TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MUHAMMAD SHOAIB

ANALYSIS OF TIME LAPSE MICROSCOPIC IMAGES TO UNDERSTAND REAL TIME GENE EXPRESSION KINETICS DURING CELL CYCLE.

Master of Science Thesis

Examiner: Assistant Professor,
Meenakshisundaram Kandhavelu.
Examiners and topic approved by
the Council of the Faculty of Natural
Sciences on 5th March 2014.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Science and Bio-engineering
STUDENT, MUHAMMAD SHOAIB: Analysis of time lapse microscopic images to understand real time gene expression kinetics during cell cycle.
Master of Science Thesis, 48 pages, 6 Appendix pages
November 2014
Major: Bio-imaging
Examiner: Assistant Professor Meenakshisundaram Kandhavelu
Keywords: transcriptional bursts, ON/OFF kinetics, downstream and upstream promoter regions, transcription delayed time, single mRNA detection method.

Series of transcriptional events are one of the central processes that define activity of the cell and population behavior with response to changing environment. The duration of events consisting of active (ON) and inactive (OFF) states regulates the transcriptional burst. However, the exact mechanism of ON/OFF states kinetics over the cell phases in *Escherichia coli* remains to be fully elucidated. We use single mRNA detection method in live cells to study the ON/OFF mechanism in first transcriptional (T_F) and consecutive events (T_C) regulated by P_{lac} and $P_{lac/ara1}$. A programme is written that can measure the time duration after cell division and appearance of first RNA molecule. Further the code is extended to determine burst size and probabilities. We determine that the duration of T_F ON/OFF is inversely proportional and are opposite to T_C. Both states behave differently over the cell cycle in T_F resulting in reverse exponential trend. Transcription initiation from downstream and upstream promoter regions affect the dynamics of ON/OFF states in fast and slow dividing cells correspondingly. Duration of T_F ON state predefines the behavior of T_C by modulating number and frequency of mRNA being produced. Moreover, we show that delayed OFF time in T_F affects dynamics of both states in T_C. Delay in production of T_F burst is ruled by upstream promoter region. We further independently validated the transcription delayed time by elongation arrest experiments which shows that mRNA noise in T_C is determined by delayed OFF period in T_F. Finally we observed that noise modulation depends on transcription initiation region. Thus, we claim that T_F regulates the future cell to cell diversity of the population.

PREFACE

This master's thesis work was conducted at the department of signal processing, Tampere university of Technology. First of all I would like to thank God almighty for giving me the courage to carry out this research work, and then I would like to express my deepest gratitude to my supervisor Meenakshisundaram Kandhavelu and assistant professor Andre S. Ribeiro for giving me the opportunity to work in one of the vibrant lab of computational system Biology.

I also want to acknowledge my team members for their continuous support during the entire research period. It has been an incredible experience to work in molecular signaling lab that actually expanded my scientific knowledge and research experience.

This thesis work was funded by the department of signal processing therefore I am extremely grateful to the higher authorities.

Last but not the least, I would like to thanks my parents and family for their unconditional support. They have been a source of encouragement and motivation throughout my life.

Tampere, Finland, November 11, 2014

Muhammad Shoaib

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

Acc	Acceleration
aTc	Anhydrotetracycline
Au	Arbitrary unit
cAMP	Cyclic adenosine monophosphate.
CAP	Catabolite activator protein
CLSM	Confocal laser scanning microscopy
CRP	CAMP receptor protein.
CV^2	Coefficient of variation square
D_0	Cutoff frequency
DNA	Deoxyribonucleic acid
DRS	Downstream regulatory sequence
E-coli	Escherichia coli
Exp	Exponential
FCS2	Focht Chamber System 2
FISH	Fluorescence In Situ Hybridization
HPF	High pass filter
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LAC OPERON	Lactose operon
LB medium	Luria broth media
LPF	Low pass filter
mRNA	Messenger ribonucleic acid
MS2 FP method	MS2 Fluorescence protein method
NTPs	Nucleoside triphosphates
OD_{600}	Optical density measured at 600 nanometre wavelength
OFF_D	Delay in transcription inactive period

ON_D	Delay in transcription active period
PCA	Principal component analysis
P_{lac}	Lactose promoter
$P_{lac/ara1}$	Lactose variant promoter
Retd	Retardation
RNA	Ribonucleic acid
RNAP	Ribonucleic acid polymerase
RPM	Revolutions per minute
Stat	Stationary
Std	Standard deviation
T_C	Consecutive transcription event
T_F	First transcription event
Toff	Transcription inactive period
Ton	Transcription active period
URS	Upstream regulatory sequence
URS/DRS	Upstream and downstream regulatory sequences

Chapter 1

INTRODUCTION

Living cells possess stochastic nature of gene activation-inactivation (Golding et al. 2005; Raj and van Oudenaarden 2009). The kinetics of activation-inactivation are dependent on several mechanism such as activator and repressor, DNA structural confirmations, RNAP binding and fall off (Choi et al. 2008; Malan and McClure 1984). All these events are spanning transcripts production. The sequence of the transcriptional events in a cell is one of the central mechanisms that define activity of the cell over the phases and is an important factor for regulation of cell population behavior under altering environment (Kandhavelu et al. 2012c; Raj and van Oudenaarden 2008; Elowitz et al. 2002; So et al. 2011). Recently, we have shown that the mRNA production time is a crucial step in determination of cell to cell diversity (Kandhavelu et al. 2012c, 2012a). Moreover, the distribution of first production event is different from the consecutive ones, which is highly dependent on inducer intake time (Mäkelä et al. 2013). On the other hand, it has been reported that the first transcriptional event is delayed due to the molecular memoryless process compared to the consecutive events, explained as a transcriptional memory process (Yurkovsky and Nachman 2013; Wolf and Arkin 2003; Ribeiro 2010). It indicates that the flow of sequential transcriptional events depends on primary event. Thus, studying the molecular mechanisms of first mRNA production is a crucial in understanding behavior of the cells in population over the phases.

In *Escherichia coli*, gene active and inactive periods, otherwise called ON and OFF respectively, are studied using single mRNA detection methods, FISH and other biochemical methods (Golding et al. 2005; So et al. 2011; Malan and McClure 1984; Lutz and Bujard 1997). In case of *Lac* operon, Contesse *et al* reported that this operon has the periodicity of 40 seconds (Contesse et al. 1969). Later, Baker and Yanofsky observed that ON follows after the long OFF period (Baker, R.F., and Yanofsky 1968). In the same report it is noted that OFF time is longer with factor of five. Later, single molecule time series experiments on lac variant gene reported that the mean OFF period is six times greater than ON during which multiple mRNAs are being produced in burst following a Poisson process. More recently, the study on the mechanism of transcriptional bursting in bacteria reported that mRNA copy number distribution of lac and other 19 genes are well fitting with the Poisson distributions (So et al. 2011). It is also noted that kinetics of mRNA production determine the protein burst (Choi et al. 2010). The fluctuation in ON/OFF, otherwise called two-state model, determines the kinetics of mRNA production and copy-number statistics which leads cell to cell diversity (Golding and Cox 2004; Golding et al. 2005; So et al. 2011; Sanchez and Golding 2013).

Individual cells must activate their genes based on the need in a given environment while not all the genes are being activated in a given moment. It might be explained by the lack of macro molecules such as RNAP, NTPs and Mg^{2+}, involved in transcription activation (Kandhavelu et al. 2012b; Klumpp and Hwa 2008; Gaal et al. 1997). For example, immediate after the division cells have to use partitioned RNAP besides preparing enough RNAP molecules to transcribe the necessary genes (Klumpp and Hwa 2008). In eukaryotes it has been reported that birth or death of each molecule involves several small steps which create the memory between individual events and delay in events has consequences on the mRNA burst size (Pedraza and Paulsson 2008). The similar behavior is also observed for *lac* promoter, in *E.coli* (So et al. 2011; Golding et al. 2005; Choi et al. 2010). Using MS2-FP and FISH it has been reported that changes in ON and OFF kinetics determines the cell to cell diversity (So et al. 2011). However, we do not have a biophysical understanding of the nature of the first transcriptional event and the role of ON and OFF states over the cell phases of bacteria. Thus, we aim to define the effect of ON and OFF states in first transcriptional event onto consecutive productions of lac and lac variant promoter and how they are determined by upstream and downstream gene sequences.

From this study, we address ON and OFF kinetics transition over the phases using single mRNA detection method, MS2-FP method (Golding and Cox 2004; Golding et al. 2005). Advantageously, this method allows the observation of single mRNA production and burst events in live cells. Using this method, we recently reported that activation of upstream and downstream sequence of the promoter with different concentrations of inducers can have different function in the transcription kinetics (Kandhavelu et al. 2012c). In the present work we used lac promoters to study the kinetics of ON and OFF. The lac promoter which we used consists a Catabolic Activator Protein-CAP site in the upstream for the binding of CAP and the Operator 1 site in the downstream for the binding of lac repressor (Golding and Cox 2004). CAP and Lac repressor are inducible by cyclic AMP (cAMP) and Isopropyl -D-1-thiogalactopyranoside (IPTG) respectively. To generalize our observations we also used another promoter, lac/ara1, a variant of lac (Lutz and Bujard 1997; Lutz et al. 2001; Golding et al. 2005). Upstream activation of this promoter depends on Ara C/L-arabinose complex instead of CRP/cAMP complex in wild type. We also observe burst from time series experiments and show how first burst event is not consistent with the consecutive events. Finally, by quantifying mRNA numbers produced from every event we extend our conclusion by explaining influence of ON and OFF kinetics source in mRNA variation over the phases.

More precisely, the main objective of this thesis is to analyze the time series data obtained from the microscopic images. In order to study the first and consecutive transcription events at single cell level we applied image processing methods such as segmentation and thresholding for the detection of cells and RNA spots. The intensity value of single mRNA molecule is determined and by using jump detection method we determined RNA numbers at different time instants. Furthermore, we develop a code

that can be used to determine the time duration after the cell division and the appearance of first RNA molecule. For finding burst size and probabilities another Matlab code is written and it is described in the method part of thesis. The ON/OFF kinetics during cell phases was calculated using production interval information.

Chapter 2

FUNDAMENTALS OF DIGITAL IMAGE ANALYSIS

2.1) Digital image

Image can be defined as a two dimensional function f(x,y) where x and y are the spatial coordinates and the amplitude of function, f(x,y) indicates the intensity value at a particular point. More specifically, in a digital image the spatial coordinates and the amplitude of function are finite and discrete, therefore the digital image is produced by the process of digitization of a continuous optical image. The optical image is sampled on a rectangular grid and then the sampled values are quantized to produce integer values in a rectangular array. Digital image consists of finite points termed as pixels with specific location in spatial coordinate system and a nonnegative integer value. Here it is worth to mention, most of the content of this chapter is adopted from the book of, Rafeal C. Gonzalez, Digital image processing.

2.2) Image enhancement techniques

Image enhancement is a process to improve the overall look of the image so that certain attributes of image are visible and clear to the observer. This technique is quite effective when the data is too noisy and difficult to visualize. The improved and enhanced version of original image is given as an input to the image processing algorithms for further analysis.

The method of image enhancement can be broadly classified into two major categories.

2.2.1) Spatial domain methods

2.2.2) Frequency domain methods

2.2.1) Spatial domain methods

In spatial domain techniques the mathematical operations are applied directly on the pixels to obtain desired enhancement. Whereas, in frequency domain methods we first compute the Fourier transform of image and convert the data from time domain into frequency domain. Further the enhancement operations are applied in frequency domain to improve the visual characteristics of image and by computing the inverse Fourier

transform, original image with improved attributes is obtained. Simply image enhancement can be represented by the following expression.

Output image= Transformation (Input image)

$$S = T(r) \tag{2.1}$$

Here ‘S’ and ‘r’ indicates pixel values in output and input image respectively and T is the transformation method used.

