Elements of protein structure

Proteins are described in terms of the sequence (primary structure), their short-range conformations and building blocks (secondary structure), tertiary structure (which is how the different secondary structures elements are packed together), and quaternary structure that describes how separate domains or individual chains of proteins are packed together.

Proteins are one-dimensional polymers made of monomers – amino acids. There are twenty types of amino acids all of them share the same backbone structure. The amino acids are different in the side chains that contribute to the observed diversity in protein function and shapes. Without this diversity it is unlikely that protein would have a unique three-dimensional structure. Some of the side chain diversity is connected (but not limited) to residues that are either well solvated or poorly solvated in water (residues -- yet another name for the side chains).

The side chains are considerably shorter that the protein backbone. Glycine is the smallest (only one hydrogen for a side chain). which allows significant flexibility. Proline is the most unusual since both ends of the side chain are covalently linked to the backbone.

To quantitatively discuss protein conformations we start with a detailed description of the backbone of the protein. In the usual chemical notation an amino acid contains carbon (C), nitrogen (N), oxygen (O) and hydrogen (H) atoms. A bond between the atoms is a dash. We have

$$NH3 - C_{\alpha}^{(1)} HR_1 - CONH ... NH - C_{\alpha}^{i} HR_i - CONH ... COO^{-}$$

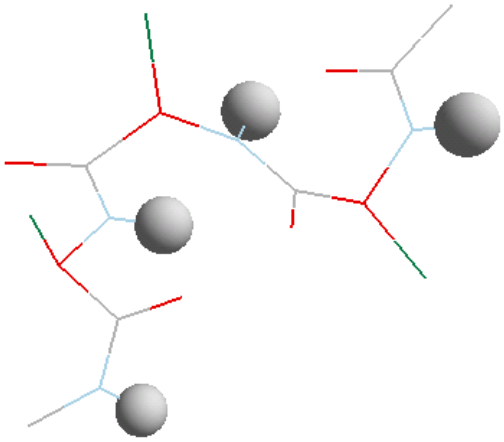
Below we write the chemical formula for the two exceptional amino acids as well as of a "generic" amino acid (xxx) where R denotes the side chain.

$$-gly-: \quad - \quad NH \quad - \quad C_{\alpha}H_2 \quad - \quad CO \quad -$$

$$-pro-: \quad - \quad \begin{matrix} N \\ | \\ CH_2 \end{matrix} \quad - \quad \begin{matrix} C_{\alpha}H \\ | \\ CH_2 \end{matrix} \quad - \quad CO \quad -$$

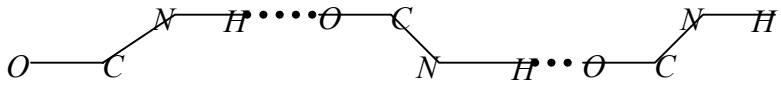
$$-xxx-: \quad - \quad NH \quad - \quad C_{\alpha}HR \quad - \quad CO \quad -$$

A more colorful image of short protein chain (3 alanines) is attached below. The image was prepared with the cmoil program, a part of a larger package for molecular dynamics simulations (http://cbsu.tc.cornell.edu/software/moil/index.htm)



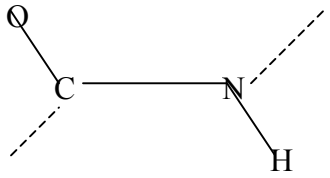
The gray spheres are the hydrogen atoms of the amide planes. There is no real significance of making sphere only for them. The red/green sticks are bonds between the C_α and the side chain center C_β.

A single protein chain varies in length from tens of amino acids to roughly 1,000. The polymer is never branched but is frequently cross-linked by hydrogen bonds and more rarely by sulphor bonds between cysteine residues. Backbone hydrogen bonding is the prime contributor to the so-called secondary structure, and is made between the amide planes (the hydrogen H and the oxygen O of the $NHCO$ groups).

