BB 592: M.Sc. Project Stage II

“Information encoding via chromatin packaging”

Project report submitted by:

Ritika Aggarwal
195300008

In partial fulfilment of the requirements for the award of the degree of Master of Science (Biotechnology)

Guide: Prof. Ranjith Padinhateeri

Department of Biosciences and Bioengineering
Indian Institute of Technology Bombay
Mumbai – 400076

May, 2021

LETTER OF CONSENT

The work reported in this project stage I entitled “Information encoding via chromatin packaging” has been carried out by Ritika Aggarwal (195300008) under my guidance in my laboratory. I hereby approve the submission of project report.

(Approved by email)
Guide - Prof. Ranjith Padinhateeri

Date – 02 May 2021

Plagiarism Undertaking

I, **Ritika Aggarwal**, roll no. **195300008** understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.).
2. Uncredited improper paraphrasing of pages or paragraphs (by changing a few words or phrases, or rearranging the original sentence order).
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what.

I have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the other authors of this paper, or guide of the thesis / dissertation may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Signature:

Date: **7.12.2020**

Name: **Ritika Aggarwal**

Roll No: **195300008**

Acknowledgement

I would like to express my gratitude towards my lab members Kiran, Sangram, Vinod, Rakesh, Shantanu and Nithya. They all have helped me somewhere or the other in my project and each and every conversation was highly valuable to me. Next, I want to thank my sister ***Priyanka****, who helped me in programming, finding the mistakes in small codes and debug them was not easy for me in the beginning, she helped me get an eye for those which immensely helped me improve my coding skills. I would like to thank* ***Sachin Gupta****, who is like a mentor to me, for explaining me even the smallest doubts I ever had.*

I would also like to thank ***Anirudh Jairam****, for taking his time every week just to help me in my project and coursework, it wouldn't be easy without his constant help. I also want to thank my* ***parents*** *who supported me and trusted me to take care of myself while I was away from them in the time of pandemic of COVID 19.*

I want to thank my guide professor ***Prof Ranjith Padinhateeri****, who was very patient with me and helped me understand the basics related to me project along the way. I also want to thank* ***Prof Rohit Manchanda*** *and my best friend* ***Parul****, for being there for me to tackle the pressure I faced during this time to perform well in exams and as well as in project. The small conversations I had with each of them boosted me a lot to keep going. Last but not the least, I want to thank* ***IIT Bombay****, for giving me this wonderful opportunity to complete my masters under guidance of great minds. I also want to thank myself, for being dedicated to my work and being committed to this project assured me that I am ready for upcoming challenges in my life.*

Abstract

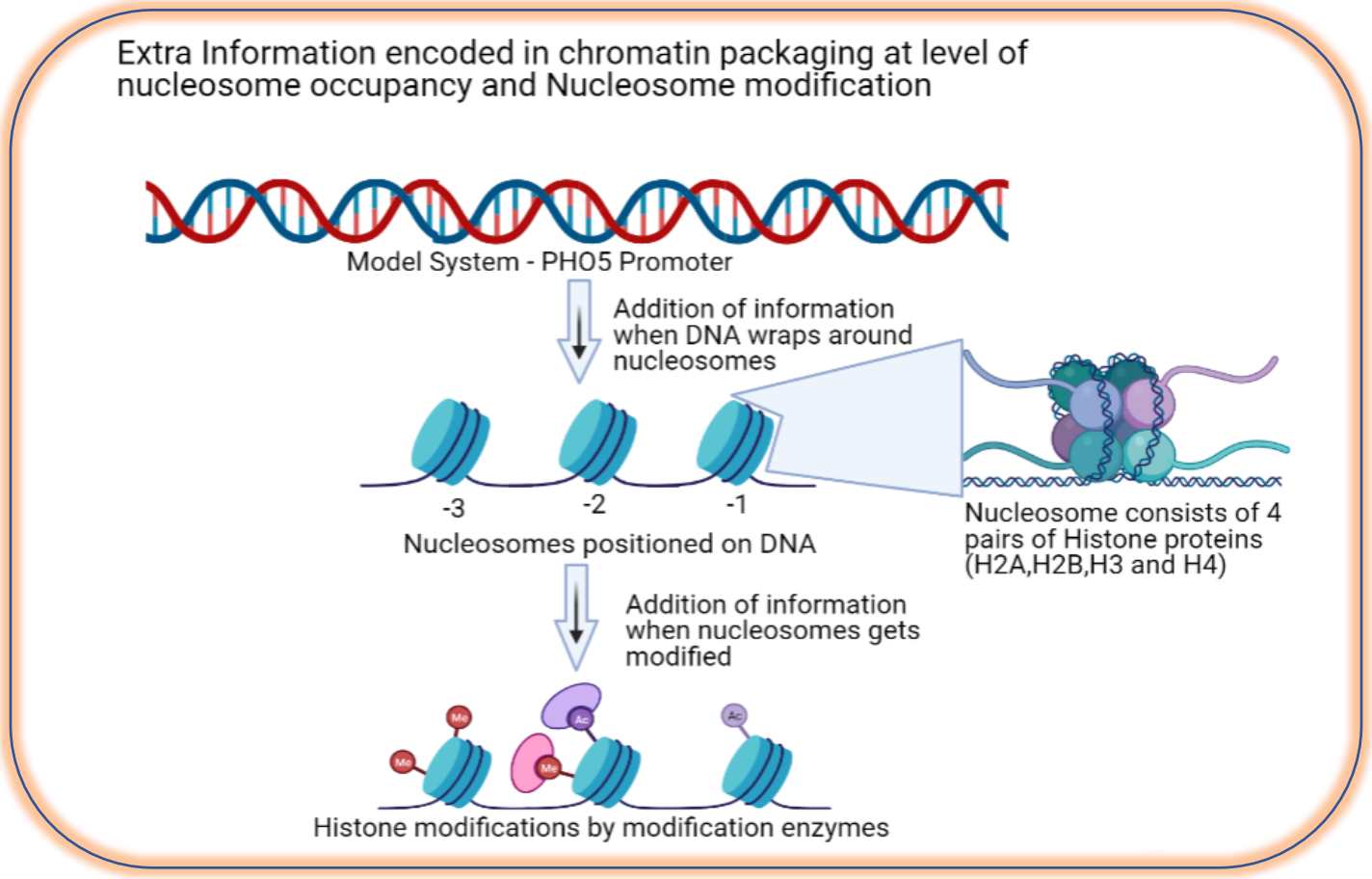
*The concept of **epigenetics** came from studies of embryology and differentiation. Today we think of it in molecular terms. The word "epigenetics" literally means "above the DNA," and, in general, refers to modifications to gene expression that do not involve changes in DNA sequence. These changes take place in histone and non- histone proteins which greatly effects the organisation of chromatin inside the cell and thereby plays an important role in gene expression. According to our current understanding we know that even when each and every cell of our body contains the same genetic material, the expression of cell present in different tissues is different, for example an eye cell is different than a liver cell even when they both contain the same set of genomes, this happens because of the difference in expression of genes in every cell which is greatly influenced by histone and non-histone proteins binding to chromatin under different conditions. Chromatin organisation plays a major role in the expression of genes.*

We know that DNA is packaged in the form of chromatin but the story does not end there, there is extra information coded in the DNA due to presence of DNA binding protein which play a major role in DNA packaging, replication and transcription, every cell has encoded this information differently. As chromatin gets packaged, cell has the information when to switch ON and when to switch OFF a particular gene. This information is present at different levels, when the DNA gets packaged with the help of nucleosomes, one extra layer of information is added with the nucleosome organisation on the DNA, these nucleosomes are having certain modifications at different amino acids which further regulate their role in conjugation with expression and repression of the gene, this adds another extra layer of information to the DNA.

***Entropy is essentially the information** which is determined by the number of possible conformations which can be achieved by a segment of DNA like Promoter of a gene, computing entropy will essentially give us the estimate of how much information can be encoded by that particular region of the gene. In this context, we can use basics of **information theory** and **statistical thermodynamics** to study information encoded by chromatin packaging, which can improve our current understanding of how one cell gathers all the required elements in organization of chromatin. This estimation of information may vary in different cells depending upon organisation and modification of the nucleosomes on DNA. One important part of this information is change in gene expression as mentioned above, information is a way to quantify the arrangement of nucleosomes and their*

modification which affects the gene expression at different levels of chromatin packaging, we have estimated this information using ***Kinetic Monte Carlo Simulations****. These studies can help in our understanding of epigenetics at level of chromatin packaging and organization.*

Graphical Abstract



Contents

List of Figures

Introduction

1. Genetics and Epigenetics

1.1 Gene, chromatin and its organisation

Gene is basically defined as the segment of DNA which is responsible for carrying out certain functions of the cell, in more scientific language one can say that gene is a segment of DNA which can be either translated into RNA or translated and transcribed into protein (building blocks of life), since it is the sole carrier of information for the cell, its organisation is highly compacted and taken care by the cell. Chromatin is the highest level of organisation in the packaging of DNA inside the cell. Many different proteins are involved in the process of chromatin packaging which involve both histone and non-histone proteins (1) .Two of the classic models include the solenoid model and zig zag model of nucleosome organisation. Nucleosome is octamer of histone proteins, homo dimer of tetramer consisting of H2A, H2B, H3 and H4 histone proteins (2) and information in the nucleosome arrangement on DNA is known to affect the 3 D organisation of the chromatin, but how this information is stored in the cell is still bit of a puzzle which ultimately affect the expression of chromatin by regulating the gene activity (3). The expression of gene is also affected by different nucleosome modification, such as H3K4me3 (trimethylation modification of histone 3 at 4^{th} amino acid which is lysine), H3K4ac (acetylation modification of histone 3 at 4^{th} amino acid which is lysine) and many more. There are 26 nucleosome modifications and each modification affects the gene expression by controlling the winding and unwinding of DNA from the nucleosome.

1.2 Histone modification

A histone modification is usually a covalent post-translational modification (PTM) to histone proteins which include attachment of functional groups to amino acids of histone protein such as methylation, phosphorylation and acetylation. These modifications made to the histone protein alter the chromatin structure thereby impacting the gene expression. These modifications contribute to major biological processes including transcription activation and inactivation, packaging of chromosomes and DNA damage or repair. In most of the organism's histone H3 is primarily acetylated at lysine 9, 14, 18, 23, and 56, methylated at arginine 2 and lysine 4, 9, 27, 36, and 79, and phosphorylated at serine 10, serine 28, Threonine 3, and Threonine 11. Histone H4 is primarily acetylated at lysine 5, 8, 12 and 16, methylated at arginine 3 and lysine 20, and phosphorylated at serine 1.

the most important region which is influenced by these modifications is promoter region as this region contain the upstream sequences and TATA box which acts as a hallmark for binding of RNA polymerase. Most of the proteins involved in the regulation of gene expression with respect to external

signals are also known to bind to the promoter region of the gene. The modifications mostly target the histone tails as they are most easily accessible to the modification enzymes.
Therefore, information encoded by the nucleosome modification will help us better understand the epigenetic regulation of cellular processes (4). Many techniques can be used to identify the level of histone modifications in a particular genome, one of the most common technique is ChIP seq (chromatin immunoprecipitation).

1.3 ChIP Sequencing

ChIP-Sequencing delivers genome-wide profiling with massively parallel sequencing reads which generates millions of counts across multiple samples for precise and unbiased investigation of epigenetic patterns (5). Advantages of ChIP seq includes –

- It captures DNA targets for transcription factors or histone modifications across the entire genome of the organism taken for analysis.
- It defines transcription factor binding sites
- It reveals gene regulatory networks in combination with RNA sequencing and methylation analysis.
- It offers compatibility with various input DNA samples

For this project, we worked with the normalized histone modification values for different histone modification from chip seq data obtained from Weiner at al. (6)

1.4 Epigenetics and Gene Expression

Most of the differences among cell types are due to differences in expression of genes, i.e., which ones are actively making RNA to produce proteins, and which ones are inactive, or silent. **Variation in gene expression makes a liver cell liver, and not neuron.** As a simplified example, in the nerve cells gene X is active, producing the neurotransmitter protein X, while gene Y is silent, whereas, in the liver cells, gene X is silent, but gene Y actively produces protein Y, a metabolic enzyme. (7)

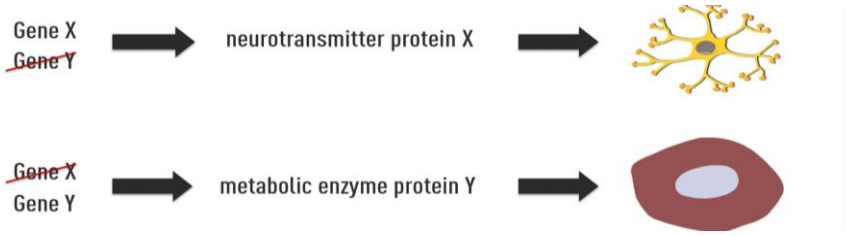


Figure 1 Epigenetics leads to different gene expression in different cells (neuron and eye cell)

Protein "packaging" of DNA is an important factor in regulation of cell-specific gene expression. The proteins packaging the linear DNA can either fold it tightly or fold it in more accessible

conformation as seen in this simplified diagram. The most abundant proteins in chromatin are known as histones, which are small basic proteins that associate with each other and can form a "spool" around which DNA is wound, thus compacting and reducing the length of the DNA (8)

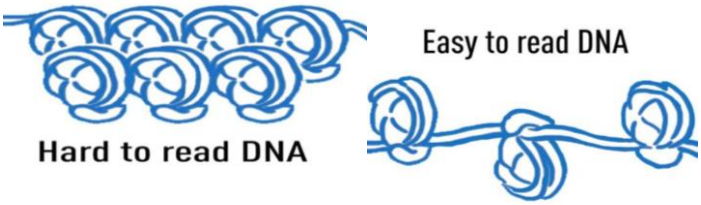


Figure 2 Heterochromatin (hard to read DNA) and Euchromatin (easy to read DNA)

2 Nucleosome organisation studies on PHO5 promoter

2.1 Gene Expressional "ON" and "OFF" state

In "conventional" model of gene expression, it was assumed that when there is complete unspooling of DNA, then only the Gene is said to be "active" and when the gene is completely wrapped around histone proteins, then the gene is said to be "inactive". Gene is active when the transcription machinery is able to access the gene and produce mRNA whereas when transcription machinery is not able to access the gene, it will not be able to produce mRNA and hence will said to be in inactive state. Normally one would think that if that is the case then the predicted conventional model should be right as it considers that when the DNA is fully occupied by the nucleosome, it would not be able to get accessed by RNA polymerase and other activators leading to gene inactivity but from the research experiments done (9, 10), researchers were able to contradict the "conventional" method with "stochastic" method and tried to answer the fundamental question of nucleosome chromatin structure changes when the gene is active.