The spatial domain methods can be further classified as,

1) Point processing

2) Thresholding transformation

3) Dynamic range compression

4) Grey level slicing

5) Histogram equalization

1) Point processing

This method is based on the intensity of one pixel, not on the neighborhood pixel values. Creating negative of an image is one example of point processing enhancement methods. It’s the most simple and basic technique in digital image processing. Each pixel intensity value is subtracted from the maximum intensity value. The negative of an image can be computed as,

$$O(r,c) = 255 - I(r,c) \tag{2.2}$$

Here O(r,c) is the negative of an image which is output and I(r,c) is the input image.

2) Thresholding transformation

This method is useful when we want to separate an object from the background. In order to perform this method we need to select certain threshold value. The pixel values which are greater than threshold value are retained whereas values less than threshold are set to zero.

$$s(r) = 1, \text{ if } s(r) > t \tag{2.3}$$

$$s(r) = 0, \text{ if } s(r) < t$$

3) Dynamic range compression

When the range of input grey values is very small, log function maps grey values into a wide range of output values. Inverse log transformation is used in the opposite situation, that is mapping wide range into narrow range intensity values, mathematically,

$$S = c \times \log (i+r) \quad (2.4)$$

4) Grey level slicing

It is equivalent to the band pass filtering in frequency domain. The enhancement method is applied only the region of interest or certain range of intensity values. Following figure is showing grey level slicing.

5) Histogram equalization

The histogram of a digital image gives us information, how the intensity values are distributed. If the intensity values are concentrated in one region, it shows image has less contrast. High contrast images have histograms which are uniformly distributed. In histogram equalization our objective is to enhance the contrast of an original image. Suppose we have a dark image and the intensity values are skewed towards the left side of the histogram.

2.2.2) Frequency domain methods

The idea of image enhancement in frequency domain is simple and straightforward. Firstly fourier transform of the image to be enhanced is computed and then the result is multiplied with the filter (Fig 1.1). Finally, the inverse fourier transform gives us the enhanced image. The transfer function of the filter can be designed and implemented directly on the fourier transformed image,

$$G(x,y) = H(u,v) \cdot F(u,v) \quad (2.5)$$

Frequency domain methods can be classified as, Low pass filtering and high pass filtering

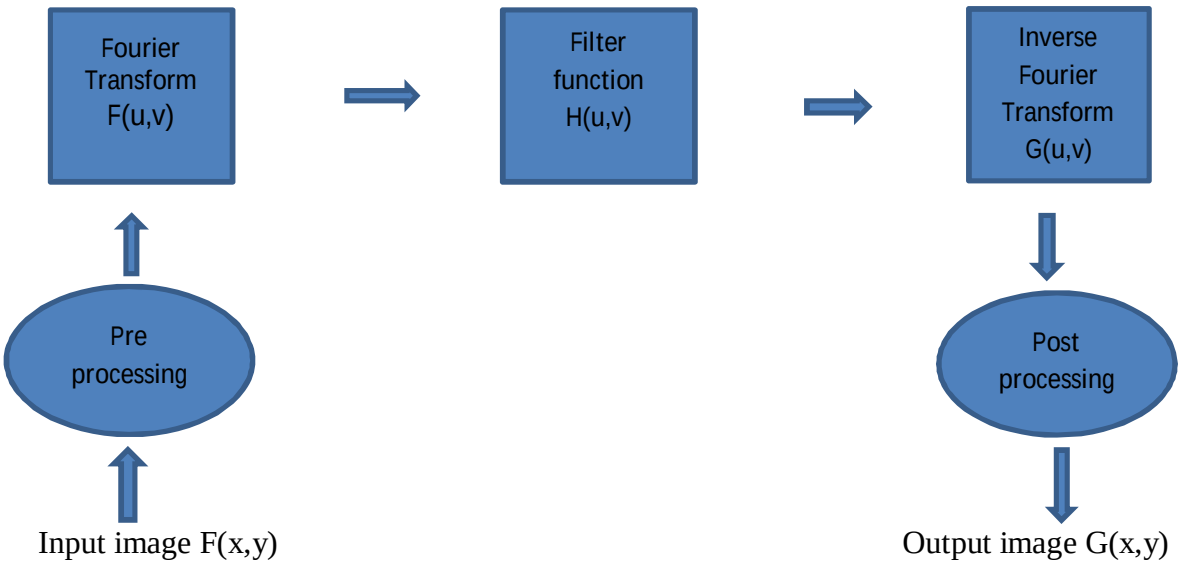

Figure 1.1. *Flow chart showing basic steps of image enhancement.*

1) Low pass filtering

Low pass filtering removes high frequency components of image and allows only low frequency values. The edges and sharp transitions of the intensity values in spatial domain corresponds to high frequency signals in the frequency domain, whereas uniform gray values represents low frequency components of the image. The purpose of low pass filter is to smoothen the image by attenuating high frequency values. For the sake of simplicity we are considering only real and symmetric filters.

The transfer function of ideal low pass filter with cutoff frequency D_O can be given as,

$$H(u,v) = 1, \text{ if } \sqrt{u^2 + v^2} \leq D_O \quad (2.6)$$

$$H(u,v) = 0, \text{ if } \sqrt{u^2 + v^2} > D_O$$

The amount of frequency values passed by the ideal LPF is determined by the cutoff frequency D_O. The value of cutoff frequency is selected in such a way that the signal which is of interest is passed through filter while others are discarded or set to zero. In order to establish standard value of D_O we need to compute circles which enclose a specific fraction of total image power.

$$P_T = \sum_{V=0}^{N-1} \sum_{u=0}^{m-1} P(u,v) \quad (2.7)$$

Let's consider a circle having radius $D_O(\alpha)$ as a cutoff frequency value and α is the threshold value, such that $\sum_u \quad \sum_v P(u,v) = \alpha P_T$.

After adjusting the threshold value we can find the appropriate value of cutoff frequency. Butterworth lowpass filter and guassian filter are commonly used in digital image processing,

The transfer function of two dimensional butterworth low pass filter is,

$$H(u,v) = 1/(1 + [\sqrt{u^2 + v^2}\ / D0]\ )^{2n} \quad (2.8)$$

Here the transitions in the frequency response are not sharp as compared to the ideal low pass filter and it's more suitable for image smoothing as it avoid the problem of ringing.

Another example of low pass filter is Gaussian filter. The main objective of using this filter is to blur the image and removal of noise. In two dimensions the transfer function of this smoothing filter is given by

$$H(u,v) = 1 - e^{-D^2(u,v)/26^2} \quad (2.9)$$

Here D(u,v) is the distance in the frequency plane from the origin and the value of ϭ determines the standard deviation of the Gaussian curve. The cutoff frequency would be greater for larger values of ϭ, hence there would be less filtering.

2) High pass filtering

High pass filter blocks low frequency signals and permits high frequency components of image. The output of a high pass filter is the sharper image with enhanced details. As low pass filter smoothen the image and reduces noise, the HPF does the opposite and amplify the noise.

An ideal high pass filter with cutoff frequency D_0 has transfer function,

H(u,v) = 0, If $\sqrt{u^2 + v^2} \leq D_0$ (2.10)

H(u,v) = 1, If $\sqrt{u^2 + v^2} > D_0$

Butter worth high pass filter with filter order 'n' and cutoff frequency D_0 can be defined as,

H (u,v) = $1/(1 + [D0/\sqrt{u^2 + v^2}\,])^{2n}$ (2.11)

There are no sharp transitions in the frequency response as in the ideal high pass filter. It is more appropriate for image sharpening as it does not introduce ringing effect.

Second example of high pass filtering is Gaussian high pass filter, In two dimensions it can be defined as,

H (u,v) = 1- $e^{-D^2(u,v)/26^2}$ (2.12)

The ϭ values measures the dispersion of Gaussian curve. There would be higher filtering with greater value of ϭ and cutoff frequency.

2.3) IMAGE SEGMENTATION

The categorization of an image into meaningful objects or structures is termed as segmentation. It is an essential step in the image analysis and visualization of data. If segmentation is performed well and desired results are obtained then the further steps of image processing would be easier. Segmentation algorithms can be applied directly on the original images before any preprocessing step or after the application of filters and

enhancement techniques. In general, image segmentation can be classified into three types, threshold based, edge based and region based segmentation methods

2.3.1) Threshold based segmentation

It's the most commonly used technique for segmentation of an image. Based on the threshold value each pixel is assigned a category of either 0 or 1.

$$G(v) = \begin{cases} 0 \; if \; v < t \\ 1 \; if \; v > t \end{cases} \tag{2.13}$$

Here 'v' is the grey value and 't' is the predefined threshold value. After thresholding, the image is converted into binary image with two pixel values 0 and 1. This method is most effective when we have to segment an object from the background.

a) Selection of threshold

One of the simplest methods of selecting threshold is with the help of image histogram. The maximum frequency values or the histogram peak enable us to find appropriate threshold value. The threshold value 't' lies in the middle of two peaks with grey values g_1 and g_2

$$t = (g1 + g2)/2 \tag{2.14}$$

Also, it may be the minimum grey value between two peaks. Most of the times we need to smooth the histogram for finding global extrema. For smoothing histogram the commonly used methods are Gaussian convolution or discrete averaging kernel.

By considering uniform grey values of pixels to be segmented, the threshold value can be selected so that variance is minimum within pixels of object. Otsu thersholding method is based on the same principle.

b) Otsu algorithm

According to this method the image to be segmented consists of two pixel classes such as foreground and background. The value of threshold is selected so that intraclass variance is minimized or in other words the interclass variance is maximum. This methods is best to use when we have bimodal distribution of grey values. The total variance can be expressed with the following mathematical formula.

$\sigma^2 = \sigma^2_w(t) + \sigma^2_b(t)$

$\sigma^2_w(t)=q1(t)+ \sigma_1^2(t)+q2(t)+ \sigma_2^2(t)$

$\sigma^2_b(t) =q1(t) [1-q1(t)] [\mu_1(t)- \mu_2(t)]^2$ (2.15)

Here,

σ^2_w is within class variance and σ^2_b is variance between classes.

$q_{1,2}(t)$ are class probabilities and $\mu_{1,2}(t)$ are class means.

There are few drawbacks of Otsu thresholding method, for example it assumes the image consists of only two pixel classes therefore, it cannot perform well when values are in wide range and consists of many classes or the image with variable illumination.

c) K means clustering algorithm

This method partitions the image into K segments by minimizing class variance within each segment. The variable K is defined before running the algorithm. The variance within each segment of image can be defined as,

$\sigma_w^2 = \sum_{i=0}^{k-1} h_i \sigma_i^2$ (2.16)

Here h_i is the probability of the random pixel belonging to a particular image segment. The variance of the grey values of the segment is represented by σ_i^2.

2.3.2) EDGE BASED SEGMENTATION

This method is based on finding pixels of object boundaries and then linking these edges to form contours. The edge based segmentation methods are preferred in many cases due to less complexity of algorithms. The edges are important features of object and corresponds to large changes in the intensity values. At the boundary regions of objects the gradient in the intensity values is quite noticeable and this variation of intensity values forms the basis of edge detection step. Firstly, the edges of the object are determined and these edges are then connected to form close boundaries. In the last step, the inner regions of boundaries are filled that represents the required segmented image. The second step of edge connection is difficult because it requires the removal of those edges which are caused by artifacts and noise. In those rare cases where edges are easy to separate, the following algorithm can work best for segmentation,

1) Suppose we have an image *f*, first we can compute the boundaries of object, delta f, with the help of a gradient operator.

2) Next, based on the threshold values only the edges are preserved and now the image is binary, other pixels other than edges are set to 0, whereas edge pixels are set to 1.

3) Further, the laplacian of image Δf is computed from image *f*.

4) Finally the segmented image is obtained by computing, g = delta $(f)_t$. sgn(Δf). Here the operator of sgn return the sign of its argument.

Edge linking is very important and difficult step and most of the time it does not give perfect boundaries that can be directly used for segmentation. Most of the time, false edges are detected also with the problem of gaps. In case of incomplete edges, edge linking is essential to perform to get close boundary of object.

