further developments

Pete Rigas, Lambert Lab

Wednesday, May 5

this week's progress

less substantial change: adding "boundary" term to the Hamiltonian, as an analogy to the external magnetic field $h \sum_{i \in \partial\lambda} \sigma_i$ in the Ising model

more substantial change: examining the distribution of growth rates to determine whether the growth rate of the protein as it undergoes binding can be reflected in the partition function as we calculate the transition probability at the position of binding

less substantial change

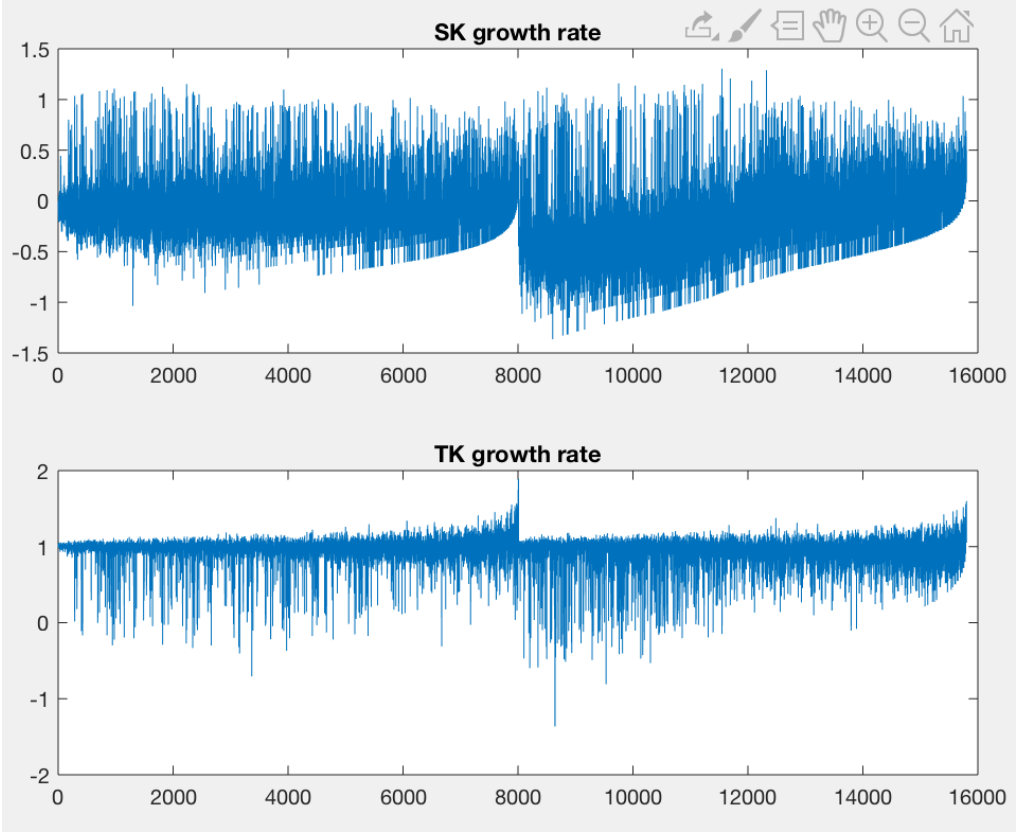
with an additional term to the Hamiltonian so that the transition probability at the position of binding does not vanish altogether, the general formula is of the form,

$$\mu^i = \frac{\prod_{i\sim j}\exp\left(-\frac{w(x_i)}{(|\{\text{mat}\}|+2)^3}(\mathcal{J}_{ij}\sigma_i\sigma_j + N\sigma_i)\right)\cdots}{1+\mathbf{1}_{|\text{mat}|=\text{mat}}\left(\frac{i}{N_{\text{bind}}}\right)\lambda_c e^{-\beta\epsilon_{\text{PAM}}}+\cdots}$$
$$\frac{\cdots\prod_{i\not\sim j}\exp\left(-\frac{w(x_i)}{(|\{\text{mis}\}|+2)^3}\mathcal{J}_{ij}(1-\sigma_i\sigma_j)\right)}{\cdots\mathbf{1}_{|\text{mis}|=\text{mis}}\sum_{\text{mis}}\lambda_{\text{mis}}e^{-\lambda_{\text{mis}}x_{\text{mis}}+\ln(N_{\text{mis}})}}$$

