bounds on N_{mis} parameter from measure inequalities

Pete Rigas, Lambert Lab

Wednesday, July 15

Review I

Recall,

$$\mu(N, N_{\rm mis}, |X_{\rm mis}|, |{\rm mis}|, \lambda_{\rm mis}) = \frac{\prod_{i\sim j} \exp\left(- \frac{w_i}{(|\{{\rm mat}\}|+2)^3} \mathcal{J}_{ij}\sigma_i\sigma_j \right) \cdots}{1 + \lambda_c e^{-\beta\epsilon_{\rm PAM}} + \cdots}$$
$$\frac{\cdots \prod_{i\not\sim j} \exp\left(- \frac{w_i}{(|\{{\rm mis}\}|+2)^3} \mathcal{J}_{ij}(1-\sigma_i\sigma_j) \right)}{\cdots (N - X_{\rm mis})^3 \sum_{\rm mis} \lambda_{\rm mis} e^{-\lambda_{\rm mis} X_{\rm mis} + \ln(N_{\rm mis})}} .$$

Review II: manipulating the inequality

Rearranging, and comparing, the values of the measure at neighboring positions along any sequence gives,

$$\frac{\prod_{i\in \mathrm{Mat}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}}{1+\lambda_p e^{-\beta\epsilon_p}+\frac{|\mathrm{Mat}_i|}{N}\lambda_c e^{-\beta\epsilon_c}+(N-X_{\mathrm{mis}})^3\sum_{\mathrm{mis}}\cdots}$$

$$< \frac{\prod_{i\in \mathrm{Mat}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}}{1+\lambda_p e^{-\beta\epsilon_p}+\frac{|\mathrm{Mat}_{i+1}|}{N}\lambda_c e^{-\beta\epsilon_c}+(N-X_{\mathrm{mis}})^3\sum_{\mathrm{mis}}\cdots}$$

$$\Updownarrow$$

$$\frac{\prod_{i\in \mathrm{Mat}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}}{\prod_{i\in \mathrm{Mat}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}} = e^{-\frac{w_{i+1}}{\mathcal{N}}\mathcal{J}_{i+1,j+1}(1-\sigma_{i+1,j+1})}$$

Review III: cancellation from terms of matches and mismatches in neighboring positions

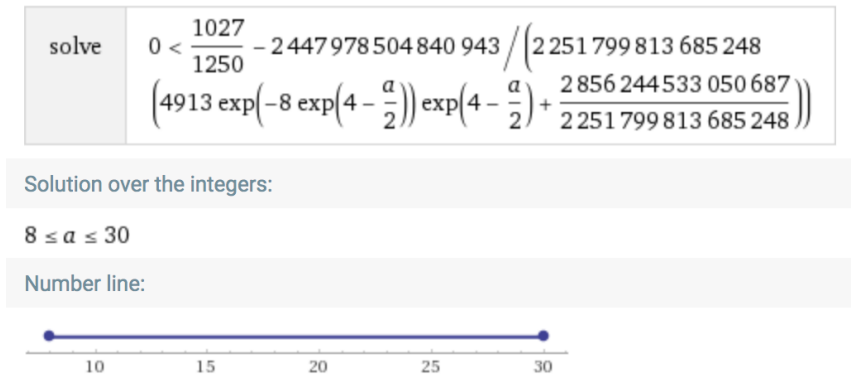
Strategy: Make use of the values that we have previously computed for the measure vectors to determine the mismatch parameter for which the sequence of transition probabilities will be strictly decreasing

$$\Rightarrow \mu_i^N - \frac{\prod_{i\in \mathrm{Mat}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}}{1+\frac{|\mathrm{Mat}_i|}{N}e^{-\beta\epsilon_c}\lambda_c + (N-X_{\mathrm{mis}})^3 \sum_{\mathrm{mis}} \lambda_{\mathrm{mis}} e^{-\lambda_{\mathrm{mis}}X_{\mathrm{mis}}+\ln(N_{\mathrm{mis}})}} .$$

