Second Presentation

Pete Rigas, Lambert Lab

Wednesday, October 9

introduction: presentation overview

- for binding across different proteins, we have been studying how to use the number of visits of a random walk to N as a query for protein occupancy
- in what follows, we will present a formula for transition probabilities of a "fresh" random walk X'_k, which through differing from an "experimental" random walk X_k that is able to achieve binding at N through a sufficient integer number of visits to N, will reveal how probable it is for a certain binding energy level to be achieved
- we will start out by defining a partition function and probability measure, from which we will then study different formulations for obtaining transition probabilities for base pair mismatches, and matches, closer to the binding site N

incorporating a modification to the Ising Hamiltonian into the model

slightly changing terms from the Hamiltonian in the Ising model, to reflect base pair mismatches with opposing charges σ_i, σ_j from the space $\{+1, -1\}$ where $i, j \in \{A, C, T, G\}$, gives an expression of the form

$$\mathcal{H} = \sum_{i \sim j : i,j \in A,C,T,G} -\mathcal{J}_{ij}\sigma_i\sigma_j - \frac{\mathbf{1}_{\sigma_i=\sigma_j=+1}}{|j-N|}\left(\sum_j \sigma_j + \mathbf{E}(\mathbf{1}_{X_n=j}) - \mathbf{E}(\mathbf{1}_{X_n=j+1}) \right),$$

which is useful for constructing a measure μ that will be used to evaluate the change of $\mathbf{E}(\mathbf{1}_{X_n=N}) \geq c$ with $c(x_i, p_i, p_i', N)$, and the coupling constant satisfying $\mathcal{J}_{ij} = 1$ if $\sigma_i = \sigma_j = +1$ and 0 otherwise

defining μ

in addition to a term that measures the extent to which the base pair mismatch, from a lower number of visits of X_n to position $i+1$, it is also important that, from the Hamiltonian and the randomized grand partition $\mathcal{Z}_\delta$, of the form

$$\mathcal{Z}_\delta = 1 + \lambda_p e^{-\beta\epsilon_p} + \lambda_c e^{-\beta\epsilon_{\mathrm{PAM}}} + \lambda_c e^{-\beta\epsilon_c} + \mathcal{X}_{X+\delta} \ ,$$

with $\mathcal{X}$ exponential distributed, we construct a probability measure for different spin configurations of base pairs for different sequences, with the probability measure μ of the form

$$\mu = \frac{e^{-\mathcal{H}}}{\mathcal{Z}_\delta} \ .$$

assigning Ising spins to base pairs

before introducing a formula to calculate how transition probabilities of X_n will depend on base pair mismatches at some point further along in the sequence, our algorithm from the last presentation will

- inspect base pairs of the guide and DNA sequences in pairs of 2, and from such "neighborhoods" of inspection, will assign the spins $\sigma_i = \sigma_j = +1$ for agreeing base pairs (ie C-G, A-T), and disagreeing spins otherwise
- after the spins have been assigned, linearly inspect forward in the sequence past an experimentally observed base pair mismatch, which is protein dependent
- quantify, across all positions in the sequence from the experimentally observed base pair mismatch and N, the expected number of visits of different random walks with transition probabilities that we can calculate from a *transition probability* formula

also, we will

- exploit the mapping $\varphi : V \longrightarrow \{+1, -1\}^V$ for the Ising spin arrangement to calculate sequences of transition probabilities to compare against the steps, and number of visits, for the **experimental** random walk X_k
- formulas for computing the transition probabilities will first be presented for the disagreeing spins, and then for agreeing spins
- dependent on the location of the DNA strand, changing a match to a mismatch allows us to make use of **either** the disagreeing, or agreeing, spin formulas to construct another walk X_k^j, for base disagreement at j, with transition probability p_j that has a sharper rate of decay than $p')_j$ had the bases been in agreement

the Fourier Transform: towards an inspiration for a formula of the transition probabilities for $j > i$

Goal: model changes in momentum, for a sufficient class of proteins, with base pair mismatches to obtain the energy due to base pair mismatches later in the DNA sequence, sufficiently far from binding position N, but beyond those experimentally observed at some $i < N$

