International Journal of Engineering &Technology, 5 (2) (2016) 33-37

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v5i2.6098
Research paper

Modeling *in vivo* dynamics of RNA polymerase II meeting Nucleosomes

Roozbeh Abedini-Nassab *, Xu Zhang

Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27707, USA
**Corresponding author E-mail: ra121@duke.edu*

Abstract

Nucleosomes are shown to be barriers for RNA Polymerase II elongation along DNA, and their entry site behaves as the major obstacle. In this work, based on recent available *in vivo* data, we introduce a mathematical model for RNA Polymerase II reads. Moreover, as an alternative way, we use Radial Basis Function Network to predict RNA Polymerase II reads. Results of our models are in good agreement with experimental data. Furthermore, we introduce a random walk model which includes stalling, backtracking, and elongation phenomena. This model can predict and simulate the RNA Polymerase II trajectory on DNA, when it meets various nucleosomes.

Keywords: *Modeling; Nucleosome; Radial Basis Function Network; Random Walk; RNA polymerase.*

1. Introduction

In eukaryotic, RNA Polymerase II moves along DNA and copies it into RNA. During this process, called transcription, RNA Polymerase II meets nucleosomes, in which DNA is wrapped 147bp around a histone octamer (See Figure 1a). Studies show that RNA Polymerase II elongation is influenced and regulated by nucleosomes. It has been shown that nucleosomes mostly behave as physical barriers *in vitro* [1], [2], and *in vivo* [3]; and the strength of this barrier is not constant across one nucleosome or among different nucleosomes. Some of the key elements affecting this barrier strength are believed to be histone tails, histone-DNA contact points, and their connections [4].

In a recent work, Weber and his coworkers [3], by sequencing 3'end of nascent RNA, determined the precise location of RNA Polymerase II, *in vivo*. They mapped it on DNA strands overlaid on a nucleosome landscape derived from MNase-seq. They found that for almost all genes, the first nucleosome is a major barrier for the elongation. This fact has been also observed *in vitro* by other people [1], [2] and mathematical model based on these *in vitro* results are also introduced [4]. However, *in vivo* studies show barrier positions that differ from *in vitro* results [3].

In this work, we use *in vivo* extracted data from [3] and based on that we defined a mathematical model to predict RNA Polymerase II reads at each position. We also use Radial Basis Function Network, as an alternative method, to predict RNA Polymerase II reads. Based on these models, we extract another model to predict the elongation, stalling, or backtracking probability at each point of nucleosomes. Since we include the nucleosome positions in our model, it can be used in understanding the effect of nucleosomes on elongation or stalling of RNA Polymerase II.

2. Methods

To extract experimental data from [3], first of all, we find the position of each gene from [5]. Then we consider three facts: First, the entry of the nucleosome is a major barrier for transcription; second, the consensus barrier position for nucleosomes is at about -7bp from nucleosome entry [3]; and third, nucleosome length is about 147bp. Thus, in next step, we use the data provided by [3] for reads of 3'end of nascent RNA, which is used as RNA Polymerase II reads (RNAP II reads) and find the highest peak position at the beginning of each gene. Then, we assume that 7bp downstream that point would be the nucleosome entry, and that nucleosome continues 147bp further downstream. RNAP II reads of this area, which is the first (+1) nucleosome is what our model predicts. Then we find the next peak in RNAP II reads, and again repeat steps mentioned above to find second and third (+2 and +3) nucleosome positions and model RNAP II reads on those spots.

RNAP II reads for all genes drops very fast before the peak (upstream); however, it decreases gradually after that (downstream) until it reaches the nucleosome exit. That means, to model RNAP II reads, we need an asymmetric function that falls at a faster rate on one side with respect to the other side. Thus normal distribution, which is usually a good candidate for modeling some natural processes, can't be used here. However, a good candidate, having these properties, is the log-normal (Galton) distribution [6], which can be written as:

$$f\left(x|\mu,\theta\right)=\frac{M}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}\right) \quad (1)$$

where x is position on DNA (distributed random variable), M is the magnitude, μ is the mean, and σ is the standard deviation. In order to find optimal μ and σ, to fit the log-normal distribution on RNAP II reads, objective function [7], is used to find the error and it should be minimized:

$$E=\left[f\cdot\left(f-f^*\right)\right]^T\left(f-f^*\right) \quad (2)$$

where f and f^* are the model and experimental reads for RNAP II reads and T=2 in our modeling. We use a computer code (See supplementary materials) to find these optimal μ in [0.01,0.6] and σ in [1,60] for nucleosomes on different genes (e. g.μ_1 and σ_1 for all +1 nucleosomes, μ_2 and σ_2 for all +2 nucleosomes, and so on). We employ magnitude (M) as another degree of freedom which is unique for any individual nucleosome. Since x in (1) should always be positive, and because we assigned negative numbers to positions upstream nucleosome entry, a positive shift in x before using (1) and a negative shift after that is required.

