Discussion on the effect of rotation of the coordinates on the method

In this supplementary section, we discuss the effect of rotation of the coordinate data on the message lengths and the resulting segmentation. (It is easy to see that the method described in the paper is invariant to translation.)

A protein is specified in some arbitrary orientation using (x, y, z) coordinates. In our method, the hypothesis (*i.e.,* the collection of end points of the segmentation) is transmitted using $\log^* k + k \log V$ bits, where V is the volume of a bounding box (see Section 4.1 in the main text). However, when the protein is rotated, the bounding box changes, and hence V, resulting in different message lengths. This can be trivially avoided by defining the bounding box in some canonical way.

Again, in our method, each intermediate point corresponding to a line segment is transmitted as spatial deviations, *i.e,* in $(\Delta s, t, u)$ coordinates, where Δs is along the line segment, and t and u are in the plane that contains the point, and is normal to the line segment. For the receiver to decode the coordinate (to a reasonable precision) from the spatial deviations, the t and u axes must lie in the perpendicular plane but their choice is otherwise "arbitrary".

The intermediate point is transmitted as Δs, t and u deviations from a predicted location. This location is the centre of a probability distribution, the spread of which shows the uncertainty.

For a protein, it is straightforward to model Δs as independent of t and u and the choice of the perpendicular plane, as the lateral distances along the line, Δs's, are invariant to rotation. However, t's and u's change under rotation of the coordinates.

The simplest assumption with t and u deviations is to also model them independently of each other (as in the main text), since there is no obvious reason to believe that any observed deviations in the t and u directions are correlated, thus leading to 3 independent standard deviations as described in the main text: $\sigma_{\Delta s}, \sigma_t$, and σ_u.

Since the choice of the t and u axes is arbitrary except that they must lie in the perpendicular plane arbitrarily chosen, if σ_t and σ_u differ, a random rotation of the axes, (t,u) to (t', u') will result in a different probability distribution, making the method sensitive to rotation of the coordinates. A principled scheme to overcome this sensitivity is to choose the *rotationally invariant bivariate normal distribution* where the model assumes $\sigma_t \equiv \sigma_u$. Under this assumption, if the coordinates are rotated by an arbitrary angle, the distribution still remains exactly the same, making this procedure completely rotation-invariant.

Alternatively, 'less principled' strategies to overcome the arbitrariness of the choice of t and u axes is to choose these spatial deviations in some rotation-independent way. For example, this can be done with respect to the previous transmitted segment, or by transmitting in the second part of the two part

message, a reference point for every line segment, based on which deviations t and u can be computed.

In practice, based on the experiments we conducted, the choice of the strategy resulted in the message lengths to change, if only slightly (by up to 2%) between the methods. This indeed can cause minor differences in the endpoints of the resulting segmentation.