The space

$$\{e_{b_i} + \int_{x_i}^{x_{i+1}} f(x)e^{-2\pi x}\mathrm{d}x, 1 \leq x_i < x_{i+1} \leq N\} \ ,$$

with $f(x)$ smooth, is composed of Fourier transforms. Because we do not expect that there are any periodic effects in the binding phenomena, we can look to define ...

remarks

... a formula for the transition probabilities from an expression similar to a contribution from the Fourier transform to the binding energy mismatch from the Boltzmann factor at i, because heuristically,

- the FT, given protein tolerance to mismatches, can be of use in generating more base pair mismatches, and the associated binding energy costs with such mismatches within closer distance to N
- moreover, it is possible to easily quantify a decrease in occupancy of some protein at N by penalizing transition probabilities at sides of the DNA sequence along which there are base pair mismatches

but first: defining the random walk entropy

However, before providing a formula that is dependent on the Ising spins that we have assigned for agreement amongst DNA base pairs, we define the following

Definition: The random walk entropy $\mathcal{E}$ as

$$\mathcal{E} = \sum_{k=1}^{n} \left(\mathbf{E}(1_{X_k = N}) - \mathbf{E}(1_{X'_k = N}) \right) ,$$

for $n > N$

- with $\mathcal{E}$, we are interested in measuring a different in the number of visits to N, which can be evaluated against μ, which will be shown later

clarification of notation

- for the notation on the previous slide, we have that X_k and X'_k are random walks whose transition probabilities differ at arbitrary position j of mismatch
- past this additional base of mismatch at j, all of the transition probabilities, and therefore the number of visits to N, will decrease in comparison to the visits that X_k has at N

decomposing $\mathcal{E}$ for base pair mismatches at i

Across different positions of each sequence, altering transition probabilities at arbitrary positions gives a decomposition of $\mathcal{E}$ as,

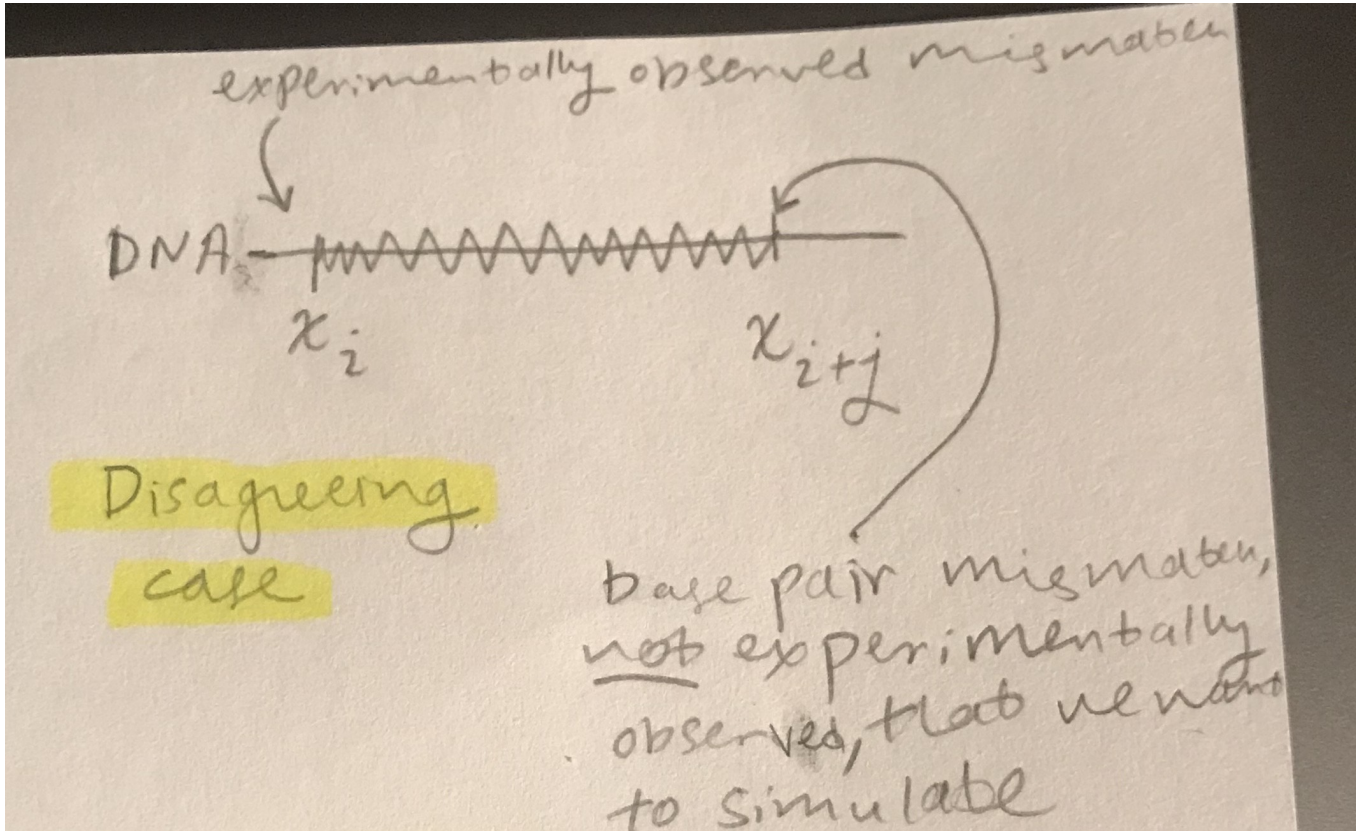
$$\sum_{\gamma \cap e_i} \mathbf{E}(\mathbf{1}_{X_k=N}) - \underline{\mathbf{E}(\mathbf{1}_{X_i=N})} - \sum_{\gamma' \cap e_i'} \mathbf{E}(\mathbf{1}_{X_k'=N}) - \underline{\mathbf{E}(\mathbf{1}_{X_i'=N})} \ ,$$