We observe that the solution space for the inequalities given on this slide and on the previous slide are equivalent.

mismatch parameter bounds

I am still trouble shooting why the Matlab solve function, which does not return the solution set to the inequality given on the previous slide. Inputting the same exact expression from the Matlab command line into Wolfram Alpha gives,

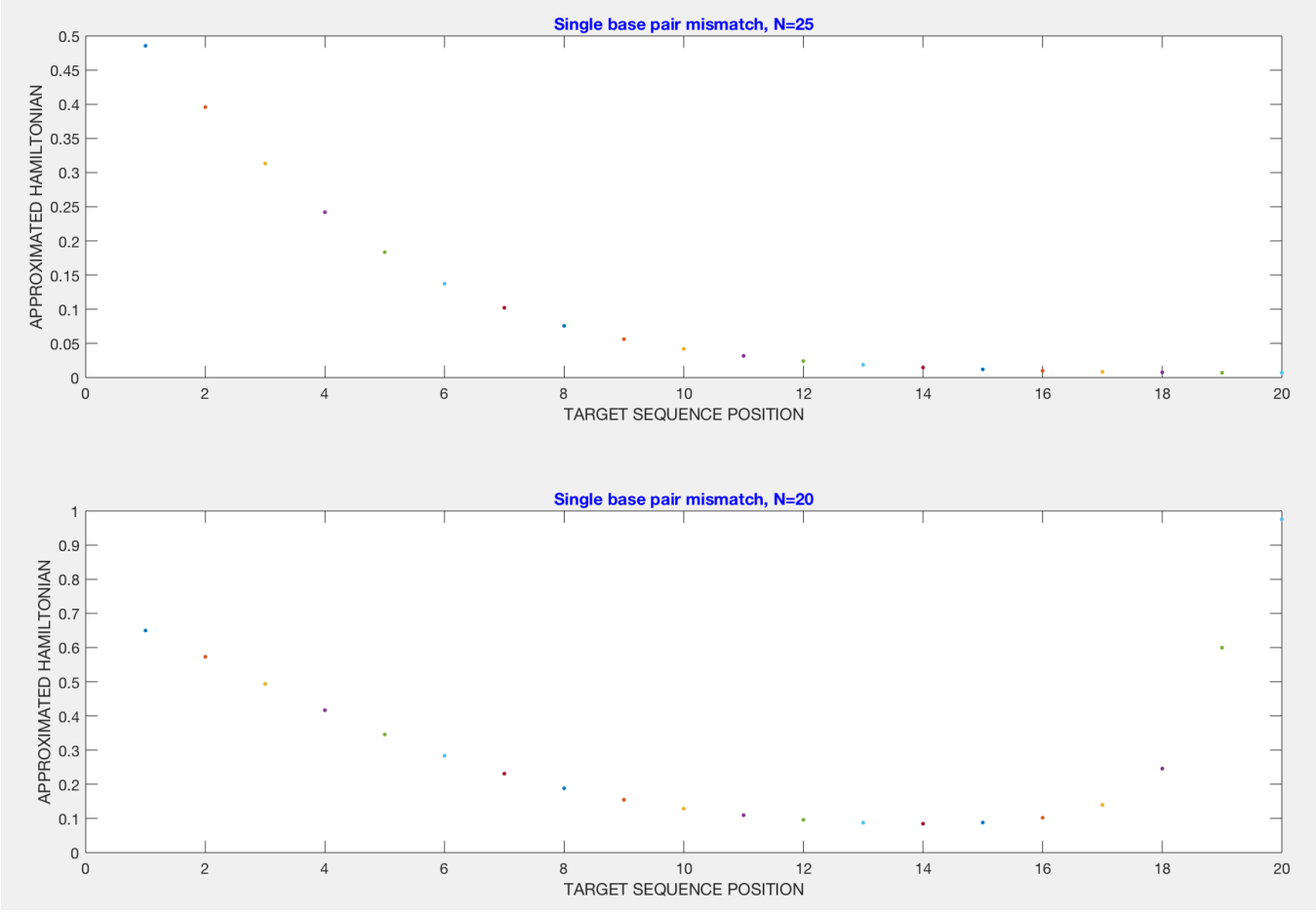


while by comparison Matlab gives,

```
eqn =

0 < 1027/1250 - 2447978504840943/(2251799813685248*(4913*exp(-8*exp(4 - a/2))*exp(4 - a/2) + 2856244533050687/2251799813685248))

>> S = solve(eqn,a)
Warning: Unable to find explicit solution. For options, see help.
> In solve (line 317)

S =

Empty sym: 0-by-1
```

connection with plots from April 8 presentation

such a choice of parameters from the ineqaulity that we have enforced is consistent with previous plots that I have discussed in the past (to be specific, N denotes the mismatch parameter)