2.2 Nucleosome organisation in stochastic model – Transcriptional bursting

Gene organisation can be seen as successive molecular transitions, meaning that one cannot be certain about the nucleosome chromatin structure when the gene is active or inactive. Reason being the dynamic association and dissociation of nucleosome from DNA at all times. Thus, we only can consider the probabilistic distribution of the states at the time of gene activation and inactivation. "Stochastic" model of nucleosome organisation states that DNA nucleosome configuration is a process of nucleosome assembly, disassembly and position specific sliding (nucleosome is displaced from its fixed position but not completely removed from the DNA (11)This model supports the idea of "Transcriptional bursting".

Transcriptional bursting simply means that formation of transcription is not a continuous process but a

process which occurs in “bursts”. In other words, one can say that the process of transcription is interrupted by gene inactivity which leads the process to occur in sets rather than in continuous motion. In “conventional” model, the transcriptional activity was considered to be affected by only rate of transcription, meaning that the process is solely dependent on the efficiency of transcription machinery and nothing else. But in “stochastic” model, transcription was considered to be affected by the burst frequency and rate of transcription as we considered that transcriptional activity was interrupted by transcriptional inactivity. The following figure helps in visualizing the transcriptional bursting along with the dynamic process of molecular transitions (9).

Note – α, β, η, Ɛ, δ and ζ are the rates of the processes as shown in the figure.

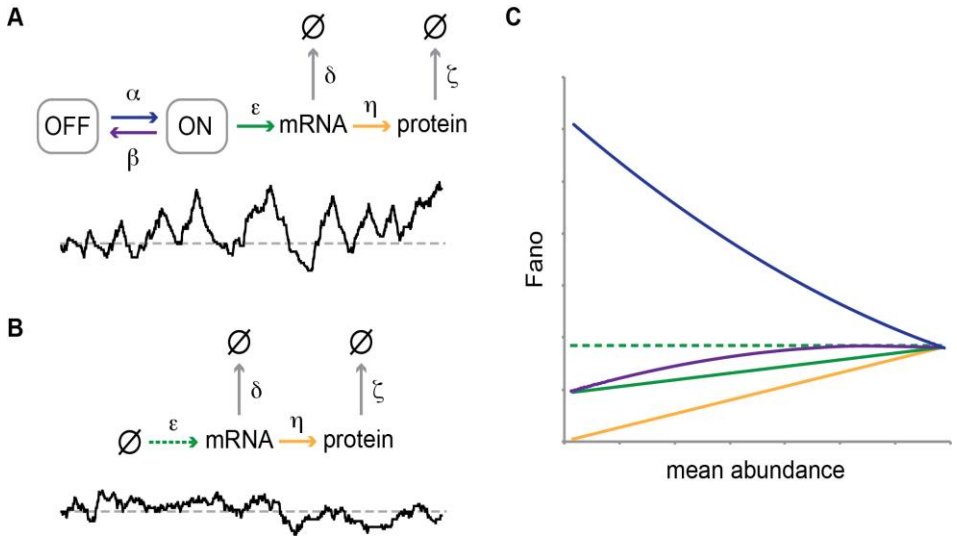


Figure 3 Different modes of gene regulation predict distinct expression noise profiles (9)

Description of figure 3 - (A) The ‘‘two-state model’’ of gene expression. The model takes into account the dynamic process of gene expression transition, ON (transcriptionally active), and OFF (transcriptionally inactive). Transitions -> Ø here means degradation of the product formed. Greek letters signify probabilities of transition per unit time and molecule (‘‘kinetic parameters’’). Also, a time trace just below the model (shown in black curve) refers to the production of mRNA which is then compared with steady state mean (dashed grey line). (B) ‘‘Conventional model’’ of a transcriptionally active gene. The black curve below the model shows a typical time trace of mRNA fluctuation about the same mean (dashed grey line) as in (A). (C) Fano factor for rate of formation of mRNA and protein of both the model (“conventional model” and “Stochastic model”), along with α and β were plotted against mean abundance in order to study the change in process with change in mean abundance. Fano factor is given by ratio of variance with mean. Steady-state Fano factor values (Fano) were calculated as a function of a single kinetic parameter (the ‘‘regulatory parameter’’), with all other kinetic parameters held constant. The dashed green line indicates the expected Fano profile for the modulation of ε for the deterministic model B (9)

2.3 PHO5 gene and its Function

PHO5 encodes the gene responsible for production of acid phosphatase in budding yeast (pho5 paper) in response to the unavailability of phosphate in surrounding environment. Acid phosphatase catalyzes the conversion of orthophosphoric monoester and H_2O to alcohol and phosphoric acid and thus will make phosphoric acid available for the growth of yeast. It is highly expressed under low-phosphate conditions and repressed when phosphate is abundant. PHO5 regulation has been extensively researched (12)Under high-phosphate conditions, PHO5 is inactive. When phosphate is depleted, Pho4p, a basic helix-loop-helix transcription factor, and Pho2p, a homeodomain protein, act cooperatively to bind the PHO5 promoter and activate PHO5 transcription (13)

2.4 Role of Pho4, Pho2 and PHO80 in PHO5 gene expression

PHO5 gene expression is regulated by many proteins, some of which are Pho4, Pho80, Pho81 and Pho2. Just removal of phosphate from extracellular medium is not enough to switch on the transcription of the PHO5 gene. The signal transduction pathway is evolved in such a way that multiple proteins are involved to take care of even a slight change of event in the cell. Pho4 is a basic helix loop helix protein that specifically binds to the E box (CAGGTG), present upstream of the Pho4 regulated genes. These elements are present close to Pho2 binding site which binds cooperatively to PHO4 site at the PHO5 promote, this cooperative binding increase Pho4 binding efficiency. Pho4 when present in unphosphorylated state will be able to bind to the PHO5 promoter and enhance the transcription.
Pho4 is further regulated by phosphorylation via cyclin dependent-kinase pair of Pho80-Pho85. Pho80/Pho85 is inhibited to phosphorylate Pho4 under phosphate starvation condition so as to allow the unphosphorylated Pho4 binding to the promoter, thereby increasing the transcription of the PHO5 gene.(14)

In case cell is having mutated/deleted Pho4 protein, the transcription of PHO5 gene is negatively influenced. If Pho80 is deleted/mutated, it will no longer phosphorylate the Pho4 protein which will lead to transcription of PHO5 gene even in the presence of Phosphate in the environment. But if both the proteins are nutated or deleted, transcription will be negatively influenced as Pho4 will not be able to bind in the first place. Since Pho2 cooperatively binds to Pho4, if Pho2 gets mutated or deleted, there will be less transcription as compared to when Pho2 is not mutated or deleted.

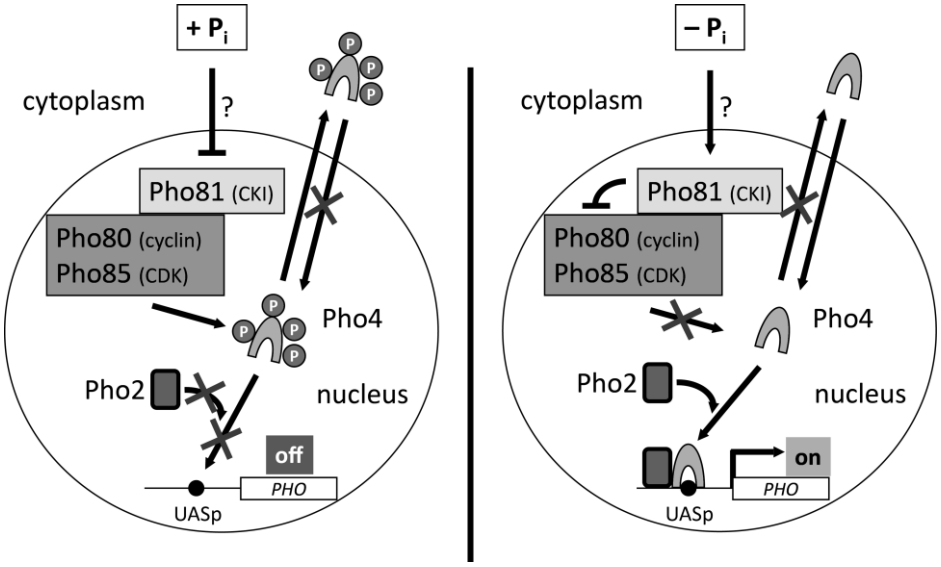


Figure 4 Schematics of PHO regulation via signal transduction

2.5 Single Molecule Analysis of PHO5 Chromatin Structure

In conventional designed experiments, the nucleosome chromatin was studied by taking the population of cell and then studying how the change of stimulus is effecting the gene expression, and even for a specific locus, nucleosome positioning was determined on the basis of information averaged from millions of cells (small et al.) but it was predicted that this approach might misguide us as there are multiple cells in a population and if the concept of transcriptional bursting is true then some cells might be forming the transcripts at one point and some cells might not be forming transcripts at the same time but when the ensemble of molecular states are taken into account at different time points, we will not see the difference in the sum of all the states in a population and our results will be biased towards only one state which will be probabilistically high, justifying the "conventional" model of the nucleosome DNA configuration at the time of gene activity and inactivity. Therefore, in order to know how the nucleosome are assembling and disassembling on the DNA when given a cue to switch "on" the gene expression, it is important to study the configuration at single cell level, one of the studies explained in small et al. (7) (see appendix S1 and S2) obtained 8 microstates when they studied PHO5 promoter region in both phosphate rich and phosphate lacking environment but in different quantities, suggesting that no one state in present 100% at any time, rather the cues to switch "on" or "off" decides the dominating microstate.

2.6 Model of Promoter Nucleosome Dynamics

On successfully isolating the promoter region of Pho5 gene, researchers were able to observe 8 major nucleosome DNA configurations as shown in the figure. These 8 states are the macrostates which were visible under electron microscope, on modelling this using probability matrix, it was inferred that there

is a possibility of large number states based on length of DNA, and no. of nucleosome to be arranged on DNA, this probability distribution explained the future configurations and lifetime of those states probabilistically. Transitions in this matrix assumed that the nucleosome DNA configuration is impacted by nucleosome assembly, disassembly and sliding. With the help of this, researchers were able to find the probability of the states and compared it with the probability of the states observed under electron microscope. It was found that values were highly similar and indeed we have transcriptional bursting taking place when gene expression is turned. Figure 4 shows the nucleosome DNA configurations as found under electron microscope and figure 5 shows the comparative results of probability of different states as found by the matrix considering the stochastic model and as found using EM technique. The results of this study are highly valuable to our project ad it helps us understand the mechanism of gene expression at the cellular level. (9)

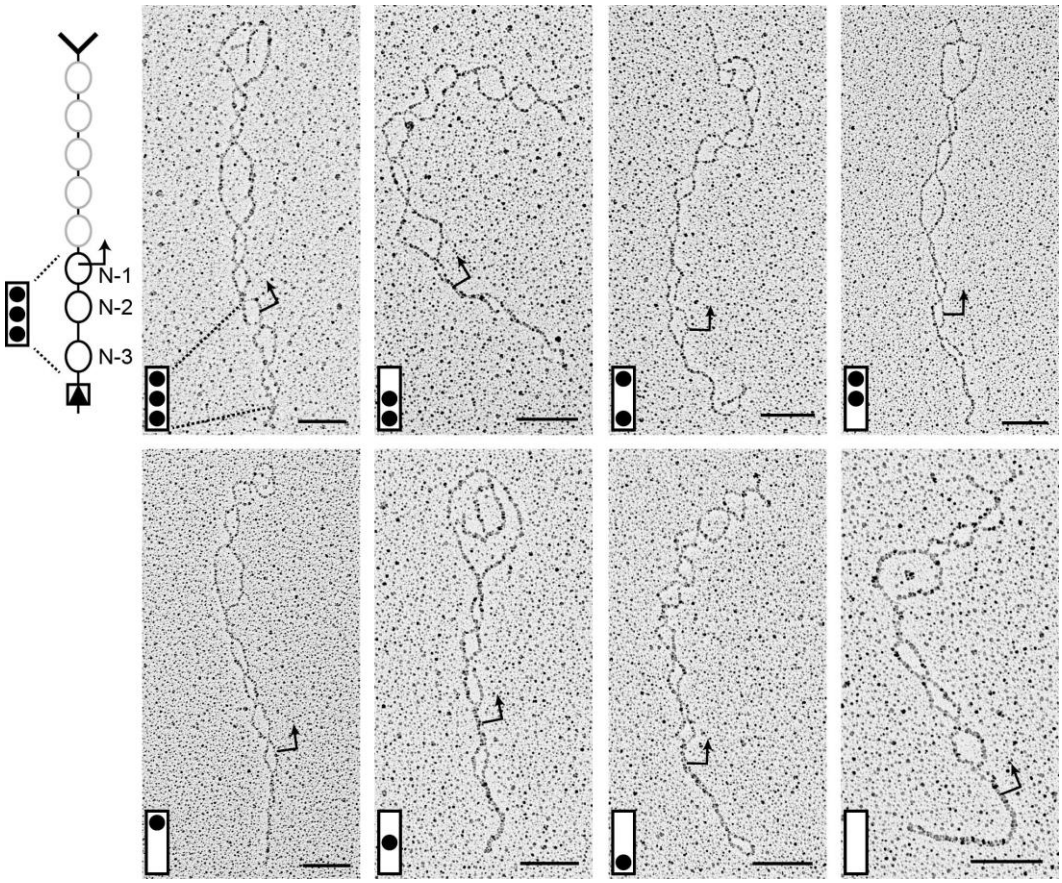


Figure 5 Nucleosome configurations of ‘‘activated’’ promoters. (9)

Description of Figure 5– the above figure shows the 8 macrostates of activated promoters differing on the basis of nucleosome configurations, found with the help of Electron Microscopy studies.