where γ, γ' are piecewise defined paths defined along the DNA sequences, and $\gamma \cap e_i$, $\gamma' \cap e_i'$ are paths constructed along the same sequences with exception to the i^{th} edge, where another random walk X_k' is sampled with transition probability $p_j' < p_j$. Specifically, the "sequence configurations" are constructed as follows,

- if base pair matches "persist" in pairs of 2, ie $\mathbf{1}_{\{\sigma_i = \sigma_j, \sigma_{i+1} = \sigma_{j+1}\}} = 1$, then an edge is filled between i and $i+1$
- if an edge along γ or γ' is empty, the transition probability at $j+1$ will be calculated by associating a **higher** energetic cost for the next position along the sequence

illustration for first formula

- we integrate over the sequence from i to $i + j$
- we subtract the exponential factor that we compute from the transition probability at i, which gives the transition probability at $i + j$



defining transition probabilities for disagreeing spins

Claim: From base pair mismatches experimentally observed at i, we can define energies associated with base pair mismatches at neighboring positions $i + j$ from,

$$e^{-}_{b_{i+j}} = e^{-}_{b_i} - \int_{x_i}^{x_{i+j}} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} \frac{e^{-(x-N)^2} w(x) \mathrm{d}x}{E_N + \mathcal{X}_{x_i+j}} \ ,$$

