Kinetic Proofreading and Dynamic Nucleosome Structure in *S.cerevisiae*

By Nathaniel Wolff

Abstract:

In this analysis, we posit the existence of a kinetic proofreading step that increases the specificity of transcription. We propose that this step involves the cyclic disassembly and reassembly of nucleosomes. Since it is known that nucleosome assembly occurs in two steps, we propose that the reassembly substep of kinetic proofreading does also. As such, we intend to verify the existence of the kinetic proofreading step by observing evidence of the reassembly substep.
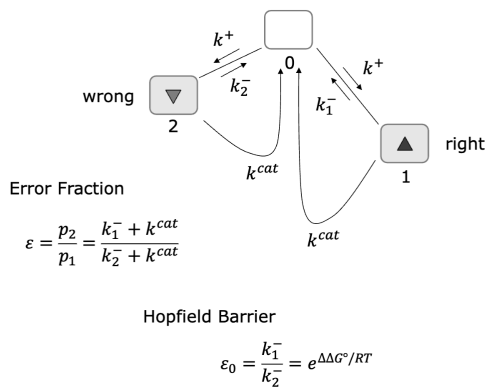
Here, we propose a method for doing so that uses transmission electron micrographs of chromatin isolated from *S.cerevisiae*, spanning the PHO5 gene. We use a Python analysis suite that resolves a set of transcriptionally repressed or active micrographs of isolated chromatin into regions. These regions include bubbles that span nucleosomes. Finally, we discuss statistical analysis additions to the suite that interrogate the distributions of the bubble region sizes, and their implications in the context of the proposed reassembly substep.

Introduction:

Transcriptional regulation ensures that the correct genes are transcribed in response to a cellular signal. Effective regulation requires the binding of the proper transcriptional activator that facilitates the recruitment of other factors and binding proteins that properly initiate transcription. Transcriptional activators are structurally homogeneous, with differences between correct and incorrect activators sometimes depending on a single hydrogen bond. This makes most activators nearly indistinguishable on

a thermodynamic basis. Under these conditions, the minimal error rate of transcriptional activation is governed by the Hopfield barrier, which is the ratio of unbinding reactions constant for the correct and any incorrect activator.
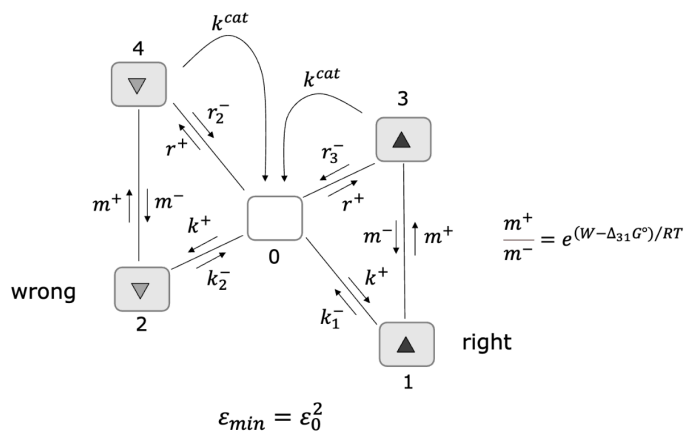
Figure 1: Hopfield Graph describing kinetics of catalysis of an enzyme based on standard Michaelis-Menten kinetics. (16)

Error Fraction

$$\varepsilon = \frac{p_2}{p_1} = \frac{k_1^- + k^{cat}}{k_2^- + k^{cat}}$$

Hopfield Barrier

$$\varepsilon_0 = \frac{k_1^-}{k_2^-} = e^{\Delta\Delta G^\circ/RT}$$

Incorrect and correct activators are termed "wrong" and "right". In this case, $\Delta\Delta G^0$ refers to the difference in the standard free energy change of binding the correct and incorrect activators. Figure courtesy of Hinrich Boeger, 2024. (36)

Yet, the proper functioning of many enzymes require much lower error rates, and thus high specificities. Many enzymes surpass the Hopfield Barrier by introducing a kinetic proofreading step prior to catalysis. This step involves coupling of a biological system with low specificity to a chemical work reservoir, normally ATP/ADP. Kinetic proofreading steps dissipate free energy to increase the energy of the activator-enzyme complex, delaying catalysis. The input of free energy provides the activator-enzyme complex a second opportunity to unbind the incorrect activator before catalysis. In this way, it lowers the error rate of the reaction. An example of this kinetic proofreading step exists in DNA polymerase I, which uses its exonuclease domain to hydrolyze incorrect bases added to a growing replicated strand, dissipating the free energy in the phosphoester bonds(14).

