we have:

$$\mathcal{R}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k}) = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = (\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}') \tag{101}$$

   Since the starting vectors are orthogonal, the resultant vectors must be also orthogonal which are the columns of the rotation matrix.

2. *The transpose of a rotation matrix is its inverse*

   Premultiplying a rotation matrix with its transpose we have (form equation 1.36)

$$\mathcal{R}^T\mathcal{R} = \begin{pmatrix} \boldsymbol{i}' \\ \boldsymbol{j}' \\ \boldsymbol{k}' \end{pmatrix} \begin{pmatrix} \boldsymbol{i}' & \boldsymbol{j}' & \boldsymbol{k}' \end{pmatrix} \tag{102}$$

   Since $\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}'$ are orthogonal unit vectors we have $\boldsymbol{i}' \cdot \boldsymbol{i}' = 1, \boldsymbol{i}' \cdot \boldsymbol{j}' = 0$ etc. (see equations 1.12 and 1.13). We have:

$$\begin{aligned} \mathcal{R}^T\mathcal{R} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \Rightarrow \mathcal{R}^T\mathcal{R} &= \boldsymbol{I} = \mathcal{R}^{-1}\mathcal{R} \\ \Rightarrow \mathcal{R}^T &= \mathcal{R}^{-1} \end{aligned} \tag{103}$$

3. *The determinant of a rotation matrix is equal to* $+1$

   The determinant of a matrix is given by the triple product $\boldsymbol{r}_1 \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3)$ where $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$ are its column vectors. The is also the volume of the parallelopiped given by these vectors as edges. Since for a rotation matrix, the column vectors are orthogonal the determinant can have a value of ± 1. The volume of the parallelopided given by the unit matrix is 1, hence the determinant of $\mathcal{R}$ also must be equal to 1.[60]

---

59. The order of the rotation and translations are important. Rotating a vector and then translating it is not the same as first translating it and then rotating the translated vector.

60. An orthogonal matrix with determinant -1 corresponds to an inversion. Such a matrix can be obtained by interchanging any two columns of the rotation matrix.

One can describe a rotational transformation uniquely in terms of the rotation matrix. Alternatively it can be defined in terms of a rotation matrix and an angle of rotation about this axis. A third approach to define rotational transformations is in terms of three angles of rotaion also called *euler angles* about the coordinate axes. The euler angles can be chosen in two different ways viz., (a) The first and the third rotations are about the same axes i.e., first rotate around the z-axis, then by the new (rotated) y-axis and then by the new z-axis again or (b) the rotations are about different axes each time i.e., first rotate about z-axis, then around the new y-axis and lastly around the new x-axis.

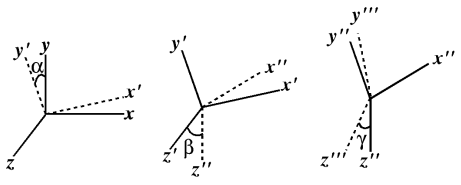


Fig. 35. Diagrammatic representation of the effect of the three eulerian rotations on the coordinate axes.

The matrices used for rotations around the principal axes x, y and z by an angle θ are as follows:

$$\mathcal{R}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \tag{104}$$

$$\mathcal{R}_y = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \tag{105}$$

$$\mathcal{R}_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{106}$$

Using the elementary rotation matrices described above, one can derive a general rotation matrix in terms of the euler angles $\alpha, \beta,$ and γ. Thus:

$$\mathcal{R}_z^{\alpha}\mathcal{R}_y^{\beta}\mathcal{R}_z^{\gamma} = \mathcal{R}$$

and

$$\mathcal{R} = \begin{pmatrix} \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma & -\cos\alpha\cos\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta \\ \sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma & -\sin\alpha\cos\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta \\ -\sin\beta\cos\gamma & \sin\beta\sin\gamma & \cos\beta \end{pmatrix} \tag{107}$$

Perhaps a somewhat more convenient formulation of the rotational transformation is in terms of a rotation axis and an angle of rotation about this axis. The rotation axis may be specified by its direction cosines or alternatively, a pair of angles[61], the so called *polar angles*. Figure 36 pictorially describes the polar angles.

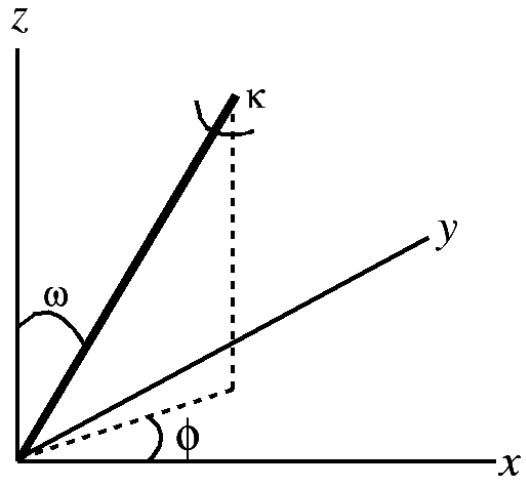


Fig. 36. Definition of polar angles. The rotation is by an angle κ about an axis that makes angles of ω and ϕ with the z and x axes respectively.