N-1. N-2 and N-3 are the positions of nucleosomes 1, 2 and 3 respectively. The bent black arrow shows the transcription start point. Lower left corner of each image shows the nucleosome configuration predicted with the help of image. (9)

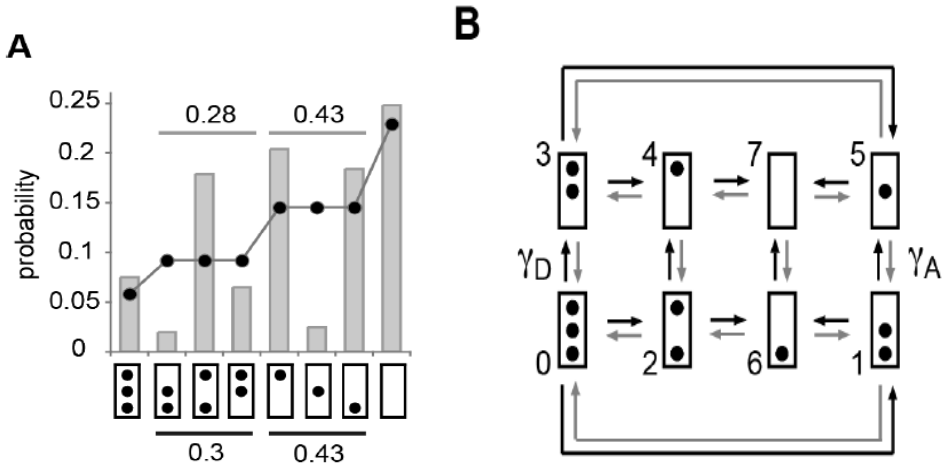


Figure 6 Probabilities of promoter nucleosome configurations (A) probability of different nucleosome states by EM studies and as predicted by "stochastic" model. (B) transition between states of nucleosome configurations. (9)

Description of figure 6 - Configurational probability distribution of PHO5 promoter nucleosomes in activated cells (PHO4 pho80D). Numbers above and below horizontal lines refer to the sum of probabilities for 2- nucleosome and 1-nucleosome configurations, determined by EM (above light grey line), or model calculation (below dark grey bar). Predictions were based on the transition topology in (B). (B) shows the considered model of transition states of nucleosomal configurations numbered from 1 to 8. This figure shows that the model calculations are very much in accordance with the EM determined values and thus considering stochastic model for calculations proved that indeed there is a dynamic interplay between "ON" and "OFF" state during the process of transcription.

2.7 Different mutants of PHO5 promoter have different nucleosome dynamics

PHO5 promoter nucleosome dynamics is found to be very different in different mutants of PHO5 promoter gene (figure 6), this difference is observed due to difference in expression of PHO5 which is regulated by difference in "Transcription burst" frequency. The dotted lines show the probability of occurrence of microstate predicted using theoretical markov model and the bar plot shows the probability of occurrence of microstate from experimental predictions. The theoretical model used nucleosome assembly, disassembly and sliding. The sliding rate does not change the macrostate but will only change the microstate and therefore, there is not much difference in nucleosome dynamics due to introduction of sliding rates into the theoretical model.

1. **Mutant 1 (shown in figure (5A))** – PHO4 pho80D tata (Pho80 is deleted, TATA box is mutated, Pho4 is normal and PHO5 is activated), as in the previous sections, we discussed

Pho80 deletion mutation will not allow the PHO4 to get phosphorylated, due to which even in the presence of phosphate, PHO4 will bind to the PHO5 promoter, leading to greater transcription frequency, even with mutated TATA box. This implies that Pho4 is extremely important in the regulation of PHO5 transcription.

2. **Mutant 2 (shown in figure (5B))** - pho4[85-99], pho80Δ (deletion of pho4 binding region from 85 to 99 bp, Pho 80 is deleted, mutated TATA box and partially activated PHO5 state), since in this case, Pho4 binding site is partially deleted, this means Pho4 will now not able to bind strongly which will lead to less transcription frequency, even with deletion of Pho80 protein, transcription frequency can also be reduced due to mutated TATA box, implying that for reduction of PHO5 transcription, both Pho4 and TATA box plays a important role.

3. **Mutant 3 (shown in figure (5C))** - pho4Δ, pho80Δ (deletion of both Pho80 and Pho4 with mutated TATA box and repressed PHO5 state), since both the regulatory proteins are mutated, we can observe that the PHO5 is repressed, which might be because of deletion of Pho4 because now it won't be able to bind the PHO5 promoter to switch on the transcription and mutated TATA box, as RNA polymerase will be no longer able to recognise the binding site to carry on transcription.

4. **Mutant 4 (shown in figure (5D))** - PHO4, PHO80, pho2Δ (having phosphate in the cytoplasm and wild type TATA box), since Pho2 is involved in recruiting the Pho4 to its binding site and activation of transcription, in this case Phosphate is present, therefore, The PHO5 promoter will have activated Pho80 which will phosphorylate the Pho4 and the PHO5 promoter will be majorly in repressed state.

5. **Wild type strain (shown in figure (5E))** - PHO4, PHO80 (wild type TATA box and repressed PHO5 promoter, also phosphate is present in the cytoplasm), in wild type strain when phosphate is present in the cytoplasm, the PHO5 promoter is expected to be in repressed state as observed in experimental and theoretical predictions.

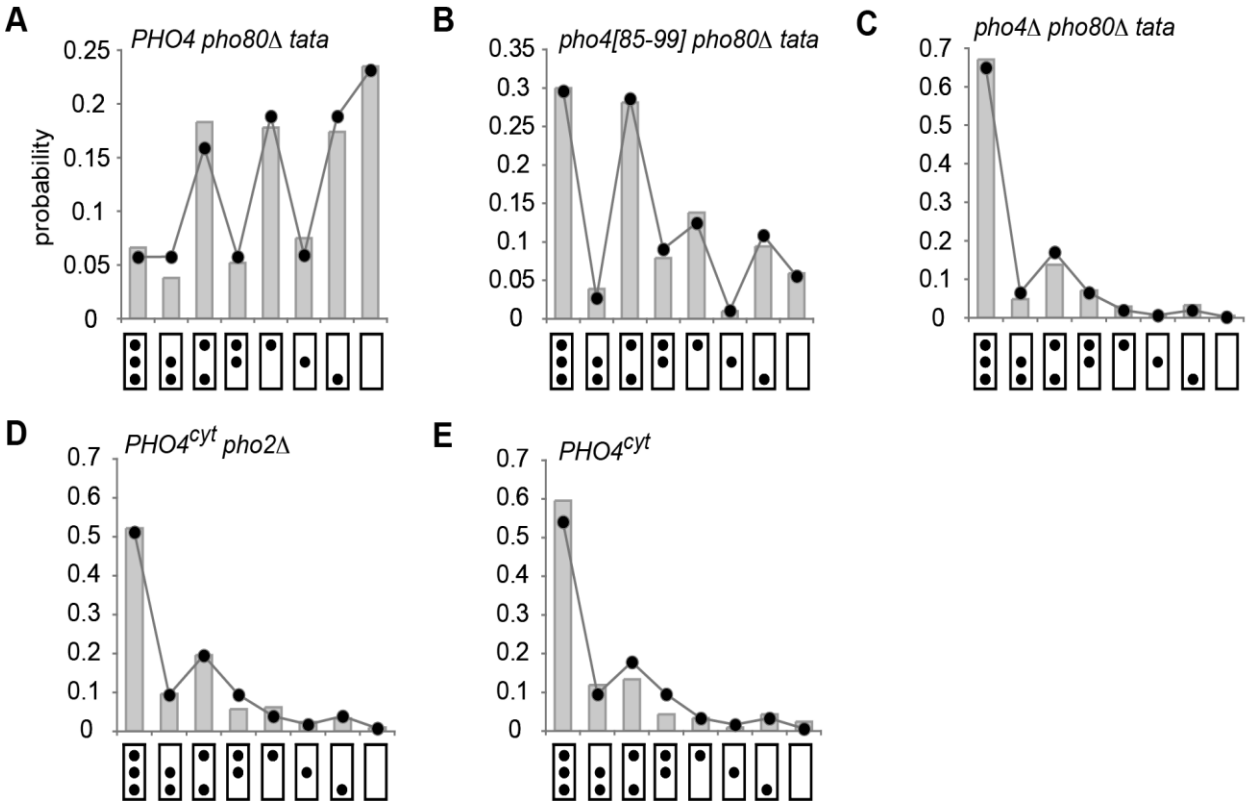


Figure 7 - Configurational probability distributions in activator and promoter mutants.

Description of Figure 7 - The dots represent the probabilities calculated using theoretical predictions whereas bars indicate the probabilities calculated using experimental predictions (A) PHO4,pho80D with mutated TATA box and activated PHO5 state (indicated by high probability of no nucleosome state),(B) pho4[85-99], pho80Δ with mutated TATA box and partially activated PHO5 state (indicated by high probabilities of intermediate microstate),(C) pho4Δ, pho80Δ with mutated TATA box and repressed PHO5 state (Indicated by high probability of 3 nucleosome state) (D) PHO4, PHO80, pho2Δ with wild type TATA box and inactivated PHO5 promoter (indicated by high probability of 3 nucleosome state) (E) PHO4, PHO80 TATA (wild type strain, having wild type TATA box sequence and repressed PHO5 state (indicated by high probability of 3 nucleosome microstate)

2.8 Role of Transcription factor mediated nucleosomal Disassembly in PHO5 gene expression

In the earlier studies, researchers only focused on only nucleosome assembly, disassembly and sliding as a criterion for studying how nucleosome organisation affects the gene expression. In case of Constitutive genes, promoters are having a nucleosome free region (NFR) of approximately 50 to 100 bp upstream the Transcription start site. However, in case of inducible genes, promoters can be switched ON and OFF according to the requirement of the transcript by the cell. In such promoters, there are multiple factors which are involved in assembly and disassembly of the nucleosomes thereby affecting

the switching ON and OFF of the gene, but how exactly these factors are involved in the interplay of nucleosome organisation is a topic of interest for many researchers.

Among the many factors affecting the nucleosome organisation, Transcription factors play a major role in governing the positioning of nucleosome when bound to the upstream sequences. Not only that, presence of certain nucleotide stretches i.e., AA/AT/TT and GC dinucleotides in oscillatory placement also affects the nucleosome arrangement (15)

Considering that Transcription factor (pho4p) can bind at upstream sequence 1 or upstream sequence 2 or both, and nucleosome can bind at three positions on promoter region as mentioned earlier (N-1. N-2 and N-3), Total number of possible states of promoter region increases significantly as compared to the states found when only nucleosome organisation was considered. It was found that there are total 24 explicit states when Nucleosome arrangement (shown in circles) and Transcription factors (shown in triangle) were assumed in affecting the gene expression (also, T represents the region of TATA box of promoter) whereas only 8 implicit states were found when only nucleosome arrangement was considered as shown in figure 6. (16)

TATA box plays a crucial role in recruitment of RNA polymerase and following the TATA box, after 25-35 bp, there is transcription start site where the actual transcript formation begins. Therefore, TATA box must be nucleosome free in order to activate the gene expression. Also, assuming that pho4p (Transcription Factor: TF) binds independent of each other to upstream sequences and binding of TF's recruit active remodellers which are responsible for removing or displacing the nucleosomes which helps in activation of gene expression. Binding of atleast one pho4p is required for the gene to become active along with nucleosome free TATA box, which constitutes 8 Activated states from the explicit states of promoter region as shown in figure 6. Complement of these activated states gives the inactivated states.

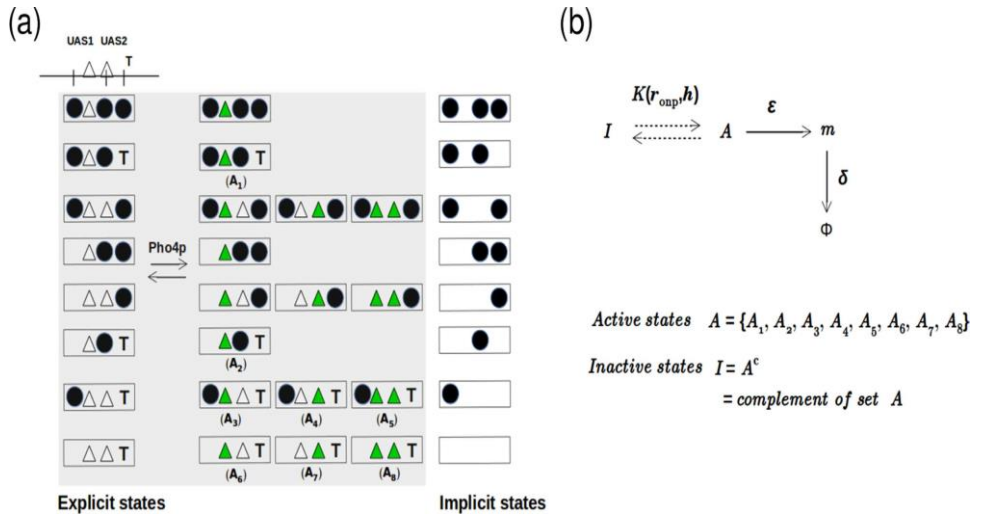


Figure 8 Transcription Model (A) Number of nucleosome-promoter states when transcription factors are considered. (B)formation of mRNA and protein from activated state and dynamic interconversion of I and A state. (16)

Description of figure 8 – (A) the model describes the possible number of states when considering the binding of transcription factors (shown in triangles) and binding of nucleosomes (shown in circle) to affect the activation of gene expression, the competition between the pho4p transcription factor proteins and nucleosome binding give rise to different states of promoter region of PHO5 gene. Transcription factor can only bind to the upstream sequence 1 or 2 or both. There are total of 24 states found as compared to 8 states when only nucleosomes are considered in affecting the promoter states.
(B) I represent the "OFF" state of the promoter and A represents the "ON" state of the promoter. K (r_{onp}, h) represents the dynamic rate of interconversion of these states where r_{onp} represents protein binding rate and h represents local chromatin remodelling parameter. E represents the rate of formation of mRNA and δ represents the rate of formation of protein from mRNA.

