PHYS 414 Problem Set 1: Aliens and amino acids

Problem 1: Are we alone in the universe?

In this problem we will see how Bayesian analysis can help us estimate model parameters even in the extreme case of a single datapoint: life had to arise on Earth earlier than 3.5 Gyr (gigayears) ago (see Fig. 1 for the oldest fossilized evidence currently known). As of now we have no other datapoints of life existing anywhere in the universe (though according to a study published in January 2015 there are tantalizing indications that the Curiosity rover on Mars may be on the verge of adding another datapoint; see part f of this problem for an actual calculation of what this would imply). In general, can we say anything about the likelihood of life arising from non-living matter, a process known as *abiogenesis*? Life began early in the Earth's history: the Earth is 4.5 Gyr old, and life arose within the first 1 Gyr of its existence, though almost certainly not within the first 0.5 Gyr because conditions on the very early Earth were inhospitable. This fact seems to support the idea that abiogenesis is a typical occurrence in the universe, fueling optimism about life existing on many Earth-like exoplanets in habitable zones around Sun-like stars. The current estimate based on data from the Kepler spacecraft is that there could be roughly $\approx 10^{10}$ such planets in the Milky Way alone [Petigura *et al.*, Proc. Natl. Acad. Sci. **110**, 19273 (2013)]. If they are of comparable age to the Earth, what fraction of them harbor life? Is the optimism justified?

Figure 1: Datapoint #1: fossilized evidence of microbial communities dating back to 3.5 billion years ago, discovered in western Australia [Noffke *et al.*, Astrobiology **13**, 1103 (2013)].

A more careful evaluation using Bayesian analysis was performed by David Spiegel and Edwin Turner [Proc. Natl. Acad. Sci. **109**, 395 (2012); posted on the course website]. We will derive (in simplified form) a version of their main results. The goal is to determine the conditional probability $\mathcal{P}(\mathcal{M}(x)|\mathcal{D})$. Here $\mathcal{M}(x)$ is the theoretical model for abiogenesis, which depends on some parameter(s) x (in our case it will be a single parameter). $\mathcal{D}$ is the data, which consists of humans having "measured" that life arose on earth by a time $t_{\rm emerge} \approx 1$ Gyr after the planet's formation. Since $\mathcal{P}(\mathcal{M}(x))$ can be interpreted as the probability of the model being true for a specific value of x, the conditional probability $\mathcal{P}(\mathcal{M}(x)|\mathcal{D})$ encapsulates what we can say about x given the existing data. To evaluate it, we use Bayes's rule:

$$\mathcal{P}(\mathcal{M}(x)|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}(x))\mathcal{P}(\mathcal{M}(x))}{\mathcal{P}(\mathcal{D})} \tag{1}$$

The denominator $\mathcal{P}(\mathcal{D})$ is a independent of x, so we can treat it as a normalization constant ensuring that $\int dx\, \mathcal{P}(\mathcal{M}(x)|\mathcal{D}) = 1$. To complete the analysis, we need expressions for $\mathcal{P}(\mathcal{D}|\mathcal{M}(x))$ and $\mathcal{P}(\mathcal{M}(x))$. The latter represents our prior knowledge (rough guess-work!) about x. Let us find each of these expressions in turn.

a) The first ingredient is a model for abiogenesis. We start with the assumption that conditions on a planet right after its formation will not allow life, up until some minimum time t_{min} has passed. If $t = 0$ is the time of planetary formation, we will fix $t_{\text{min}} \approx 0.5$ Gyr, assuming it is comparable for all Earth-like planets. Though abiogenesis is a complex series of chemical events, we can combine them all into a single overall "reaction", which happens at an unknown constant rate λ (a Poisson process) for all times $t \geq t_{\text{min}}$. More precisely, λ is the probability per unit time of abiogenesis, so that the probability of life arising in some short interval dt is λdt (or equivalently, $1 - \lambda dt$ is the probability that life did not arise in this interval). The probabilities in each consecutive interval (i.e. t to $t + dt$ and $t + dt$ to $t + 2dt$) are independent of each other. This model does not preclude life arising independently multiple times, but we are only interested in the first instance. Given the above assumptions, use the laws of probability (and the limit $dt \to 0$) to show that the probability that no life has arisen up to time t after a planet's formation is:

$$P_{\text{no-life}}(\lambda, t) = \begin{cases} 1 & 0 \leq t < t_{\text{min}} \\ e^{-\lambda(t - t_{\text{min}})} & t \geq t_{\text{min}} \end{cases} \tag{2}$$

