Exercise 06

Simulation data analysis II

Deadline: Please hand in your protocol by Thursday, the 6th of June 2019, 10 am to jan.joswig@fu-berlin.de and marco.manni@fu-berlin.de. We highly appreciate solutions presented as jupyter notebook (.ipynb).

In this exercise you can for example use the feature distributions of the two provided trajectories of system A and B (`login.bcp.fu-berlin.de:/home/janjoswig/MD18/Ex06/`) that you have collected in the previous exercise.

For the last task 7.4 you need to download these files: `https://www.dropbox.com/sh/88lydj4g4dubz30/AABHNfP1G9k2Fxj50Yo_2MLaa?dl=0`.

7.1 Normalised euclidean distance

The euclidean distance is the ordinary straight-line distance between two points (vectors) in euclidean space. This metric can be also applied to quantify the difference δ of two data sets (vectors) x and y (e.g. histograms) of length n.

$$\delta_{\mathrm{eu}} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Use a Python-script to compute δ_{eu} for two comparable distributions, e.g. the χ_1-angle of H294 in system A and B (normalised histograms with 360 bins in the range between -180 and 180).

7.2 Kullback-Leibler divergence

Yet another way to quantify the difference between two discrete probability distributions $p(x)$ and $q(x)$ (histograms) is to use the non-symmetric Kullback-Leibler divergence

$$\delta_{\mathrm{kl}} = \sum_{i=1}^{n} p(x_i)\ln\left(\frac{p(x_i)}{q(x_i)}\right) \tag{2}$$

This is only meaningful, if always $p(x) = 0$, where also $q(x) = 0$. In other words, this can only be computed in data regions, where $q(x) > 0$.

Use a Python-script to compute δ_{kl} for two comparable distributions, e.g. the χ_1-angle of H294 in system A and B.

7.3 Mutual information

In contrast, the mutual information of two discrete observables is a measure for the interdependent relation between them (how much does the probability distribution of one variable $p(x)$ depend on the distribution in another variable $p(y)$?).

$$I(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \ln \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \tag{3}$$

where $p(x,y)$ is the joint probability (2D histogram) of the two variables. If x and y are completely independent measures, it holds $p(x,y) = p(x)p(y)$ and therefore $I = 0$.

Since the values of I are not confined to a certain interval, for comparison it can be useful to compute the normalised mutual information as

$$\tilde{I}(x,y) = \frac{I(x,y)}{\min(h(x), h(y))} \tag{4}$$

where $h(x)$ denotes the informational entropy of the probability distribution of x.

$$h(x) = -\sum_{i=1}^{n} p(x_i) \ln(p(x_i)) \tag{5}$$

Use a Python-script to compute I and $\tilde{I}$ for two (contingently) related distributions, e.g. the χ_1- and χ_2-angle of H294 in system B.

7.4 Coupling constants

NMR ^3J-coupling constants depend on the dihedral angle between the coupling atoms and are predicted by the Karplus-equation:

$$^3J = A\cos^2(\phi - \frac{1}{3}\pi) + B\cos(\phi - \frac{1}{3}\pi) + C \tag{6}$$

where the NH–H$_\alpha$ backbone dihedral angle ϕ is given in radians and A, B and C are empirical parameters (Schmidt: $A = 7.90\,\mathrm{Hz}$, $B = -1.05\,\mathrm{Hz}$, $C = 0.65\,\mathrm{Hz}$, standard deviation of the result $\sigma = 0.78$. Bax: $A = 7.09\,\mathrm{Hz}$, $B = -1.42\,\mathrm{Hz}$, $C = 1.55\,\mathrm{Hz}$, $\sigma = 0.39$.)

Calculate the coupling constants for the given ϕ-angle trajectories (1 to 5.txt) using the Schmidt and Bax parameters. Plot the mean values with error bars and compare the result to the experimental value (experimental.dat). Which prediction is more accurate?