Problem Statement

With the help of literature survey, we understand the importance of nucleosome arrangement and how its binding to the promoter region affect the gene expression, we are questioning how nucleosome arrangement is affecting the folding of DNA, considering that Transcription factors are already bound to the promoter DNA sequence. Depending upon precise nucleosome organisation, one gets different transcriptional states. Nucleosome organisation is nothing but packaging of DNA. The problem we want to address is, by understanding how nucleosome is organized (how DNA IS folded), one can predict the transcription states of the chromosome. By folding the DNA around this protein, the cell is restricting access to the genes. In other words, the cell adds extra information to this system by folding DNA into chromatin, "the new information is when to read and when not to read the DNA".
In this project we will quantify how much information is added to this System (DNA + protein) by arranging nucleosome differently. According to information theory, the information content is quantified by computing entropy of the system. Here in this project, we will compute entropy of different chromatin states. Taking nucleosome arrangement on Promoter as our system, on adding nucleosomes, how much extra information in encoded in the DNA, by measuring the increase in entropy of the system.
Also, these nucleosomes can be modified by different functional groups such as methyl, acetyl etc., these modifications affect the transcription of the gene, thereby affecting the gene expression. After the nucleosomes are arranged, the post transcription modifications add an

extra layer of information to the system, now assuming all the nucleosomes are placed on the DNA, how much extra information is encoded in the DNA due to these modifications can also be measured.

Question 1 – How the folding of DNA is affected when we add extra information to the DNA sequence of particular length in form of nucleosomes.

In order to address the above question, we need to know more about what is Entropy and how it is related to number of microstates of a system, how we can calculate the number of microstates of a system and what other factors impacts the entropy of a system.

Question 2 – How does the gene expression influenced by the addition of nucleosome modification.

To answer this question, we can predict the Information for nucleosome modification using probability of nucleosome occupancy for our model system (PHO5 promoter) and using these probabilities, we can do monte carlo simulation with which we can get the modification information of nucleosomes.

1.Entropy and Microstates

1.1 Free energy and Entropy

A general concept which we are familiar with is relation of free energy and entropy. This relation can be depicted as the following equation

$$Free\ energy = energy - Temperature * entropy$$ -------(1)

Where entropy is a measure of possible number of states of the system taken into consideration

And as states by rob Philips, “the equilibrium state of a system is that choice out of all states available to the system that minimizes the free energy”, (physical….). and ideally, a system will try to stay in the state of lowest energy in equilibrium. But is is always true, can we experimentally find out which state is preferable by the system and under what conditions system will not oblige this rule? these questions are still fundamental to the field of research. (17)

1.2 Microstates and Macrostates

When we use the term **microstate, we consider all the possible number of states of arrangement of a particular system**. Let us take an example of ligand binding to a receptor. We imagine the solution

as a series of tiny boxes within which we can place our ligand molecule, let us also consider the receptor in this solution, microstates of this system can be defined as possible ways of arranging all the ligand molecules in the solution and on the receptor. Since all the ligands are undistinguishable molecules, therefore, we cannot distinguish between the microstate where two ligand molecules are interchanged because we cannot differentiate between the ligand molecules. We can calculate the total number of microstates with the help of following formula which allow us to calculate the possible ways of arranging the L ligand molecules in the total of n boxes provided that all the ligands are identical.

$$Number\ of\ microstate = n! \div (L! \times (n - L)!) \quad \text{------(2)}$$

Number of microstates allow us to get the total number of ways of arranging L indistinguishable objects in n boxes without allowing any box to have more than one object. When we put the first object in a box, we have n choices for where it can be placed. Once this first object has been placed in one box, now there are only n-1 boxes left to put one of the remaining L-1 objects. When this is repeated L times with the result that we have n(n-1) (n- 2) × ・ ・ ・× [n-(L − 1)] ways of distributing the L objects. This can be written more simply as number of arrangements as shown above in equation (2). **Macrostate is the sum of those microstate which have same energy**. For example, in this case, all those microstates where ligand is not bound to the receptor, the energy of all those microstates is same and equivalent to each other, whereas the microstate where ligand is bound to the receptor, the energy of the microstate is less than that of other microstates as one ligand is now bound and no more free in the solution. Therefore, it is another macrostate.

The particular cases in the figure show several different microstates in which all of the ligands are free in solution and one microstate in which the receptor is occupied. (18)

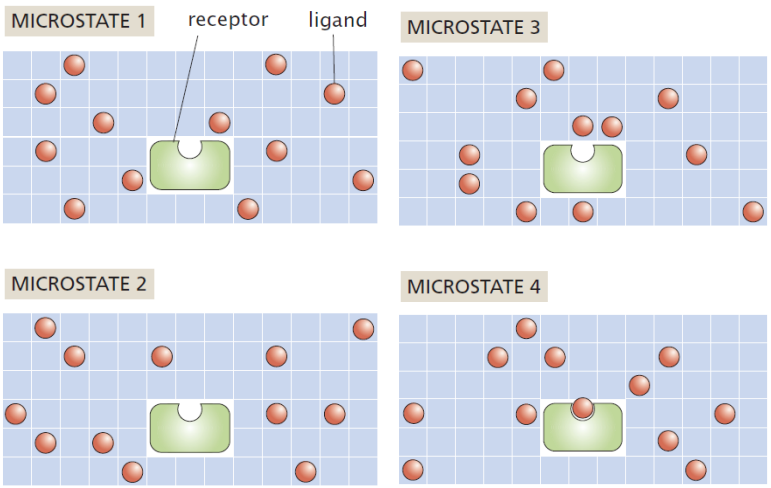


Figure 9- The lattice model of ligand-receptor binding, Microstate 1, 2 and 3 are part of one macrostates where the ligand is not bound to the receptor and microstate 4 is another macrostate where ligand is bound to the receptor (18)

1.3 The Entropy is a Measure of the Microscopic Degeneracy of a Macroscopic State (Boltzmann theory of entropy)

In general Entropy is the degree of randomness or one can say it provide a measure of the total number of microstates of a system. It can be written as:

$$\boldsymbol{S = K_B \ln W} \text{ ---(3)}$$

where W is the number of microstates which are part of a macrostate of interest and K_B is the Boltzmann constant

Consider a DNA molecule which has a total of N number of binding sites, out of which n sites are occupied by Binding proteins. Also provided that energy of binding of protein to DNA is same at every site. Now entropy is the measure of total number of microstates of a macrostate. In this case it depends upon the number of binding sites and number of protein molecules bound to the DNA, which can be estimated with help of following equation (3)

$$\boldsymbol{S = K_B \ln W\,(n, N)} \text{ ---(4)}$$

Where W (n, N) are the number of ways of arranging n number of protein molecules on N number of binding sites

Using equation (2), we can calculate W (n, N), which is equal to:

$$W(n, N) = N!\ \div [n! * (N - n)!] \text{ ------(5)}$$

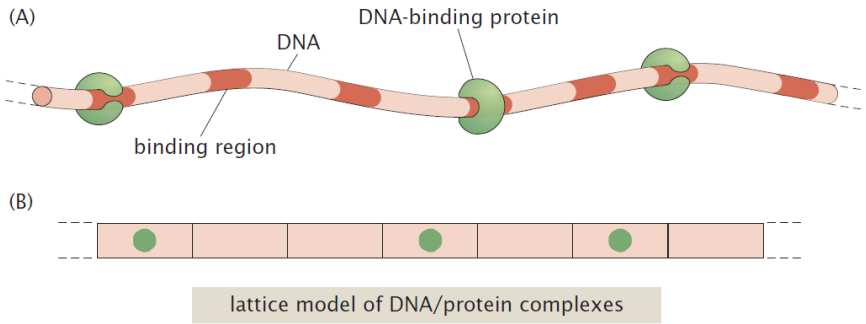


Figure 10 Possible arrangements of DNA binding proteins on a DNA molecule. (18)

Description of figure 10- (A) Diagrammatic representation of a DNA molecule on which there are number of binding sites, which are shown in dark orange. The DNA-binding proteins can occupy any of these sites. (B) The lattice model represents a picture in which we can imagine the DNA molecule as a series of boxes into which we can place the DNA-binding proteins (Adapted from Physical Biology of the cell, chapter 5, section 5.5)

Now to compute this entropy, we will evaluate the equation which we got from equation (4) and (5):

$$S = k_B \ln \frac{N!}{(N-n)!\,n!}$$

Using Stirling's approximation of $\ln N! \approx N\ln N - N$ in above equation, we have

$$S = k_B(\ln N! - \ln(N-n)! - \ln(n)!)$$
$$\Rightarrow S = k_B(N\ln N - N - (N-n)\ln(N-n) + (N-n) - n\ln n + n)$$
$$\Rightarrow S = k_B(N\ln N - (N-n)\ln(N-n) - \ln n)$$

On dividing and multiplying by N,

$$S = k_B N\left(\frac{N\ln N}{N} - \frac{(N-n)}{N}\ln(N-n) - \frac{n\ln n}{N}\right)$$
$$\Rightarrow S = k_B N\left(\ln N - \left(1-\frac{n}{N}\right)\ln(N-n) - \frac{n\ln n}{N}\right)$$
$$\Rightarrow S = -k_B N\left(\left(1-\frac{n}{N}\right)\ln(N-n) + \frac{n\ln n}{N} - \ln N\right)$$

Considering C to be the ratio of Number of proteins to the number of sites for binding on DNA

We can Put $C = \frac{n}{N}$,

$$S = -k_B N(\ln(N-n) - C\ln(N-n) + C\ln n - \ln N)$$

On rearranging the terms,

$$S = -k_B N(\ln(N-n) - \ln N + C\ln n - C\ln(N-n))$$

$$\Rightarrow S = -k_B N\left(ln\frac{(N-n)}{N} - Cln\frac{(N-n)}{N_P}\right)$$
$$\Rightarrow S = -k_B N\left(ln(1-C) - Cln\left(\frac{1}{C}-1\right)\right)$$
$$\Rightarrow S = -k_B N\big(ln(1-C) - Cln(1-C)\big) - ClnC)$$
$$\Rightarrow S = -k_B N\big(ln(1-C) - (Cln(1-C)\big) - ClnC))$$
$$\Rightarrow S = -k_B N(ln(1-C) - Cln(1-C) + ClnC)$$
$$\Rightarrow S = -k_B N\big((1-C)\,ln(1-C) + ClnC\big)$$
$$\Rightarrow S = -k_B N(ClnC + (1-C)\,ln(1-C))$$
$$\Rightarrow \boldsymbol{\frac{S}{k_B} = -N(ClnC + (1-C)\,ln(1-C))}$$ Equation A

Now, we know that N will always be greater or equal to n

$$N > n$$

$$\frac{N}{n} > 1$$

$$\frac{n}{N} < 1$$

$$C < 1$$

Hence C will always be less than or equal to 1, therefore value of C can range from 0 to 1

When we calculate value of C for different values of N and different value of n, we will notice than when C is equal to 0.5, we will get the maximum value of entropy. Which means that Entropy is maximum when half of the sites are occupied by proteins.

On plotting a graph of S/K_B on y axis and C on x axis, we can observe the increase in entropy until the value of C reaches 0.5 and decrease in entropy from value of c ranging from 0.5 to 1.0

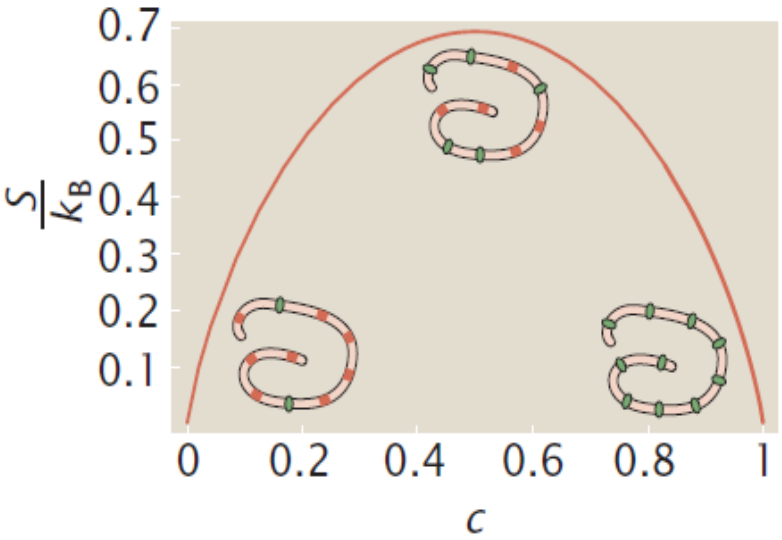


Figure 11 Entropy as a function of Concentration of DNA binding protein ((17)

Description of figure 11 – this schematic representation of protein on DNA within the graph shows Green color as protein bound on the DNA, Red color represents the binding site available for the protein to bind, as we move along from left to right, the concentration of protein increases affecting the entropy as shown in the graph above.

1.4 Shannon's Theory of Information

In Shannon theory of information (20), we can consider C as the probability of finding the nucleosome on one binding site of the DNA, therefore, now we have two possible states depending upon the value of C. Shannon described this in mathematical terms by calculating the entropy in binary units using two probabilities of C and (1-C):

$$H = -K((C \times \log_2 C) + ((1 - C) \times \log_2(1 - C))$$

Where k amounts to a choice of a unit of measure

If there are N such sites of protein binding on the DNA, each site will have p probability of protein binding. The expression will therefore get multiplied by N to get the total information of the system.

$$H/K = -N((C \times \log_2 C) + ((1 - C) \times \log_2(1 - C))$$

The above equation of information resembles Equation A, except that Boltzmann used $\log_e$ and Shannon used $\log_2$, the only difference in using different logarithmic base is in defining the unit of information Choice of logarithmic base corresponds to the choice of a unit of measuring information. Therefore, Boltzmann's entropy measures the information of a system in natural units whereas, Shannon's entropy measures the information of a system in binary units or bits. Plot of H as a function p (as shown in figure 12) resembles that of figure 11.