Hence the probability that life has arisen (at least once) before time t is $P_{\text{life}}(\lambda, t) = 1 - P_{\text{no-life}}(\lambda, t)$. This will be our main model, governed by a single parameter λ which we would like to pinpoint. (As we will see in part c, we will do this by estimating $x \equiv \log_{10} \lambda$, the overall order of magnitude.)

b) To get a sense of the physical meaning of λ, show that the above model predicts the mean time at which life arose as $\langle t \rangle = t_{\text{min}} + \lambda^{-1}$. *Hint:* Which probability distribution do you use to evaluate $\langle t \rangle$? Do not just plug in $P_{\text{life}}(\lambda, t)$, since this is a cumulative distribution: it measures the probability of life emerging at *any* time before t. How do you find the probability of life emerging just during some small interval t to $t + dt$?

c) If you assume λ is set by fundamental chemistry and is the same throughout the universe, let us get a feel for the consequences of its scale. Find the different numerical values of λ (in units of Gyr^{-1}) that would imply the following facts are true for Earth-like planets of comparable age to ours ($t_0 = 4.5$ Gyr):

- λ_1: on average, we are the only such planet at the present time in the entire observable universe where life has emerged (out of $\approx 10^{20}$ Earth-like planets of similar age in the universe)
- λ_2: on average, we are the only such planet at the present time in the Milky Way where life has emerged (out of $\approx 10^{10}$ Earth-like planets of similar age in our galaxy)
- λ_3: on average, life emerges 1 million years after t_{min}. This would virtually guarantee that every Earth-like planet of comparable age in the universe has life.

From top to bottom, these give you a sense of the immense breadth of possible λ values. Since we do not even have a grasp of its order of magnitude, our prior probability distribution $\mathcal{P}(\mathcal{M}(\lambda))$ should reflect this. Let us define $x = \log_{10} \lambda$ and say that all orders of magnitude between $x_{\text{min}} = \log_{10} \lambda_1$ and $x_{\text{max}} = \log_{10} \lambda_3$ are equally probable. Writing $\mathcal{M}(x)$ instead of

$\mathcal{M}(\lambda)$ we will choose our prior probability distribution to be:

$$\mathcal{P}(\mathcal{M}(x)) = \begin{cases} \frac{1}{x_{\text{max}} - x_{\text{min}}} & \text{if} \quad x_{\text{min}} \leq x \leq x_{\text{max}} \\ 0 & x < x_{\text{min}} \quad \text{or} \quad x > x_{\text{max}} \end{cases} \tag{3}$$

d) The implications of our single datapoint $\mathcal{D}$ are more complicated than just specifying an upper bound on Earth's abiogenesis. What $\mathcal{D}$ really states is that: "an intelligent life form on Earth was able to gather evidence at the present time ($t_0 = 4.5$ Gyr) showing that life started before a time $t_{\text{emerge}} = 1$ Gyr in the Earth's history." This presupposes that enough time has passed between t_{emerge} and the t_0 for evolution to produce a scientifically-advanced species capable of investigating fossil evidence of abiogenesis. If life on Earth emerged at $t = 4.0$ Gyr, there almost certainly would not be enough time for evolution to produce a species to collect the datapoint $\mathcal{D}$ at t_0. Let us specify a minimum time delay δt_{evolve} for the evolution of an intelligent species after abiogenesis. Then only abiogenesis events where $t_{\text{emerge}} < t_0 - \delta t_{\text{evolve}} \equiv t_{\text{required}}$ could have any possibility of being measured. Let us choose $\delta t_{\text{evolve}} = 1$ Gyr to set a rough time scale (probably on the short side) for the evolution of intelligence, so $t_{\text{required}} = 3.5$ Gyr is the cutoff for measurable abiogenesis required by evolutionary constraints. Let E be the statement "abiogenesis occurred between t_{min} and t_{emerge}", and R be the statement "abiogenesis occurred between t_{min} and t_{required}". Then we will take $\mathcal{P}(\mathcal{D}|\mathcal{M}(x))$ to mean $\mathcal{P}(E|R, \mathcal{M}(x))$, or the probability that E is true given that R and the model $\mathcal{M}(x)$ are true. Using the laws of probability and the result of part a, argue that for any measured value of t_{emerge},

