In this homework, we will compute a rudimentary differential geometric measure of protein shape. You may not use any loops for this assignment.

Since a protein’s backbone can be represented as a sequence of $C_\alpha$ coordinates, it can be made into a smooth curve by interpolating these coordinates. Differential geometry tells us that every curve has a curvature function that measures how the curve bends. The details are as follows. Let

$$\vec{r}(t) = (x(t), y(t), z(t)), \quad a \leq t \leq b$$

be a parameterized curve defined on some interval $[a, b]$. Then

$$\dot{\vec{r}} = (\dot{x}, \dot{y}, \dot{z})$$

is the velocity of $\vec{r}$ and

$$\hat{e} = \frac{\dot{\vec{r}}}{|\dot{\vec{r}}|}$$

is the unit tangent vector of $\vec{r}$. Keep in mind that $\dot{\vec{r}}$ and $\hat{e}$ are still functions of $t$. The curvature of $\vec{r}$ is then defined as

$$\kappa(t) = \left| \frac{d\hat{e}}{dt} \right|$$

When $\kappa$ is large, $\vec{r}$ bends sharply; when $\kappa$ is small, it bends little.

**Part (a).** Show that

$$\frac{d\hat{e}}{dt} = \frac{\ddot{\vec{r}}}{|\hat{e}|^3} - \frac{\dot{\vec{r}}(\dot{\vec{r}} \cdot \ddot{\vec{r}})}{|\dot{\vec{r}}|^3}$$

To use these concepts on a protein $P$, we build a curve $\vec{r}$ out of $P$ by interpolating. If $P$ has $N$ residues, and if $\vec{s}_i$ is the position vector of the $i$-th residue ($i = 1, \ldots, N$), then we need $\vec{r}(i) = \vec{s}_i$ for each $i$, and we need $\vec{r}(t)$ to pass through all the $\vec{s}_i$ smoothly. We will use cubic spline interpolation to do this for PDB protein 1A6Z.

---

1Recall that $f \equiv \frac{df}{dt}$ for any function $f(t)$.
Part (b). Download 1A6Z from the PDB and use pickCA to read the C\textsubscript{\textalpha} coordinates into a Matlab matrix ca.\textsuperscript{2} You should verify that size(ca) = [272 3]. The matrix ca represents our protein \( P \). The \( i \)-th row is \( \vec{s}_i \). (Nothing to show for this part).

Part (c). Use Matlab’s \texttt{spline} function to compute a cubic spline interpolation of ca. You will need to compute three splines, one for each component (column) of ca. These three splines represent the \( N - 1 \) cubic interpolants

\[
\tilde{r}_i(t) = \vec{a}_i + \tilde{b}_i t + \tilde{c}_i t^2 + \tilde{d}_i t^3, \quad t \in [0,1], \quad i = 1, \ldots, N - 1
\]

so the “concatenation” of the \( \tilde{r}_i \) gives \( \tilde{r} \). Display the vectors \( \vec{a}_{13}, \vec{b}_{13}, \vec{c}_{13}, \vec{d}_{13} \), which are the vector coefficients of the thirteenth cubic. Show Matlab code and results.

Part (d). Let \( \kappa_i(t) \) denote the curvature of the protein between residues \( i \) and \( i + 1 \) \( (t \in [0,1]) \). For this problem, we are interested only in the curvature at residues; we won’t care about the curvature between them. Let \( k_i \) denote the curvature at residue \( i = 1, \ldots, N - 1 \). Express \( k_i \) in terms of \( \kappa_i \). Then, using your derivation of part (a) and referring to the expression for \( \tilde{r}_i(t) \) above, show that

\[
k_i = \frac{2}{|\vec{b}_i|} \sqrt{\frac{|\vec{c}_i^2 - (\vec{b}_i \cdot \vec{c}_i)^2|}{\vec{b}_i^2}}
\]

Part (e). Compute the \( k_i \) from your splines using the formula in part (d) (show only the Matlab code). Also show a plot of \( k_i \) vs. \( i \) for \( i = 1, \ldots, N - 1 \).

Part (f). Does curvature correlate well with secondary structure? To answer this question, use the structure info on 1A6Z in the PDB. The link path from the 1A6Z PDB page is \texttt{Geometry->Motif Summary}. (Remember that we are looking only at chain A.) Look up the residue numbers of all \( \alpha \)-helices that are at least 10 residues long and label them with a 0. Then look up the residue numbers of all \( \beta \)-strands that are at least 10 residues long and label them with a 1. Label all other residues 2. Now do a scatter plot of the curvature values vs. labels. That is, put these labels into a vector \( l \) and plot \( k \) vs. \( l \). In the third argument to Matlab’s \texttt{plot} command, you will need to specify points instead of lines. Show the plot and your Matlab code. Answer the question posed at the top of this paragraph.

\textsuperscript{2}Remember that Matlab filenames are case sensitive.