# The effects of cell asynchrony on time-series data: an analysis on gene expression level of *Plasmodium falciparum*

Wei Zhao[1,2], Justin Dauwels[1,2], and Jianshu Cao[1,3]
[1]Singapore-MIT Alliance for Research and Technology, Singapore;
[2]School of Electrical and Electronic Engineering, Nanyang Technological University;
and [3]Department of Chemistry, Massachusetts Institute of Technology

***Abstract*— To investigate the intraerythrocytic cycle of *Plasmodium falciparum*, time-series gene expression data are measured of infected red blood cells. However, the observed data is blurred due to cell asynchrony during experiments. In this paper, the effects of cell asynchrony are investigated by conducting numerical experiments. The simulation results suggest that cell asynchrony has varying effects on different intrinsic expression patterns. Specifically, the intrinsic patterns with high expression around the late life stage are more likely to be affected by cell asynchrony. It is also investigated how the effects of cell asynchrony are influenced by the experimental conditions. Certain parameters are identified that have a strong effects on cell asynchrony, and these parameters should be measured during biological experiments in order to deblur time-series gene expression data.**

## I. INTRODUCTION

Approximately 207 million people are infected by malaria, and in 2012, about 627,000 people died from this disease [1]. *Plasmodium falciparum* (*P. falciparum*) is the most fatal *Plasmodium* species which cause human malaria. In many efforts to understand the blood stages of *P. falciparum* infection, time-series gene expression data are measured over the 48-hour intraerythrocytic cycle (IDC) [2], [3]. Although the experiment starts with synchronized parasites, the parasite cultures gradually lose synchrony. Consequently, the intrinsic gene expression patterns are blurred in the observed gene expression data. In our earlier work, we developed a linear system to model the superposition across cells over the IDC [4]. In particular, the decay of cell synchrony is described as a cell age distribution which changes over the course of the experiment. The cell asynchrony in other cells has been studied earlier, such as yeast [5] and *Caulobacter crescentus* [6]. However, so far a quantitative analysis of the effects of cell asynchrony has not been conducted. In this paper, we analyze the linear model proposed in [4] to better understand the effects of cell asynchrony. There are two questions that we are specifically interested in:

1) Are there specific shapes of intrinsic expression patterns that are more likely to be affected by the effects of cell asynchrony?
2) How does the effect of cell asynchrony depend on the experimental conditions?

In Section II, we review our linear model of cell asynchrony. In Section III, we analyze the effect of different parameters in that model on the cell asynchrony, and in Section IV, we discuss our results. Conclusions are drawn at the end of paper.

## II. MODEL

Here we briefly review the model for cell asynchrony proposed in our earlier work [4]. Gene expression levels are measured at discrete time points over 48-hour IDC in the experiments of *P. falciparum* [2], [3]. The resulting observed expression data $e_i(t)$ at time point $t$ can be modeled as an integral over one life span of infected red blood cells (iRBCs); this integral can be approximated as the following sum [4]:

$$e_i(t) \approx \sum_{\ell_{\mathrm{re}}=1}^{L} N(t,\ell_{\mathrm{re}}) f_i(\ell_{\mathrm{re}}) \triangle \ell_{\mathrm{re}}, \tag{1}$$

where $\{N(t,1), N(t,2), ..., N(t,L)\})$ denotes the cell age distribution of iRBCs at the time point $t$, and $\{f_i(1), f_i(2), ...f_i(L)\}$ denotes the intrinsic gene expression pattern of specific protein $i$. Along this line, a linear system can be derived to model the relationship between intrinsic pattern $f_i(\ell_{\mathrm{re}})$ and observed expression data $e_i(t)$:

$$\underbrace{\begin{pmatrix} N(1,1) & \dots & N(1,L) \\ N(2,1) & \dots & N(2,L) \\ \vdots & \ddots & \vdots \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} f_i(1) \\ f_i(2) \\ \vdots \\ f_i(L) \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} e_i(1) \\ e_i(2) \\ \vdots \end{pmatrix}}_{\mathbf{b}}. \tag{2}$$

The constant vector $b$ denotes the observed gene expression data $e_i(t)$. The unknown variable vector $x$ stands for the intrinsic expression pattern $f_i(\ell_{\mathrm{re}})$. The element of the observation matrix $N(t,\ell_{\mathrm{re}})$ denotes the relative number of iRBC that stays at the rescaled cell age $l_{re}$ at time point $t$, which is calculated as [4]:

$$\begin{aligned} N(t,\ell_{\mathrm{re}}) = & \int_{t}^{+\infty} S(t') p_{\widetilde{L}} \left(\frac{t'-t}{L-\ell_{\mathrm{re}}}\right) \frac{t'-t}{(L-\ell_{\mathrm{re}})^2} \mathrm{d}t' \\ & + \int_{-\infty}^{t} R(t') p_{\widetilde{L}} \left(\frac{t-t'}{\ell_{\mathrm{re}}}\right) \frac{t-t'}{{\ell_{\mathrm{re}}}^2} \mathrm{d}t' \\ & + \int_{-\infty}^{t} R_f(t') p_{\widetilde{L}} \left(\frac{t-t'}{\ell_{\mathrm{re}}}\right) \frac{t-t'}{{\ell_{\mathrm{re}}}^2} \mathrm{d}t'. \end{aligned} \tag{3}$$

We refer to our earlier paper for more details [4].