$$\mathcal{P}(\mathcal{D}|\mathcal{M}(x)) = \begin{cases} \frac{P_{\text{life}}(10^x, t_{\text{emerge}})}{P_{\text{life}}(10^x, t_{\text{required}})} & \text{if } t_{\text{min}} \leq t_{\text{emerge}} \leq t_{\text{required}} \\ 0 & \text{if } t_{\text{emerge}} < t_{\text{min}} \text{ or } t_{\text{emerge}} > t_{\text{required}} \end{cases} \tag{4}$$

Hint: Think about the definition of conditional probability. Also note that if $t_{\text{min}} \leq t_{\text{emerge}} \leq t_{\text{required}}$, then R is definitely true if E is true.

e) Putting the result of parts c and d together, use Bayes's rule to determine the posterior probability $\mathcal{P}(\mathcal{M}(x)|\mathcal{D})$. Make sure to normalize by choosing some appropriate numerical value for $\mathcal{P}(\mathcal{D})$. Plot $\mathcal{P}(\mathcal{M}(x)|\mathcal{D})$ versus x to see how the probability behaves. Using numerical integration, figure out the probability that x is between x_{min} and $x_{\text{mid}} = \log_{10} \lambda_2$. Let us call this probability p_{L}, where L represents extreme loneliness (we are surely alone in our galaxy, and possibly the observable universe). On the other extreme, figure out the probability p_{M} that 99% or more of Earth-like planets of comparable age to ours have seen life emerge. M represents "the more the merrier." How do you like these odds? While p_{M} is greater, p_{L} is still significant, making choosing between these options a tossup. *Hint:* you may find your numerical integrator (Mathematica!?!) gives nonsense when you try to extend the integration range down to

Figure 2: Datapoint #2 (hypothetical): the Gillespie lake outcrop on Mars exhibiting potential signs of microbial structures.

x_{min}. To resolve this, use the $\lambda \to 0$ limit of Eq. (4) (it goes to a simple constant) when integrating below x_{mid}. Use the full expression above x_{mid}.

f) Nora Noffke, the geobiologist responsible for discovering the oldest fossils on Earth (Fig. 1) published an article recently analyzing photos taken by the Curiosity rover on Mars (Fig. 2; see the write-up at: http://shar.es/1bNqS7). She makes a case that Mars exhibits structures remarkably similar to fossilized microbial mats seen on Earth. If these speculations are proven to be true, we would have a second datapoint. What would be the consequences? The Gillespie lake outcrop on Mars where these photos were taken is 3.7 Gyr old, so $t_{\text{emerge}}^{\text{Mars}} = 0.8$ Gyr (Mars has the same age as Earth). Assuming t_{min} is unchanged for Mars, and that life arose there independently of Earth, how would $\mathcal{P}(\mathcal{D}|\mathcal{M}(x))$ change with two datapoints? Recalculate p_{L} and p_{M} from part e (be careful to find the new normalization constant of the distribution first). That's a big pretty big difference, no? Stay tuned: searching for fossilized microbial mats is a major target for the upcoming Mars 2020 rover.

Note: a more complete Bayesian analysis would have allowed the other parameters like t_{min} and δt_{evolve} to vary, with appropriately chosen prior probabilities. This would be significantly more complex, beyond the scope of the problem set. If you are overly bothered by these limitations, feel free to do the analysis and write a research article!

Problem 2: The Levinthal paradox of protein folding

A chain of amino acids can assume an astronomical number of configurations, of which one corresponds to the folded, biologically functional state of a protein. (In reality the folded state fluctuates to a degree and does not necessarily have to be a unique configuration, but for simplicity let us assume there is only one folded configuration.) The Levinthal paradox poses the following question: how is it possible for the protein to find the correct configuration in a short amount of time if it does a random search among all the possibilities? We will explore and resolve this paradox, retracing the path of Zwanzig, Szabo, and Bagchi in a classic paper [Proc. Natl. Acad. Sci. **89**, 20 (1992); on the course website].