where

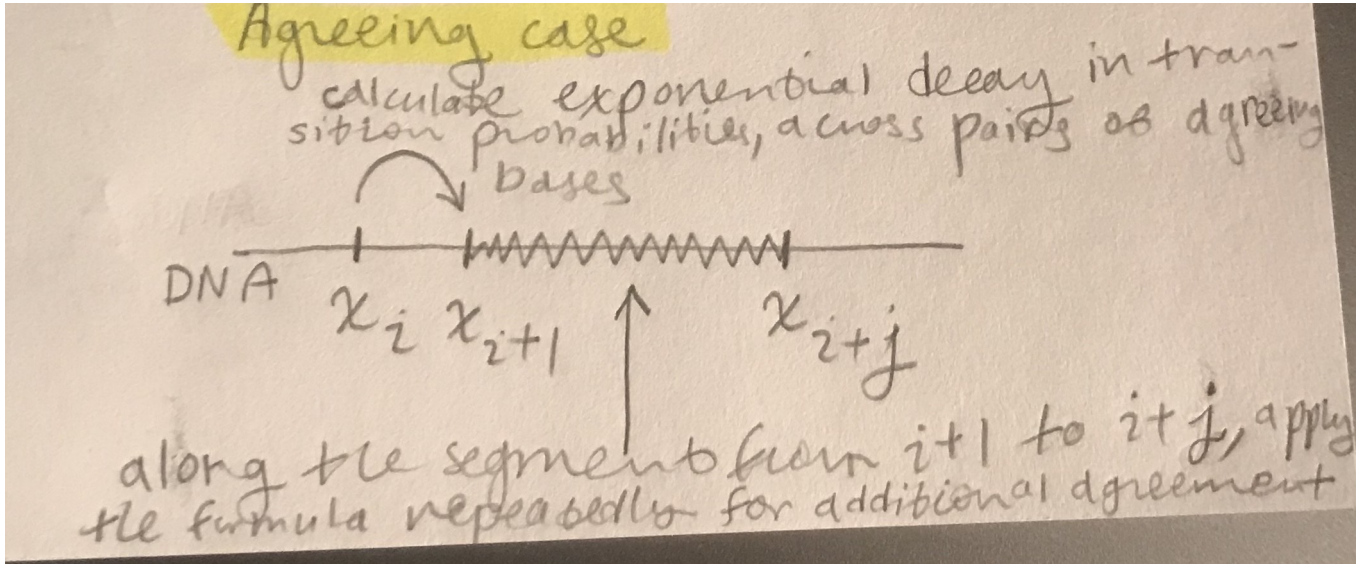
- under sufficiently small perturbations, base pair mismatches that we encounter within closer proximity to the binding site N give a smaller fraction of time, and therefore occupancy, for any protein binding at N
- the indicator results in a non-zero contribution from the Hamiltonian...

continued...

- ... if Ising-*like* spins that we define for the DNA base pairs are in agreement
- we can easily interpret the contributions from base pair matches and mismatches, by first checking whether the indicator from Ising spins that we have assigned to DNA base pairs is satisfied, from which we would expect there to be a non-zero contribution from the integral as it is computed over a portion of the DNA sequence before N, by quantifying the number of visits of different random walks X_n and X'_n, where X_n and X'_n differ in transition probabilities at position j of base pair mismatch, with $j > i$
- the "damping" $w(x)$ contains information as to how significant a base pair mismatch is as $x_i \longrightarrow N$, with the experimentally observed binding energy E_N at N; in terms predictions, we would expect that the binding energy at N, if we were to energeticaly perturb gRNA and DNA sequences, per base pair mismatch, would energetically make it more difficult for binding to occur at N

illustration for the second formula

- we inspect whether the bases are in agreement at position i
- assign an transition probability at $i+1$, and all other positions leading up to $i+j$, by applying the formula given on the next slide, given that all of the bases between $i+1$ and $i+j$ are in agreement



defining transition probabilities for agreeing spins

similarly, because the transition probabilities for each subsequent position closer to N are strictly decreasing,

$$e^{+}_{b_{i+1}} = e^{+}_{b_i} - \mathbf{1}_{\{\sigma_i=\sigma_j=+1\}} \frac{e^{-(x_i-N)^2} w(x_i)}{E_N} \ ,$$

with differences including that

- we calculate the transition probability for $i+1$ given base agreement at i, to which we assign a decreasing transition probability that is **greater** than the transition probability that would have been assigned to $i+1$ if the spins at $i+1$ were in disagreement

outline of the approach

For our strategy,

- we begin by ordering binding energies based on base pair mismatches, that is, determining, for the specific parameters of binding events of each protein, the maximum number of base pairs, in addition to the location of base pair mismatches, that still permit binding to occur as well as the associated binding energy
- from sequence dependent binding energies, we proceed by to determine "energetically unfavorable" sequences, which in the case of different proteins is dependent on the protein tolerance to base pair mismatches before N
- future slides will describe how different options for the damping $w(x)$ influence our interpretation of how a suitable binding level energy is achieved at N

note

So far, what has been shown can be applied towards adding in more base pair mismatches, from which the difference in number of visits to N can be determined. Observe that,