In addition to our log-normal model, we use Radial Basis Function (RBF) Network to predict RNA Polymerase II reads, by defining a hidden layer of kernel nodes:

$$g(x) = \sum_{i=1}^{N} w_i \rho(x - c_i) \quad (3)$$

where N is the number of kernels in the hidden layer, w_i and c_i are the center vector and weight for kernel node i, respectively, and ρ is the activation function, which characterizes the kernel shape [8]. This function is included in most of programming software, and by defining the number of kernels in the hidden layer, one can model a distribution. We found optimal N to be 15, for achieving minimal error and minimizing over-fitting for each set of nucleosomes, explained in Supplementary Materials. We use experimental data of RNA Polymerase II reads on several nucleosomes, from [3], to train our RBF Network model, which later is used to predict the data for reads on other nucleosomes on other genes.

Based on our model for RNAP II reads, we derive a random walk model to simulate RNAP II trajectory. Figure 1b illustrates RNAP II "random walk" along DNA. In this figure, RNAP II is on position x, and in next time step it may either go further to position x+1 (elongation), or stop at position x (stall), or go backward to position x-1 (backtracking).

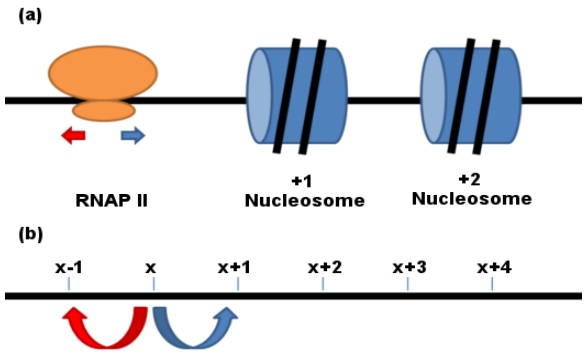

Fig. 1: Cartoon Illustration of RNAP II Movement on a Gene. (A) RNA Polymerase II (Orange) Approaches the Nucleosomes (Blue Cylinders). Blue and Red Arrows Represent Elongation And Backtracking Probabilities. (B) RNA Polymerase Positions along DNA. X Stands for the Current RNAP II Position, while X+1 Is One Step Downstream, and X-1 Is One Step Upstream. Blue and Red Arrows Represent Elongation and Backtracking Probabilities.

Here, to simplify our model and to be able to use available experimental data, first we assume that RNAP II either stalls or elongates along the DNA, neglecting backtracking:

$$P(x+1|x) = P_{Elongation} \quad (4)$$

$$P(x|x) = P_{Stall} \quad (5)$$

$$P_{Elongation} + P_{Stall} = 1 \quad (6)$$

where P (x|x) stands for probability of RNAP II staying at position x in next time step, given that it is at position x, and P(x+1|x) stands for probability of RNAP II moving to position x+1 in next time step, provided that it is at position x. Normalizing RNAP II reads leads us to stalling probability (P_{Stall} (x)) at each point of the gene. Using this model, we assign a start point to RNAP II, and then predict whether it goes forward (elongates) or stops at that point (stall). Then, having that result, we predict RNAP II position at next time step, and continuing this process yields to RNAP II trajectory simulation.

In next step, to include backtracking in our model, we assume that if RNAP II starts stalling (P_{stall}>0.008) the elongation probability is twice the backtracking probability. Thus, in addition to equations (4, 5), we can write (7):

$$P(x-1|x) = P_{Backtrack} \quad (7)$$

Where $P_{Backtrack}$ stands for backtracking probability, and equation (6) is replaced by (8, 9):

$$P_{Elongation} + P_{Stall} + P_{Backtrack} = 1 \quad (8)$$

$$P_{Elongation} = 2P_{Backtrack} \quad (9)$$

3. Results

Results for deriving fitting parameters μ and σ for our log-normal model for the first three nucleosomes are listed in Table 1. Similarly, results for deriving the third fitting parameter, M, in this model is shown in Table 2. Based on these parameters, we model the results in Figures 2, 3, and 4 for +1, +2, and +3 nucleosomes, respectively. These figures show that our model is in a good agreement with experimental data.