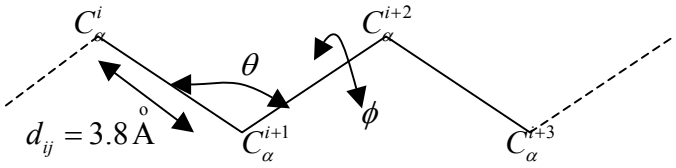


Where the dotted lines denoted a hydrogen bond. A hydrogen bond is significantly weaker than a covalent bond so under normal circumstances (room temperature) hydrogen bonds can be broken. Unfolding of a protein usually includes breaking a very significant fraction of the total number of hydrogen bonds. The cost of not forming hydrogen bonds when the protein is folded can be very high, and they are typically satisfied when the protein accepts a compact state. The condition of satisfying as many as possible hydrogen bonds for folded proteins limits the search space and "helps" the protein (and us) find the right conformation. Note that in the unfolded state the hydrogen bonding between amide groups does not play a significant role since the amide groups form alternative hydrogen bonding to water molecules.

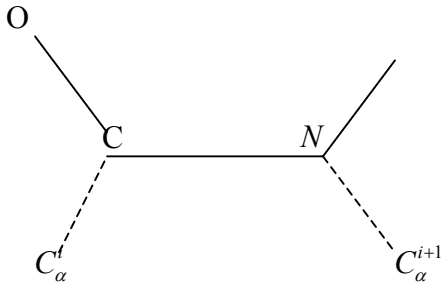
The amide group, $CONH$, is rigid and planar and the typical geometry is trans



The dotted lines denote the continuous direction of the rest of the protein backbone. Since this unit for all practical purposes is rigid and since the dotted bonds to the C_α of the previous and next residues have fix lengths and orientations, the distance between C_α of sequential amino acids is fixed at 3.8 angstroms (angstrom $=10^{-8}$ centimeters). A useful reduced representation of a protein chain is based on the C_α only.

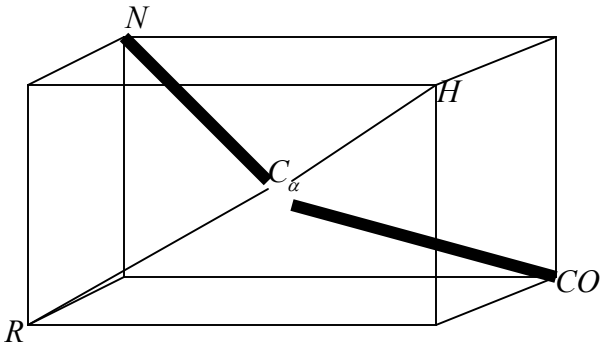


In which the only degree freedom are the angles θ and ϕ. The index i, is the index of the amino acid. There is only one exception to the picture above and this is the amino acid proline. Proline may have amide planes in the cis configuration



The cis configuration has significantly shorter distance than 3.8 angstrom between the alpha carbons.

Let us have a deeper look into the geometry of a single amino acid (L amino acid)



The alpha carbon is at the center of the cube. The bonding so create is of a tetrahedral geometry with angles of (roughly fixed at 109 degrees). The thicker line is the direction of the backbone and R denotes the position of the side chain. The above picture is of an L amino acid. A mirror image isomer is the D amino acid that is not synthesized directly in the ribosome but by special enzymes for that purpose. A D isomer is made from the L isomer by flipping the position of H and R.

Note that there is no continuous transformation that is leading from L to the D structure. It is therefore important (since after all they are quite similar) to have a computational measure that will differentiate between the two. At the least we wish to avoid simulating a protein and accidentally constructing D amino acids instead of L.