Hough transform is the fundamental approach for finding objects with simple shapes such as lines and circles. Finding complex objects demands more effort and time, therefore its major disadvantage is its inability to tackle complex and variable objects. Another way of edge linking is neighborhood search method. The starting edge pixel is first selected and then the second pixel is attached based on the closest edgeness value direction. The searching process and linking continues until the image boundary is completed.

2.3.2.1) WATER SHED SEGMENTATION

As the name suggests, this method segments the different image regions into catchment basins. Let's say function 'f' represents continuous height function over an image domain, then the catchment basin can be defined as the set of point whose path meet at the local minimum of 'f'. The underlying idea of this method comes from the geography; there is some uneven landscape region which is flooded by water. Watersheds are the partitioning lines which are dividing the landscape into different regions. Catchment basins starts filling up from the local minima. After some time, a point is reached where water from separate basins meet, here dams are built. The process is stopped when the water is reached to the maximum level. The expression watershed transform express the labelling of the points of each catchment basin. Each label is distinguished from the labels of other catchment basins. In practice, it is not possible to apply watershed transform directly on the original image, but can be applied to the morphological gradient of image. Watersheds are formed at the grey level discontinuities which is required for image segmentation.

2.3.3) REGION BASED SEGMENTATION

In region based segmentation approach objects are isolated by establishing the regions they covered. Its main goal is to subdivide an image *I* into regions *R*. Theoretically,

finding an object using edge based or region based segmentation approach gives us the same result. Boundary and region are two different illustration of same object. However, in practical situations, using edge based segmentation process may give different results as compared to region based technique. This is due to the fact; the images and procedures we are dealing with are imperfect and are not ideal. There are five formulations for region oriented segmentation, suppose we have image *I*, and it is segmented into regions R_1, R_2, R_3, R_n, the following conditions must be satisfied.

Each pixel should belong to some region $\bigcup_{i=1}^{n} Ri = R$

All the points of region Ri should be connected in some sense, where *i*= 1,2,3…n

All of the regions should be disjoint, $R_i \cap R_j = \emptyset$, *for* $i \neq j$

Pixels of a region should satisfy atleast one common property *P*, $P(R_i) = True$ for i= 1,2,3…n

Two different regions should be inhomogeneous as compared to each other, both have different properties. $P(R_i \cap R_j) = False$

Region based segmentation are based on two simple operations, i-e, splitting and merging. In merging, initial segmentation is performed first and then segments are merged based on similar grey scale values or close pattern of segments. Whereas in splitting, after the initial segmentation each segment is splitted, if there is high variance among pixel values or inhomogeneity exists within each segment. The process is repeated until uniform segments are attained.

Region growing

This is one of the merging method in which adjacent pixels having higher similarity are merged together and form segments. At first a set of starting pixels known as seeds are formed. The seeds are selected by human or automatically excluding high intensity pixels. Taking one seed at a time and neighborhood search is performed, Out of four neighborhood pixels, those which are very close in intensity values to the seed are merged. Further the first seed is removed from the set and same procedure is implemented on other seeds. At the end the seed set becomes null with no values.

This method is fast and based on a simple concept. There are also few disadvantages associated with it, such as its sensitivity to noise and time consumption. Also, this method is local therefore global view of problem does not exist.

2.4) PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a statistical technique used to find patterns in data and representation of information in such a way that the similarities and differences are

highlighted. The major advantage of this technique is data compression by reducing the number of dimensions once the patterns are identified.

Method

We have applied this powerful tool on the microscopic images to find the size and orientation of bacterial cells. As we are dealing with images, the information is in the form of pixel intensities which are stored in the matrix form.

1) Normalization of data

For the proper working of PCA first of all we need to normalize the data. Let's say we have three dimensional data x, y, and z. After the calculation of mean values for each dimension, we subtract them from the corresponding data points.

2) Covariance matrix

The measurement of variance between two dimensions is the covariance. For the three dimensional data the covariance matrix would be of 3x3 dimension.

$$\begin{matrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{matrix}$$

The formula for covariance is,

$$\mathrm{cov}(x,y) = \sum_{i=1}^{n} \frac{xi-\bar{x}}{yi-\bar{y}} \qquad (2.17)$$

3) Eigen vectors and eigen values of covariance matrix

Eigen vector 'e' when multiplied by a square matrix 'X' results another vector which is the integral multiple of same vector 'e'.

$$X.e = k.e \qquad (2.18)$$

Here 'e' is the eigen vector and 'k' is the eigen value.

Next step in PCA is the calculation of eigen vectors and eigen values of the covariance matrix.

4) *Assigning of principal components*

The eigen vector with the greatest eigen value indicates data points with greater importance, i.e, PRINCIPAL COMPONENT, and represents certain feature or pattern

of data. The eigen vectors are then ordered based on eigen values from highest to lowest. Data compression takes place at this step by ignoring data points with least eigen values. The feature vector 'V' is simply the matrix formed by the combination of eigen vectors.

'V'= [e1 e2 e3 …..en] eig value1 > eig value 2 > eig value 3 (2.19)

5) Obtaining new data values

This is the last step of PCA. After the formation of feature vector, 'V', the final data is obtained by multiplying the transpose of feature vector and original data set.

New data = V^t x $(\text{original data})^t$ (2.20)

6) Retrieval of old data

The old data with less dimensions is obtained back by taking inverse of feature vector and multiplying it with new data,

Original data = $(V^t)^{-1}$ x New data (2.21)

Chapter 3

MATERIALS AND METHODS

This part of thesis includes the experimental procedure and image analysis techniques that were used to study the real time gene expression kinetics of bacterial cells. All the experiments were performed in the computational system biology lab, department of signal processing, TUT. The sequence of steps to measure ON/OFF kinetics of first and consecutive transcription events are described below.

3.1) Selection of strains and media

Escherichia coli Dh5α-PRO strain was used as a host to study the gene ON/OFF kinetics over the phases. For the generation time and phases determination, *E.coli* DH5a-PRO strain was grown overnight in an orbital shaker (Labnet) at 30°C with aeration at 250 RPM. Following overnight culture, cells were diluted in fresh LB media to reach an optical density (OD_{600}) of 0.05 using a spectrophotometer (Ultraspec 10-Amersham Biosciences). Cells were grown at 37°C with aeration at 250 RPM and cell growth was monitored every 30 minutes. OD_{600} values were used to calculate the generation time as described in earlier literature (POWELL 1956). Cell phases were determine as suggested from Jacobs Monod study based on the cells division time, i.e. lag, acceleration, exponential, retardation and stationary phases (Monod 1949).

3.2) Confocal laser scanning microscopy

With the help of confocal microscopy we observed live bacterial cells and it is also possible to create three dimensional images with this microscopic method. Another advantage of using this technique is that, the dynamic properties of cells can also be observed. The term confocal means the arrangement of two lenses in order to focus the same point. Due to the presence of optical pinholes only focused light can reach the detector and out of focus light is removed by using pinhole at suitable position (Figure. 2.1).

With the confocal principle, laser light scanning system was employed. The laser beam scanned the sample in x and y directions. The gray scale image is displayed on computer screen with the intensity values ranging from 0 to 255. When the focus is shifted by a fixed amount the object is scanned at different z position. The data is stored to reconstruct the final 3D image. In order to provide some form of contrast

between different regions of the object, fluorescence proteins were used. Here in our experiment we used green fluorescent proteins. The major advantages of confocal microscopy are better resolution, greater sensitivity as compared to other microscopy methods and 3d reconstruction of images.

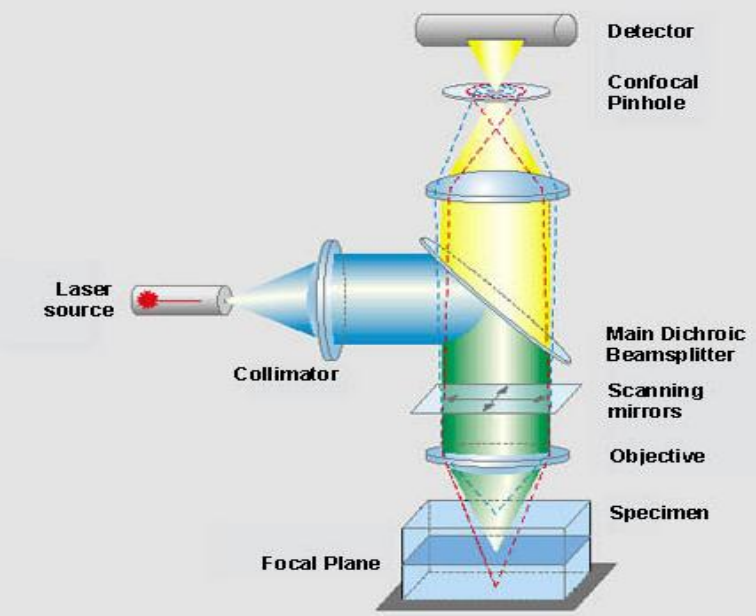

Figure 2.1. *Schematic diagram of light path in CLSM, (Laboratory of biosensors and bioelectronics, University of Zurich, UZH)*

3.3) Fluorescence microscopy of live cells

For real time monitoring of mRNA production, reporter and target promoters were activated. Cells were then placed between a 1 % LB-agarose gel pad and a coverslip in a FCS2 temperature-controlled perfusion chamber (Bioptechs), maintained at 37 ˚C during the measurement. The pad was supplemented with required inducers to maintain the induction conditions during time-lapse microscopy. When the reporter and target RNA are coexpressed, MS2d-GFP binds with RNA and bright spots are formed. For the image acquisition of the produced mRNA molecules in the cells, inverted fluorescence microscope (Nikon, Eclipse Ti-E) was used with 100x N.A. 1.49 oil immersion objective. The microscope was equipped with a hardware autofocus module, motorized z-drive and Nikon's Perfect Focus System to maintain the cells in focus during the image acquisition. Images were acquired by interval of one minute using inbuilt microscope software (Nikon, NIS-Elements C). Fluorescence was measured using a 488 nm laser (Melles-Griot) and a 515/30 nm detection filter. Images were captured after each one minute for about two hours.

From the JPG format, images were converted into TIF format and then analyzed in Matlab 2013-b.

3.4) Image analysis

Time series images were analyzed in several steps which are explained as follows.

1) Drift correction

There was slight change in the position of same cells in the adjacent frames and was corrected by using cross correlation of the adjacent frames. This method enabled us to find the number of shifted pixels in x and y axis. As this process computes integer number of pixel, round off error can accumulate. This error can be removed by the comparison of each frame with the previous frames and calculating the average shift. The process is performed for all the time series images and those pixels which are common to all frames are computed. By considering the amount of drift which occurred from the beginning, the frames are cropped based on the common region. This preprocessing technique of drift correction made the tracking of cells effortless over the time.

2) Cell detection

After drift correction cells were manually masked with solid and unique colors using Paint.net. The time series pipeline considered the information of only masked cells. Cell division was also taken into account by masking daughter cells with different colors. The statistical procedure of principal component analysis (PCA) was further applied to obtain the dimensional characteristics of bacterial cells. The method of PCA assumes that the intensity of pixels is uniform across the cell region and null in the background. PCA enabled us to calculate the centroid and area of each cell in time series data. Principal components of each masked region were obtained using the intensity distribution as a probability distribution function. The major axis of the ellipse was considered as the first principal component whereas the minor axis of the ellipse was taken as the second principal component.

3) Spot detection

After cell detection, MS2-GFP tagged RNA molecules were segmented using Kernel density estimation also known as parzen windows as proposed in [Chen et al, 2008]. It’s a non-parametrical statistical method used to estimate the underlying distribution of a continuous random variable. It is non parametric because it is not based on any presumed probability density function. From the estimated density function each pixel of the cells was classified as either spot or cell background. Based on the preselected threshold classification was performed.