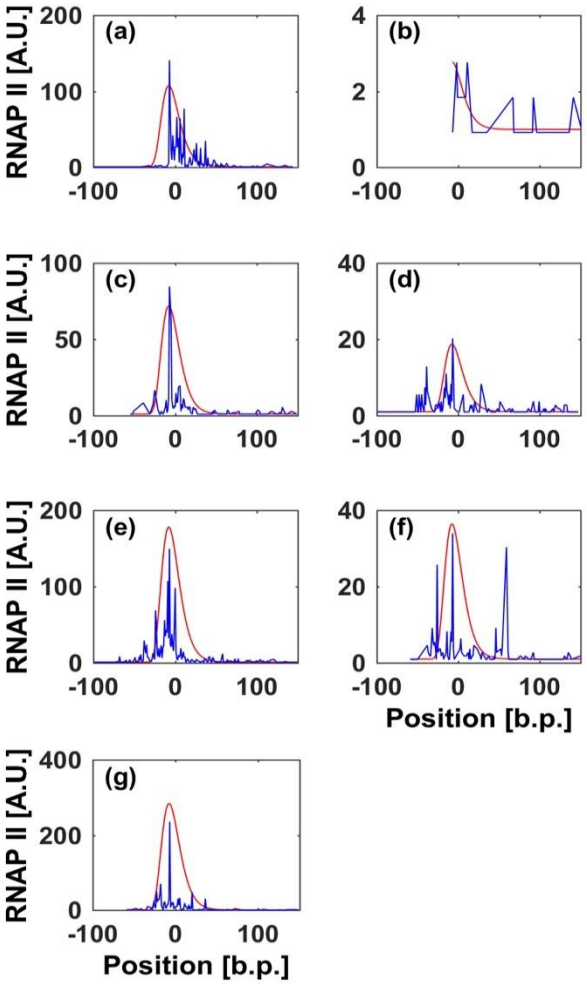

Fig. 2: Modeling RNA polymerase II Reads When It Approaches +1 Nucleosome, Based on Log-Normal Distribution. Experimental (Blue) and Simulation (Red) Results for +1 Nucleosome for (A) Acon, (B) CG31627, (C) CG9246, (D) Mcm10, (E) Bur, (F) CG9243, and (G) CG9247. Experimental Results are based on [3]. X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

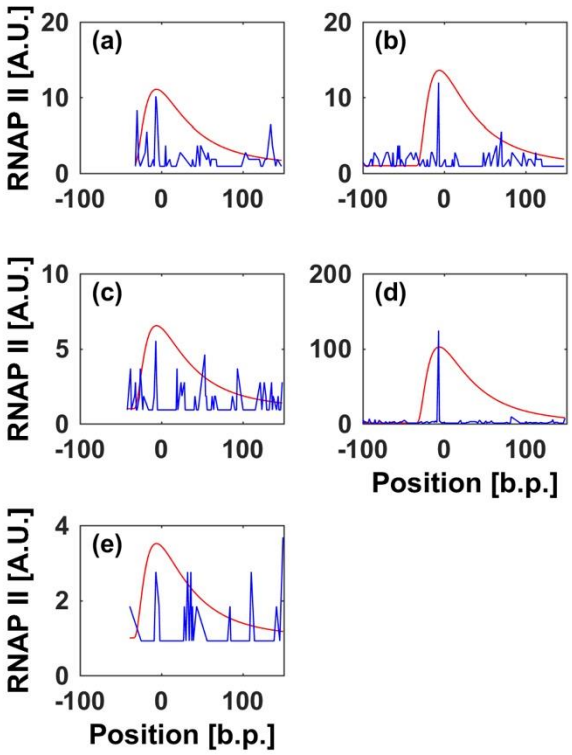

Fig. 3: Modeling RNA polymerase II reads when It Approaches +2 Nucleosome, Based on Log-Normal Distribution. Experimental (Blue) and Simulation (Red) Results for +1 Nucleosome For (A) Acon, (B) CG9246, (C) Mcm10, (D) Bur, and (E) CG9243. Experimental Results Are Based on [3]. X-Axis Stands For RNAP II Position With Respect To The Entry Site of the Nucleosome.