Three generations of iRBCs appear in the experiment over the 48-hour IDC. The first generation stands for the late-stage iRBCs which are used to infect fresh RBCs and initialize

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

the experiment. These fresh RBCs infected at the beginning of experiment constitute the second generation. Due to the diversity of growth rate, few fast-growing iRBCs of second generation will burst and infect additional RBCs. As a results, the third generation of iRBCs appear at the end of experiment. As shown in (3), the three integrals respectively stands for the iRBCs from three generations. To be specific, $S(t)$ denotes the number of first generation iRBCs burst at time $t$; $R(t)$ stands for the number of second generation iRBCs infected at time $t$; $R_f(t)$ means the number of third generation iRBCs infected at time $t$. These three functions are essential to calculate the element of the observation matrix $N(t, \ell_{\text{re}})$. In the rest of this section, we review the key parameters used to describe $S(t)$, $R(t)$, and $R_f(t)$.

### *A. Burst rate in infection period*

The number of first generation iRBCs which burst at time $t$ is denoted as $S(t)$. We derive the expression of $S(t)$ based on the percentage of first generation iRBCs which burst in the two-hour infection period. According to the experimental specifications [4], $r\%$ of the first generation iRBCs burst in the two-hour infection period prior to the experiment. The remaining $1 - r\%$ iRBCs remain alive and continually infect fresh RBCs till around $h$ hours after the two-hour infection period. Therefore, $S(t)$ is approximated as a piecewise function:

$$S(t) = \begin{cases} c, & \text{if} - 1 \leq t < 1, \\ at + b, & \text{if } 1 \leq t \leq h, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

which satisfies the equations:

$$\begin{cases} S(1) = c, \\ S(h) = 0, \\ \frac{\int^{1}{}_{1} S(t)\mathrm{d}t}{\int_1^h S(t)\mathrm{d}t} = \frac{r}{100-r}. \end{cases} \tag{5}$$

Hence, (4) can be written as a function of $r$:

$$S(t) = \begin{cases} c, & \text{if} - 1 \leq t < 1, \\ \frac{rc}{4(r-100)}t + \frac{3r-400}{4(r-100)}c, & \text{if } 1 \leq t \leq \frac{400}{r} - 3, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $c$ stands for an arbitrary positive value.

### *B. Infection factors*

The number of second generation iRBCs infected at time $t$ is denoted as $R(t)$. Since the second generation iRBCs are infected by the first generation iRBCs, $R(t)$ is proportional to the number of first generation iRBCs bursting at time $t$:

$$R(t) = \begin{cases} a_{\text{in}} S(t), & \text{if t} \in \text{[infection period]}, \\ a_{\text{af}} S(t), & \text{if t} \in \text{[after infection period]}. \end{cases} \tag{7}$$

where the average number of RBCs infected by one iRBC during and after the infection period are respectively denoted as the parameters $a_{\text{in}}$ and $a_{\text{af}}$.

### *C. Distribution of normalized life span*

Individual iRBCs grow at different growth rates. The normalized life span of iRBCs $\widetilde{L}$ is assumed a Gaussian random variable: $\widetilde{L} \sim N(1, \sigma^2)$. Given the probability density function $p_{\widetilde{L}}(l)$ of normalized life span $\widetilde{L}$, the number of third generation iRBCs infected at time $t$ can be derived from $R(t)$ as [4]:

$$R_f(t) = \frac{a_{\text{af}}}{L} \int_{-\infty}^{+\infty} R(t') p_{\widetilde{L}} \left( \frac{t - t'}{L} \right) \mathrm{d}t'. \tag{8}$$

## III. Analysis

In this section, we conduct simulations on synthetic intrinsic gene expression patterns. The observed expression pattern $b$ is obtained by substituting the intrinsic pattern $f_i(\ell_{\text{re}})$ into the linear system described in (2). The difference between observed pattern $b$ and intrinsic pattern $f_i(\ell_{\text{re}})$ is calculated to investigate the effect of cell asynchrony. As discussed in the previous section, the linear system is dominated by three groups of parameters: the burst rate in infection period $r\%$, the infection factors $\{a_{\text{in}}, a_{\text{af}}\}$, and the standard deviation of growth rate $\sigma$. Consequently, the parameters of the linear system $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are also changed to model different experimental conditions.