Further we calculated background corrected intensities by subtracting the background intensity from the spot intensity. After this, these intensity values were summed for each cell in time series analysis.

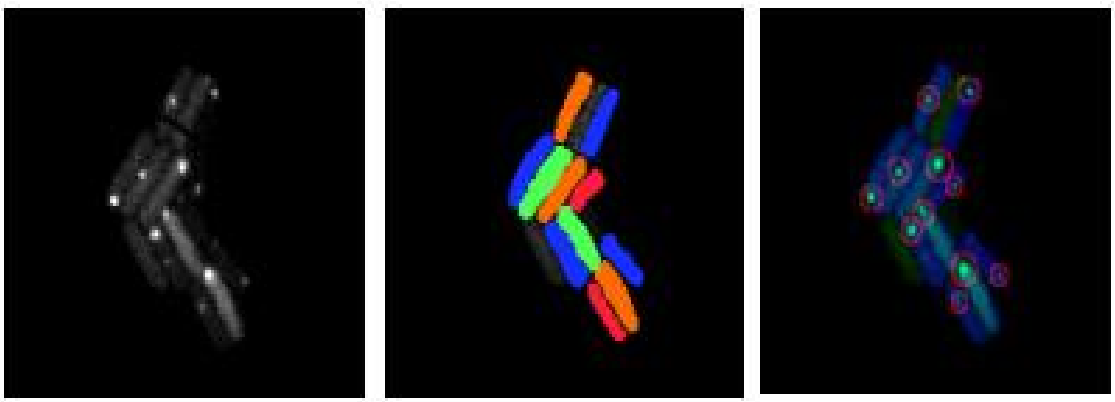

Figure 2.2. *MS2-FP-tagged mRNA molecules in E. coli. Unprocessed image of individual cells shows the produced mRNA molecules as a bright spot in gray scale (left). The corresponding cells are segmented by manual painting (center).mRNA molecules were segmented (red circle) using image processing method as reported in methods section (right).*

4) Jump detection

One of our objectives is to find the production interval after cell division and first transcription event and consecutive events in both parent and daughter cells. Here we assumed the fact that MS2d-GFP tagged RNA molecules are not degrading with the passage of time or in other words RNA molecules are not decreasing while measuring them.

By using this assumption, the new RNA molecules were calculated by fitting monotonic piece wise function on the background corrected spot intensities [Kandhavelu M et al, 2012]. In this procedure the curve is fitted by finding the minimum squared difference between the measured intensity and fit intensity values. F test was used for finding number of terms in piece wise function. With the threshold of 0.01 we rejected lower order curves and accepted higher order as it is required in F test. The process results of this method are explained with the help of figure in result section.

3.5) Quantification of single mRNA

To quantify the single mRNA intensity we tracked the production of first mRNA in weak induction. For this, cells were grown at 37 °C at 250 RPM, with the aTc for reporter activation and without inducers for the target mRNA production. From the time

series images, we monitored the production of first mRNA spot for fifty random cells. Intensity value of first RNA spot and mean value for fifty observations were then calculated which represents the intensity value of one RNA (Golding and Cox 2004). The mean intensity of single *m*RNA is found to be,

Intensity of single *m*RNA = 1.056 au

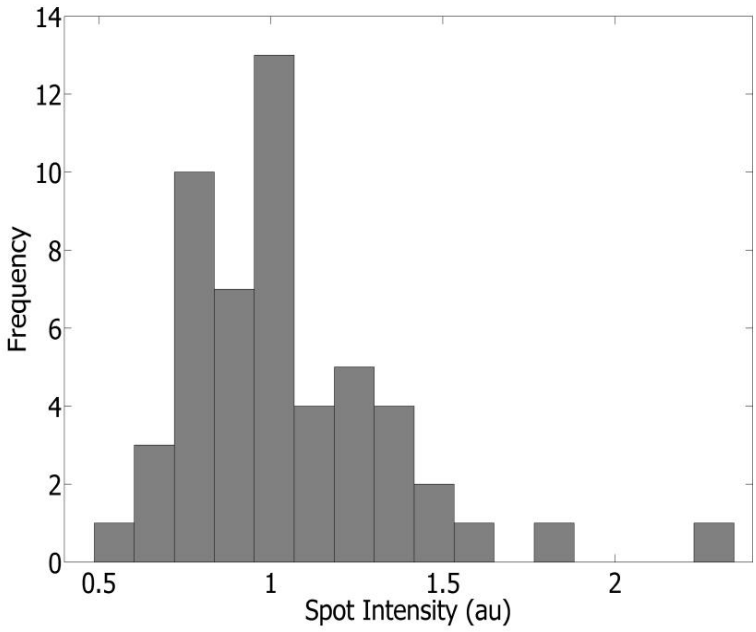

Figure 2.3. *Fluorescence intensity distribution of single mRNA (P lac -LacZα - 96xbs) molecule coated with fluorescence proteins (MS2-FP). The distribution represents the frequency of fluorescence intensities of fifty individual spots of single mRNA produced inside E.coli cells at low expression level.*

Combining information from jump detection method, cell background-corrected spot intensities and mean intensity of a single mRNA molecule allowed us to calculate the number of mRNAs produced from the transcription events. To study the transcriptional bursting behavior over phases we investigated the transcription events of individual cells. Number of RNA in each burst was calculated by,

Fit intensity (t_{n+1}) - Fit intensity (t_n) / 1.056 , where n= 1,2…….120 (3.1)

Further we rounded off RNA numbers to the integer values by using the following rounding off scheme,

```
for i=1:maximum no of cells
    for j=1:maximum RNA number
        if RNA(i,1)<= j+0.49
            RNA(i,2)=j
        break
        end
    end
 end
roundedRNA=RNA(i,2)
```

Programme 3.1. *A matlab script for rounding RNA numbers into integer values.*

The production interval, between the burst was also estimated from the intensity jump. From the data we calculated the burst size (number of mRNA produced in a burst) and frequency (occurrence of event in a population).

3.6) MATLAB SCRIPT FOR FINDING BURST SIZE AND PROBABILITIES

After using automated pipeline of time series analysis we have the information of background corrected spot intensities and fit intensities. Following code was written in Matlab 2013-b, for finding burst size at each transcription event and burst probabilities.

1) For extracting fit intensity values per cell from t=1 to t=121

```
for j=1:e %cells
    for i= 1:a %time
        fitint(i,j)=cells(i,j).Fit_Intensity
    end
end
```

2) For calculating difference between intensity values

```
for j= 1:e
    for i=1:a-1
        diff(i,j)=fitint(i+1,j)-fitint(i,j)
    end
end
z=diff;
```

3) For removing Ist intensity value

```
for j=1:e
    for i= 1:a-2
        if z(i,j)>0
            z(i,j)=0
            break
        end
    end
end
```

```
4)For finding unique intensity values
unique=find(z>0);
r=z(unique);
% intervals coming from all cells and their corresponding
intensity
% ist column is ival and second column is intensity
ival= get_intervals(times, cells);
ival(:,2)=r;
```

```
5)ival and their corresonding RNA (Given,1 RNA= 1.05
intensity value)
  now 3rd column is RNA
ival(:,3)=r/1.056604868; round(:,1)=r/1.056604868;
```

```
6) For rounding off and finding burst size
m = max(round);
n = 256;  %ceil(m);
for i=1:n      % max raw for std from round
    for j=1:max(round)+1
        if round(i,1)<= j+0.49
            round(i,2)=j
        break
        end
    end
end
round1=round(:,2);
for i=0:max(round1)
 count(i+1)=sum(sum(round1==i+1));
end
count=count';
```

```
7) Probability
count(:,2)=count(:,1)/5;
```

Programme 3.2. *A matlab code for finding burst size and probabilities.*

3.7) Calculation of T_F and T_C, T_{ON} and T_{OFF}

We calculated the time intervals of T_F and T_C in all bacterial cells phases and conditions. Here, T_F represents the time duration between the cell division moment and appearance of first burst while T_C stands for the production interval between two consecutive transcription events. For T_F we tracked only daughter cells and considered first RNA burst while discarding others. To calculate ON time from the mRNA burst we used single mRNA elongation rate, ~25 nucleotide/sec in vivo, as reported in previously (Golding and Cox 2004; Golding et al. 2005). Given the length of target mRNA, about 4000 bases, transcription of one message is completed in 150 seconds.

T_{ON}=mean RNA/Burst*time required to transcribe single mRNA (150 seconds) (3.2)

T_{OFF} = Mean burst interval – T_{ON} (3.3)

This allows us to follow the dynamics of P_{lac} and $P_{lac/ara-1}$ over the induction and cell phases.

3.7.1) Calculation of the delayed ON/OFF time between T_F and T_C over the phases

We determined delay as a difference between TF and TC production intervals over the phases. We considered delayed time in ON/OFF periods (ON_D/OFF_D) in genes under different inductions so that

Delay OFF state in independently induced upstream/downstream genes

$$OFF_D=OFF_{TF}-OFF_{TC}, \quad (3.4)$$

where OFF_{TF} and OFF_{TC} are periods of TF and TC respectively.

Delay ON state in independently induced upstream/downstream genes

$$ON_D=ON_{TF}-ON_{TC}, \quad (3.5)$$

where ON_{TF} and ON_{TC} are periods of T_F and T_C respectively.

Delay ON state in fully induced genes compared to ON delay time in independently induced genes

$$ON_D=ON_F-ON_I, \quad (3.6)$$

where ON_D is delayed ON state between fully and independently induced genes, ON_F is delayed ON period in fully induced genes, ON_I is delayed ON period in independently induced genes.

3.7.2) MATLAB SCRIPT FOR T_F,T_{ON} and T_{OFF}

We removed the intensity information of parent cells and kept only the daughter cells intensity values in the matrix fitint. Moreover the procedure of taking difference in intensity values and converting them into RNA number was performed similarly as described in section 4.1,

```
a=fitint;
for j=1:size(a,2)    % size(a,1) for finding maximum rows
s=find(a(:,j)>0)     % index of non zero elements
k=find(a(:,j)==0)    % index of zero elements
Tf(1,j)=s(1)-k(1)    % diff between first zero and non zero
                     value that will give time of appearance
                     of Ist RNA
end
j=find(Tf<0)
Tf( Tf < 0 ) = [];
meanTf=mean(Tf)
Ton=round1*150;       % In round1 matrix rounded values of
RNA are saved.
Toff=ival-Ton;        % Production interval values are saved
in ival matrix.
```

Programme 3.3. *A matlab script for finding transcription active and inactive periods.*

Chapter 4

RESULTS AND DISCUSSION

4.1) Jump detection output

Jump in the fit intensity values corresponded to the appearance of new transcripts. The production interval is calculated by taking the time difference of consecutive intensity jumps. For instance, the first jump is occurring at time t_1=18 min and the second jump is appearing at t_2= 27 minutes, so the production interval would be t_2-t_1, that is 9 minutes. The cells were tracked before and after division. After cell division the daughter cells were considered as a new cells and same jump detection method was used to calculate the production interval. The difference in the fit intensity values helped us to quantify RNA numbers. In figure (4.1), the fluctuating curve indicates the corrected intensity values which is fitted by monotonic piecewise curve (solid line) as elaborated in the method section. The black curve is showing parent cell whereas blue and magenta curves are suggesting daughter cells. At t= 60m, cell division is occurring and intensity values of the daughter cells are falling down due to the partitioning of RNA molecules. Cells are tracked for about two hours.