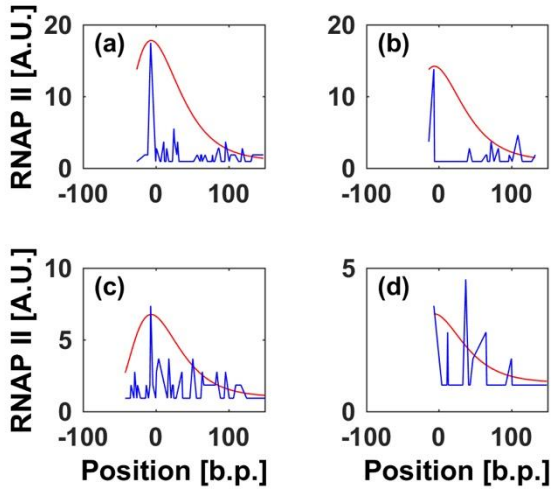

Fig. 4: Modeling RNA polymerase II reads when it Approaches +3 Nucleosome, Based on Log-Normal Distribution. Experimental (Blue) and Simulation (Red) Results for +1 Nucleosome for (A) CG9246, (B) Mcm10, (C) Bur, and (D) CG9243. Experimental Results are based on [3]. X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

Table 1: Optimal Values for M and Σ. Values are used for All Genes. Results are Based on Experimental Data from [3] and Optimization Code, Described in Supplementary Materials.

	μ	σ
+1 Nucleosome	0.29	3.7
+2 Nucleosome	0.8	4
+3 Nucleosome	0.4	4.5

Table 2: Optimal Values for Magnitude (M). Results are based on Curve Fitting Using Experimental data from [3]. NA is used for Nucleosomes, for which we could not extract data.

	Acon	CG31627	CG9246	Mcm10	bur	CG9243	CG9247
+1 Nuc	3000	50	2000	500	5000	1000	8000
+2 Nuc	800	NA	1000	440	8050	200	NA
+3 Nuc	NA	NA	700	550	240	100	NA

As another modeling method, we use RBF Network, in which we first use experimental data of RNAP II reads, from [3], to train our model, and then predict that for other nucleosomes on other genes. In Figures 5, 6, and 7, model training (red) and prediction (black) results are provided in comparison to experimental data (blue), for +1, +2, and +3 nucleosomes, respectively. In this model, we use 15 kernel nodes, which is a good choice for N, to reduce errors.

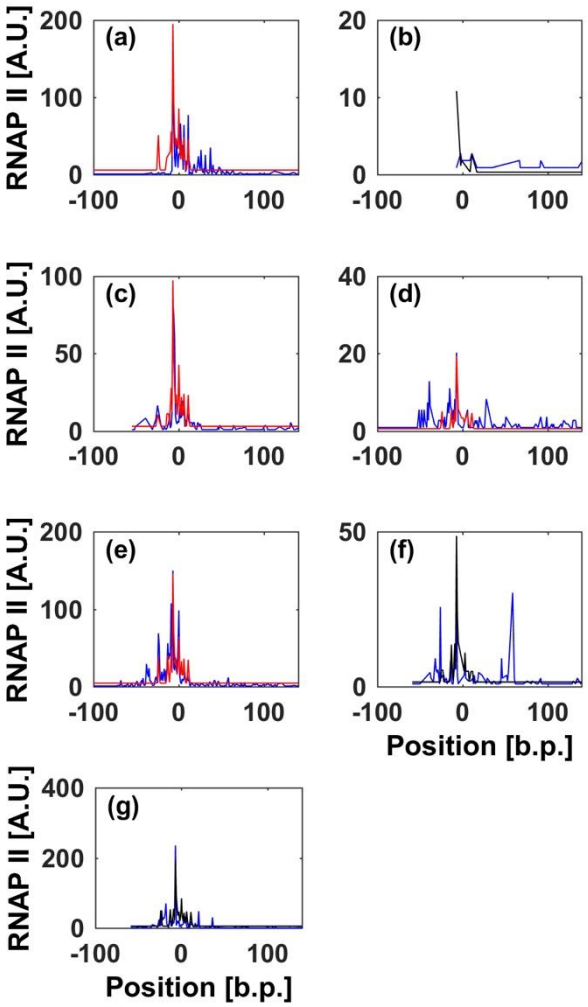

Fig. 5: Modeling RNA Polymerase II Reads when it Approaches +1 Nucleosome, Based on RBF Network (N=15). Results of Model Training (Red) and Prediction (Black) are Compared with Experimental Data (Blue) from [3].(A) Acon, (B) CG31627, (C) CG9246, (D) Mcm10, (E) Bur,(F) CG9243, and (G)CG9247.X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

