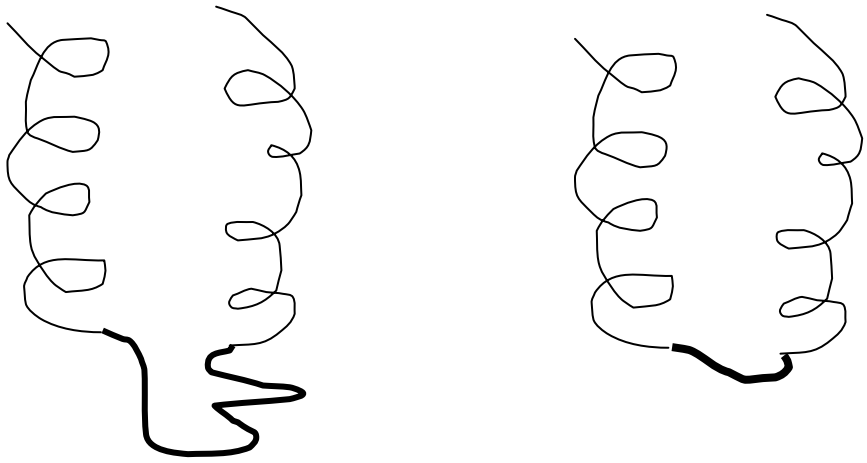
Structural Overlap

Consider the two shapes below



We should be able to detect the similarity between the fragments (helices), and point to the loop as the structural segment that deviates the most, regardless of the sequence. Since the identities of the amino acids are not used (only the C_α positions) highly remote evolutionary connections may be observed. This is our final goal. We start however with the (much) simpler case of overlapping proteins of the same length (no alignment is necessary just proper measure of their distance).

Computing the distance between protein structures

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 $\det(U) = 1$ the usual condition on a rotation matrix that we discussed earlier.

Let us start with the simpler problem, that of translation. We wish to determine a vector of translation t that will be added to each of the atom in protein A so that D^2 is minimal. This is trivial

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

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

$$t_\eta = \frac{1}{N}\sum_{n=1}^{N} \left(r_{\eta n}^B - r_{\eta n}^A \right) = \frac{1}{N}\sum_{n=1}^{N} r_{\eta n}^A - \frac{1}{N}\sum_{n=1}^{N} r_{\eta n}^B = r_{\eta(gc)}^A - r_{\eta(gc)}^B$$

η=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 $r_{\eta(gc)}^A$ and $r_{\eta(gc)}^B$. 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. The unknown below is the rotation matrix U. The structures are known and are presumed rigid. The distance between the two structures is a function of the rotation matrix, and we need to pick such a rotation that makes the distance as small as possible (minimal). As we shall see this problem has a unique solution that will be extremely useful for further analysis. Of course the rotation matrix U cannot be any matrix it must satisfy the obvious conditions we stated earlier. It must keep the overall shape of the protein the same (hence the proteins must be rotated with respect to each other as rigid bodies). We therefore must have $UU^t = I$ to preserve all the internal distances in the protein . We also

must avoid reflection $\left(\det\left(U\right)=1\right)$ since reflection changes the so-called "chirality" of proteins and their chemical identity. We shall deal with distance conservation first ($UU^t = I$) and only later return to the reflection problem $\left(\det\left(U\right)=1\right)$.

After the lengthy introduction here is the optimization task that we are facing: Minimize D^2 as a function of the matrix U . U is a rotation matrix.

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

subject to the constraint: $UU^t = I$

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

We have used the notation $u_{ki} = \left(U\right)_{ki}$ and $\delta_{ij} = \begin{Bmatrix} 1 & i = j \\ 0 & i \neq j \end{Bmatrix}$

Note that the condition $UU^t = I$ is a constraint on a matrix, or alternatively 9 different equations (an equation for each of the elements in the matrix). Some of the conditions are redundant, how many?

Using the "mechanic" of the Lagrange's multipliers we introduced earlier, we add the constraints to the target function that we wish to optimize.

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

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 3. This must be the case since are we argued earlier a rotation is completely determined once three parameters (three rotations angles) are given.

To find the minimum of D^2 subject to the constraint of unitary matrix U (another way of saying $UU^t = I$), 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

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