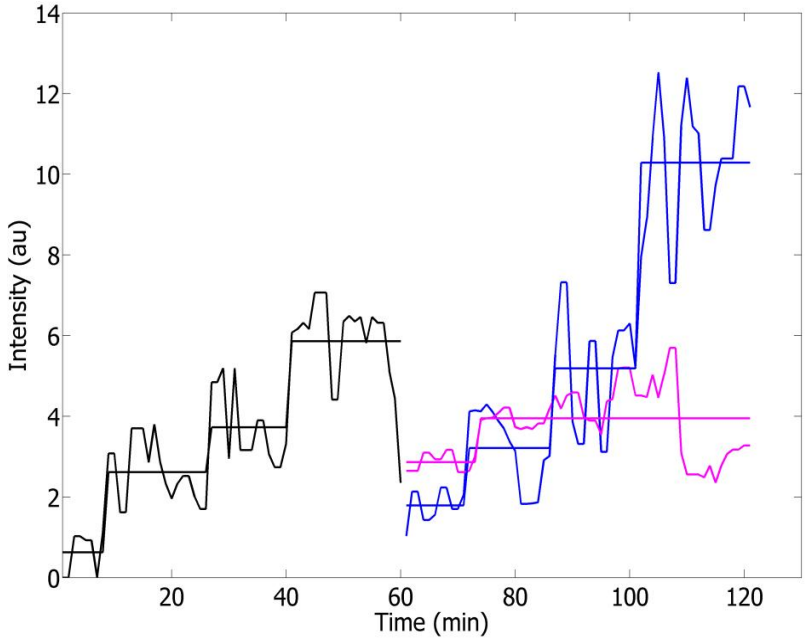

Figure 4.1. *Measured intensity values of parent and daughter cells for two hours. Fluctuating intensity values are fitted by least squares.*

By combining the information of single mRNA intensity value and jump detection method we calculated RNA numbers by taking the difference of fitted intensity values. Time traces of mRNA production and cell division time are extracted from time-lapse fluorescence microscopic images. In the following figure 4.2, the vertical solid lines indicate the number of mRNAs newly synthesized after the interval. The time resolution is 1 minute. The dotted line mark is the cell division time. The time traces show that mRNA production occurs in random bursts, with varying number of mRNA molecules are generated.

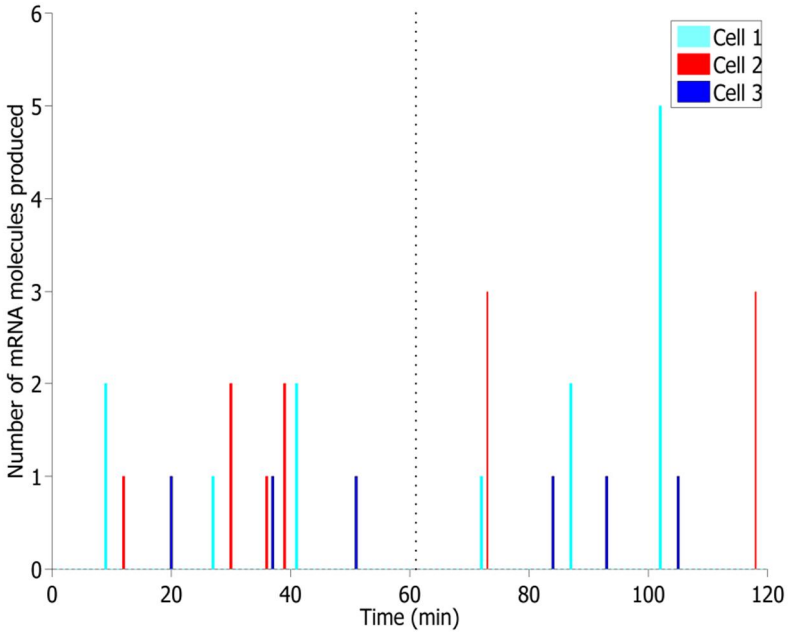

Figure 4.2. *Time traces of mRNA production and cell division of three random cells. Division is occurring at t=61min.*

4.2) Production interval of T_F and T_C

In live bacterial cells the kinetics of RNA production was studied. The production interval after cell division and appearance of first transcript, T_F is reported in Table 4.1. Similarly the consecutive transcription time duration, T_C is also shown. Table 4.1 also shows mean, standard deviation and coefficient of variation square of production interval in all phases and four different induction schemes. The coefficient of variation is calculated by taking the square of the ratio of standard deviation and mean. The value of CV^2 is smaller than one in all the cases (medium induction and full induction) which indicates the process of RNA production is sub-poissonian.

Table 4.1. *In vivo kinetics of first and consecutive transcripts events in five phases and four induction regimes.*

Conditions	Phases	T_F (sec)	std	CV^2	T_c (sec)	std	CV^2
$P_{lac/ara1}$ URS	LAG	2487.6	1366.4	0.301	1767	1337.604	0.573
	ACC	1772	1208.5	0.465	1244.364	1211.292	0.947
	EXP	1594.7	1530	0.92	1190.412	965.574	0.657
	RETD	2006.5	1208.9	0.363	1238.694	1142.952	0.851
	STAT	3029	1793.2	0.350	1496.4	1248	0.695
$P_{lac/ara1}$ DRS	LAG	1726.2	930.85	0.290	823.944	569.298	0.477
	ACC	1633.1	1080.2	0.437	698.802	569.046	0.663
	EXP	2263.8	1054.3	0.216	1158.6	961.806	0.689
	RETD	2588.4	1263	0.238	1902.558	1648.056	0.750
	STAT	2044.8	1307.2	0.408	1820.868	1686.726	0.858
$P_{lac/ara1}$URS/DRS	LAG	2524.2	1279.7	0.257	1660.212	1197.366	0.520
	ACC	1959	1239.7	0.40	1525.224	1181.124	0.599
	EXP	1294.8	1020.2	0.620	1552.2	1276.626	0.676
	RETD	1716.2	1313.2	0.585	1725.618	1498.104	0.753
	STAT	1887.9	1660.6	0.773	1222.962	1205.04	0.970
P_{lac} URS/DRS	LAG	1122	743.04	0.438	1381.5	1085.496	0.617
	ACC	1564.6	1136.7	0.527	1123.662	994.332	0.783
	EXP	2045.8	1219.7	0.355	976.296	760.8	0.607
	RETD	1635.9	926.13	0.321	1288.53	973.266	0.570
	STAT	1589.4	974.76	0.376	792.066	740.208	0.873

4.3) Active and inactive period calculation of T_F and T_C

Active and inactive periods during transcription process were calculated with the help of formulae described in method section 3.7. The ON/OFF periods for first and consecutive transcription events for fully and partially induced cells are show in Table 4.2.

Table 4.2. *Active and inactive periods of T_F and T_C*

$P_{lac/ara1}$ URS	Phases	$T_F$on (sec)	$T_F$off (sec)	$T_C$on (sec)	$T_C$off (sec)
	LAG	180	2307.6	213.45	1553.55
	ACC	190.5	1581.5	193.05	1051.314
	EXP	270.6	1324.1	184.65	1005.762
	RETD	286.65	1719.85	165.15	1073.544
	STAT	154.5	2874.5	150	1346.4
$P_{lac/ara1}$ DRS	LAG	161.4	1564.8	165.9	658.044
	ACC	210.9	1422.2	196.95	501.852
	EXP	202.65	2061.15	175.2	983.4
	RETD	219	2369.4	209.25	1693.308

	STAT	162	1882.8	202.2	1618.668
$P_{lac/ara1}$URS/DRS	LAG	219.75	2304.45	270.6	1389.612
	ACC	202.5	1756.5	330	1195.224
	EXP	354.75	940.05	600	952.2
	RETD	260.25	1455.95	366	1386.6
	STAT	160.35	1727.55	207	1015.962
P_{lac} URS/DRS	LAG	281.85	840.15	298.8	1082.7
	ACC	222.3	1342.3	218.25	905.412
	EXP	248.55	1797.25	382.65	593.646
	RETD	229.5	1406.4	269.25	1019.28
	STAT	205.8	1383.6	237.45	554.616

4.4) Distribution of the ON/OFF durations in T_F and T_C shows opposite behavior

Even though ON/OFF two states model is well-accepted mechanism in transcription, the molecular mechanism of ON/OFF kinetics in T_F over the cell phases remains unclear (Golding et al. 2005; So et al. 2011). To model its dynamics, we observed single mRNA molecules after the induction of the promoter as described in methods (Fig. 4.3 *A*). From the monitoring of production intervals in single cells, we observed the distribution of first and following sequence of transcriptional events. To distinguish gene activity further, we calculated the ON and OFF duration from the production interval and mRNA numbers as explained in methods (Fig 4.3*B*). Figure 4.3B shows the distributions of ON/OFF duration of T_F and T_C. To find the relationship between ON and OFF in T_F and T_C we have plotted the estimated frequency. Interestingly, we observe that intervals of T_F and T_C are close to the inverse proportion. Moreover, short OFF period (Fig. 4.3*C*) is followed by long ON period (Fig. 4.3*D*) and vice versa. Such counter relationship between two states might affect the dynamics of the gene and thus cells behavior over the phases.

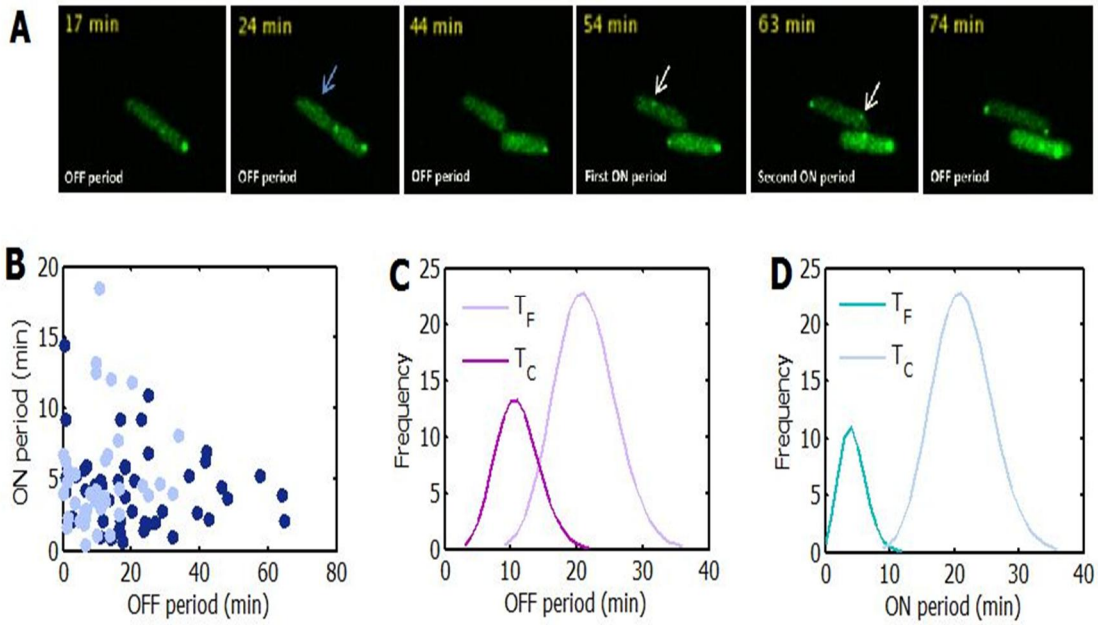

Figure 4.3. Observation of first transcriptional event and ON/OFF kinetics. *A) Time-lapse microscopy images show the production of transcriptional events. Produced mRNAs appear as a spot (indicated by arrow). B) The distribution of first transcription (TF) and consecutive transcription (T_C) events show generally longer inactive (OFF) state followed by shorter active (ON) state. C) Estimated frequency of OFF states distribution as a function of time reveals longer mean duration of T_F than mean T_C. D) Fit defining first and mean of sequential ON states of transcriptional events. Faster T_F event corresponds to longer mean of T_C.*

4.5) Variation of T_F ON/OFF states in fast and slow dividing cells depending on activation mechanism

To estimate the effects of the cell division rate in the dynamics of ON and OFF duration of T_F over the gene activation in fast and slow dividing cells, we followed the mRNA production interval over the time. Figure 4.4 shows the trend of duration of ON and OFF periods in fast (Lag, Acc and Exp phases) and slow (Ret and Sta phases) dividing cells. In Lag phase, OFF time period is the longest and behaves similarly when activated from URS, DRS and both URS/DRS (Fig. 4.4*B*). It is followed by the fastest ON period reaching maximum in fully induced cells (Fig. 4.4*A*). At the Acc phase, shortening OFF period, corresponds to the increase of ON state time. Both states have the highest time point in fully induced cells (Fig. 4.4*A and 4.4B*). Interestingly, cells under full induction at the Exp phase are characterized by the shortest OFF period that overall is further decreased compared to the previous phases (Fig. 4.4*B*). Following ON state of this phase is the longest among the group of fast dividing cells (Fig. 4.4*A*). Inconsistent OFF state behavior in Exp phase may indicate that division activity starts to switch from fast to slow dividing rate by showing the longest OFF time period in upstream induced cells. The reduced OFF time state in Exp phase is determined by upstream induction as its

activation requires less time than downstream induction. This may influence the changes in division rate, or in other words cells switch from fast to slow division in exponential phase. We also observed increased mRNA production from the upstream activation later in a cell cycle that would result in decreasing trend of ON time in the following phases.