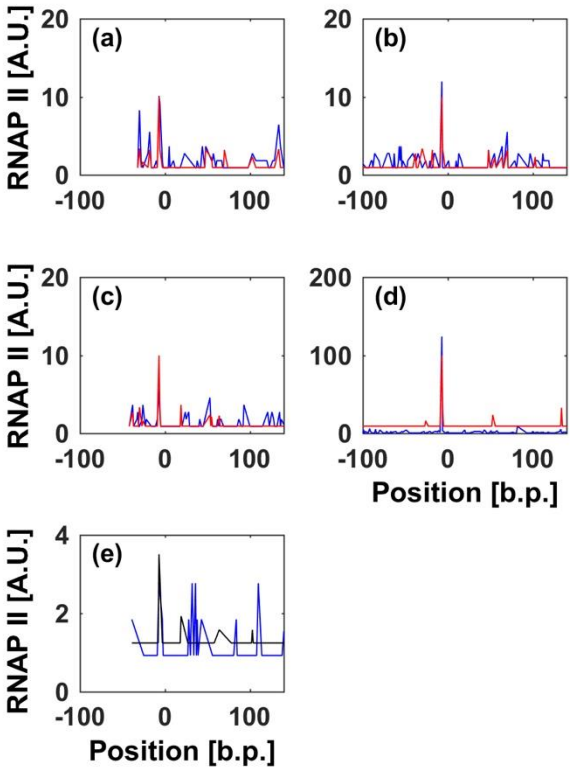

Fig. 6: Modeling RNA Polymerase II Reads when it Approaches +2 Nucleosome, Based on RBF Network (N=15). Results of Model Training (Red) and Prediction (Black) are Compared with Experimental Data (Blue) from [3]. (A) Acon, (B) CG9246, (C) Mcm10, (D) Bur, And (E) CG9243.X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

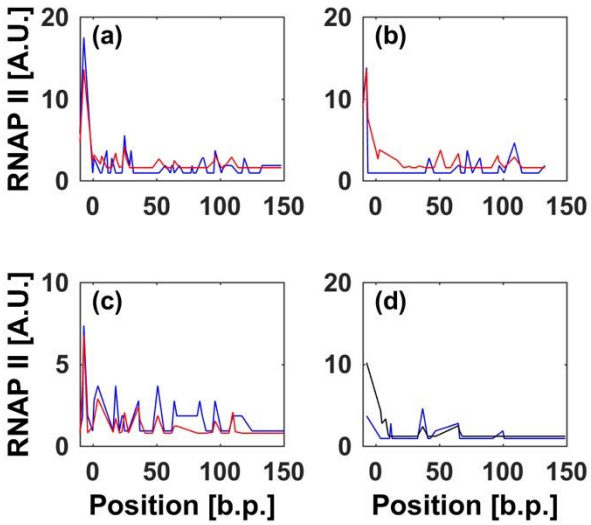

Fig. 7: Modeling RNA Polymerase II Reads when it Approaches +3 Nucleosome, Based on RBF Network (N=15). Results of Model Training (Red) and Prediction (Black) are Compared with Experimental Data (Blue) from [3]. (A) CG9246, (B) Mcm10, (C) Bur, and (D) CG9243.X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

We further simulate the trajectory of RNAP II along the first three nucleosomes of different genes. To do this, we use our random walk model, when we neglect backtracking, and the results are plotted in Figure 8. Supplementary movie S1 illustrates RNAP II movement simulations, when it meets +1 Nucleosome on Acon gene.