Figure 2: Hopfield Graph describing the kinetics of catalysis of an enzyme that employs a kinetic proofreading step with a rate constant of m^+. The other direction, while possible, is extremely unlikely; the step is effectively irreversible.

$$\varepsilon_{min} = \varepsilon_0^2$$

Incorrect and correct activators are termed "wrong" and "right". In this case, W refers to the work invested into the system from the chemical work reservoir in the proofreading step. $\Delta_{31}G^0$ refers to the difference in the standard free energy change of the kinetic proofreading step. Figure courtesy of Hinrich Boeger, 2024.

Transcription also achieves error rates well below the Hopfield barrier, despite the aforementioned structural homogeneity of activators. This suggests that transcriptional molecular machinery may also employ kinetic proofreading to lower their error rates. A candidate for the enzymes that conduct the kinetic proofreading step are ATP dependent chromatin remodelers. These enzymes use the ATP chemical work reservoir to alter the chromatin structure of eukaryotes in response to a number of known molecular stimuli. In particular, some remodelers are used to assemble nucleosomes after DNA replication and before mitosis. Previous biochemical literature has shown that this remodeler mediated nucleosome assembly has been known to occur via a two step process[1]. This process is responsible for assembling the

---

[1] Source 4

nucleosome core particle, an octamer of histone proteins, around which ATP dependent chromatin remodelers wrap DNA.

Experiments[2] described in Boeger et.al, 2003[3] have determined that several, but not all, nucleosomes present over *S.cerevisae'*s PHO5 promoter unfold completely during transcription. This suggests that nucleosomes may be disassembled and reassembled during transcription. As this requires ATP expenditure, we propose that this cycle of disassembly and reassembly of nucleosomes might constitute a kinetic proofreading step. Furthermore, since the assembly of nucleosomes occurs in two steps, so too might reassembly. As it occurs in two steps, we also posit the following. A certain population of nucleosomes belonging to transcriptionally active chromatin will contain half the DNA length wrapped around a transcriptionally repressed nucleosome. These nucleosomes we will refer to as a pre-nucleosome. Said DNA length has already been determined in *S.cerevisiae* via experiments conducted in Kornberg, et al, 1974[4], approximately 147 BP on average.

Thus, we first assert that distributions of the DNA lengths wrapped around nucleosomes would vary between transcriptionally active and repressed populations of *S.cerevisiae*. In particular, the lengths wrapped around repressed populations would tend to be larger than those stochastically chosen from an activated population. Following this assertion, we assert also that the modalities of these distributions would differ, as the DNA wrapped around pre-nucleosomes would present as mode half the size of the repressed nucleosome. So, the distribution of DNA lengths of transcriptionally active chromatin would approach bimodal, whereas the repressed population's distribution would be unimodal. We now turn to our methodology for obtaining these DNA length distributions in order to test these two assertions. We turn to analysis of electron micrographs of isolated PHO5 chromatin circles, over the promoter, to do so. Our approach is using Transmission Electron Microscopy (TEM) of chromatin isolated from *S.cerevisiae* (16). Analysis of such micrographs is made possible with the GitHub repository GeneRing Python suite

---

[2] Source 1
[3] Source 1
[4] Source 8.

originally authored by R. Shelanky, Ph.d (18). The DNA length distributions are then tested using Mann-Whitney U to test the first assertion, and modeled using a Gaussian Mixture to test the subsequent one.

Methods:

I. _S.cerevisiae_ strain building

Several strains of _S.cerevisiae_ were constructed to be used in analysis (12). Surveyed in this analysis are several strains constructed from a diploid base strain. Information about the base strains' parents are not available at this time. Homologous recombination plasmid transformation was to form PHO80 knockout strains (pho80Δ); this strain then had constitutive transcriptional activity of PHO5[5]. This was compared to a repressed strain control with normal PHO80 activity.

Table 1: Populations of _S.cerevisiae_ surveyed.

Strain Index	Population	Description
1; Repressed	PHO80; GHY2630_GC_Pho80_hht215r_ REPRESSED	Negative Control; PHO5 is transcriptionally inactive
2; Activated	pho80Δ; GHY2625_GC_pho80_hht215r_ ACTIVATED	Experimental Group; PHO5 is constitutively active

(Figure is 62 words)

II. Isolation, Purification, and Cross-Linking of _S.cerevisiae_ nuclear chromatin

Chromatin molecules spanning PHO5 were isolated from _S.cerevisiae_ using the same procedures as in Source 1[6] (18). These fully double stranded molecules, known as "chromatin circles", were cross linked using the reagent psoralen and denatured in basic solution. Parts of the circle that were not