Indeed, Figure 4.4*C* demonstrates the trend of ON time states in slow dividing cells as monotonically decreasing in all the conditions. According to the above shown tendency, OFF time periods is harmoniously increasing up to the end of the cell cycle. Such counter relationship between two states within the T_F determines mRNA production kinetics of the cells during the cell cycle. Moreover, shown exponential tendency of time flow in two-state model agrees with previous findings (Golding et al. 2005), which show that periods of ON and OFF are fluctuating exponentially. Here we extend the previous observation by studying the behavior of both states in cells under different induction conditions over the phases in lac variant.

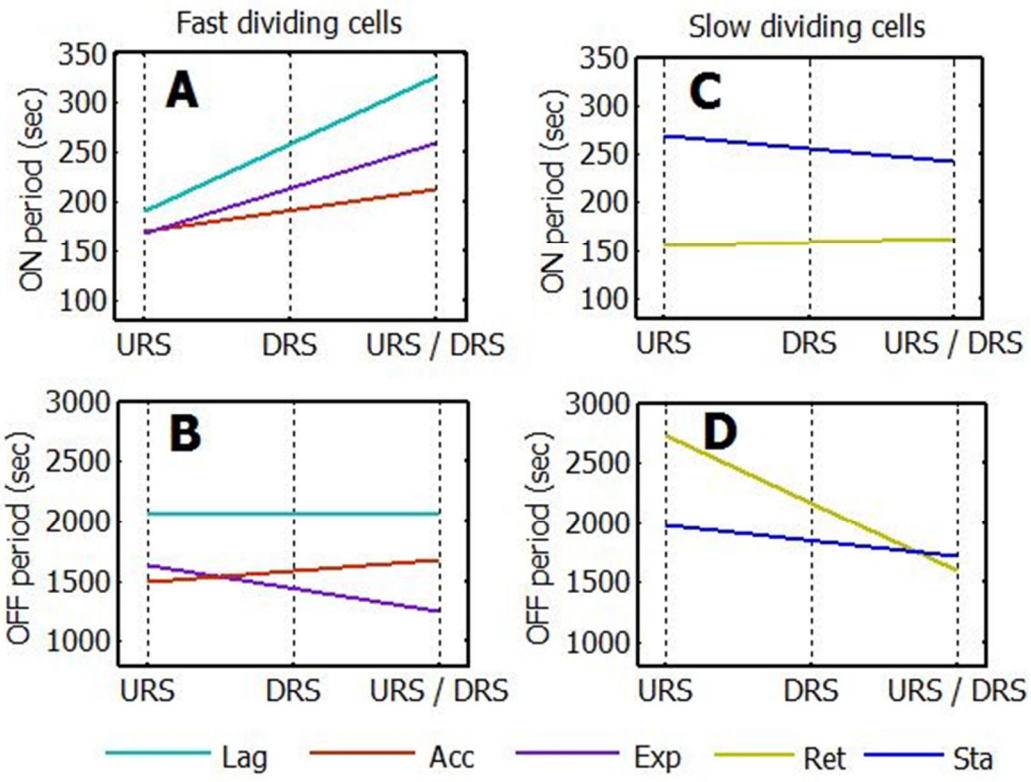

Figure 4.4. ON/OFF kinetics of T_F under different gene activation conditions over the phases. *$P_{lac/ara-1}$ induced from upstream. downstream and both using 0.1 % arabinose and 1 mM IPTG. A) ON state duration. in fast dividing cells under different activation conditions. gradually increases from Lag to Exp phase. B) OFF state duration in fast dividing cells under different induction OFF decreases from Lag to Exp phase. C) The trend of ON state duration in slow dividing cells is decreasing over the*

induction. D) Duration of OFF state period in slow dividing cells shows decreasing trend.

4.6) Fluctuation of ON periods according to OFF duration in T_F and T_C

To generalize the above findings we used wild type lac promoter (P_{lac}) in the kinetics studies of ON/OFF of T_F over the phases. During the first transcriptional event, long OFF period increases over the phases (Fig. 4.5*B*), preceding the fast progressively declining ON state (Fig. 4.5*A*). As a result, rapid production of high mRNA number is continuously decreasing over the phases (Fig. 4.5*A*). Notably, high molecule number at the beginning of the cell cycle makes population more stable (discussed later) which determines population homogeneity at the early stages. To support our finding we calculated the mean number of mRNA per burst within the first transcriptional event that fluctuates similarly to the ON period over the phases (Fig. 4.5*A*). Besides, tendency of mRNA production over the phases in lac promoter and lac variant are not significantly different from each other (Fig. 4.5*A*).

Consecutive transcriptional events display opposite behavior by displaying fast decreasing OFF period (Fig. 4.5*B*) followed by slow increasing ON state (Fig. 4.5*A*), causing cells to raise mRNA production. Consequently, mRNAs diversity is constantly reducing from Lag to Sta phase. Such a different behavior between T_F and T_C determines the behavior of the cell population over the phases in lac and lac variant. Thus, dynamics of the first transcriptional event under changing environment is a crucial step that rules future of the cell population.

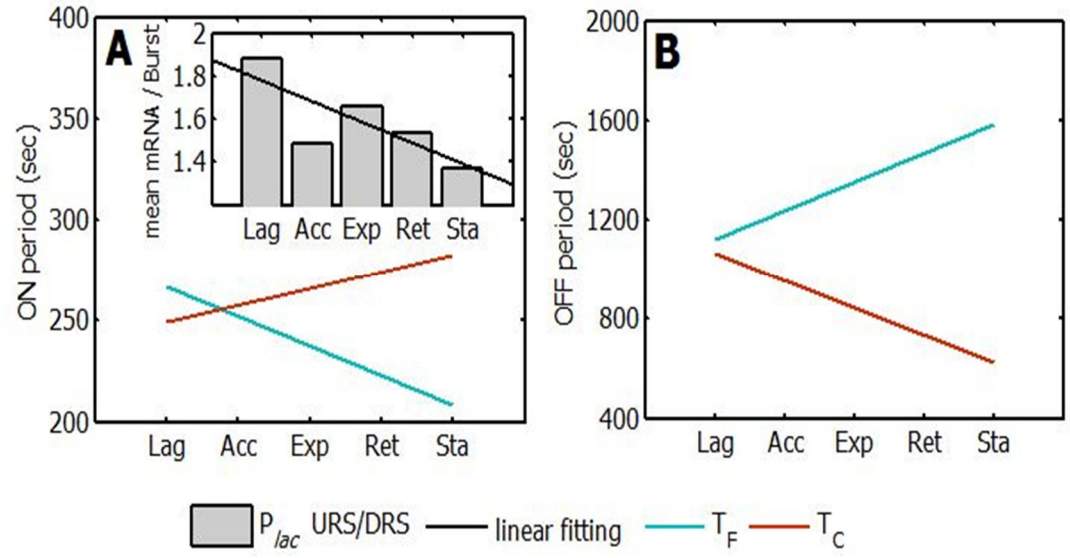

Figure 4.5.ON/OFF kinetics of wild type lac promoter over the phases*. A) OFF duration is monotonically increasing in T_F over the phases while it decline in T_C. B)*

Estimated time of ON periods shows increased trend in T_F and decreased time in T_C. C) mRNAs per burst of T_F is fluctuating similarly to the ON period along the phases.

4.7) Behavior changes in ON/OFF periods in T_F and T_C over the phases

Further, we investigated the role of upstream/downstream inductions of the promoter and their effect on mRNA production through the regulation of ON and OFF dynamics. To study this, we followed the mRNA production in T_F and T_C under microscope. From the data analysis, we observed that both T_F and T_C OFF period duration follow the same trend when activating upstream promoter sequence (Fig. 4.6*A*). OFF period duration decreases in the early phases of cell growth, toward exponential phase, followed by increasing duration up to stationary phase. Statistical tests show that both trends are correlating, thus our data suggests that state of the cell has similar effect on all transcription events in the cell growth. When the transcription is activated by the downstream gene sequence (Fig.4.6*B*), similar trend is observed between T_F and T_C OFF periods. Duration of this state decreases between Lag and Acc phases, followed by an increasing trend up to retardation to finally decrease in stationary phase. Under full induction (Fig.4.6*C*), both T_F and T_C OFF trends show decreasing duration time in early phases up to exponential phase while retardation and stationary phases show greater duration time than at exponential phase.

In case of ON periods, T_F and T_C show negative moderated correlation with OFF period. It indicates the dynamical changes of the transcription process, where higher activity results in shorter OFF and longer ON period with greater number of mRNA being produced (Fig.4.7). We have noticed that when the gene is activated from the URS alone, the delay of ON period appears to be longer than in DRS and full induction. Independent activations of URS and DRS in fast dividing cells shows that DRS determines the OFF period while in the slow dividing cells. URS has more impact. It infers that the gene is activated by independent RNAPs, either simultaneously from URS and DRS (synergistic effect), or by competing each other (antagonistic effect) (Nickerson and Achberger 1995; Malan and McClure 1984; Choi et al. 2010). Our results also agree with the previous observations (Malan and McClure 1984). Interestingly, we observed delay in the T_F compared to the T_C in both lac and lac variant In the previous studies it has been shown that single transcript could determine the population switch (JACOB and MONOD 1961; Monod 1949; Novick and Weiner 1957). Later, evidence suggested that single transactional events are the determent of protein numbers controlled by the architecture of the promoter (Choi et al. 2010. 2008). Since we observed the delayed first event we were interested to see how T_F and T_C frequencies affect the production of mRNA number by modulating ON/OFF kinetics. We analyzed the frequency of mRNA produced per burst in each phases as a function of the induction scheme. We observed opposite behavior between OFF duration and burst frequency in T_F and T_C (Fig 4.7) and a correlated behavior between ON and burst size. Also we noticed that OFF period take longer time to activate in T_F compared to T_C.

Thus we were interested to study the dynamics of delayed ON/OFF over the phases in T_F.