alternative approach for ranges of Mis parameter

alternatively, we could examine the complement of the interval that we obtained from the previous size, in order to study the quantity which involves the difference between the maximum and minimum admissible Mis parameters (precisely given by the L1 norm of the line over which there are no real solutions to the inequality constraint)

solve	$0 > \frac{1027}{1250} - 2\,447\,978\,504\,840\,943 \Big/ \left(2\,251\,799\,813\,685\,248 \left(4913 \exp\left(-8 \exp\left(4 - \frac{a}{2}\right)\right) \exp\left(4 - \frac{a}{2}\right) + \frac{2\,856\,244\,533\,050\,687}{2\,251\,799\,813\,685\,248}\right)\right)$

Solution over the integers:

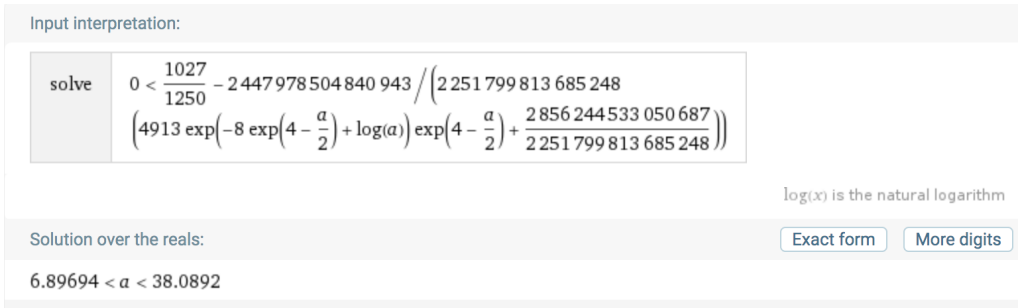
$a \leq 7$

$a \geq 31$

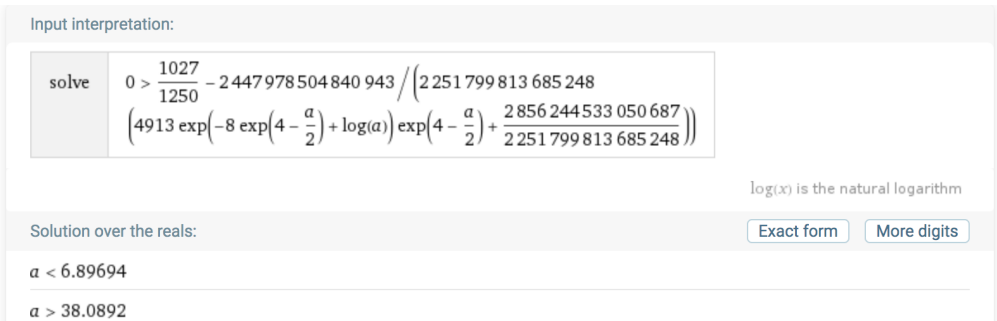
Number line:

factoring in the natural logarithm of the Mis parameter

modifying the formula to obtain the precise closed form that we have discussed over the past several weeks gives upper and lower bounds on the Mis parameter that are numerically comparable to those given in the previous slides, with intervals of the form



or equivalently,



potential issues with the logarithm term

Input interpretation:

solve $0 < \frac{1027}{1250} - 2\,447\,978\,504\,840\,943 \Big/ \left(2\,251\,799\,813\,685\,248 \left(4913 \exp\left(-8 \exp\left(4 - \frac{a}{2}\right) + \log(a)\right) \exp\left(4 - \frac{a}{2}\right) + \frac{16\,635\,416\,881\,964\,253}{2\,251\,799\,813\,685\,248}\right)\right)$

$\log(x)$ is the natural logarithm

Result:

(All values of a are solutions)

I am thinking about how we would be able to still incorporate the log term, without being able to encounter the situation above, in which the inequality is satisfied for any value of the parameter

quantifiable change in the magnitude of transition probabilities

without ln, we have poor modulation of expected binding energy

λ_{mis}	resulting transition probability
50	$\approx 0.120313\cdots$
200	$\approx 0.14715462\cdots$
$10,000$	$\approx 0.14715462\cdots$

remarks

numerically we observe a change in the magnitude of the transition probability past a mismatch in the sequence when the power of the exponential factor in the λ_{mis} parameter vanishes (to be made more precise for the numerical choices that I used above in **2** slides)

numerically motivating the other mismatch parameter

with ln term, we have much stronger modulation of binding energy, holding one mismatch parameter fixed and one constant, while also varying both mismatch parameters simultaneously

$\lambda_{\text{mis}}, N_{\text{mis}}$	transition probability
1 , 1000	$\approx 0.0006566660595\cdots$
1 , $50,000$	$\approx 0.00001319100776\cdots$
1 , $100,000$	$\approx 6.595799\cdots \times 10^{-6}$

remarks

- due to an additional multiplicative factor, in addition to the power λ_{mis} of the exponential, the factor $\lambda_{\text{mis}} N_{\text{mis}} \exp(-\lambda_{\text{mis}} X_{\text{mis}}) \equiv \lambda_{\text{mis}} \exp(-\lambda_{\text{mis}} \lambda_{\text{mis}} + \ln(N_{\text{mis}}))$ is the other parameter that is included
- including N_{mis} is preferable because of the inverse proportionality between the second Mis parameter and the magnitude of the probability, which monotonically decreases

a further on the parameter λ_{mis} of the exponential random variable in the partition function

in the numerical experiments from the past 2 slides, we set the multiplicative λ_{mis} to be of the form,

$$\lambda_{\text{mis}} \equiv e^{50-\frac{100}{2}} \ ,$$

which is responsible for the change in the transition probabilities **without** the ln of the N_{mis} term, because the power of the exponent for a particular mismatch vanishes

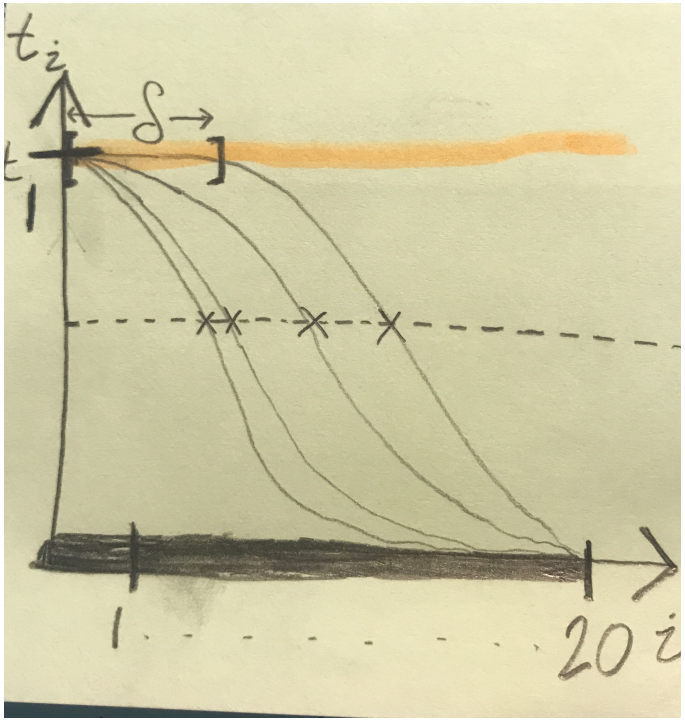
updates in the implementation from plots given in previous presentation