---

[5] Source 8 contains the research justifying this
[6] Source 1

cross-linked span the DNA lengths wrapped around nucleosomes, termed "bubbles". In between bubbles,

cross-linked DNA lengths are termed linkers. Both ends of the molecule have what are termed "forks"

where the 3' and 5' ends are joined together (19). An example of such a molecule is given in Figure 3.
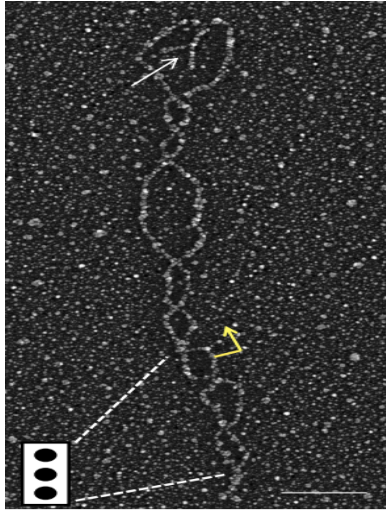

Figure 3: Electron micrograph of a chromatin ring, from Brown Et.Al, 2013[7]. Arrow is pointing from the

molecule's fork (5' end) to the 3' end. Bubbles are visible where nucleosomes are located, separated by

linkers. (36 words)

III.   Imaging and Tracing of Chromatin Circles

Transmission Electron Microscopy (TEM) was used to image the isolated chromatin circles (12).

133 chromatin circles from Strain 3, 170 from Strain 2, and 80 from Strain 1 were imaged. Each image

was traced using a software about which more information is not currently available. Each trace was then

converted to a .txt containing a list of all coordinates in the molecule, termed "raw traces".

IV.   Analysis of Raw Traces via the GeneRing Github Repository[8]

In order to analyze and view raw traces, the Github repository GeneRing was used. The

repository makes use of the following modules: Scipy, Numpy, Glob, Os, Argparse, and Bisect. GeneRing

consists of a set of front and back-end modules (10). The universal backend module, Trace.py, contains
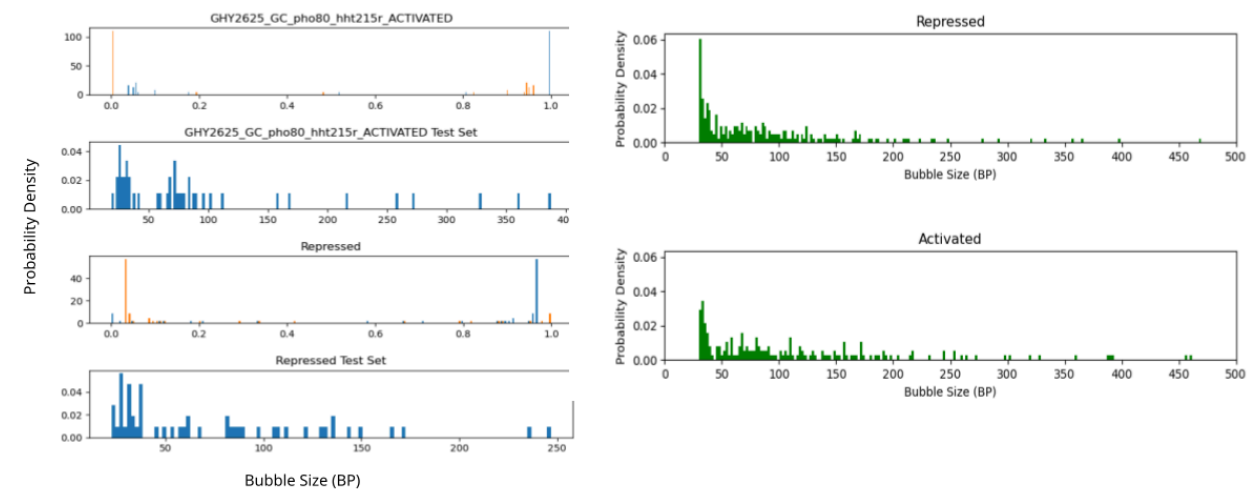
---

[7] Source 2
[8] Source 3

the Trace class, which is initialized by a reference to a raw trace (13). At initialization, an array containing the distances between adjacent points, termed "linked difference" (_ld) is calculated, as is its sum, "linked length" (_ll). Linked length corresponds to the length of the whole molecule in coordinate space. Then, a matrix of cumulative distances from one end of the molecule to the other, (_cld) is computed. Finally, an array of distances between every point and every other is computed as a squareform matrix, with self-distances masked, called "difference matrix" (_d).

