Ramachandran plots with Gaussian Mixture Models

2019-07-24

Contents

1 Aim

Our goal is to identify kinks in alpha-helices and compare helices from different structures. For this we want a metric that captures how "helix-like" a particular residue is as we travel along the helix.

A simple way to analyze this is just to look at the ϕ and ψ angles along the backbone. Alpha helices are known to have restricted conformations, with the majority of residues clustered around about (-300,-40).

Although we sometimes don't think of it as such, a Ramachandran plot is actually a 2D histogram, where the z axis corresponds to the probability of finding a residue with that (ϕ, ψ) pair.

So, the "helixness" of a residue is something like it's distance to that central peak. Note that this ignores other possible indicators of the kink, such as disruptions to the hydrogen bonding network.

2 Gaussian Mixture Model

To approximate the density of points in the Ramachandran plot we approximate it with a gaussian mixture model (GMM). This fits a set of 2D gaussians to the plot which together sum to the density at each point. This is convenient, because we can then interpret each component gaussian as a type of secondary structure.

To fit the model, ϕ/ψ angle for 137888 residues were downloaded from the Protein Geometry Database [2]. This contains residues from high-quality structures ($\leq$ 1.2Åresolution, $\leq$ 0.25 R-factor, $\leq$ 0.3 R-free). PGD 1.0.2 was used.

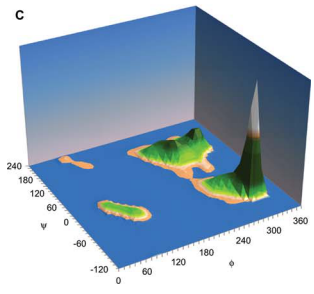


Figure 1: "Geo-style 3D Ramachandran plot" from Hollingsworth and Karplus [1]

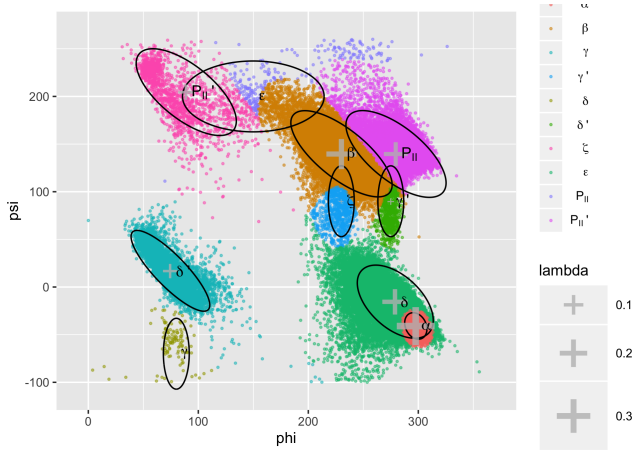


Figure 2: **Initial Classifications**. Initial rough classes based on [1]

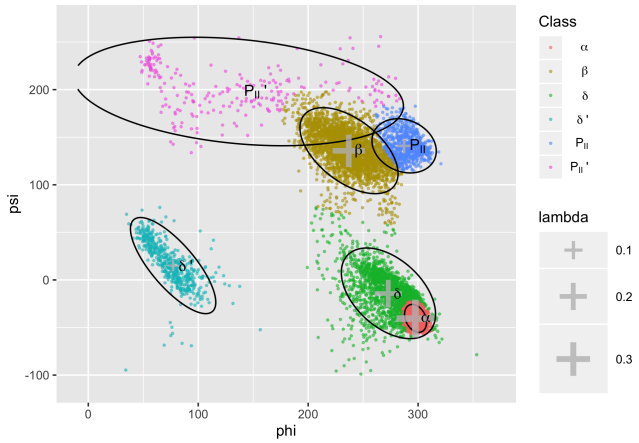


Figure 3: **First round classifications**. Classes after training with a subset of residues.

Fig 2 shows the data, labelled with the secondary structure classifications suggested in [1]. We will use these as the starting point for our model.

We reduce the number of classes somewhat (mostly for performance, but also because minor classes are poorly constrained and make it harder to converge). Then we run some iterations of GNN with a subset of the data to get some rough classes (Fig 3).

Now expand to the full dataset and do additional training (Fig 4).

There are a few indicators that this is not a great model of Ramachandran space. The classes have changed significantly from the first round (particularly P'_{II}), and the likelihood had not yet converged during training. In our case this suggests that we don't have enough gaussians to accurately model the non-gaussian tails of the density. Even well-populated gegions like δ' have some empty regions, indicating that a single gaussian might not be a good fit for those. However, the α region that we're particularly interested in are fairly stable and well-fit, so I won't extend the training further.

