Reporting errors on fingerprint summaries

D. Provasi

(Dated: 12 March 2019)

If calculate interaction fingerprints on a MD trajectory, each fingerprint bit is a series $f_i(t) \in \{0, 1\}$. How can we estimate the probability that a contact is formed, and its error? What should we conclude if, e.g., different trajectories j have different averages $\langle f_i \rangle_j$?

Physically motivated, we can describe the dynamics of each contact by a simple two-state Markov model, that specifies the probability that a contact stays formed $\theta_{11} = p(f_i(t) = 1 | f_i(t - \tau) = 1)$ or is broken $\theta_{10} = p(f_i(t) = 0 | f_i(t - \tau) = 1)$, and all other entries of the 2×2 transition matrix are determined by these two numbers. We notice that all 2×2 transition matrices satisfy the detailed balance, and thus it's simple to estimate the model from the counts

$$c_{ij} = \sum_t \delta_{f(t),j} \; \delta_{f(t-\tau),j} \tag{1}$$

If c_{ij} are the observed transition counts between states $i, j \in \{0, 1\}$ the likelihood of a given transition matrix is

$$p(c|\theta) = \prod_{ij} \theta_{ij}^{c_{ij}}$$

and using a Dirichlet prior[1] on θ, $p(\theta) = \prod_{ij} \theta_{ij}^{b_{ij}}$, (which we specialise for the case $b = -1$ that gives scale-invariance), we have that the posterior, which we can express as a function of the two independent parameters θ_{10} and θ_{01} is:

$$p(\theta_{10}, \theta_{01}|c) = \frac{\theta_{01}^{c_{01}-1}(1-\theta_{01})^{c_{00}-1}}{B(c_{01}, c_{00})} \frac{\theta_{10}^{c_{10}-1}(1-\theta_{10})^{c_{11}-1}}{B(c_{10}, c_{11})}$$

the denominators are beta functions that normalise the posterior:

$$B(a, b) = \int d\theta \; \theta^{a-1}(1-\theta)^{b-1}$$

Given that θ satisfies the detailed balance, the probability for the contact to be formed is given by the relation

$$\frac{\pi_1}{1 - \pi_1} = \frac{\theta_{01}}{\theta_{10}}$$

i.e.

$$\pi_1 = \frac{\theta_{01}}{\theta_{10} + \theta_{01}}$$

and its posterior distribution is therefore:

$$p(\pi_1 = x|c) = \int d\theta_{10} \; d\theta_{01} \delta(x - \frac{\theta_{01}}{\theta_{10} + \theta_{01}}) p(\theta_{10}, \theta_{01}|c) \tag{2}$$

We can summarise $p(\pi_1|c)$ with the median and $(\alpha, 1-\alpha)$ credible intervals. Notice that the model also provides an estimate of the stability of the bond in terms of, e.g. survival time (how long on average, the bond is formed) from the posterior of $\tau_{\text{off}} = \theta_{10}^{-1}$.

[1] F. Noé, C. Shütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. 106, 19011 (2009)