Consider four atoms (in that order) $C_\alpha - N - CO - R$. The location of each atom is determined by a Cartesian vector $r = (x, y, z)$. We have

$$r_{C_\alpha - N} = r_{C_\alpha} - r_N$$
$$r_{N-C} = r_N - r_C$$
$$r_{C-R} = r_C - r_R$$

to determine the isomer (mirror image) state of the amino acid we consider the angle between the planes, the plane defines by the three atoms $C_\alpha - N - CO$ and the plane define by $N - CO - R$. It is useful to compute vectors perpendicular to each of the planes $r_\perp^{(1)}$ and $r_\perp^{(2)}$

$$r_\perp^{(1)} = r_{C_\alpha - N} \times r_{N-C}$$
$$r_\perp^{(2)} = r_{N-C} \times r_{C-R}$$

The scalar products of $r_\perp^{(1)}$ and $r_\perp^{(2)}$ is $r_\perp^{(1)} \cdot r_\perp^{(2)} = \left|r_\perp^{(1)}\right|\left|r_\perp^{(2)}\right|\cos(\Phi)$, where Φ is the angle between the planes that we are after. The difference between L and D isomers of an amino acid will be in the sign of Φ. Unfortunately $\cos(\Phi) = \cos(-\Phi)$. This means that we cannot get information on the sign of the angle from the scalar product. The product

$r_{\perp}^{(1)} \times r_{\perp}^{(2)} \cdot r_{N-C} = V \sin(\Phi)$ is proportional to the volume enclosed by the three vectors and to the sine of the Φ angle. If the result is positive we have the L isomer, if it is negative we have the D isomer.

Another task that we will be interested in is how to measure similarity between two protein structures. In the same spirit to sequence comparison method in which we defined first a mean to measure similarity (the sum of similarities scores between aligned amino acids), we need to define a similarity measure between two structures. While it is possible to compare the position of all atoms in the two proteins, we shall be mostly interested in comparing the positions of the C_α only. The reason being that we may compare proteins that differ in their side chain (but still have the same backbone).

We consider two proteins A and B with the same number of amino acids n (the question of alignment of two structures with different number of amino acids will follow the simpler case of overlap). The coordinate vectors of protein A and B are denoted by X_A and X_B respectively. Each of these vectors is of length $3n$ including the (x,y,z) (Cartesian) positions of the C_α-s of the amino acids. The rank 3 vector of amino acid i in structure A is denoted by r_i^A. The distance between the two structures D is defined (and written explicitly as)

$$D^2 = \sum_{i=1}^{n} \left(r_i^A - r_i^B \right)^2$$

Hence we think on the two proteins as a collection of points, or alternatively as a point in $3n$ space for which we compute norm two of the vector difference $\left\| X_A - X_B \right\|_2$

Since the coordinates are defined in Cartesian space, it is possible to translate or rotate one of the structures with respect to the other without changing any of the internal distances between the points that belong to the same object, the protein. That is, maintaining its rigid shape. For simplicity we will always move structure A.

We will consider the translation and the rotation separately. A translation is defined by adding to each of the r_i^A vector a single constant vector t. A rotation is defined by multiplying a coordinate vector by a 3x3 matrix U (e.g. Ur_i^A). U satisfies $UU^t = 1$ and $\det(U) = 1$ (** What are the conditions on U good for?)

Let us start with the simpler problem, that of translation. We wish to determine t so that D^2 is minimal. This is trivial

$$D^2 = \sum_{n=1}^{N} \left(r_n^A + t - r_n^B \right)^2 = \text{minimum}$$

$$2\frac{dD}{dt_\eta} = 2\sum_{n=1}^{N} \left(r_n^A + t - r_n^B \right)_\eta = 0$$

$$t_\eta = \sum_{n=1}^{N} \left(r_{in}^B - r_n^A \right) \qquad \eta = \text{x,y,z}$$

Hence, all we need to do is to correct the position of r_i^A by the difference in the geometric centers of the two proteins. After doing this we will be ready to consider the more interesting problem of overlapping two structures, the problem of rotation.

In fact, to make sure that the next item on the agenda is pure rotation we will set the two geometric centers of the two proteins to zero. In the following derivation we assume that this was already done. We will keep the same notation of r_i^A and r_i^B for the vectors with the adjusted translation.