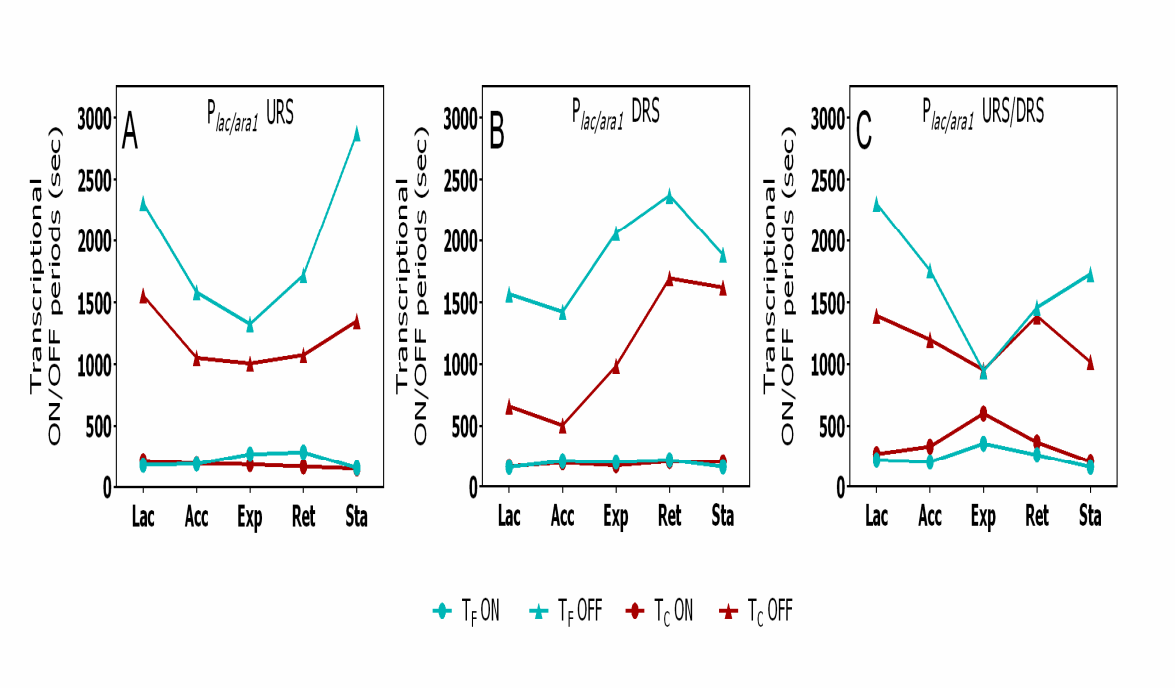

Figure 4.6. Activation dependent ON/OFF kinetics changes between T_F and T_C: *A. Mean period time for ON and OFF for TF and Tc when upstream only is activated. The same trends ON/OFF of T_F and T_C are similar. ON period trend for both event is opposite to OFF trends. B. Activation of $P_{lac/ara1}$ from downstream region shows the trend of T_F and T_C OFF is similar. The trends of ON periods are similar as in (A). C. $P_{lac/ara1}$ under full induction shows the similar trend in both T_F and T_C OFF. The trend of ON period is opposite to OFF*

4.8) Dynamic behavior of transcription events

Burst size and probability distribution were obtained by matlab code as described in method section 4.1. The exponential trend can be observed when the upstream and downstream genes are activated individually as well as in simultaneous activation (Fig. 4.7) Also it shows that the frequency of mRNA production from DRS is greater than URS. The trends of P_{lac} and $P_{lac/ara1}$ distribution are similar.

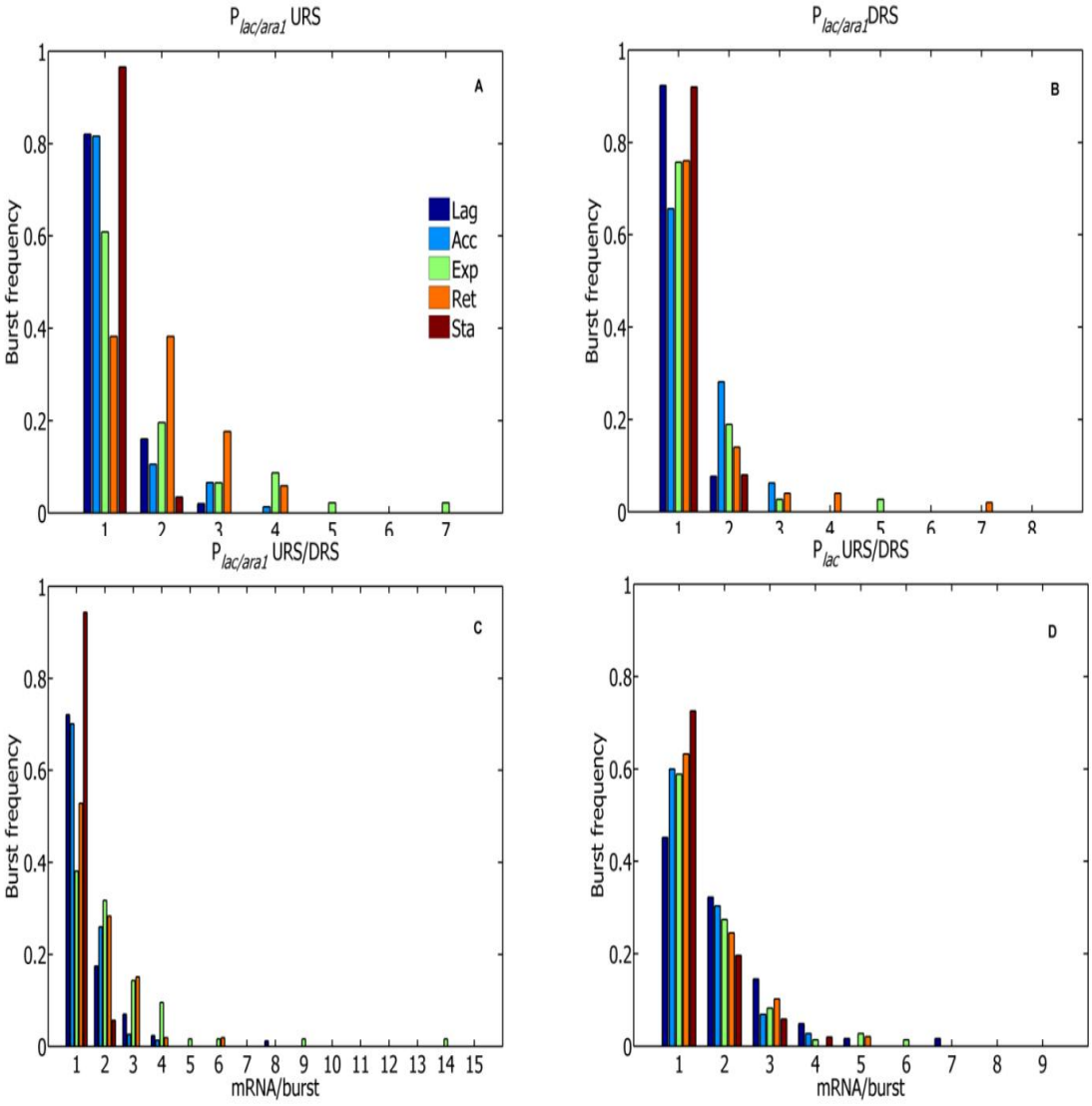

Figure 4.7. *A) Probability distributions of mRNA per burst and their frequency for $P_{lac/ara1}$ under upstream (URS) B) downstream (DRS) and C) activation from both regions D) Wild type P_{lac} regulated mRNA per burst and their frequency. All results show the exponential trend.*

4.9) Mean RNA/burst for all phases

Higher mean mRNA per burst was noticed under full induction of $P_{lac/ara1}$ in comparison with URS and DRS activation (Fig.4.8). Mean mRNA production per burst is higher in URS induction than DRS induction. However, the frequency of mRNA produced from the DRS is greater than URS (shown in Fig 4.7).

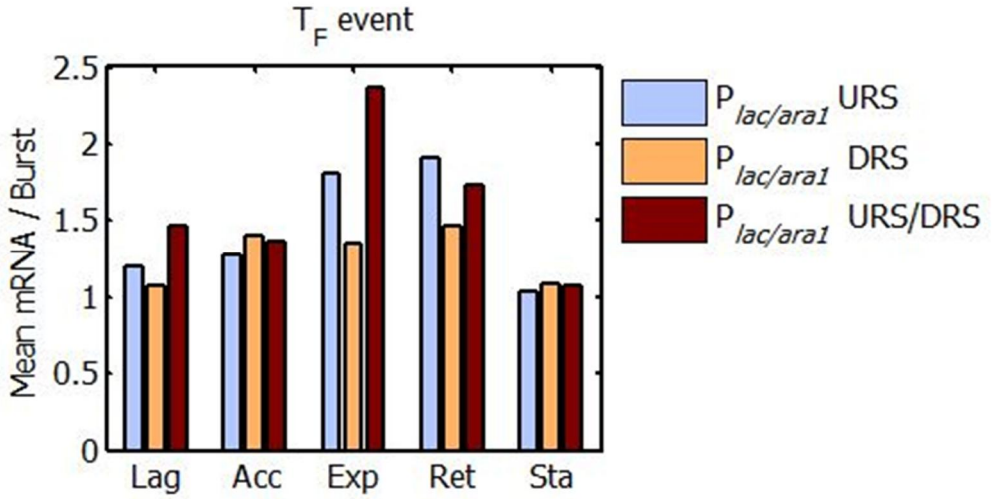

Figure 4.8. *Mean mRNA/burst of T_F for all phases.*

4.10) Transcription delay of the first event reduces mRNA noise

To determine the delayed time we calculated the difference between ON/OFF in T_F and T_C, under different induction. We observed that the delay of ON/OFF periods is ruled by URS activation in the fast dividing cells, as it takes less time to be initiated (Fig. 4.9*A*. *B*). Slow dividing cells combine the effect of URS and DRS inductions to determine OFF state in Ret and Sta phases respectively (Fig. 4.4), whereas ON state is regulated by DRS activation (Fig. 4.9*B*). Such distribution of gene activation types along with the division time in the cell cycle shows compensative and cooperating strategy of both states during transcriptional events. Nevertheless, changes in lac variant over the phases are closer to downstream gene activation as it is considered as a stronger gene activator resulting in higher production of mRNA number and similar trends (Fig. 4.9*C*). Overall, such multifaceted regulation is substantially important in T_F and thus mRNA numbers determination in the cell population.

Next, we measured rate of transcription delay in live *E.coli* cells using transcription inhibitor. We added rifampicin to the exponential phase cell after the full induction $P_{lac/ara-1}$. at time 5 min onwards up to 25 min. We consider the mean production interval

time of first transcription event is about 22 min (Fig. 4.9*D*). Moreover, the mean OFF period is 15 min, indicating that the gene starts to produce transcripts after this time. Thus, adding rifampicin below 30 min is considerable time to observe the delay. The results show that the normalized signal of protein abundant is decreased immediately upon the addiction of rifampicin. Signals of the cells induced with rifampicin after 5 min is similar to the leaky condition (no induction). From 10 min onwards up to 25 min the increasing signals were observed and the signal at 25 min is similar between cells with and without rifampicin. Error bars show the overlap of signal between each condition. The signal variation from cells incubated with rifampicin at 15 min and 25 min shows no overlap. As we expected, if the RNAP is not been blocked before the 15 min most population of the cells are producing first transcription. It is also known that rifampicin block the promoter clearance i.e. transcription initiation (Campbell et al. 2001). These independent observations further support our results calculated from the single molecule detection method. Since the observed signals are the output of translation, it indicates changes in transcription affect the protein concentrations of the cells.