- the disagreeing base formula has an integral to reflect the long range distance changes in energy from disagreeing pairs
- the agreeing base formula decreases the transition probability p_{i+1} from p_i but not as sharply

From this expression of the energy base mismatch at $i+1$, we know that differences in transition probabilities of X_n and X'_n at j can be determined so as to be proportional to the magnitude of the energetic contribution of a base pair mismatch that is experimentally observed at i. Furthermore, with a suitable basis of exponentials, it is possible to choose, depending on the parameters N and "base pair spins" σ_i and σ_j assigned for base pairs at each position of the sequence, transition probabilities p_{i+1} and p'_{i+1}, respectively corresponding to the transition probabilities of x_n and X'_n, so that the conditional probability, taken with respect to the suitably defined measure μ, which is of the form

$$\mu\left(\left|\mathbf{E}(\mathbf{1}_{X_n=N}) - \mathbf{E}(\mathbf{1}_{X'_n=N})\right| < n\middle|p_1, p'_1 > 0)\right) \ ,$$

... with $n \in \mathbf{N}$. The event that is being evaluated against the measure μ reflects the extent to which either one of the random walks, amongst the distribution of all random walks with transition probabilities "modified" at some position i of the sequence with $i < N$, result in an integer different of visits at n, that is sufficient to prevent sufficient occupancy from occurring.

Altering either a single transition probability, or multiple transition probabilities from single, or multiple, base pair mismatches that is,are, sufficiently far from N demonstrates properties of a phase transition that one can study, from the perspective that the expected number of visits at $i + 1$, given the earlier base pair mismatch in the DNA sequence at i, will smoothly vary with respect to the transition probability p'_{i+1} of X'_n.

proceeding with a computation

To determine admissible upper bounds for the largest magnitude of change in transition probabilities from i to $i + 1$, we return to a previous expression, in which a possible upper bound, albeit not the most tight, can be calculated by taking a summation over nonempty edges of sequence configurations that we introduced earlier, along which transition probabilities do not exhibit as sharp an exponential decay in comparison to empty edges of the sequence configurations, for which neighboring transition probabilities between empty and nonempty edges, if in parallel and within the first 6 positions of the sequence, make binding at any later position before N impossible.

alternative definitions for the damping $w(x_i)$

nevertheless, defining different exponential "penalties" for base pair mismatches is achievable in several ways

- *exponential:* define a basis of exponentials, proportional to base pair mismatches, that are energetically proportional to compatible, or incompatible base spin pairs
- *polynomial:* reflect energetically unfavorable base pair sates with higher degree polynomials, upon choosing a basis of coefficients $\{c_i(x, N)\}_i$, so that $\sum_{i \in \mathbf{Z}: N - x_i \geq 0} c_i x^i$
- for the polynomial case, once such a suitable polynomial is determined, *ii* can reduce to an upper bound in *i* by choosing suitable $\alpha(x)$ so that $\exp(-\alpha(x)N) = \sum_i c_x N^i$, and also that $\exp(-\alpha(x, N)N) \geq \sum_i c_x N^i \quad \forall i$, or $|\exp(-\alpha(x, N)N) - \sum_i c_x N^i| < \epsilon$ for suitable ϵ that is not too small

first case of damping: polynomial $w(x_j)$

setting $x_f \equiv x_j$, in addition to incorporating a polynomial weight for base pair mismatches gives

$$\int_{x_i}^{x_f} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} \frac{e^{-(x_i-N)}}{E_N + \mathcal{X}_{x_f}} x_j^{N-j} \mathrm{d}x_j \ ,$$

which, because $N - i \in \mathbf{Z}$, integrating by parts a fixed number of times $\leq N - i$ times gives, taking a suitable energy $E \geq E_N + \mathcal{X}_{x_j}$, an expression that can be simplified, which is of the form