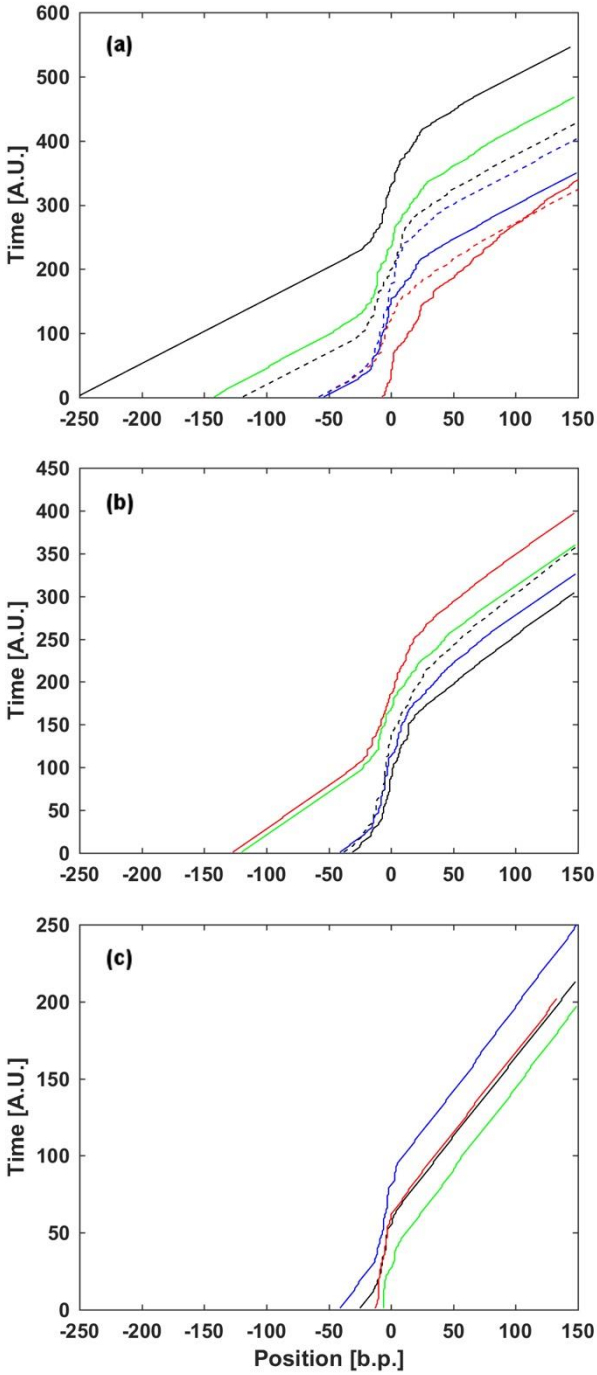

Fig. 8: RNAP II Trajectory Simulation. Simulation Results are Demonstrated by Black, Red, Blue, Green, Dashed Black, Dashed Red, and Dashed Blue Curves for Acon, CG31627, CG9246, Mcm10, Bur, CG9243, and CG9247, Respectively for (A) +1 Nucleosome, (B) +2 Nucleosome, and (C) +3 Nucleosome. X-Axis Stands for RNAP II Position with Respect to the Entry Site of the Nucleosome.

4. Discussion

RNA Polymerase II movement along the DNA is influenced by nucleosomes. We derive models for this effect and simulate RNAP II movement. The process of extracting data from available experimental results is not always easy. Our definition for nucleosome positions is based on RNAP II reads. We do not use MNase-seq data [3], because assuming peaks in MNase-seq data as nucleosome positions, places big RNAP II read peaks downstream nucleosome dyad. This is not consistent with what we expect and experimental findings of [3] (see Figure S1). Another problem, prohibiting us from using MNase-seq data is the existence of dou-

ble peaks, and difficulty in finding the exact position of the nucleosomes.

Several curve fitting methods, such as polynomial or exponential curve fitting are available. However, we chose to use log-normal distribution as a mathematical model to repeat experimental results extracted from [3]. The benefit of this method is simplicity and small number of parameters. To find optimized μ and σ, various methods such as Matlab functions, Matlab statistics toolbox, and minimizing the error (the difference between data and model) can be used. But; error minimization is one of the best methods for our case, in which several data sets need to be modeled by a single fitting parameter set.

There are some points on genes at which unexpected peaks in experimental RNAP II reads are available. Moreover, since our goal is to have a single model for all nucleosomes on different genes, at some points, the error of our log-normal based model is increased comparing with other points. However, in general, the output is in a relatively good agreement with experimental data (See Figures 2, 3, and 4).