### *A. Synthetic gene expression patterns*

The gene expression level of *P. falciparum* is expected to peak just before the encoded protein is needed [2]. Therefore, synthetic gene expression patterns $f_i(\ell_{\text{re}})$ are generated by utilizing the bell curve of normal distribution. Each of them simulates a synthetic gene with high expression level at different stages in the life span. The mean and standard deviation of normal distribution $\{\hat{\mu}, \hat{\sigma}\}$ respectively indicates the position of the peak and width of the bell curve.

The gene expression patterns $f_i(\ell_{\text{re}})$ present the change of gene expression level over one life span. Once the iRBCs reach the end of its life span, they will burst and start the next life cycle. Hence the expression level at the first data point $f_i(1)$ is highly correlated to the expression level at the last data point $f_i(L)$. Therefore, we simply assume that $f_i(\ell_{\text{re}})$ has the same value at $\ell_{\text{re}} = 1$ and $\ell_{\text{re}} = L$. The synthetic gene expression patterns are generated as follows:

$$f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}} = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \left[ e^{\frac{(\ell_{\text{re}}\ \ \hat{\mu})^2}{2\hat{\sigma}^2}} + e^{\frac{(\ell_{\text{re}}+L\ \ \hat{\mu})^2}{2\hat{\sigma}^2}} + e^{\frac{(\ell_{\text{re}}\ \ L\ \ \hat{\mu})^2}{2\hat{\sigma}^2}} \right],$$
$$\ell_{\text{re}} = 1, 2, 3, \ldots, L. \tag{9}$$

### *B. Effects of cell asynchrony*

We assess the difference between synthetic intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ and observed expression data $e_i(t)$ by a measure $D(\hat{\mu}, \hat{\sigma})$ defined as follows:

$$D(\hat{\mu}, \hat{\sigma}) = \int_0^L \left| \frac{e_i(t)}{\int_0^L e_i(t)\mathrm{d}t} - \frac{f_i(t)|_{\hat{\mu},\hat{\sigma}}}{\int_0^L f_i(t)|_{\hat{\mu},\hat{\sigma}}\mathrm{d}t} \right| \mathrm{d}t. \tag{10}$$

Since only the trend of the expression data is of interest here, the values of $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ and $e_i(t)$ are normalized across the cell life span. The observed expression data $e_i(t)$ is measured

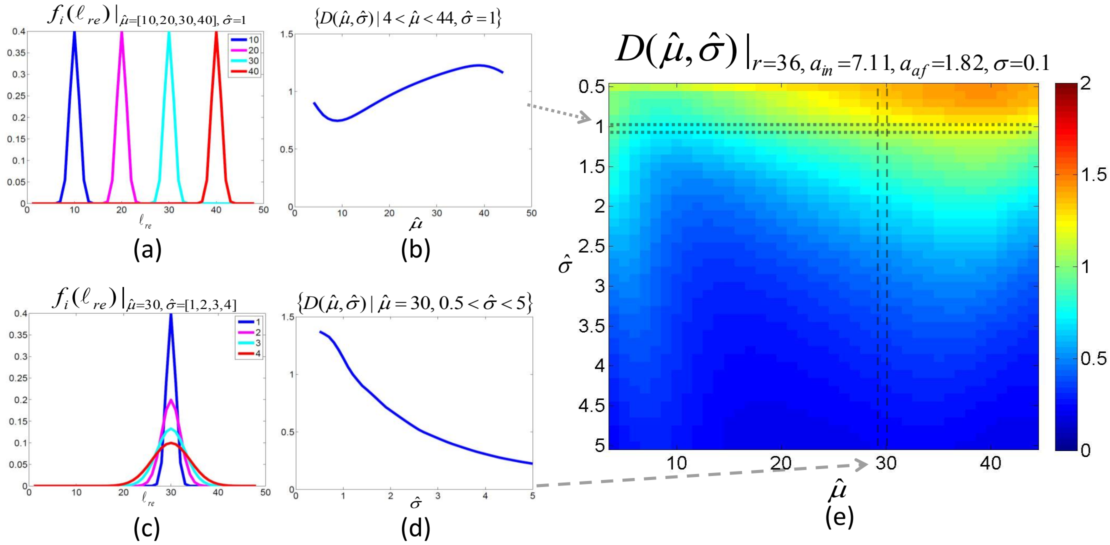

Fig. 1. (a) Synthetic intrinsic patterns $f_i(\ell_{\text{re}})$ with fixed shape and different position of the bell curve. (b) The row vector of the $D(\hat{\mu},\hat{\sigma})$ with $\hat{\sigma}=1$. (c) Synthetic intrinsic patterns $f_i(\ell_{\text{re}})$ with different shape and fixed position of the bell curve. (d) The column vector of the $D(\hat{\mu},\hat{\sigma})$ with $\hat{\mu}=30$. (e) The 2-D plot of $D(\hat{\mu},\hat{\sigma})$ which is calculated with parameters $\{r=36, a_{\text{in}}=7.11, a_{\text{af}}=1.82, \sigma=0.1\}$.