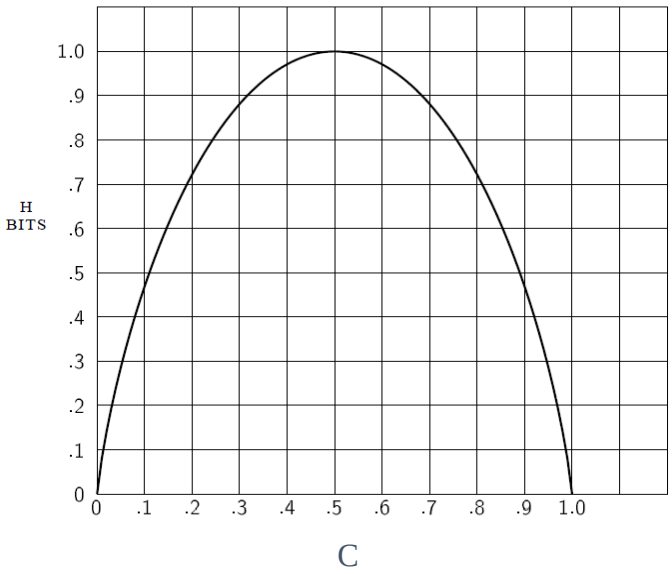


Figure 12 Entropy in the case of two probabilities C and (1-C)

2. Entropy of n nucleosome positioned along the DNA

Let us consider a set of **n Nucleosomes**, A_1, $A_{2...}$. A_n having a length σ, constrained to move along a length of DNA of length l. Let the position of A_k be x_k, Now the possible values of x_i, are restricted, both by the length of DNA and by the finite size of the other nucleosomes of diameter σ. The conditions imposed on the nucleosome positions are (19)

$$\frac{\sigma}{2} \leq x_k \leq l - \frac{\sigma}{2}$$

Assuming the uniform density of ρ, the total number of positions acquired by the nucleosome on the DNA of length l can be written as:

$$W = \rho \int_{(2n-1)\frac{\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{3\sigma}{2}}^{x_3-\sigma} \int_{\frac{\sigma}{2}}^{x_2-\sigma} dx_1\, dx_2 \dots dx_n$$

Neglecting the term of uniform density for our case, we can solve the above equation to get the total number of microstates W

$$W = \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{5\sigma}{2}}^{x_4-\sigma} \int_{\frac{3\sigma}{2}}^{x_3-\sigma} \int_{\frac{\sigma}{2}}^{x_2-\sigma} dx_1 dx_2 dx_3 \dots dx_n$$

$$= \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{5\sigma}{2}}^{x_4-\sigma} \int_{\frac{3\sigma}{2}}^{x_3-\sigma} \left(x_2 - \frac{3\sigma}{2}\right) dx_2 dx_3 \dots dx_n$$

Let $\left(x_2 - \frac{3\sigma}{2}\right) = a$. Then,

$$W = \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{5\sigma}{2}}^{x_4-\sigma} \int_{0}^{x_3-\frac{5\sigma}{2}} a\, da\, dx_3 \dots dx_n$$

$$= \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{5\sigma}{2}}^{x_4-\sigma} \frac{(x_3 - 5\sigma/2)^2}{2} dx_3 \dots dx_n$$

Let $(x_3 - 5\sigma/2) = b$. Then,

$$W = \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{7\sigma}{2}}^{x_5-\sigma} \int_{0}^{x_4-\frac{7\sigma}{2}} \frac{b^2}{2} db\, dx_4 \dots dx_n$$

$$= \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \dots \int_{\frac{7\sigma}{2}}^{x_5-\sigma} \frac{(x_4 - 7\sigma/2)^3}{3.2} db\, dx_4 \dots dx_n$$

Proceeding in the same way, we have

$$W = \int_{\frac{(2n-1)\sigma}{2}}^{l-\frac{\sigma}{2}} \frac{((x_n - (2n-1)\sigma/2))^{n-1}}{(n-1)!} dx_n$$

Let $(x_n - (2n-1)\sigma/2) = c$. Then,

$$W = \int_0^{l-n\sigma} \frac{c^{n-1}}{(n-1)!} dc$$

$$\Rightarrow W = \frac{(l-n\sigma)^n}{n!}$$

Similarly, we can mimic this system in case of arrangement of nucleosome on promoter and calculate the number of microstates (W) provided that length of DNA is l bp, size of each nucleosome is σ bp and total number of nucleosomes are n using the following formula:

$$\boldsymbol{W = \frac{(l - n\sigma)^n}{n!}}$$

Given the number of microstates, we can calculate the Entropy of the system which is given by

$$S = K_B \ln W$$

$$S = K_B ln \frac{(l-n\sigma)^n}{n!}$$

Work done in phase 1

Calculation of number of microstates and Entropy

Number of microstates refer to the all the possible number of arrangements for a system,

From the experiments done previously on PHO5 promoter (nearly of length 600 bp), we know that nucleosomes majorly bind on 3 positions i.e. N-1, N-2 and N-3,

If I have to arrange three nucleosomes on the DNA sequence of 600 base pair provided that each nucleosome can cover only 146 base pair on the DNA. To calculate the number of microstates, we will use equation (6)

Here, n = 3, l = 600 and σ = 146

$$\boldsymbol{W = \frac{(l - n\sigma)^n}{n!}}$$

$$\boldsymbol{W = \frac{(600 - 3 * 146)^3}{3!}}$$

$$\boldsymbol{W = \frac{(600 - 3 * 146)^3}{3!}}$$

$$\boldsymbol{W = \frac{(600 - 438)^3}{3!}}$$

$$\boldsymbol{W = \frac{(162)^3}{3 \times 2}}$$

$$\boldsymbol{W = \frac{4251528}{6}}$$

$$\boldsymbol{W = 708,588}$$

$$\boldsymbol{S = K_B \ln W}$$

In Information theory, K_B is equivalent to 1 bit, putting the value of K_B in the above equation, we will get

$$\boldsymbol{S = \ln(708588)}$$

$$\boxed{\boldsymbol{S = 13.47}}$$

This value of entropy means that we are adding **13.47 natural units** of extra information to the 600 bp long DNA fragment which was earlier not bound by the nucleosomes but after getting bound by 3 nucleosomes, it has 13.47 joules/K of extra information. This information determines when the gene

will switch ON and when the gene will switch OFF. For a larger segment of DNA and greater number of nucleosomes, one can make this calculation easier by writing a small piece of code in any programing language like Python or R

Where on giving the input of length of DNA, length covered by nucleosome and number of nucleosomes to be placed on DNA, the output will give the number of microstates.

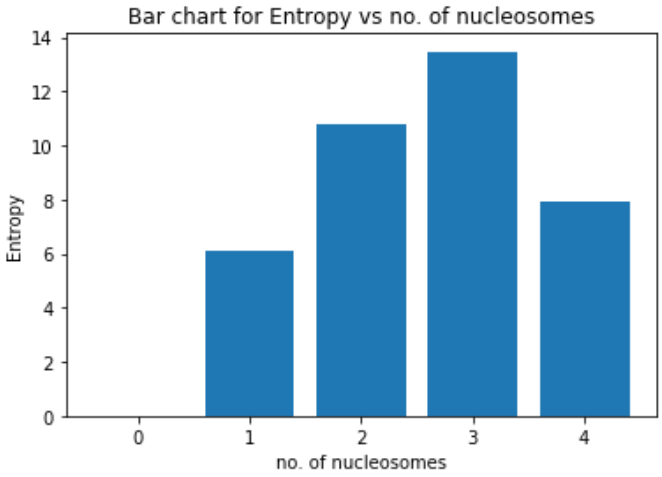


Figure 13 Bar plot for entropy vs no. of nucleosomes arranged on PHO5 promoter

For a particular system of 600 bp DNA on which nucleosomes are placed, each of size 146 bp, we can observe the difference in value of entropy for different number of nucleosomes placed on this DNA segment.

We can observe that when we place 3 nucleosomes on this DNA of 600 bp, maximum entropy of 13.47 Joules/K is observed, means that maximum information that can be added to our system is 13.47 natural units.

Revised objectives for Phase II

Objective 1

To study the extra information encoding by the chromatin packaging, using the basics of information theory and **statistical thermodynamics** when nucleosomes are added to PHO5 promoter.

Objective 2

To predict the changes in gene expression at different level of chromatin packaging (post translational modification of nucleosomes of PHO5 promoter) with help of modification information using **Monte Carlo Simulation (MCS) and ChIP seq data.**

Objective 3

1. To study the trend of change in modification information with change in association and dissociation rate of the nucleosome modification.
2. To predict the range of Association and dissociation rate of nucleosome modification for the modification information calculated in objective 2

Methodology

1. Information of different mutant strains of PHO5 model (taken from Brown et al.) using experimental predictions, theoretical predictions and monte carlo simulations.

1.1 Information from experimental and theoretical predictions

There are 5 mutant strains and one wild type strain (taken from brown et al.), these strains have different configurational probabilities of microstates depending upon the mutations and deletions of the strain (as explained in the introduction section 2.5). The data of configurational probability is taken from supplementary material to calculate the information using Shannon's theory. For monte carlo simulation, nucleosome assembly and disassembly rate were also taken from supplementary material of brown et al.

According to Shannon theory of information, the entropy of a particular system defines the information of that system, given the total number of microstates and probability of occurrence of each microstate. As given in Brown et al., PHO5 promoter have 3 nucleosome and hence, the total no. of microstates for this model are 8. Therefore, Shannon's information is given by:

$$S = -\sum_{1}^{8} \frac{1}{8} log_2 \frac{1}{8}$$ Equation 1

Since, we knew the configurational probabilities for experimental and theoretical predications, we calculated the information (in binary units) using the above expression.

1.2 Information from Kinetic Monte Carlo Simulations

- To calculate the information using nucleosome assembly and disassembly rate, we used Kinetic Monte Carlo (KMC) simulation to get the configurational probability by building an array of binary digits (1 and 0) of rows n and column 3, where n is the number of loops in the simulation.
- We then calculated the number of occurrences of individual microstates.
- Further, we calculated the probability of occurrence of each microstate
- We then used the expression of Shannon's information (Equation 1) to calculate the information of different strains using the probabilities obtained from the simulation.

2. Changes in gene expression at nucleosome modification level of PHO5 promoter with help of modification information

- For calculation purposes, the probability of occurrence of nucleosome modification for -3, -2 and -1 nucleosome of PHO5 promoter were derived from the ChIP coverage data. The data was taken from S3 table containing Log2 modification level values relative to input (unmodified levels) from weiner et al. (6)
- Processing of the data contain the following steps:
  1. Filtering the dataset - Values of all the modifications were only taken for PHO5 promoter occupying -3, -2 and -1 nucleosomes (for time = 0)
  2. Took the antilog (2^x) to get the normalized values (x = V1, V2, V3 for nucleosome -1, -2 and -3 respectively)
  3. Calculated the probability of the nucleosome occurrence at each position relative to the total (for all the 3 nucleosomes).
  4. Calculated the modification information using the following expression.

$$S = -\sum_{i=1}^{3} Pi \times log_2 Pi$$

3. Modification information with change in association and dissociation rate of the nucleosome modification.

3.1 Nucleosome modification information of PHO5 promoter using Kinetic Monte Carlo Simulation.

- Using rate of association of modification (Kon) and rate of dissociation of modification (Koff), it is possible to do Kinetic Monte Carlo simulation (KMC) and determine the information for PHO5 model system at different values of Kon and Koff.
- The results can be verified by simultaneously calculating the average rate of modification. Note that average rate of modification should be 0.5 when both the association and dissociation rate of nucleosome modification are same. The same should be obtained in the results.
- Heat map was made in this case to observe the values of information and to observe the average rate of modification for different rates of association and dissociation rate of nucleosome modification.
- A heat map is a data visualization technique which shows magnitude of a phenomenon (in our case, modification information or average rate of modification) and as color in two dimensions (the two dimensions in our case has Koff on x axis and Kon on y axis). The variation in color intensity gives visual cues about how the phenomenon is varied for the different values of Koff and Kon.

3.2 How the modification information changes with the change in ratio of association and dissociation rate of nucleosome modification for PHO5 promoter

We have assumed that rate of association for nucleosome modification (Kon) is 0.5 modification/sec. Using KMC simulations, we calculated the information for a range of Koff values and plot of modification information vs Koff/Kon to understand how the information increase or decrease with the change in ratio of Koff/Kon.

Results and Discussion

1. Information of different mutant strains of PHO5 model (taken from Brown et al.) using experimental predictions, theoretical predictions and monte carlo simulations.

Table 1 Information calculated using Shannon theory of information for the different mutants of PHO5 model system (brown et al.), using experimental predictions, theoretical predictions and monte carlo simulations

Figure no.	PHO5 state	TATA box	Relevant Genotype	Information using Experimental predictions	Information using Theoretical Predictions	Information using Monte carlo simulations
7E	Repressed	Wild type	PHO4, PHO80 TATA (wild type)	1.946	1.996	1.93
6A	Activated	Wild type	PHO4, pho80Δ TATA	2.639	2.628	2.413
7D	Repressed	Wild type	PHO4, PHO80, pho2Δ TATA	2.117	2.108	2.007
7C	Repressed	Mutant	pho4Δ, pho80Δ tata	1.648	1.562	1.58
7B	Partially activated	Mutant	pho4[85-99], pho80Δ tata	2.529	2.523	2.59
7A	Activated	Mutant	PHO4, pho80Δ tata	2.761	2.734	2.49

From table 1 and figure 14, we can observe the similarity in the values of information calculated for nucleosome positioning on PHO5 promoter, using experimental predictions,theoretical predictions and using monte carlo simulations. This implies that Kinetic Monte Carlo simulation, with only assembly and disassembly rates of nucleosome positioning will also give us approximately same information as experimentally obtained and theoretically predicted by the authors of brown et al.

Also, the theoretical model designed by the authors containing only assembly and disassembly rates, gave the same probability distribution for the microstates of different strains when found using monte carlo simulation, one of which was also shown by the authors (figure 15).