reminder: the couplings $\mathcal{J}_{ij} = N - j$ vanish when $j \equiv N$ so introducing the boundary term allows for an extra transition probability in the sequence from which visits of the random walk can still be computed as shown in last week's update

more substantial change I

we plot the distribution of SK and TK growth rates below



more substantial change II

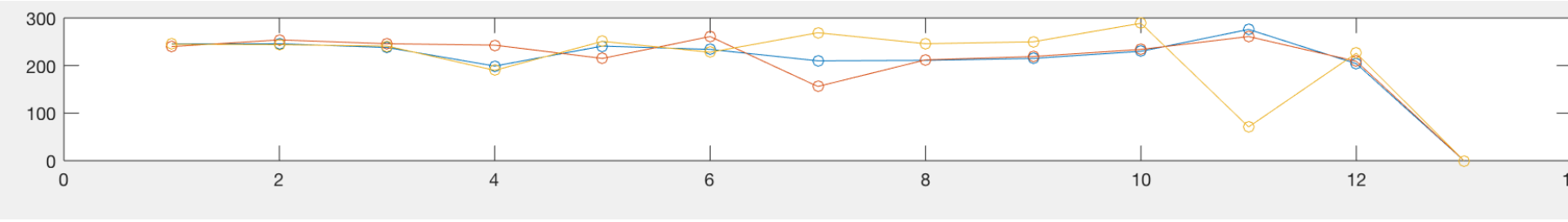
after plotting the growth under SK which is more informative for occupancy, we attempt to introduce a multiplicative factor to the growth rate at the position of binding which is:

- dependent on the distance from the maximum and minimum SK growth rates from the experiments,
- in addition to the location of base pair mismatches in the sequence

Tentatively, although the growth change certainly does not occur at the binding position only, as a starting point I have incorporated the multiplicative factor only at the position of binding. If the decay in transition probabilities at the position of binding due to this additional exponential in the partition looks promising, I can introduce adjustments to the protein fold energy as it takes place in inspection across a sequence of bases rather than only **one** at the position of binding.

recap of plots from last week

the boundary term helps with plots from last week that had no visits to the position of binding, as shown below