$$(N(x_j)^{N-j} e^{-(x_j-N)})|_{x_i}^{x_f} \mathbf{1}_{\{\sigma_i=\sigma_j=+1\}}$$
$$- \int_{x_i}^{x_f} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} e^{-(x_j-N)} \frac{-(x_j - N)(x_j)^{N-i-1}}{E} \mathrm{d}x_j \leq$$
$$\int_{x_i}^{x_f} \mathbf{1}_{\{\sigma_i=\sigma_j=+1\}} \frac{e^{-(x_j-N)}}{E_N + \mathcal{X}_{x_j}} (x_j)^{N-i} \mathrm{d}x_j \ .$$

case one

The terms from polynomial weights that we have assigned in our transition probability formula, for disagreeing Ising spins, do not present contributions to the binding energy at N. Evaluating other terms from the integral above represents the energy, again for disagreeing spins, that have incompatible base pairs. Altogether, obtaining a suitable lower bound for

$$\int_{x_i}^{x_{i+j}} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} \frac{e^{-(x_j-N)}}{E_N + \mathcal{X}_{x_{i+j}}} x_j^{N-j} \mathrm{d}x_j \ ,$$

can be achieved by maximizing, over all base pair mismatches, and with E from the previous slide, the expression

$$\int_{x_i}^{x_{i+j}} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} \frac{\max_j \prod_{\text{mismatch}} e^{-x_j-N} x_j^{N-j}}{E} \ .$$

case one

That is, for base pair disagreements before N, assigning polynomial weights to $w(x_j)$ helps in forming interpretations of the change in binding energy, but may not be completely realistic in capturing the magnitude of base pair disagreements closer to N. We take the maximum of the product of base pair mismatches to obtain an upper bound.

second case of damping: exponential $w(x_j)$

this time, setting and incorporating an exponential weight for base pair mismatches gives

$$\int_{x_i}^{x_{i+1}} \mathbf{1}_{\{\sigma_i=-1,\sigma_j=+1\}} \frac{e^{-(x_i-N)}}{E_N + \mathcal{X}_{x_{i+1}}} e^{\alpha} \mathrm{d}x_i \text{ , for } \alpha \in \mathbf{R}$$

which can be interpreted as

- more realistic for base mismatches closer to N, because the energy cost of a base mismatch does not linearly vary as $x \to N$, higher order terms are necessarily involved

case two

With exponential weights, the magnitude of exponential decay α could be adjusted to characteristic binding energies of different proteins, so that mismatches closer to N are not energetically undermined.

interpreting the transition probabilities formula with an inequality for the expected number of visits of the random walk

the quotient

$$\max_A \frac{\mathbf{E}(\mathbf{1}_{X_k^\alpha = N}) - \mathbf{E}(\mathbf{1}_{X_n^k = N})}{|x^\alpha - x^k|} \ ,$$

with $A = \{\alpha \neq k : a, k \in \mathbf{Z}, N - j \leq \alpha, k \leq j\}$, gives an expression for the maximum discrepancy in the number of visits to N between the random walks X_k^α, X_n^k

flavors of the phase transition

furthermore, the binding phase transition, as a phase transition, can be interpreted as

- does there exist suitable weights $w(x_j)$, as given in the transition formula, so that for parameters β past a threshold β_c, for the number of visits to N, would there not be a sufficient occupancy for binding?
- does there exist a critical transition probability p_c, such that an exponential decrease to p_c will prevent binding at N, even if all bases after the base mismatch with transition p_c are all matches?

questions, future thoughts

it is important that we figure out

- what are the appropriate range of parameters, and average binding energies at a fixed site of the sequence for different proteins
- determine the appropriate parameter from which the exponential $\mathcal{X}_{x_j}$ should be drawn

conclusions

the approach given here

- is more self contained than the arguments in the first presentation, and gives complete steps on how to process binding energy data
- able to, from the transition probability formulas, correlate changes in binding energy to the difference in number of visits of random walks
- can be generalized to base pair mismatches that are not directly next to each other