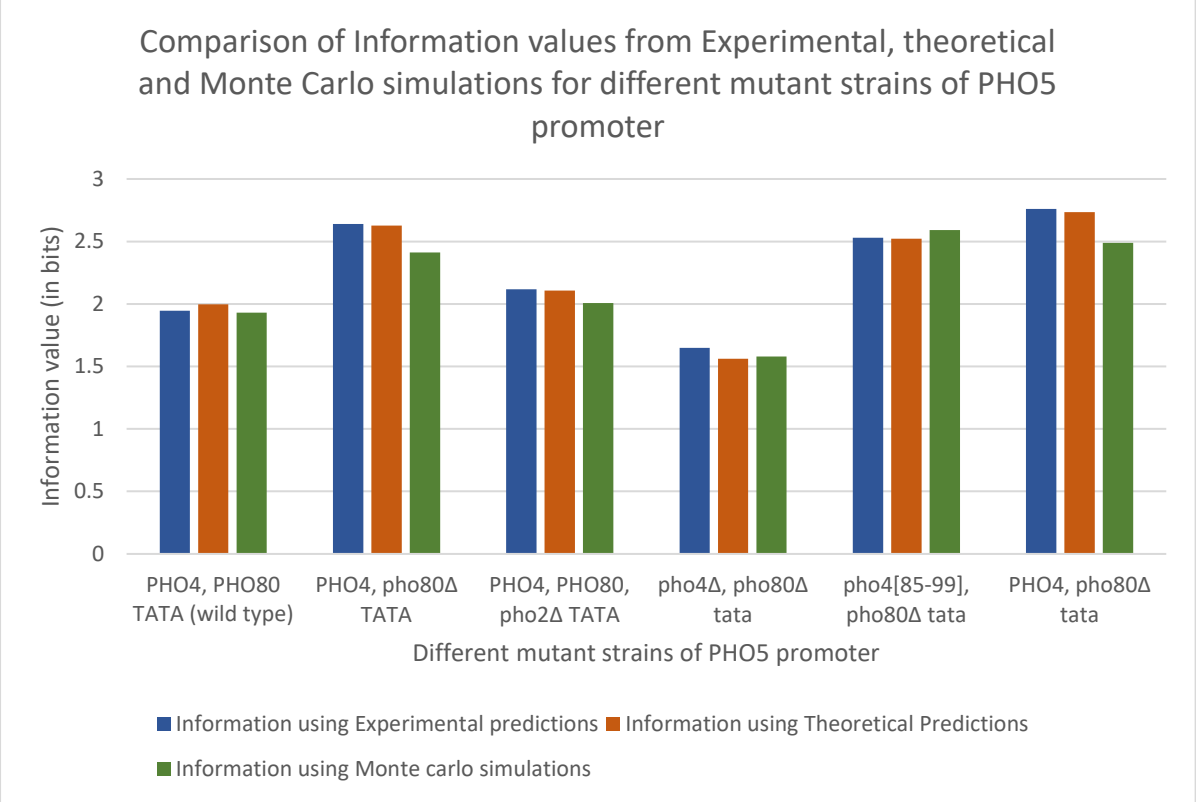


Figure 14 Comparison of Information values from Experimental, theoretical and Monte Carlo simulations for different mutant strains of PHO5 promoter, the information obtained using all three predictions are nearly same.

On comparing the information obtained for different mutant strains with the wild type strain of repressed PHO5 promoter having no mutation and wild type TATA box sequence (Fig 16), we were able to draw the following conclusion:

- The mutant strain (refer 7C) having Pho4D and Pho80D mutations and mutated TATA box sequence (repressed PHO5 state) have less information than the wild type strain also having repressed PHO5 state (refer 7A).

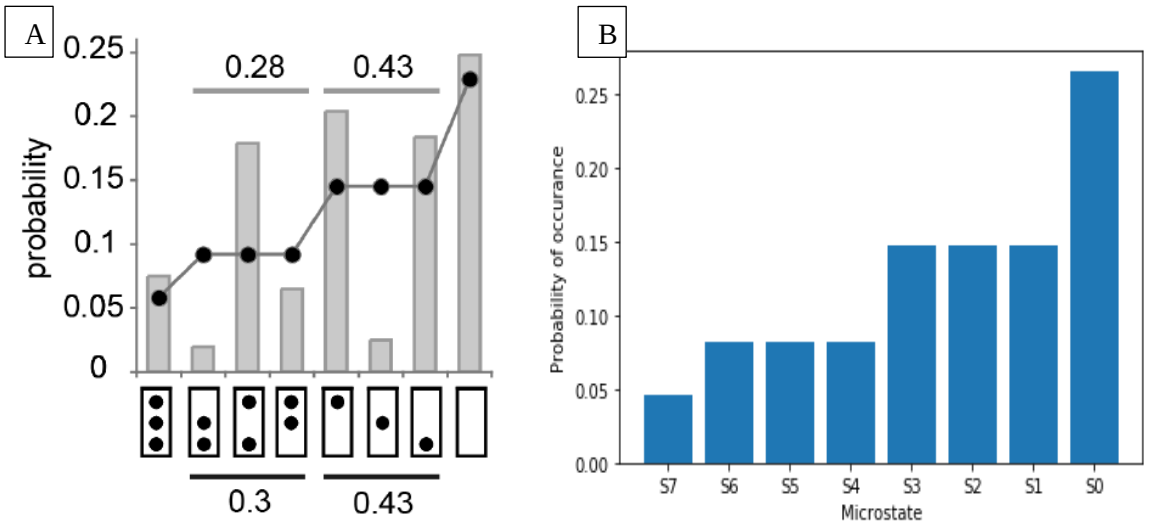


Figure 15 Configurational probability distribution of nucleosomes on PHO5 promoter DNA, (A) Probability distribution for different microstates for PHO5 promoter nucleosomes in activated cells (PHO4 pho80D), black dots indicate the theoretical predictions and bar chart shows experimental predictions. (B) Probability distribution of different microstates using KMC simulation with nucleosome assembly and disassembly rates.

- The mutant strain (6A) with wild type Pho4 and TATA box, but mutation in Pho80 (activated PHO5 state) is having slightly less information as found for the mutant strain (7A) with wild type Pho4 but mutation in Pho80 and TATA box (activated PHO5 state). This implies that there is less effect of mutation in TATA box for governing the transcription burst frequency when Pho4 protein in active, as TATA box provides a hallmark for RNA polymerase to bind to the promoter and activate transcription. One explanation to this observation can also be drawn from the role of Pho4 in activating the transcription, as Pho4 is known to bind to Upstream Activating Sequence (UASp) and initiate the transcription, since UASp is also involved in this regulation, it reduces the dependence (not completely) of the cell on TATA box for activation of transcription.

- The mutant strain (7E) having wild type Pho4, Pho80 and TATA box (repressed PHO5 state) is having less information than the mutant strain (7D) having wild type Pho4, Pho80 and TATA box, but mutation in Pho2 (repressed PHO5 strain), which implies that Pho2 mutation will lead to more transcription burst frequency. Since Pho2 cooperatively binds to Pho4 and increases the probability of "ON' state, deletion or mutation in Pho2 tend to increase the probability of occurrences of intermediate states, thereby increasing the information of the mutant strain.

2. Changes in gene expression at nucleosome modification level of PHO5 promoter with help of modification information

The data was taken from S3 table containing Log2 modification level values relative to input (unmodified levels) from weiner et al. (6), assuming that all the three nucleosomes are placed on the PHO5 promoter DNA, we have calculated the nucleosome modification information. Just like the case of nucleosome positioning, if all the three nucleosomes are already placed on the promoter, there are 8 microstates for the nucleosome modification model system, which can be picturized as figure 16

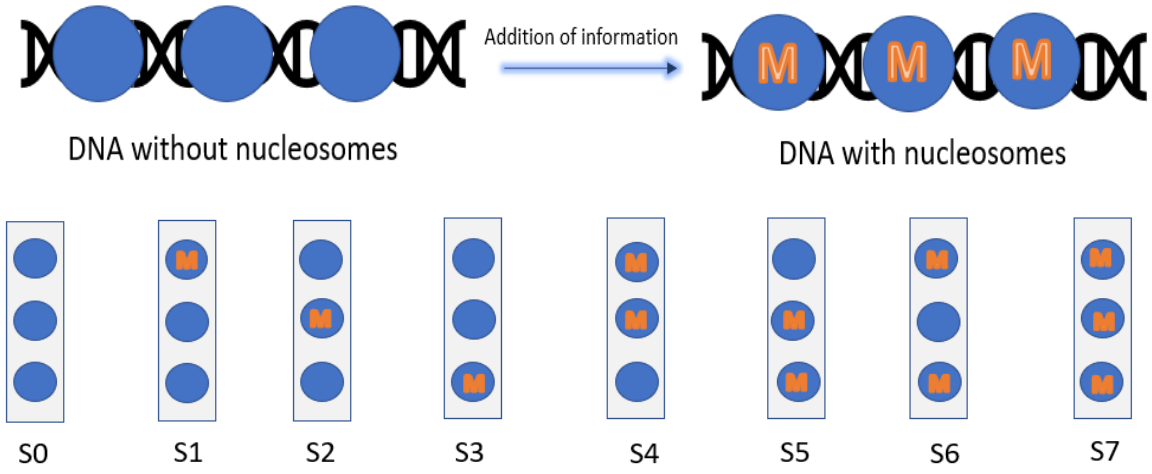


Figure 16 microstates for nucleosome modification model for PHO5 promoter region, assuming that all the nucleosomes (-3, -2 and -1) are already placed on their respective positions.

More information means more uncertainty of finding modified nucleosomes on PHO5 promoter, different modification on nucleosomes can therefore increase or decrease the gene expression according to their role, which can be predicted by this information. More information means more fluctuations in the microstates of modified nucleosomes, if there are more fluctuations, this means that these epigenetic modifications are transient in nature. Since, different modifications have different information values. It is possible to compare among these modifications and comment on which modifications are more transient in nature as compared to the other modifications.

It is already known that histone methylation causes nucleosomes to pack tightly together which does not allow the transcription factors to bind to the DNA and gene is not expressed, if these methylation modifications are highly transient in nature, for example – H3K79me, have highest information of 2.37 bits (Figure 17) this means that they are not stable and therefore can also lead to greater transcription burst frequency as compared to other histone methylation modification.

Also, Histone acetylation leads to lose packing of the histone, allowing the transcription factor to bind to DNA and increase in gene expression, however it can be seen from the figure that histone acetylation information is less as compared to other modifications. For example – H3K4ac is having information of 2.23 bits, which shows that they are least transient in nature.

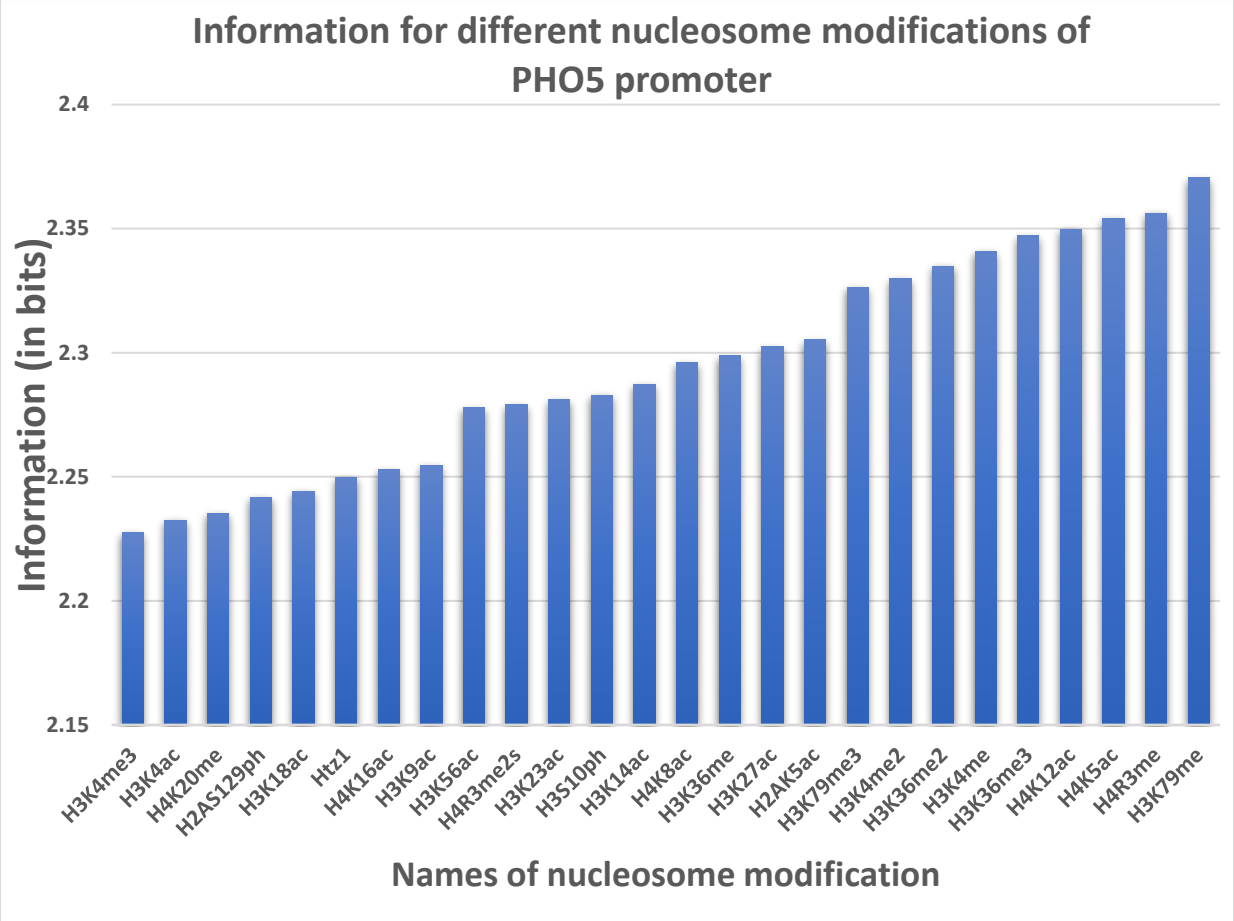


Figure 17 Information for different nucleosome modifications of PHO5 promoter (data taken from Weiner et al.) (6)

3. Modification information with change in association and dissociation rate of the nucleosome modification.

Using Kinetic Monte Carlo simulations, we determined the variation of information for range of association rate (0.1 to 0.9 modification/sec) and disassociation rate (0.1 to 0.9 modification/sec) of nucleosome modifications for PH05 promoter. Figure 18 shows a heat map with color bar showing the change in intensity of color for different values of information in bits, x axis shows Koff (dissociation rate of nucleosome modification) and y axis shows Kon (association rate of nucleosome modification). Following observations can be made from the results obtained:

- When Kon/Koff = 1, the information is approximately 2.6 bits (lies in range of 2.4 to 2.8), which means that information is high (as maximum information is 3 bits) when the association rate and dissociation rate are equal, implying higher probability of occurrence of intermediate microstates.
- When Kon/Koff = 1/3, the information is approximately 2.0 bits (which is in the middle of the color bar, shown in black color), which implies that intermediate microstates are also having some probability of occurrence.