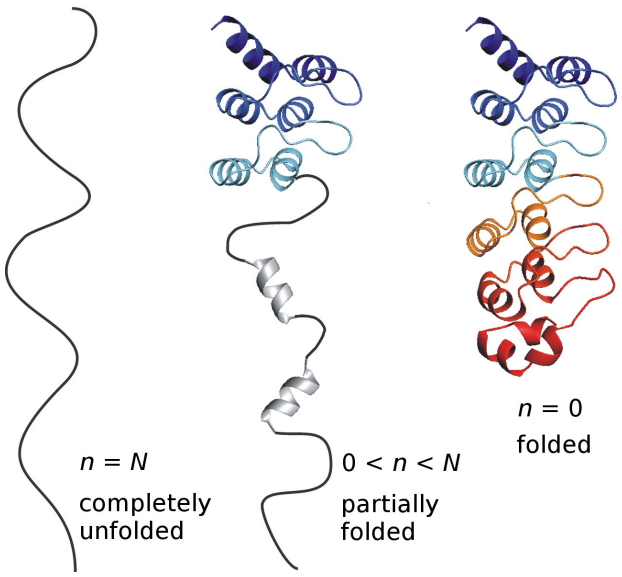

Figure 3: Different stages in the folding of a protein, denoted by the number of amino acids (n out of a total N) that are not in their correct folded structure. Image adapted from: Löw *et al.*, PNAS **105**, 3779 (2008).

a) In the simplest picture, if you have a chain of N amino acids, each amino acid has two options: it could either be in its correct folded structure (C), or it could be in an incorrect form (I). Thus you have 2^N possible configurations. Assume switching states for each amino

acid is a Poisson process (see Problem 1) with the same rate k in either direction, and occurs independently of the others. Show that the mean time between protein configurational changes (i.e. any of the N amino acids switching) is $(Nk)^{-1}$, or equivalently that the mean rate of protein change is Nk. *Hint:* think about the probability that no changes occur in *any* of the N amino acids over some time interval t. The probability that at least one change has happened is 1 minus this expression.

If we accept the above picture, the mean time it would take to find the folded state would be on the order of $2^N(Nk)^{-1}$. For typical protein values like $N = 100$ and $k = 1$ ns^{-1}, this would be around 4×10^{11} years. Yet proteins usually fold on timescales of ~ 1 ms. How is this possible?

b) The resolution is to consider that the search among configurations may be biased: the Poisson rate k_0 from C to I may be different than the reverse rate k_1 from I to C. If an amino acid can lower its energy by assuming a correct form, then it is more likely to stay in that form ($k_0 < k_1$). The size of the ratio $K \equiv k_0/k_1$ determines the degree of bias: $K = 0$ would be the most extreme bias, with no changes from C back to I allowed, and $K = 1$ would be the unbiased case discussed above. Let us label the overall state of the protein by n, the number of amino acids in the incorrect state (so $n = 0$ corresponds to folded; see Fig. 3). Show that the mean rate of going from protein state n to $n - 1$ is nk_1, and that the mean rate of going from n to $n + 1$ is $(N - n)k_0$.

c) The information from part b allows you to construct the full transition matrix $W_{nn'}$ for the protein, and think about the problem from a master equation perspective. (We assume the only allowed transitions are n to $n-1$ and n to $n+1$.) The goal will be to calculate the mean duration it takes to go from some initial state $n > 0$ and reach the folded state $n = 0$ for the first time. Since we only care about how long it takes to reach $n = 0$, and not what happens afterwards, let us employ a trick: we will set the transition rate $W_{10} = 0$, making it impossible to unfold again one you reach the totally folded state. Before we actually write down the master equation, let us sketch out the calculation with this trick. If $p_{mn}(t)$ is the probability of starting from some state n and reaching m in time t, then $U_n(t) = \sum_{m>0} p_{mn}(t)$ is the probability that if the protein started at n, it has remained always unfolded during the time interval t. (If it had reached $m = 0$ at some time before t, it would have stayed there and not be at $m > 0$ at time t.) By construction $U_0(t) = 0$, since if you start folded, you will stay folded. On the other hand, if we start at $n > 0$ at $t = 0$, the initial condition is $U_n(0) = 1$. Since there are only a finite number of states to visit, eventually the protein will fold, so $U_n(\infty) = 0$ for $n > 0$. Argue that the mean time to first reach the folded state starting from $n > 0$ can be written as:

$$\tau_n = -\int_0^\infty dt\, t \frac{dU_n}{dt}(t) = \int_0^\infty dt\, U_n(t) \tag{5}$$

d) We know from class that $p_{mn}(t)$ satisfies both the master equation and the adjoint master equation. The adjoint version has the form:

$$\frac{dp_{mn}}{dt} = \sum_{n'} p_{mn'} \Omega_{n'n}, \tag{6}$$

where $\Omega_{n'n} = W_{n'n}$ if $n' \neq n$, and $\Omega_{nn} = -\sum_{n' \neq n} W_{n'n}$. Using the definitions and results of part

c, show that Eq. (6) implies the following equation:

$$-1 = \sum_{n'} \tau_{n'} \Omega_{n'n}, \quad \text{for all } n > 0 \tag{7}$$

e) By plugging in the matrix components of Ω, show that Eq. (7) implies a recursion relation for the variable $\xi_n \equiv \tau_{n+1} - \tau_n$ of the following form:

$$-1 = g_n \xi_n - r_n \xi_{n-1}, \tag{8}$$

where $g_n = k_0(N-n)$ and $r_n = nk_1$. Note that $\tau_0 = 0$ (the mean folding time starting in the folded state is zero) so $\xi_0 = \tau_1$. Also note that $g_N = 0$, so Eq. (8) for $n = N$ gives $\xi_{N-1} = 1/r_N$.

f) Verify by substitution that the solution of Eq. (8) is:

$$\xi_n = \frac{1}{r_{n+1}} + \sum_{m=n+1}^{N-1} \frac{g_{n+1} \cdots g_m}{r_{n+1} \cdots r_{m+1}}. \tag{9}$$

g) Using Eq. (9) and the definitions of g_n, r_n, show that $\tau_1 = \xi_0$ can be written as:

$$\tau_1 = \frac{1}{Nk_0} \left[(1+K)^N - 1 \right], \tag{10}$$

where $K = k_0/k_1$ is the bias parameter. *Hint:* Remember that the binomial theorem says that:

$$(1+K)^N = \sum_{m=0}^{N} \binom{N}{m} K^m, \qquad \binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{11}$$

h) There are several remarkable things about Eq. (10): it is an *exact* solution for the mean time it takes to go from state 1 to 0. This is how long it takes for the protein to fold if it starts with only a single amino acid in the incorrect form. Even without bias ($K = 1$) you would think that τ_1 would be very fast, because the protein is starting so close to the folded state. However for $K = 1$ and large N you see that $\tau_1 \approx 2^N (Nk_0)^{-1}$, the same expression we saw before. The likelihood of making additional mistakes before fixing the one incorrect amino acid is just so high, that the protein will wander away from its nearly perfect state and spend an eternity finding it again. Adding sufficient bias makes a huge difference: for $k_0 = 1$ ns^{-1} and $N = 100$, find what value of k_1 is required to make $\tau_1 = 1$ ms. You will see that a very modest bias toward the folded state takes τ_1 from longer than the age of the universe down to biological time scales. As we will see later in the course, we can relate k_0/k_1 to energy differences: if one amino acid going from incorrect to correct lowers the protein energy by U, then $k_0/k_1 = \exp(-U/k_BT)$. Find what U is in this case in units of k_BT. Compare this to the energy of one hydrogen bond, $U_{\rm h} \approx 1.5\, k_BT$, a typical biological energy scale. Is the required bias reasonable?

Note: you could in principle calculate τ_n for $n > 1$ using the above results, but it requires more work. As it turns out, τ_1 and τ_n for any $n > 1$ have comparable orders of magnitude for large N, so long as k_0/k_1 is not too small. In other words, the folding time depends only very weakly on how far you start from the folded state. So τ_1 is a reasonable quantity to look at in estimating the general order of magnitude for the folding time.