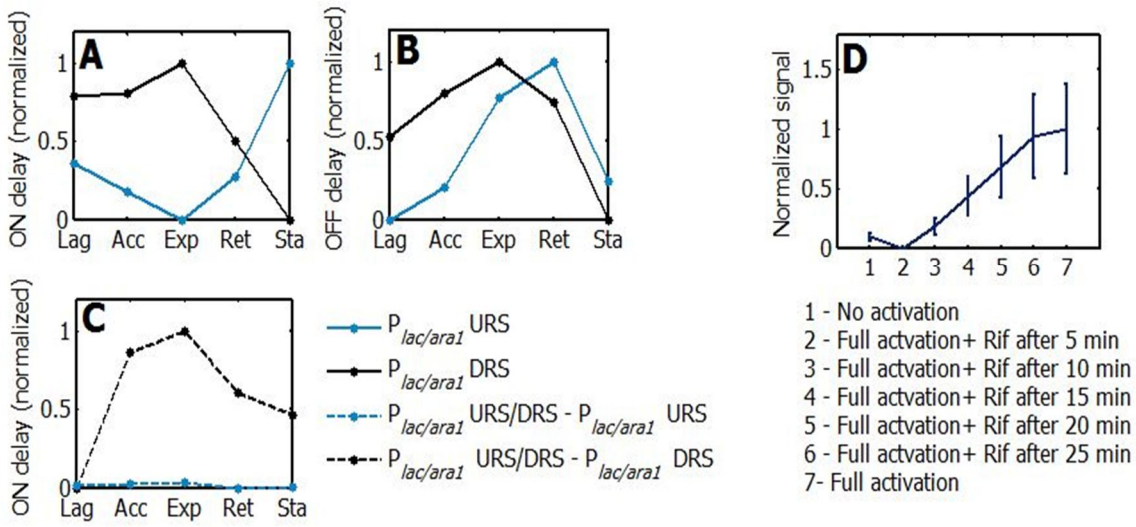

Figure 4.9. Kinetics of ON/OFF time delay over the phases*. A) Difference of OFF time has been calculated from T_F and T_C duration. OFF time delay in T_F of upstream and downstream is different. B) Normalized difference of ON time between T_F and T_C duration in upstream and downstream induction shows similar trend over the phases. C) Delayed ON time difference indicates stronger contribution of upstream than downstream gene activation when fully activated. D) Independent validation of transcription delay time is estimated by the addition transcriptional inhibitor. rifampicin. to the cells immediate after the induction of target promoter $P_{lac/ara1}$.*

Further. we studied the contribution of T_F in mRNA noise (Fano factor) over the phases. It results in less noise in T_F than T_C. Fold change shows DRS activation noisier than URS, over the cell cycle. The behavior of fold change in lac and lac variant promoter are identical and closer to the upstream noise (Fig. 4.10). Such results point out that the first event predefines the tendency of consecutive transcriptional events, depending on the gene activation condition. Finally, this concept is considerable by taking into account findings shown above (Fig. 4.6) revealing shorter T_F than T_C. It results in higher mRNA number being produced in the later transcriptional events compared to the first event. Such proportions of the produced mRNA over the production events directly influence corresponding noise level.

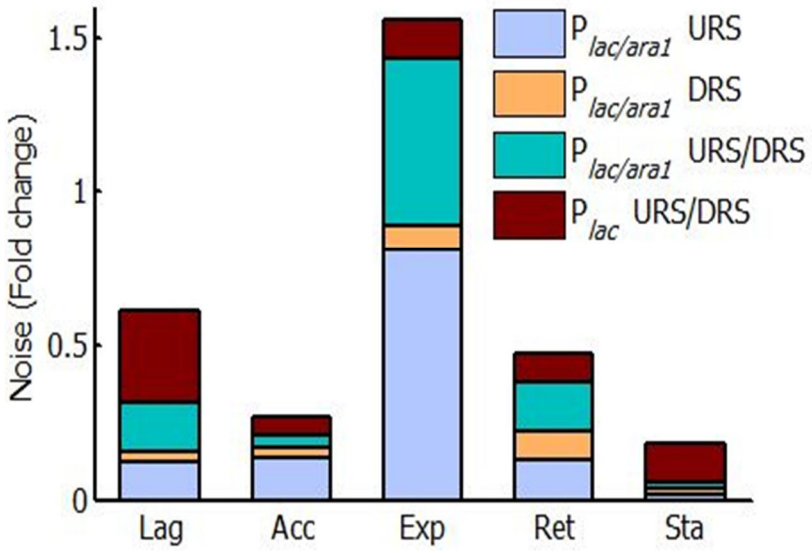

Figure 4.10. Effect of TF delay in mRNA noise. *Fold change of noise between T_F and T_C (Fano factor). It shows reduced noise in T_F and increased in T_C under different induction condition. It illustrates greater contribution of downstream activation to the noise in lac promoter and lac variant over the phases.*

In summary, the in vivo experiments demonstrate that ON/OFF kinetics determine the transcriptional delay, being regulated by the phase of the cell population. With regard to why ON/OFF is opposite in T_F and T_C, it may be a survival strategy of the population. If the cell does not produce enough mRNA molecules it affects the final protein concentration and thus delay in the T_F is compensated by the following event. We also noticed that T_C noise is greater than T_F resulting in more mRNA production. Earlier experiments have shown that in a mixed population the cells with single event produces less mRNA and protein and multiple events produces more mRNA and proteins (JACOB and MONOD 1961; Novick and Weiner 1957; Taniguchi et al. 2010). These

variations define the appearance of two mixed population. Our observation opening the future possibility to study how protein burst determines the kinetics of transcription ON/OFF. We also show that cells producing more than one event try to compensate the mRNA number by regulating the ON/OFF time. Thus, the cells with more than one event are capable of maintaining their population. Therefore we conclude that T_F regulates the future cell to cell diversity. As a perspective, it is possible to use our approach to understand the dynamics of ON/OFF periods in T_F and T_C in any specific genes at the transcriptional level.

Chapter 5

CONCLUSION

The kinetics of transcriptional events is fundamental process that defines the activity of single cells and the population behavior under changing environment. Even though ON/OFF two-state model is a well-accepted mechanism in transcription, the molecular mechanism of ON/OFF kinetics over the cell phases remains unclear. Here, we aim to study the ON/OFF mechanism in first transcriptional (T_F) and consecutive events (T_C). For this we use single mRNA detection method in live cells and develop mathematical approach to calculate the duration of both states in T_F and T_C. Wild type lac promoter (P_{lac}) and its variant ($P_{lac/ara1}$) were used to this study. It allows us to track simultaneously transcription initiation dynamics of both states from downstream and upstream promoter regions over the phases.

From this study we are able to observe the dynamics of ON/OFF states in T_F and T_C in $P_{lac/ara1}$. First, we show that duration of T_F ON/OFF is reversely proportional and they are opposite to the T_C. To understand the origin of such behavior we induced cells from both downstream and upstream promoter regions. Independent activations indicate that transcription initiation from downstream and upstream promoter region affects the dynamics of ON/OFF states in fast and slow dividing cells respectively. Secondly, we find that ON time of T_F predefines the behavior of T_C by modulating number and frequency of mRNA being produced. This analysis is further supported by observing bursty mRNA production over the phases in T_F under different inductions. Interestingly, T_F is delayed compared to T_C in all the phases. We demonstrate that delayed OFF time in T_F affects dynamics of both states in T_C. Moreover, delayed T_F production of mRNAs is ruled by upstream promoter region. Independent validation of transcriptional delay was performed by elongation arrest experiments. We observe the same mechanisms of ON/OFF states in T_F and T_C in P_{lac} which generalizes our results. Finally, we show that mRNA noise in T_C is determined by delayed OFF period in T_F and depends on transcription initiation region. These results shed light on how bacterial T_F ON/OFF states modulate the behavior of T_C.

In our understanding, our study is the first observation about real time tracking of ON/OFF dynamics in T_F over the phases that holds following significances. Dynamics of ON/OFF model in T_F is observed at single cell level. Due to the lack of detailed investigation of the dynamics of ON/OFF states over the phases, we consider this study as a significant step forward to understand magnitude of the ON/OFF states in T_F to the production events in T_C over the phases. It uncovers adaptive way on how

cells adjust the duration of both states in T_F that influence future population. Considering above points, we state that our approach and findings can be applied in understanding dynamics of ON/OFF periods in T_F and T_C in any specific genes at the transcriptional level and might be useful for researches studying molecular biophysics, genetic regulation, single molecule transcription dynamics, single cell biology, cell-to-cell variation and computational models.

REFERENCES

Baker, R.F., and Yanofsky C. 1968. The periodicity of RNA polymerase initiations: a new regulatory feature of transcription. *Proc Natl Acad Sci USA* 313–320.

Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, Darst SA. 2001. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell* **104**: 901–912.

Choi PJ, Cai L, Frieda K, Xie XS. 2008. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**: 442–446.

Choi PJ, Xie XS, Shakhnovich EI. 2010. Stochastic Switching in Gene Networks Can Occur by a Single-Molecule Event or Many Molecular Steps. *J Mol Biol* **396**: 230–244.

Contesse G, Crepin M, Gros F. 1969. Rhythmic transcription of the Escherichia coli lactose operon. *Bull Soc Chim Biol* 1445–1452.

Confocal laser scanning microscope, Laboratory of biosensors and bioelectronics, University of zurich. http://www.lbb.ethz.ch/Equipment/CLSM.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. *Science* **297**: 1183–1186.

Gaal T, Bartlett MS, Ross W, Turnbough CL, Gourse RL. 1997. Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science* **278**: 2092–2097.

Golding I, Cox EC. 2004. RNA dynamics in live Escherichia coli cells. *Proc Natl Acad Sci U S A* **101**: 11310–11315.

Golding I, Paulsson J, Zawilski SM, Cox EC. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**: 1025–1036.

JACOB F, MONOD J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318–356.

Kandhavelu M, Häkkinen A, Yli-Harja O, Ribeiro AS. 2012a. Single-molecule dynamics of transcription of the lar promoter. *Phys Biol* **9**: 026004.

Kandhavelu M, Lihavainen E, Muthukrishnan AB, Yli-Harja O, Ribeiro AS. 2012b. Effects of Mg(2+) on in vivo transcriptional dynamics of the lar promoter. *Biosystems* **107**: 129–34. http://www.ncbi.nlm.nih.gov/pubmed/22101250 (Accessed June 5, 2014).

Kandhavelu M, Lloyd-Price J, Gupta A, Muthukrishnan AB, Yli-Harja O, Ribeiro AS. 2012c. Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett* **586**: 3870–3875.

Klumpp S, Hwa T. 2008. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci U S A* **105**: 20245–20250.

Lutz R, Bujard H. 1997. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res* **25**: 1203–1210.

Lutz R, Lozinski T, Ellinger T, Bujard H. 2001. Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Res* **29**: 3873–3881.

Mäkelä J, Kandhavelu M, Oliveira SMD, Chandraseelan JG, Lloyd-Price J, Peltonen J, Yli-Harja O, Ribeiro AS. 2013. In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Res* **41**: 6544–6552.

Malan TP, McClure WR. 1984. Dual promoter control of the Escherichia coli lactose operon. *Cell* **39**: 173–180.

Monod J. 1949. The Growth of Bacterial Cultures. *Annu Rev Microbiol* **3**: 371–394.

Nickerson CA, Achberger EC. 1995. Role of curved DNA in binding of Escherichia coli RNA polymerase to promoters. *J Bacteriol* **177**: 5756–5761.

Novick A, Weiner M. 1957. Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci U S A* **43**: 553–566.

Pedraza JM, Paulsson J. 2008. Effects of molecular memory and bursting on fluctuations in gene expression. *Science* **319**: 339–343.

POWELL EO. 1956. Growth rate and generation time of bacteria, with special reference to continuous culture. *J Gen Microbiol* **15**: 492–511.

Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**: 216–26. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3118044&tool=pmcentrez&rendertype=abstract (Accessed March 21, 2014).

Raj A, van Oudenaarden A. 2009. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys* **38**: 255–270.

Ribeiro AS. 2010. Stochastic and delayed stochastic models of gene expression and regulation. *Math Biosci* **223**: 1–11.

Rafeal C. Gonzalez, Digital image Processing, third edition.

Sanchez A, Golding I. 2013. Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**: 1188–93. http://www.ncbi.nlm.nih.gov/pubmed/24311680.

So L-H, Ghosh A, Zong C, Sepúlveda LA, Segev R, Golding I. 2011. General properties of transcriptional time series in Escherichia coli. *Nat Genet* **43**: 554–560.

Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, Emili A, Xie XS. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**: 533–538.

T.been Chen, H.H-Shing LU, Y.Shien Lee, and H-Jen Lan, "Segmentation of cDNA microarray images using kernel density estimation" *Journal of Biomedical informatics*, Vol 41, 2008, pp.1021-1027

Wolf DM, Arkin AP. 2003. Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6**: 125–134.

Yurkovsky E, Nachman I. 2013. Event timing at the single-cell level. *Brief Funct Genomics* **12**: 90–98.