- When Kon/Koff = 4/1, the information is again approximately 2.0 bits.
- When Kon/Koff is highly disproportionate, kon/koff = 9/1, the information is 1 bit (very less), which implies that there is very less information, this can be explained by high probability of occurrence of fully modified microstate.

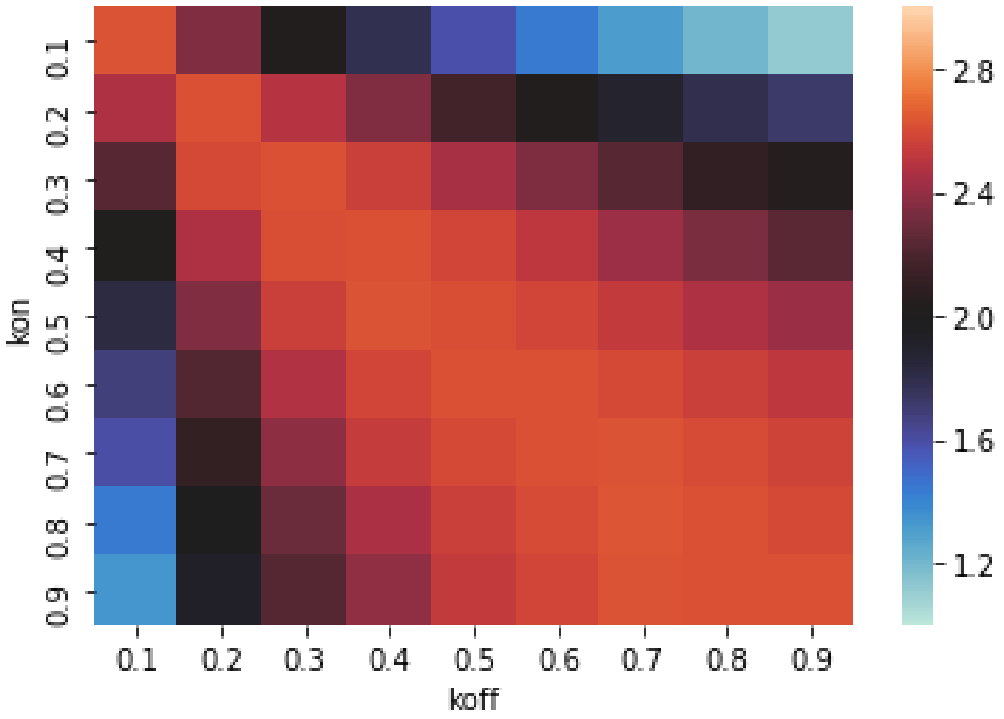


Figure 18 Heat map for value of information varying with different values of association and disassociation rate of nucleosome modification

Similarly, Average rate of modification is plotted for same range of association rate and dissociation rate of nucleosome modification shown in figure 19. Following observations were made from the results obtained:

- As predicted, for Kon/Koff = 1, the average rate of modification is 0.5, because the rate of association of nucleosome modification is equal to rate of dissociation of nucleosome modification.
- Also, when Kon/koff = 9, the average rate of modification is approximately 1 as the Kon>>Koff, therefore all the nucleosomes will be found to be modified in this condition.
- When Kon/Koff = 1/9, the average rate of modification is approximately 0 as the Koff>> Kon, therefore all the nucleosomes will be found in unmodified state in this condition.

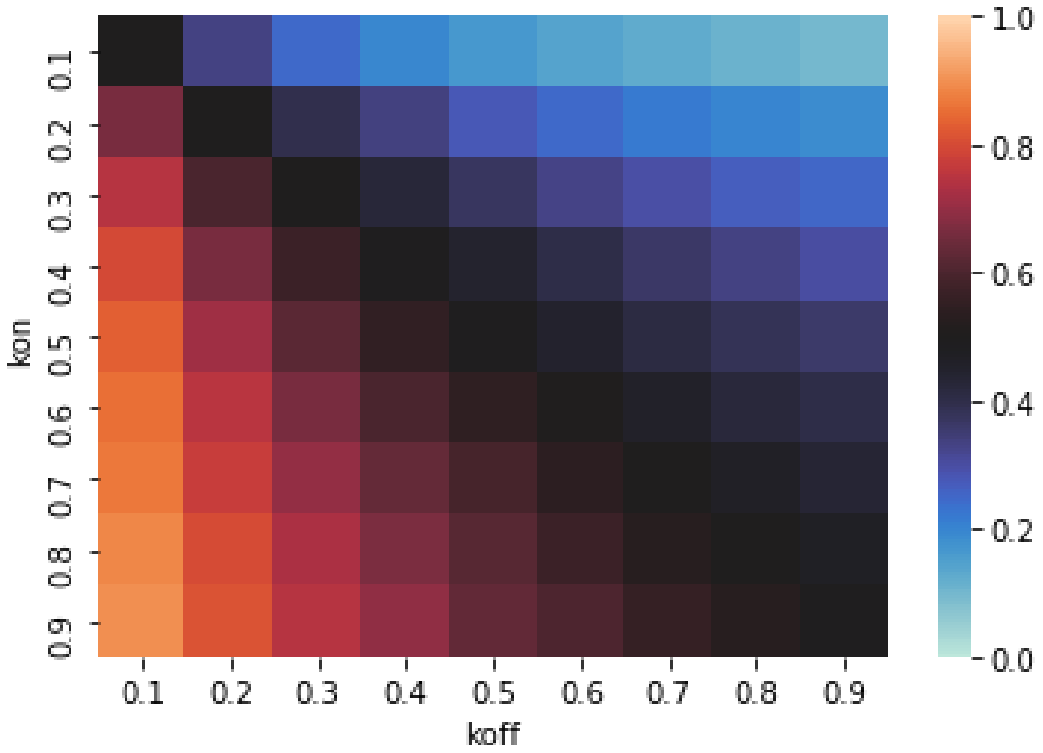


Figure 19 Heat Map for average rate of modification for range of association and dissociation rate (0.1 to 0.9 modifications/sec) of nucleosome modification

As seen in heat map of figure 18, there is a trend of information increase and decrease as a function of ratio of association and dissociation rate of nucleosome modification. To study this trend, Plot of information was obtained for ratio of Koff/Kon (Figure 20), when Kon was set to be 0.5 modifications/sec and Koff was varied from 0.01 to 5.00 modifications/sec.

Following observations were made form the results obtained:

1. Maximum information of 2.7 bits is obtained when Koff/kon is approximately equal to 1, implying that for maximum information, rate of dissociation of nucleosome modification should be equal to rate of association of nucleosome modification.

2. After the ratio of Koff/Kon crosses 1, the information value start decreasing, implying that with increase in rate of Koff will further lead to decrease in information.

Similarly, average rate of modification was plotted against ratio of Kon/Koff, and it can be observed that after a threshold value of Kon, the average rate of modification will remain 1 (Figure 21)

Also, reciprocal plot was made for the average rate of modification vs Kon/Koff, to observe the average rate of modification when the ratio of modification rates is 1, and as predicted, a straight line was obtained with 1/ (average rate of modification) = 2 for Koff/Kon = 1 (Figure 22)

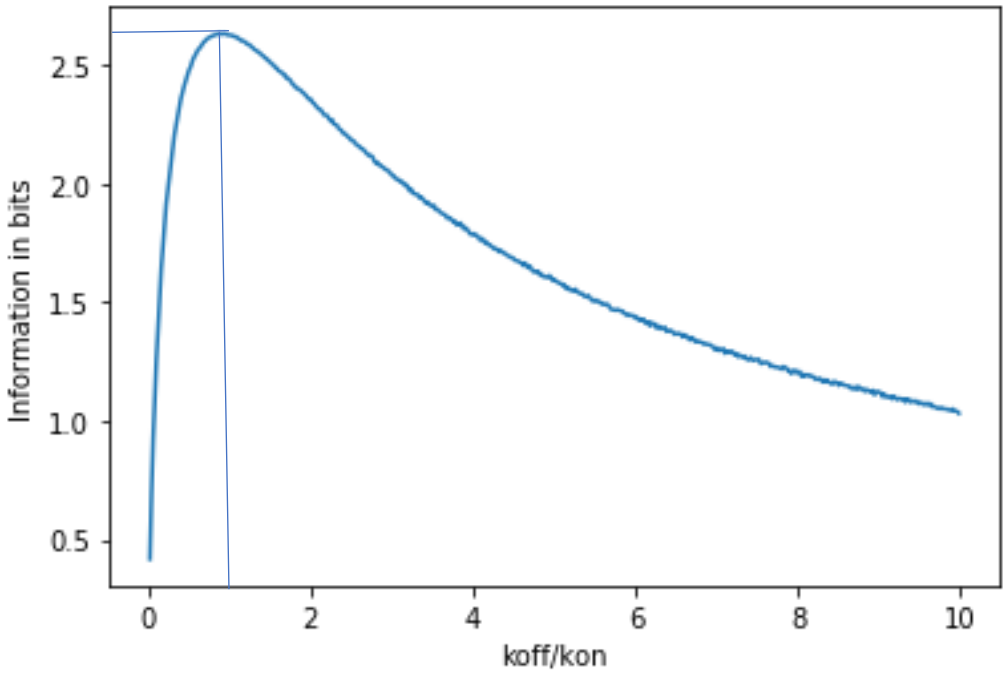


Figure 20 Plot of Information value (in bits) as a function of Koff/Kon, ratio of dissociation and association rate of nucleosome modification.

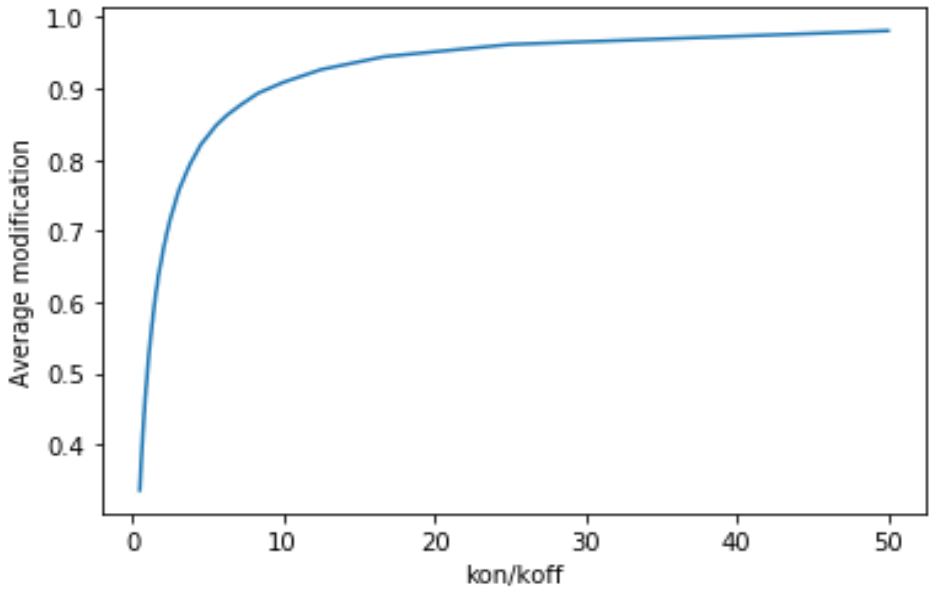


Figure 21- Plot of Average modification of nucleosomes vs Kon/Koff (ratio of association and dissociation of nucleosome modification)

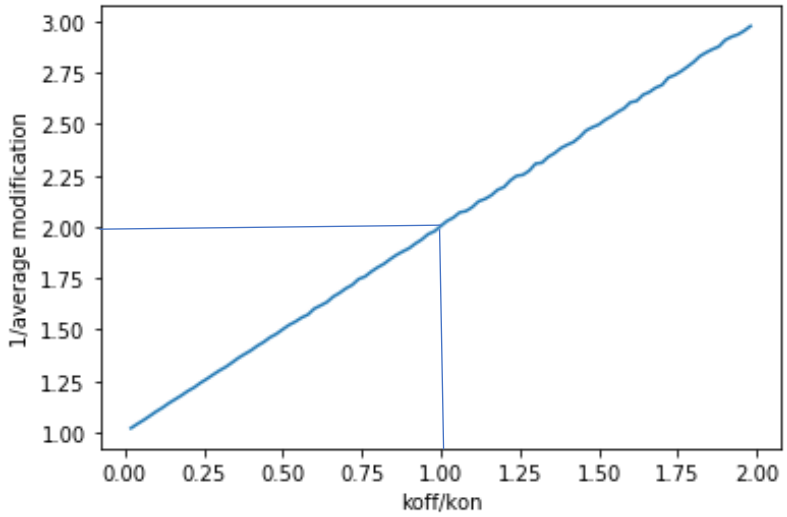


Figure 22- Reciprocal plot of average value of modification vs Kon/Koff (ratio of association and dissociation rate of nucleosome modification)

Conclusion

This project aimed to find the extra information which is added to the PHO5 promoter, at nucleosome level and nucleosome modification level. Using data from brown et al. for calculating the information at nucleosome level for PHO5 promoter model, we were able to comment on the role of different mutations or deletions of mutant strains of PHO5 promoter in increasing or decreasing the transcription burst frequency using information values predicted from experimental predictions, theoretical predictions and kinetic monte carlo simulations. Pho4 deletions were found to greatly influence the probability of occurrence of intermediate microstates of PHO5 promoter model

Using data from Weiner et al. for calculating the information at nucleosome modification level for PHO5 promoter, we derived the probability of occurrences of different modifications on the nucleosomes of PHO5 promoter, with the help of simulations, we calculated the nucleosome modification information for 25 different modifications and found that h3k37me has highest information whereas h3k4me3 has least information. The information values for all the modifications lies in range of 2.23 to 2.37 bits.