If l, m, n be the direction cosines of the rotation axis we have:

$$\begin{pmatrix} l \\ m \\ n \end{pmatrix} = \begin{pmatrix} \sin\omega\cos\phi \\ \sin\omega\sin\phi \\ \cos\omega \end{pmatrix} \tag{108}$$

To construct a general rotation matrix using the polar angles we imagine that the rotation axis is first rotated by angles $-\omega$ and $-\phi$ so as to make it coincident with the z-axis. This is followed by a rotation about the z-axis by the angle κ followed by two further rotations by angles ω and ϕ so as to restore the original orientation of the matrix. Thus we have:

$$\mathcal{R} = \mathcal{R}_z^{\phi}\mathcal{R}_y^{\omega}\mathcal{R}_z^{\kappa}\mathcal{R}_y^{-\omega}\mathcal{R}_z^{-\phi}$$

$$\mathcal{R} = \begin{pmatrix} l^2+(m^2+n^2)\cos\kappa & lm(1-\cos\kappa)-n\sin\kappa & nl(1-\cos\kappa)+m\sin\kappa \\ lm(1-\cos\kappa)+n\sin\kappa & m^2+(l^2+n^2)\cos\kappa & mn(1-\cos\kappa)-l\sin\kappa \\ nl(1-\cos\kappa)-m\sin\kappa & mn(1-\cos\kappa)+l\sin\kappa & n^2+(l^2+m^2)\cos\kappa \end{pmatrix} \tag{109}$$

61. Also called the *inclination* and the *azimuthal* angles.

An alternate form of the rotation matrix that is often useful for calculations is given below:[62]

$$\mathcal{R} = \begin{pmatrix} a^2+b^2-c^2-d^2 & 2(bc+ad) & 2(bd-ac) \\ 2(bc-ad) & a^2-b^2+c^2-d^2 & 2(cd+ab) \\ 2(bd+ac) & 2(cd-ab) & a^2-b^2-c^2+d^2 \end{pmatrix} \tag{110}$$

where:

$$a = \cos\frac{\kappa}{2} \tag{111}$$

$$b = l\sin\frac{\kappa}{2} \tag{112}$$

$$c = m\sin\frac{\kappa}{2} \tag{113}$$

$$d = n\sin\frac{\kappa}{2} \tag{114}$$

Given a rotation matrix one can always decompose it to extract the polar angles. Thus from equation 109 we have:

$$\text{Trace}(\mathcal{R}) = R_{11} + R_{22} + R_{33} = l^2 + m^2 + n^2 + 2(l^2 + m^2 + n^2)\cos\kappa = 1 + 2\cos\kappa \tag{115}$$

The direction cosines l, m, n and hence ω and ϕ can be obtained from the off-diagonal elements thus:

$$R_{32} - R_{23} = 2l\sin\kappa \tag{116}$$

$$R_{13} - R_{31} = 2m\sin\kappa \tag{117}$$

$$R_{21} - R_{12} = 2n\sin\kappa \tag{118}$$

Full range of rotations can be generated with κ, ω, ϕ either in the range $-\pi < \kappa < \pi\,, 0 < \omega < \pi/2, 0 < \phi < 2\pi$ or in the range $0 < \kappa < \pi\,, 0 < \omega < \pi$, $0 < \phi < 2\pi$ since the same matrix is generated when the angles are $(-\kappa, \pi - \omega$, $\pi + \phi)$ and (κ, ω, ϕ).If $\kappa = 0$ there is no rotation and ω and ϕ can take any value.

Note 5. **Derivation of the Pseudorotation Phase Angle Formula**

From equations 4 to 8 we have:

$$\nu_2 = \nu_m\cos P \tag{119}$$

$$\nu_3 = \nu_m\cos(P+\delta) \tag{120}$$

$$\nu_4 = \nu_m\cos(P+2\delta) \tag{121}$$

$$\nu_o = \nu_m\cos(P+3\delta) \tag{122}$$

$$\nu_1 = \nu_m\cos(P+4\delta) \tag{123}$$

62. The reader can easily verify that equations 1.71 and 1.72 are equivalent.

where ν_i are the endocyclic torsion angles, ν_m is the maximum angle of torsion, P is the pseudorotation phase angle and $\delta = 4\pi/5$.