Trace.py's methods use said attributes to resolve the raw trace into an array representation of bubbles and linkers. The helper method regionify receives a guess of the molecule's midpoint and default parameters for threshold, in coordinate space, between bubbles and linkers, smooth, the length of the ones vector used to smooth out the curve of the trace, and end, used to set the top coordinate of the end of the molecule. Using these, it divides the molecule into two antiparallel strand arrays. One array contains the beginning and end of each region on each strand and another contains each region's classification, 1 for bubble and 0 for linker. Using these two arrays, we propose an additional method that computes the size of the bubbles (in appendix) to produce the distribution of bubble sizes (DNA lengths over nucleosomes) for a certain locus of the chromatin circle. The scale method, which returns the scale factor in BP/coordinate unit, is used to produce arrays that quantify all bubble sizes in BP. For a particular population, a concatenated array of all bubble sizes is saved to a dictionary, with the strain name as the key. These bubble size arrays are plotted using Matplotlib and normalized as distributions. To interrogate whether or not the distributions differ, they are interrogated using the non-parametric Mann-Whitney U test.. Finally, to interrogate the modality of these distributions, a Gaussian Mixture Model, using the SciKit learn library, is fit to both distributions (80%/20 train/test split), and the Ashman's D[9] of the fit distributions are calculated and compared. In this analysis, regionify is called using its defaults for threshold, smooth, and end, (4,10, and 5, respectively).

---

[9] See Source 4

Results:

Left - Figure 4: Gaussian Mixture Models fit to repressed and active distributions

Right - Figure 5: Bubble Size Distributions for Strain 3 (Activated) and Strain 1 (Repressed)

Table 2: Results of Statistical Tests

Mann Whitney U	Repressed and Active: 0.06 (Not Significant*)
	*Alpha value of .05
Ashman's D of Gaussian Mixture Models	Repressed: 1.0 (Non-Bimodal*)
	Active: 1.3 (Non-Bimodal*)
	*Cutoff of D > 2

Discussion:

Our first assertion implied the activated and repressed strains' distributions are bimodal and

unimodal, respectively. This assertion is not corroborated by the Ashman's D values calculated for the

fitted Gaussian Mixture models, as neither D value exceeds the threshold for bimodality. To this end, we

conclude neither strain's bubble size distributions can be described as bimodal, within the model's

accuracy limits. As such, our assertion that the distributions of active and repressed nucleosome DNA

length may differ in modality cannot be verified. Following this, we return to our first assertion, which claims that modality aside, stochastically, bubble sizes drawn from the activated distribution would tend to be higher than those from repressed. While a significant result of the Mann-Whitney U test would corroborate this, one was not obtained. To this end, we declare that the result of this analysis constitutes a null result.

In other words, we cannot conclude that nucleosomes may be unfolding and refolding as a result of transcriptional activation. Yet, there are many considerations that could be affecting the result. In particular, the data set surveyed is rather small. The particular sample size of these data sets, at approximately 140 circles each, might be skewing the p value. Even so, increasing the size of the dataset is not feasible at this time.

However, further interrogations of the given data set could lead to equally relevant conclusions in the current context. Heretofore, we have been interrogating the exact dynamics of transcriptionally active chromatin, resulting from a theoretical kinetic proofreading step. This analysis presumes that the step's mechanism involves a two-step nucleosome refolding substep following unfolding. Nevertheless, the chromatin structure might nevertheless become dynamic by a different mechanism. As such, determining whether or not nucleosomes are lost at all in transcriptionally activated strains is also pressing, as the mechanism appears to not be straightforward. This requires a change in direction toward determining the distributions of nucleosome occupancies over a certain region of PHO5. This analysis follows the previous naturally.

Bibliography:

1. Boeger H, Griesenbeck J, Strattan JS, Kornberg RD. Nucleosomes unfold completely at a transcriptionally active promoter. Mol Cell. 2003 Jun;11(6):1587-98. doi: 10.1016/s1097-2765(03)00231-4. PMID: 12820971.

2. Shelansky. Robert. GeneRing, (2015), GitHub repository, https://github.com/rshelans/GeneRing

3. *Nucleosome Assembly during DNA Replication*, www.cell.com/current-biology/pdf/S0960-9822(95)00245-4.pdf. Accessed 11 Mar. 2024.

4. Gnedin, Oleg Y. "Quantifying Bimodality." University of Michigan, 19 Feb. 2010.

5. Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H (2013) Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. PLoS Biol 11(8): e1001621. https://doi.org/10.1371/journal.pbio.1001621

6. Roger D. Kornberg, Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* 184, 868-871(1974). DOI:10.1126/science.184.4139.868

7. Carter-Orsquo;Connell, Ian, et al. "Genome-Wide Characterization of the Phosphate Starvation Response in Schizosaccharomyces Pombe - BMC Genomics." *BioMed Central*, BioMed Central, 12 Dec. 2012, bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-697

Appendix

[Link to modified Trace backend script](#)

[Link to Statistics Script](#)

[Link to Gaussian Mixture Model fitting script](#)