Rate of association of nucleosome modification (Kon) and rate of dissociation of nucleosome modific ations (Koff) were then predicted for the information value to lie in the above-mentioned range (provi ded that the kon is 0.5) and were found to be in two range of [0.17-0.2] and [0.98-1.17] nucleosomes/ s ec. We also did Kinetic monte carlo simulations using these rates to observe the change in information with the range of Kon and Koff values, information was found to be 2.7 bits when Kon = Koff.

Future Perspectives

1. Information using Model with nucleosome assembly, disassembly and sliding rates for PHO5 promoter

In this project, we have only considered model having rate of assembly and disassembly of nucleosome for PHO5 promoter, one can also study information using the model having sliding rate to get in depth understanding of nucleosome dynamics.

The model consisting all the three rates are also shown in brown et al., but the sliding event doesn't affect the macrostate of the system, therefore, sliding is restricted to only few microstates of the system, to include this in the Kinetic Monte Carlo simulations (KMC) is out of scope for this project, but can be done as part of future research.

2. Modification Information including all the nucleosome microstates

While this project dealt with nucleosome modification information, considering only one microstate of nucleosome positioning (when all the three nucleosomes are placed on the PHO5 promoter), nucleosome modification can also be taken into account for other microstates as well which would help in better understanding of role of nucleosome modification information in regulating the gene expression.

Appendix

1. Cell to Cell diversity of Nucleosome positioning in phosphate rich media for PHO5 gene

When experiments were done based on single cell analysis of acid phosphatase inducible PHO5 gene, significant variations were observed in nucleosome positioning of the cells and it was found that shift in nucleosome position highly impact the gene expression. With single cell analysis, "fuzzy" nucleosome positions were also deciphered as these regions were known to consist nucleosomes having a larger footprint (7). The single cell analysis was done on the basis of ability of nuclease to recognize only the methylated cytosine of GC dinucleotide as shown in figure 4, when nucleosome is present, it will protect that region of DNA from getting methylated and thus get protected by MNase enzyme.

This analysis was done in Phosphate rich as well as phosphate deficient media. When the 806 single cells were observed in phosphate rich media, only 32% of the cells were found to have the same nucleosome positioning that fits the canonical nucleosome map, 11% shows slight change in nucleosome position at -1 region which may be due to the shifting of nucleosome, but nearly 91% of the cells were having nucleosomes close to TATA box and high affinity upstream sequence, whereas low affinity upstream sequence was found to be nucleosome free. This analysis shows that when single cells are examined for nucleosome positioning, considerable amount of variability is found (7)

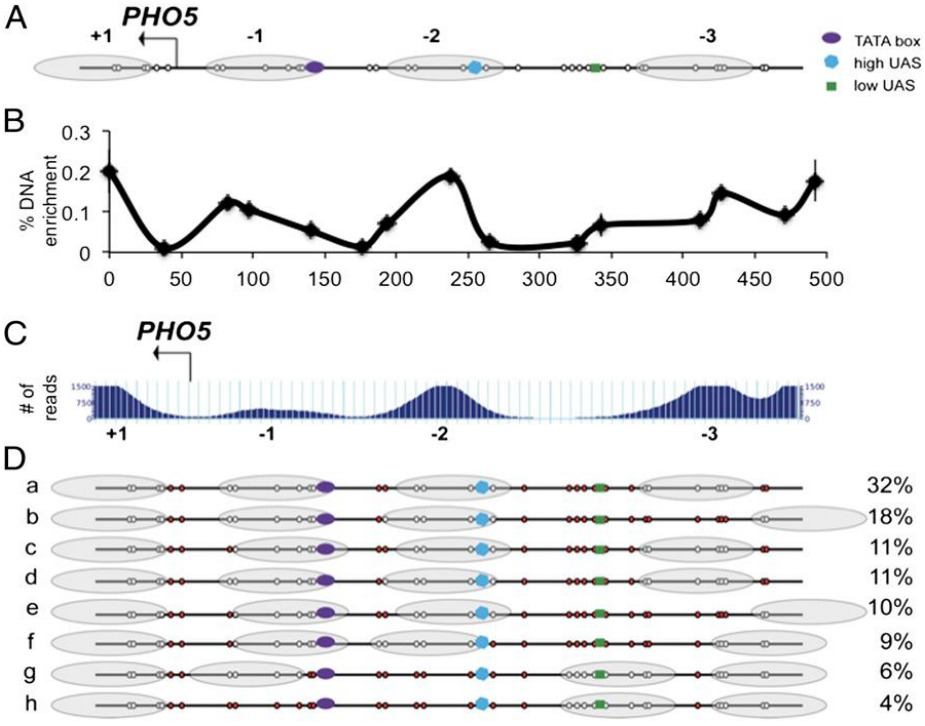


Figure S1 Nucleosome architecture of PHO5 promoter in individual cell shows heterogeneity (7)

Description of figure S1 – (A) standard position of nucleosome in PHO5 promoter, shown in grey ovals. Also, while circles represent cytosine of GC dinucleotide which were used to map the position of nucleosome using nucleosome Scanning assay. (B) mapping of PHO5 promoter observed by nucleosome scanning assay (in phosphate rich media). Nucleosome scanning assay. In this technique, micrococcal nuclease (MNase) is first used to isolate mononucleosomal DNA, and then this DNA is digested with quantitative real-time PCR to map positions of the nucleosome in chromatin. it is rapid and simple, also produces a high-resolution map of nucleosome location which helps in the analysis of a single promoter. DNA enrichment percentage is shown by the black curve where peak shows the presence of nucleosome and plateau region or broad region shows "Fuzzy nucleosome", this region is only partially protected from digestion. Fuzzy nucleosome is the term to describe the region where exact nucleosome position is not certain and nucleosome is said to occupy continuous region on DNA. (C) MNase-seq track of the PHO5 promoter from cells grown in rich media. (D) on examining nucleosome positioning in 806 cells taken from three bulk populations, a total of eight conformations of nucleosome-promoter states were found, in phosphate rich region, most of the cells were expected to be in nucleosome free state, but it was found that different states were present in different percentage. Red circles indicate methylated cytosines, and white circles indicates unmethylated cytosines that are part of GC dinucleotides. The percentage of total cells that demonstrated each nucleosome promoter state is indicated on the right.

2. Nucleosome Remodeling in PHO5 promoter Is observed upon Phosphate Starvation

When the single cell analysis was done in phosphate deprived medium, nucleosome positioning was found to be remodeled significantly and when compared to phosphate rich medium results, it was found that there were 26% additional cells having nucleosome slided away from TATA box to allow the other proteins to access the TATA box. Significant difference was found in -1 nucleosome positioning, approximately threefold less likely to be occupied in medium lacking phosphate. Also, position of -2 was shifted downstream. A total of 76% cells were having nucleosome deprived conformation as compared to only 10% cells in phosphate rich media. Nearly 50% of the cells in phosphate deprived media were lacking nucleosome at -1 position. One interesting observation was that even though significant differences were there in nucleosome positioning, there were still 25% of the cells which were in same conformation as seen in phosphate rich conditions and did not undergo any remodeling of nucleosome positioning. This might be due to underlying dynamic transition of nucleosome-promoter configurations during activation and inactivation of gene expression. (7)

When a yeast strain having C terminal of EFGP tag attached to PHO5 gene was grown in nutrient rich

media under non – permissive conditions, it was found that nearly 1% of the cells were GFP positive, showing low basal level of gene expression or leaky expression. When these GFP positive cells were separated using Fluorescence activated cell separation (FACS) and further analyzed using single cell mapping, they were found to have the same 8 conformations as found in unsorted cells. 54% of GFP-positive cells showed a loss of nucleosomes over the region of UAS and a loss or shift away of a nucleosome over the TATA box, compared with only 10% of cells in an unsorted population. (7)

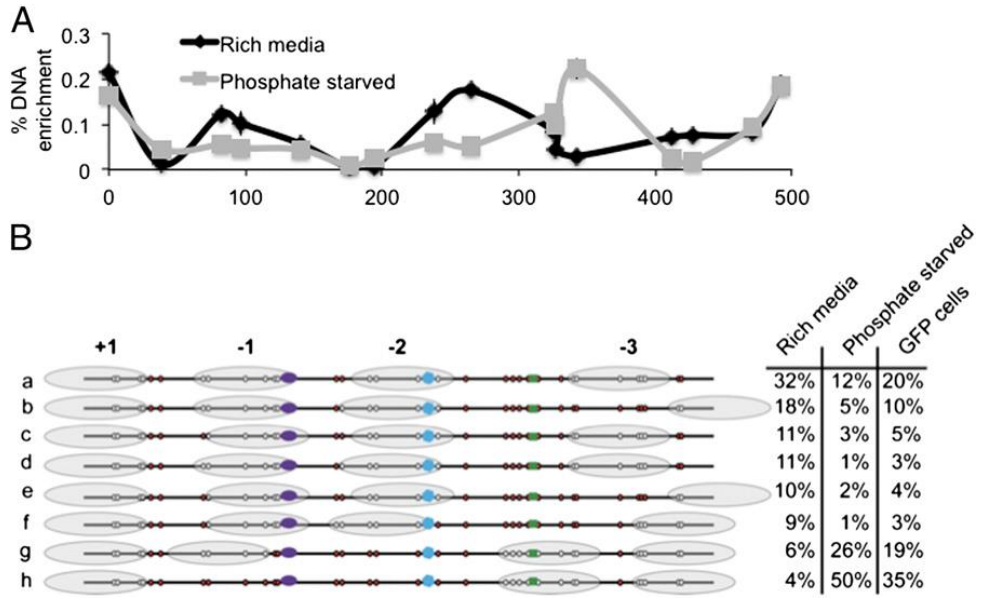


Figure S2 Remodelling of nucleosomes upon phosphate starvation correlates with an increase in gene expression. (7)

Description of figure S2 – (A) nucleosome scanning assay done for single cells in both phosphate rich media (shown in black curve) and media lacking phosphate (shown in grey curve). (B) comparison of nucleosome positioning in cells of both phosphate rich and phosphate deprived media, significant difference was observed in configuration g and h, lacking nucleosome at position -2 and nucleosome free TATA box, which increased to total of 76% in media lacking phosphate as compared to only 10% cells in phosphate rich media. (C) 1% GFP positive cells showing gene expression even in phosphate rich media were found to have the same 8 conformations.

References

1. Kate, P., Parker, A., Editor, E., Winslow, S., Project, S., Georgia, E., Hadler, L., and Blume, A. D. (2010) *Basic Principles of Heredity*
2. McGhee, J. D., and Felsenfeld, G. (1980) Nucleosome structure. *Annual review of biochemistry*. **49**, 1115–1156
3. Parmar, J. J., and Padinhateeri, R. (2020) Nucleosome positioning and chromatin organization. *Current Opinion in Structural Biology*. **64**, 111–118
4. Kuo, M.-H., and Allis, C. D. (1998) Roles of histone acetyltransferases and deacetylases in gene regulation. *BioEssays*. **20**, 615–626
5. Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., and Odom, D. T. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. 10.1016/j.ymeth.2009.03.001
6. Weiner, A., Hsieh, T. H. S., Appleboim, A., Chen, H. v., Rahat, A., Amit, I., Rando, O. J., and Friedman, N. (2015) High-resolution chromatin dynamics during a yeast stress response. *Molecular Cell*. **58**, 371–386
7. Weinhold, B. (2006) Epigenetics: the science of change. *Environmental health perspectives*. **114**, A160
8. Felsenfeld, G. (2014) A brief history of epigenetics. *Cold Spring Harbor Perspectives in Biology*. 10.1101/cshperspect.a018200
9. Brown, C. R., Mao, C., Falkovskaia, E., Jurica, M. S., and Boeger, H. (2013) Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. *PLoS Biology*. 10.1371/journal.pbio.1001621
10. Small, E. C., Xi, L., Wang, J. P., Widom, J., and Licht, J. D. (2014) Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*. 10.1073/pnas.1400517111
11. Misteli, T. (2007) Beyond the Sequence: Cellular Organization of Genome Function. *Cell*. **128**, 787–800
12. Barbarić, S., Münsterkötter, M., Svaren, J., and Hörz, W. (1996) The homeodomain protein Pho2 and the basic-helix-loop-helix protein Pho4 bind DNA cooperatively at the yeast PHO5 promoter. *Nucleic Acids Research*. **24**, 4479–4486
13. Oshima, Y. (1997) The phosphatase system in Saccharomyces cerevisiae. *Genes and Genetic Systems*. **72**, 323–334
14. Choi, J. (2013) Sensing Inorganic Phosphate Starvation by the Phosphate-Responsive (PHO) Signaling Pathway of Saccharomyces cerevisiae. *undefined*
15. Wu, B., Mohideen, K., Vasudevan, D., and Davey, C. A. (2010) Structural Insight into the Sequence Dependence of Nucleosome Positioning. *Structure*. **18**, 528–536
16. Kharerin, H., Bhat, P. J., Marko, J. F., and Padinhateeri, R. (2016) Role of transcription factor-mediated nucleosome disassembly in PHO5 gene expression. *Scientific Reports*. **6**, 1–12
17. R. Phillips; J. Kondev; and J. Theriot. *ch - 5 Phsical Biology of the cell*, Garland Science, Taylor & Francis Group, New York, (November 2008)
18. Phillips, R. (2010) *Physical biology of the cell*, Reprinted., Garland Science, New York (NY) ;;Abingdon (UK)
19. Tonks, L. (1936) The complete equation of state of one, two and Three-dimensional gases of hard elastic spheres. *Physical Review*. **50**, 955–963
20. C. E. Shannon. 2001. A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5, 1 (January 2001), 3–55. https://doi.org/10.1145/584091.584093