Recognising the periodic nature of the torsion angles we can rewite equations 119 and 123 as:

$$\nu_o = \nu_m \cos(P + 3\delta) = \nu_m \cos(P - 2\delta) \tag{124}$$

$$\nu_1 = \nu_m \cos(P + 4\delta) = \nu_m \cos(P - \delta) \tag{125}$$

From 121 and 125 we have:

$$\nu_4 + \nu_1 = \nu_m[\cos P \cos 2\delta - \sin P \sin 2\delta + \cos P \cos \delta + \sin P \sin \delta] \tag{126}$$

Similarly from 120 and 124 we have:

$$\nu_3 + \nu_o = \nu_m[\cos P \cos \delta - \sin P \sin \delta + \cos P \cos 2\delta + \sin P \sin 2\delta] \tag{127}$$

Subtracting equation 127 from 126 we get:

$$(\nu_4 + \nu_1) - (\nu_3 + \nu_o) = 2\nu_m \sin P\,[\sin \delta - \sin 2\delta] \tag{128}$$

Since $\delta = 4\pi/5$ we can write:

$$\sin \delta = \sin(4\pi/5) = \sin(\pi - \pi/5) = \sin(\pi/5) \tag{129}$$

Similarly we have:

$$\sin 2\delta = \sin(8\pi/5) = \sin(2\pi - 2\pi/5) = -\sin(2\pi/5) \tag{130}$$

From 128, 129 and 130 we have:

$$(\nu_4 + \nu_1) - (\nu_3 - \nu_o) = 2\nu_m \sin P\,[\sin(\pi/5) + \sin(2\pi/5)] \tag{131}$$

From 119, 131 and rearranging we have:

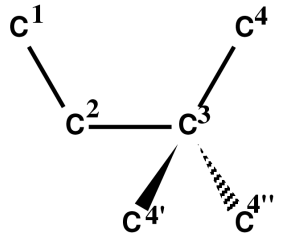
$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_o)}{2\nu_2[\sin(\pi/5) + \sin(2\pi/5)]} \tag{132}$$

which is the required formula.

Problem 1. Prove, from purely geometric considerations, that the H-C-H bond angle in methane (CH_4) is 109.5°. Assume that the shape of the methane molecule is that of a regular tetrahedron.

Problem 2. Given the external coordinates of a molecule and a list of bonds (connectivity table) construct an algorithm to generate the Z-matrix of the type shown in table 2.

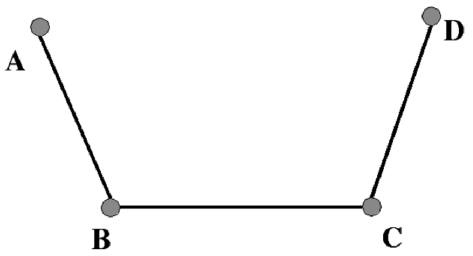
Problem 3. Consider the six atom molecule shown in the figure. Assume that all the bond lengths are identical. All bond angles with C^3 as the central atom are 109.5°. If the torsion angle defined by atoms C^1-C^2-C^3-C^4 is χ, show that torsion angles defined by atoms C^1-C^2-C^3-$C^{4'}$ and C^1-C^2-C^3-$C^{4''}$ will have values $\chi \pm 120°$ respectively.



Problem 4. It has been observed that the distance between successive C^α atoms in polypeptides is essentially constant regardless of the local or overall conformation of the polypeptide. However, there exist a few cases where this distance is very much shorter than normal. What may be the reason for this remarkble constancy in successive C^α-C^α distance and how does one interpret the few deviations that are there?

Problem 5. The value of a torsion angle is invariant to translation and rotations. Is the same also true for reflections? Discuss the reasons for your answer.

Problem 6. Consider the four atom molecule shown in the following figure. Let the bond lengths defined by atoms A-B, B-C and C-D be l_1, l_2, l_3 respectively. The bond angles defined by atoms A-B-C and B-C-D be θ_1 and θ_2 respectively, and the torsion angle defined by atoms A-B-C-D be χ. If the distance between atoms A and D is denoted by λ, derive an expression for λ in terms of $l_1, l_2, l_3, \theta_1, \theta_2$ and χ. Further assume that the bond angles θ_1 and θ_2 are free to assume any value, but the bond lengths and the distance λ remains constant. Derive an expression relating the χ with θ_1 and θ_2 under these conditions.



Exercise 1. Generate the coordinates of a pentapeptide $(\mathrm{Ala})_5$ under the following conditions: (a) All ϕ angles are equal and have a value of -60^o. (b) The ψ angle however varies from -180^o to $+180^o$ in steps of 30^o. Are all the structures feasible? If not list the atom pairs that violate the Ramachandran extreme limit. Describe the major structural features of the models that you generate.

Exercise 2. Generate the $\phi - \psi$ map for two linked peptide units both of which are in the *cis* conformation. Separately consider the cases when a C^{β} atom is present or absence.

Exercise 3. Generate a map in the space of the torsion angles ν_3 and ν_4 that allow the formation of a five-membered ring. Assume that C-C bond lengths are all 1.5 $\mathring{A}$ and the C-O bond length is 1.4 $\mathring{A}$. Bond angles where the central atom is C are all 109.5^o and where the central atom is O it is 120^o. Mark in the map the positions all positions of the pseudorotation phase angle P from 0 to 360^o in steps of 30^o.

Exercise 4. Discuss the possible reasons for natural violations of the Ramachandran stereochemical criteria. What could be the effects of pyramidalization of the peptide nitrogen atom?