Altogether, we observe that to ensure that the visits of the walk will montonically decrease given a maximum of 3 base pair mismatches in Fn Cas 12a, in addition to other Cas proteins, by modeling the dependence of the mismatch tolerance of the protein on the power of the exponential to which the transition probabilities decay. Specifically, changes to the procedure involve:

- introducing new requirements to the computation of all transition probabilities in a sequence **after** the occurrence of a base pair mismatch, in which the transition probability at the position of mismatch is the maximum probability at which the random walk can traverse all remaining bases in the sequence,
- recognizing how to implement such a change so that transition probabilities later in the sequence do not exceed the specific transition probability at a position of mismatch (see later plots),
- in addition to reliably implementing the condition on the transition probabilities at multiple instances for all mismatches

qualitatively describing how the numerical condition on the decay of transition probabilities would impact the visit distribution for different Cas proteins

- comparison of the visit distributions to the position of binding, collectively which represent the binding affinity and its rate of change of the binding energy with respect to molecular concentrations in a medium that are captured in the Hill function



extended remarks

several remarks are in order:

- in the plot given in the previous slide, the orange line represents the visit d given a perfectly matching sequence, while the solid **black** line denotes the number of visits for a sequence that has a base pair mismatch in the first position,
- the dashed line at the middle of the heuristic plot demonstrates the critical transition probability, $\frac{1}{2}$, at which the probability that a pseudo randomly generated number (which is precisely the transition probability at a particular base of the sequence) will exceed the threshold at which the random walk will either visit the next base of the sequence more **or** less times than the total number of visits that it encountered in the previous position

correcting misrepresentations of the binding energy in the model

to address the possibility of the random walk visit distribution having 0 visits for each position of the sequence if the first collection of spins σ_i, σ_j are not in alignment, it is possible to:

- introduce a nonzero penalization to the initial transition probability of the random walk at the first position of the sequence so that the algorithm does not return a null sequence of transition probabilities (ie a sequence of probabilities all of which vanish),
- after which we perform the usual steps of the algorithm, in which the exponential decrease of the transition probabilities across subsequent base pairs of inspection is modified so that the transition probabilities do not exceed the critical threshold

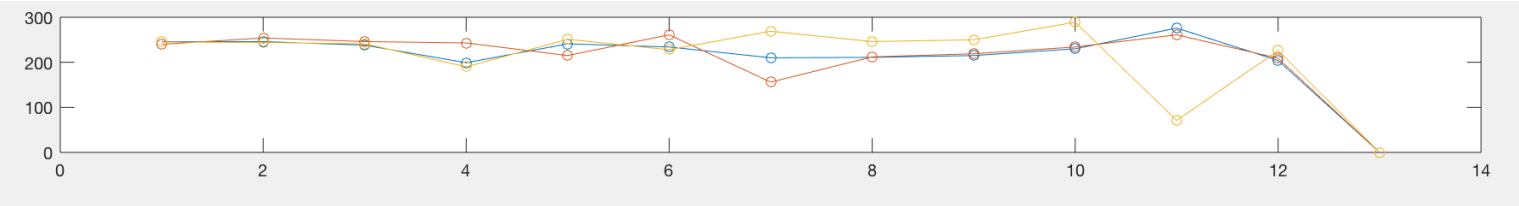
plot observations

we observe that about the dotted line, there is an equal probability of the transition probability lying below or above the line, demonstrating that this critical value, and the corresponding set of parameters for which the probabilities decay below the threshold, will satisfy:

- a condition stipulated by the inequalities given in the first few slides,
- additional conditions stipulated by the observation that we cannot have more visits to positions of the sequence given the number of visits of the walk at the most recent base pair mismatch along the binding sequence,
- from which we can infer that in sequences with multiple base pair mismatches, we would have that the transition probability at **each** base pair mismatch dictates the maximum probability of the walk for all remaining base pairs of the sequences

April 29 recap: how these observations and requirements on the numerical behavior of the transition probabilities compare with previous descriptions of the random walk visit distributions

from previous issues that I have remarked on, applying the observations mentioned in the previous slides would correct the remaining visits of the visit distributions of the walk past base pair mismatches that have already been simulated



with adjustments to the sequence of transitions beyond base pair mismatches, we can more clearly observe and tune the number of visits to the position of binding

establishing the groundwork for our generalization

we now focus on the orange term below,

$$\frac{\prod_{i\in \mathrm{Mat}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \prod_{i\in \mathrm{Mis}_i} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)}}{1+\frac{|\mathrm{Mat}_i|}{N}e^{-\beta\epsilon_c}\lambda_c+\sum_{\mathrm{mis}}(N-X_{\mathrm{mis}})^3\lambda_{\mathrm{mis}}e^{-\lambda_{\mathrm{mis}}X_{\mathrm{mis}}+\ln(N_{\mathrm{mis}})}}.,$$

in which we will obtain recursive relationships depending on the number of mismatches, which directly corresponds with the number of terms in the summation

formalism

altogether, given the restrictions on $N_{\rm mis}$, in addition to bounds that we have presented at the beginning of the slides for the inequality relationship that has been described for only **one** mismatch parameter $\lambda_{\rm mis}$ when we neglect $N_{\rm mis}$ as a multiplicative factor in the last exponential term that we have appended to the partition function, a more general version of the inequality relationship takes the form,

$$1+\frac{|\mathrm{Mat}_{i+1}|}{N}e^{-\beta\epsilon_c}\lambda_c+\sum_{\mathrm{mis}\in\mathrm{Mis}_{i+1}}(N-X_{\rm mis})^3\lambda_{\rm mis}e^{-\lambda_{\rm mis}X_{\rm mis}+\ln(N_{\rm mis})}$$
$$<\prod_{i\in\mathrm{Mis}_i\cap\mathrm{Mis}_{i+1}}e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j}\left(1+\frac{|\mathrm{Mat}_i|}{N}e^{-\beta\epsilon_c}\lambda_c\right.$$
$$\left.+\sum_{\mathrm{mis}\in\mathrm{Mis}_i}(N-X_{\rm mis})^3\lambda_{\rm mis}e^{-\lambda_{\rm mis}X_{\rm mis}+\ln(N_{\rm mis})}\right),$$

continued...

$$\Updownarrow$$

$$1+\left(\frac{|\mathrm{Mat}_{i+1}|}{N}-\frac{|\mathrm{Mat}_{i}|}{N}\prod_{i\in \mathrm{Mis}_i\cap \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j}\right)e^{-\beta\epsilon_c}\lambda_c + \sum_{\mathrm{mis}\in \mathrm{Mis}_{i+1}} (N-X_{\mathrm{mis}})^3\lambda_{\mathrm{mis}}e^{-\lambda_{\mathrm{mis}}X_{\mathrm{mis}}+\ln(N_{\mathrm{mis}})}$$

$$-\prod_{i\in \mathrm{Mis}_i\cap \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j}\sum_{\mathrm{mis}\in \mathrm{Mis}_i} (N-X_{\mathrm{mis}})^3\lambda_{\mathrm{mis}}e^{-\lambda_{\mathrm{mis}}X_{\mathrm{mis}}+\ln(N_{\mathrm{mis}})} < \prod_{i\in \mathrm{Mis}_i\cap \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j},$$

which can be interpreted as a locus in the mismatch parameter plane, with the task being to determine admissible pairs that lie within the locus or possibly on its boundary

breaking down the inequality amongst different sequence configurations

to this end, the inequality from the previous slide can be further analyzed by determining in which cases particular terms of interest would vanish from the expression, in which we are left with precisely 3 possibilities:

- one base pair mismatch, then a base pair match,
- one base pair mismatches, followed by another base pair mismatch,
- a final scenario building upon the one above, in which the binding sequence experiences a total of 3 base pair mismatches

extending the formalism to multiple base pair mismatches

along similar lines, the inequality relationship between distinct transition probabilities along unique base pairs of the sequence, we expand mismatch terms due to the λ_{mis} and N_{mis} in the summation on the LHS of the previous inequality, which gives the exponential,

$$\sum_{\text{mis}\in\text{Mis}_{i+1}} (N - X_{\text{mis}})^3 \lambda_{\text{mis}} e^{-\lambda_{\text{mis}} X_{\text{mis}} + \ln(N_{\text{mis}})}$$

$$- \prod_{i\in\text{Mis}_i\cap\text{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j} \sum_{\text{mis}\in\text{Mis}_i} (N - X_{\text{mis}})^3 \lambda_{\text{mis}} e^{-\lambda_{\text{mis}} X_{\text{mis}} + \ln(N_{\text{mis}})}$$

$$= (N - X_{\text{mis}_{i+1}})^3 \lambda_{\text{mis}} e^{-\lambda_{\text{mis}} X_{\text{mis}} + \ln(N_{\text{mis}})}$$

$$+ \sum_{\text{mis}\in\text{Mis}_i} \left(1 - \prod_{i\in\text{Mis}_i\cap\text{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}}\mathcal{J}_{ij}\sigma_i\sigma_j}\right)$$

$$(N - X_{\text{mis}})^3 \lambda_{\text{mis}} e^{-\lambda_{\text{mis}} X_{\text{mis}} + \ln(N_{\text{mis}})} \ ,$$

continued...

the relationship for more than one base pair mismatch that the protein encounters in the sequence is therefore of the form,

$$(N - X_{\mathrm{mis}_{i+1}})^3 \lambda_{\mathrm{mis}} e^{-\lambda_{\mathrm{mis}} X_{\mathrm{mis}} + \ln(N_{\mathrm{mis}})} + \sum_{\mathrm{mis} \in \mathrm{Mis}_i} \left(1 - \prod_{i \in \mathrm{Mis}_i \cap \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}} \mathcal{J}_{ij} \sigma_i \sigma_j} \right)$$

$$(N - X_{\mathrm{mis}})^3 \lambda_{\mathrm{mis}} e^{-\lambda_{\mathrm{mis}} X_{\mathrm{mis}} + \ln(N_{\mathrm{mis}})} < \prod_{i \in \mathrm{Mis}_i \cap \mathrm{Mis}_{i+1}} e^{-\frac{w_i}{\mathcal{N}} \mathcal{J}_{ij} \sigma_i \sigma_j} \ ,$$

for any number of base pair mismatches, in which incorporate the parameters that have been previously defined, in order to simulate more complicated binding scenarios with multiple base pair mismatches in succession, which can then be reduced to determining whether solutions to the inequality exist from the resulting exponential terms

upshot: controlling the magnitude of transition probabilities up to the binding position

in the near future, it would be great to have:

- more automation steps to broaden the approach to thousands of sequences,
- an idea of unique, inherent processes of binding to each protein so that I can be aware of how to appropriately adjust for terms in the partition function normalization,
- and finally, sets of visit distributions of the random walk corresponding to each binding sequence

automation steps

upcoming! (may try out the solve function in Python to see if things are easier to implement)