We noticed that different μ and σ are required for different nucleosomes. A reason for this could be the difference in RNAP II read data distribution. In particular, RNAP II reads upstream the first peak in the first nucleosome drops faster rather than the second and third nucleosomes, resulting in a smaller μ for the first nucleosome in comparison to the other two (considering the fact that, the more data we have before the first peak, the bigger shift in position we need in modeling). Moreover, the entry of the first nucleosome is a stronger barrier to RNAP II elongation, so the curve has a bigger peak and is narrower, which affects σ.

In order to predict RNAP II reads with less error, we use another model based on Radial Basis Function Networks. This model is able to repeat experimental data more accurately rather than our log-normal model. We analyzed the effect of network size on reducing/eliminating over-fitting and found optimal N for our modeling. Based on these models, we also introduced a random walk model which can predict the trajectory of RNAP II along genes, when it meets different nucleosomes. For including backtracking in our model, we made some assumptions, mentioned in Methods section. As it can be seen in Figure 8, in all genes and all nucleosomes, the slope of curves is increased, just before the entrance of the nucleosomes, represented by 0. This increase in curve slope (speed drop) is because of RNAP II stalling. However, after the time that RNAP II enters the nucleosome, stalling probability decreases and as a result RNAP II gradually moves faster. The movement of RNAP II based on this model is similar to what is expected from experimental data provided by [3]. The same behavior can be seen in supplementary movies S1 and S2.RNAP II moves with a relatively constant speed when there is no nucleosome on DNA; however, approaching nucleosome slows down the RNAP II. This speed reduction continues until RNAP II reaches about 7b.p. upstream the nucleosome entrance. At this point, RNAP II has its minimum speed and then it speeds up again, at a lower rate rather than reduction speed which happens before the nucleosome entry. The same thing happens in different nucleosomes on different genes. But the magnitude of this speed reduction and stalling is different in different nucleosomes and different genes. The strongest barrier is at the entrance of the first nucleosome (+1) and it decreases at other nucleosomes in gene body.

In our future works, we will model the case in which more than one RNA Polymerase II elongate on a single DNA molecule. We also will exclude the exon junctions in our models, to reduce the modelling errors. Another possibility is to consider the initiation rate, which was not considered in the current model, because it is based on the experimental data. Moreover, in future works, we will investigate the ability of our models to predict RNA polymerase II movement on more genes. Furthermore, we will employ some other methods to be able to involve the MNase reads and ChiP profiles into our model, so that it will include factors such as histone modifications and nucleosome composition. Finally, we will use our models for biological interpretations.

Acknowledgements

The authors are thankful to Professor Alexander J. Hartemink, Duke University, NC, USA, for helpful discussions and to Professor Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, for providing experimental data. The authors are thankful for the NIH National Institute of General Medical Sciences - Biotechnology Predoctoral Training Program (T32GM008555) for supporting R. A. N.

References

[1] F. K. Hsieh, M. Fisher, A. Újvári, et al., Histone Sin mutations promotes nucleosome traversal and histone displacement by RNA polymerase II, EMBO Reports 11 (2010) 705–710. http://dx.doi.org/10.1038/embor.2010.113.

[2] A. Újvári, F. K. Hsieh, S. W. Luse, et al., Histone N-terminal tails interfere with nucleosome traversal by RNA polymerase II, Journal of Biological Chemistry283 (2008) 32236–32243.

[3] Ch. M. Weber, S. Ramachandran, S. Henikoff, Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase, Molecular Cell53, 5 (2014) 819-830. http://dx.doi.org/10.1016/j.molcel.2014.02.014.

[4] L. Bintu, T. Ishibashi, M. Dangkulwanich, et al., Nucleosomal Elements That Control the Topography of the Barrier to Transcription, Cell 151,4 (2012) 738–749. http://dx.doi.org/10.1016/j.cell.2012.10.009.

[5] G. Dos Santos, A. J. Schroeder, J. L. Goodman, et al., FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations, Nucleic Acids Research, 43, D1 (2015) D690-D697. http://dx.doi.org/10.1093/nar/gku1099.

[6] N. L. Johnson, S. Kotz, N. Balakrishnan, Continuous univariate distributions,2nd edition, John Wiley and Sons, New York,1994.

[7] N. R. Draper, H. Smith, Applied regression analysis,3rd edition, John Wiley and Sons, New York,1998. http://dx.doi.org/10.1002/9781118625590.

[8] J. Park, I. W. Sandberg, Universal approximation Using Radial-Basis-Function Network, Neural Computation, 3, 2(1991)246-257.