To correct for possible rotations we write yet another optimization problem

$$D^2 = \sum_{n=1}^{N} \left(U r_n^A - r_n^B \right)^2 = \text{minimum}$$

subject to the constraints: $UU^t = 1$

or $\sum_{k=1}^{3} u_{kk} u_{kj} - \delta_{ij} = 0$

The constraint is inserted to the optimization using Lagrange's multipliers.

$$F = D^2 + \sum_{i,j} \Lambda_{ij} \left(\sum_k u_{ki} u_{kj} - \delta_{ij} \right)$$

The unknowns that we wish to determine are all the elements of the U matrix (9 in all). However, the constraints reduce the number of unknown (** to how many??**). To find the minimum of D^2 subject to the constraint of unitary matrix U, we differentiate with respect to the matrix element u_{ij}, we have

$$\frac{\partial F}{\partial u_{ij}} = \sum_k u_{ik} \left(\sum_n r_{nk}^A r_{nj}^A + \lambda_{kj} \right) - \sum_n r_n^A r_n^B = 0$$

We now define two matrices

$$R_{ij} = \sum_n r_{ni}^B r_{nj}^A \qquad S_{ij} = \sum_n r_{ni}^A r_{nj}^B$$

With the help of the above definition we can write $\frac{\partial F}{\partial u_{ij}}$ in a more compact form

$$U\left(S+\Lambda\right) = R$$

We have one matrix equation with two unknown matrices (!) -- U and Λ. Of course, things are not so bad since we still have the constraint equation: $UU^t = 1$
Note also that $\left(S+\Lambda\right)$ is a symmetric matrix. On the other hand R is not symmetric which makes our problem a little more interesting. The following trick will eliminate some of our problems: Multiply the last equation by its transpose:
$\left(S+\Lambda\right)^t U^t U \left(S+\Lambda\right) = R^t R$

and using $U^t U = 1$ constraint eliminates U from the equation.

$$\left(S+\Lambda\right)\left(S+\Lambda\right) = R^t R$$

The eigenvectors of $\left(S+\Lambda\right)$ - a_k are the same as the eigenvectors of $R^t R$ (assuming no degeneracy). The eigenvalues of $R^t R$ are μ_k^2. The corresponding eignevalues of $\left(S+\Lambda\right)$ are therefore

$\left(S+\Lambda\right)a_k = \pm\mu_k a_k$ (the eigenvalues of the square of the matrix are determined only up to a sign)

Recovering now the original equation we realize that

$$U(S+\Lambda) = R \rightarrow R_{ij} = \sum_k b_{ki}(\pm\mu_k)a_{kj} \rightarrow$$

$$u_{ij} = \sum_k b_{ki} a_{kj}$$

The set of orthonormal vectors b_k are obtained by rotating the set a_k with the (unknown) U. However the b_k are also the "left" eigenvectors of R. The right and the left eigenvectors, and the eigenvalues can be obtained directly from Singular Value Decomposition (SVD) of the asymmetric matrix R. Finally our optimal distance can be computed more directly without thinking on U at all (of course to make a nice plot of overlapping structures requires the rotation matrix):

$$D^2 = \sum_n \left(U r_n^A - r_n^B \right)^2 = \sum_n \left(r_n^A \right)^2 + \left(r_n^B \right)^2 - 2 \sum_n r_n^B \left(U r_n^A \right)$$

$$= \sum_n \left(r_n^A \right)^2 + \left(r_n^B \right)^2 - 2 \sum_n \sum_k \left(b_k r_n^B \right) \left(r_n^A a_k \right)$$

$$= \sum_n \left(r_n^A \right)^2 + \left(r_n^B \right)^2 - 2 \sum_k \left(b_k \right) \left(R a_k \right)$$

$$= \sum_n \left(r_n^A \right)^2 + \left(r_n^B \right)^2 - 2 \sum_k \pm \mu_k$$

Problems to think about

- What shall we do if the determinant of U is not +1, or what to do about the sign of μ?
- What may happen if both molecules are planar/linear?