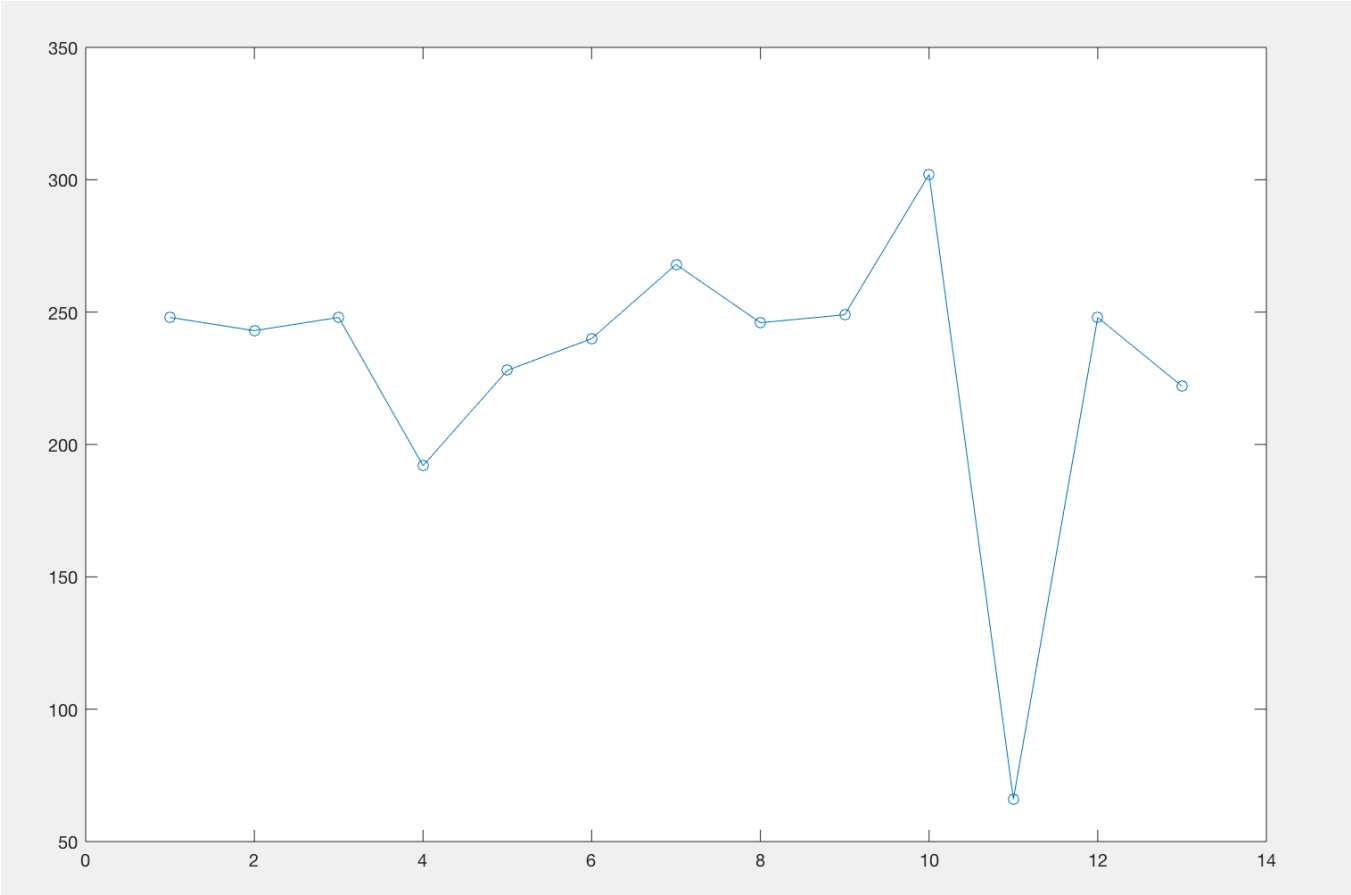
intermediate step

first, to obtain the transition probability at the position of binding, we:

- take entries of each measure vector from last week,
- introduce the "magnetic field" analogous term in the numerator of the probability measure at the position before binding (in this case, 12)
- compute measure value at the position of binding for each vector,
- identify nonzero visits of the random walk at the position of binding

comparison of visits of the random walk at the binding site

for SK growth rate in Cas 12a, $Var = max - min \approx 2.6678$, the multiplicative factor of the form $\frac{\mathrm{SKgrowth}}{Var}$ was drastically increasing $t_p rob$ at the position of binding for some sequences, while for one of them that I attempted, it kept the transition probability at the binding position nearly constant



generalizations to other proteins

for SP Cas proteins, etc, from the model it is necessary for us to know:

- the discepancy between the experimentally observed binding energy of each protein (ie, if $\mathcal{E}_1 > \mathcal{E}_2$ for proteins 1 and 2, then $\mathcal{J}_{ij} = c\mathcal{J}_{ij}$ for free $c > 1$),
- in addition to any supplementary binding processes that must be reflected through more exponential terms in the partition function,
- and finally, being able to distributionally manipulate the random walk visits to the position of binding, glimpses of which have been achieved now for all positions inclusive of the binding position, with different sets of N_{mis} free parameters

generalizations II: obtaining sets of parameters to reflect

with the analysis carried out for different proteins, we can then proceed to compare the visiting distributions of multiple proteins, with important modifications to the partition function for:

- comparatively weighing positions of mismatch differently across multiple proteins of interest,
- decreasing the transition probability substantially enough at the next position of a sequence for a protein if a mismatch occurs,
- and, consistent ranges of mismatch parameters across all proteins in the ensemble (namely, for which protein does a base mismatch adversely impact binding the **most**?)

future steps

- automat(ing) the code for all sequences
- implement more appropriate choice of N_{mis} parameters (recall unusual behavior of measure vectors from last week, the transition probability should not increase that much after a base pair mismatch)