3 Measuring "helix-ness"

With this we can measure how helix-like a particular (ϕ, ψ) measurement is. Let's just consider the α region. The probability density in this region is given by

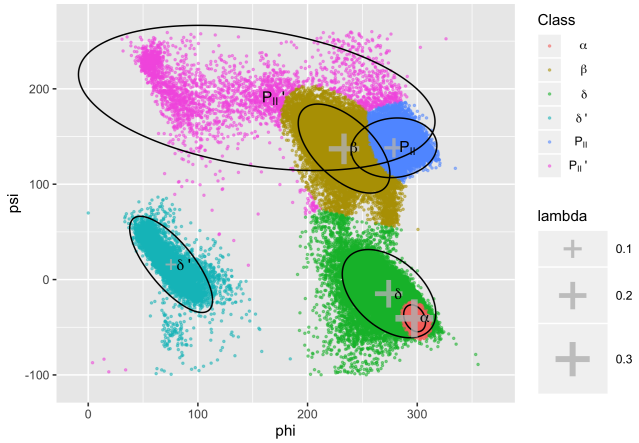


Figure 4: **Final round classifications.** Classes after training with all points (but stopped before convergence).

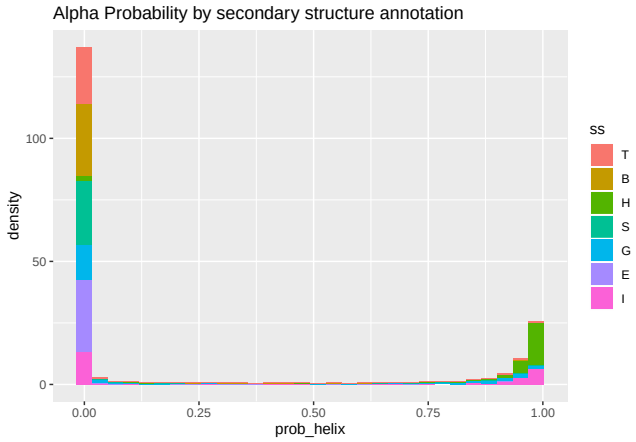
$$
\begin{aligned}
p(x \in \mathbb{R}^k) &= \frac{\lambda}{\sqrt{(2\pi)^k \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right) \\
p_\alpha(\phi,\psi) &= \frac{\lambda_\alpha}{\sqrt{(2\pi)^2(\sigma_\phi\sigma_\psi - \sigma_{\phi\psi}^2)}} \exp\left(-\frac{\sigma_\phi(\phi-\mu_\phi)^2 - 2\sigma_{\phi\psi}(\phi-\mu_\phi)(\psi-\mu_\psi) + \sigma_\psi(\psi-\mu_\psi)^2}{2(\sigma_\phi\sigma_\psi - \sigma_{\phi\psi}^2)}\right) \\
&= \frac{0.36}{184.09} \exp\left(-\frac{21.6(\phi-297.62)^2 + 17.05(\phi-297.62)(\psi+40.59) + 43.1(\psi+40.59)^2}{1716.89}\right)
\end{aligned}
$$

If desired, this can be added to δ probability to include the full region:

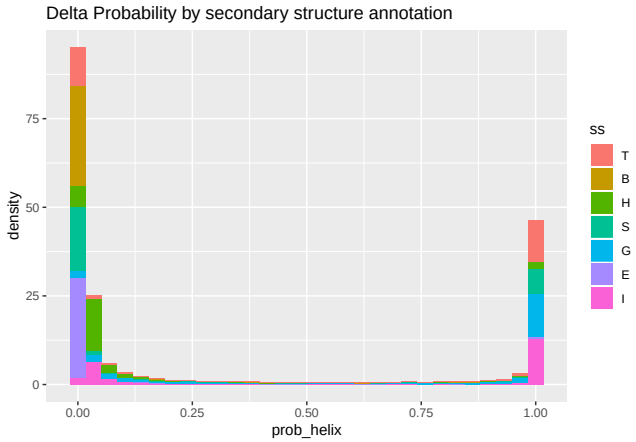
$$
p_\delta(\phi,\psi) = \frac{0.2}{2410.5} \exp\left(\frac{394.28(\phi-274.27)^2 + 378.84(\phi-274.27)(\psi+14.77) + 464.29(\psi+14.77)^2}{2.94\times10^5}\right)
$$

The dataset includes secondary structure assignments, using DSSP codes: - G = 3-turn helix (3_{10} helix). Min length 3 residues. - H = 4-turn helix (α helix). Minimum length 4 residues. - I = 5-turn helix (π helix). Minimum length 5 residues. - T = hydrogen bonded turn (3, 4 or 5 turn) - E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues. - B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation) - S = bend (the only non-hydrogen-bond based assignment). - C = coil (residues which are not in any of the above conformations).

Here is the p_α distribution by secondary structure using only p_α:



And here's the larger p_δ region:



So we see that p_α is targetting tightly the canonical α helix, while p_δ covers 3_{10} and π helices, as well as some turn residues. On the other hand, a few G and I residues are not contained in either the α or δ regions of the Ramachandran plot, indicating that hydrogen-bonding patterns provide more (or at least different) information than just looking at backbone angles.

4 Conclusion

In summary, using p_α is probably a good metric for similarity to canonical α helices. Combining this with p_δ will include other types of helix, and may be more tolerant to bends, helix caps, etc. Either metric can be easily computed based on the (ϕ, ψ) angles.

References

1. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. BioMolecular Concepts. 2010;1: 271–283.

2. Berkholz DS, Krenesky PB, Davidson JR, Karplus PA. Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucleic Acids Res. 2010;38: D320–5.