at discrete data points. By substituting (2) into (10), the expression of $D(\hat{\mu},\hat{\sigma})$ can be written in discrete form as:

$$D(\hat{\mu},\hat{\sigma}) = \text{Sum}\left(\left|\frac{Ax}{\text{Sum}\,(Ax)} - \frac{x}{\text{Sum}\,(x)}\right|\right), \qquad (11)$$

where $x$ denotes the intrinsic expression pattern $\{f_i(1), f_i(2), ...f_i(L)\}|_{\hat{\mu},\hat{\sigma}}$, and $A$ stands for the observation matrix consisting of the cell age distribution $N(t,\ell_{\text{re}})$ (2).

As described in Algorithm 1, there are three steps in our numerical experiments. First, the value of the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are chosen to model the experimental condition. Then the linear system is built based on these parameters. Specifically, the elements of the observation matrix $A$ of the linear system are calculated according to (3). Second, the bell-curved synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ are generated with values $\{\hat{\mu},\hat{\sigma}\}$, which respectively denote the position and the shape of the bell curve. Third, the observed patterns are obtained by substituting the intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ into the linear system. The effect of cell asynchrony are measured as the difference between the observed pattern and intrinsic pattern according to (11). The experimental results will be discussed in next section.

**Algorithm 1** Calculate the $D(\hat{\mu},\hat{\sigma})$ with given parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$.

```
Initialize the value of $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ and substitute them into the expressions of $S(t)$, $R(t)$, and $R_f(t)$, given by (6), (7), and (8) respectively.
Calculate $N(t,\ell_{\text{re}})$ by substituting the expressions of $S(t)$, $R(t)$, and $R_f(t)$ into (3), and next compute the observation matrix $A$ (2).
for all reasonable value of $\{\hat{\mu},\hat{\sigma}\}$ do
    Generate the synthetic gene expression pattern with $\{\hat{\mu},\hat{\sigma}\}$ according to equation (9):
        $x = \{f_i(1), f_i(2), ...f_i(L)\}|_{\hat{\mu},\hat{\sigma}}$
    Calculate the observed expression data $b$ by substituting $x$ in the linear system (2):
        $b = Ax.$
    Calculate the difference between intrinsic pattern $x$ and observed data $Ax$ according to equation (11)
end for
Return $D(\hat{\mu},\hat{\sigma})$.
```

## IV. Results

In this section, we investigate the effects of cell asynchrony on different expression profiles, and also how these effects depend on the model parameters (and hence experimental conditions).

Algorithm 1 is executed to calculated the $D(\hat{\mu},\hat{\sigma})$ on different synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$. The value of the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are fixed as $\{36\%, 7.11, 1.82, 0.1\}$, the values estimated from experimental specifications in our earlier study [4]. The synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ are generated with values of $\{\hat{\mu},\hat{\sigma}\}$ in the range of $\{\hat{\mu},\hat{\sigma}|0.5<\hat{\mu}<5, 4<\hat{\sigma}<44\}$. The difference $D(\hat{\mu},\hat{\sigma})$ is respectively calculated between each synthetic intrinsic pattern $x$ and its corresponding observed pattern $b$ according to equation (11).

To have a better understanding, we separately interpret the row vector and column vector of the 2-D plot of $D(\hat{\mu},\hat{\sigma})$ as shown in the Figure 1. The parameters $\{\hat{\mu},\hat{\sigma}\}$ respectively denote the position and the shape of the bell curve which is used to generate the synthetic intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$, as demonstrated in the Figure 1(a)(c).

The row vector $\{D(\hat{\mu},\hat{\sigma})|4<\hat{\mu}<44, \hat{\sigma}=1\}$ indicates the change of the difference $D(\hat{\mu},\hat{\sigma})$ on the intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ with different positions of peak $(4<\hat{\mu}<44)$ but with a fixed shape $(\hat{\sigma}=1)$. As shown in the Figure 1(b), the row vector $\{D(\hat{\mu},\hat{\sigma})|4<\hat{\mu}<44, \hat{\sigma}=1\}$ decreases when the value of $\hat{\mu}$ changes from 4 to 10. Then the trend reverses after $\hat{\mu}$ further moves towards $44$. The highest value of the row vector $\{D(\hat{\mu},\hat{\sigma})|4 < \hat{\mu} < 44, \hat{\sigma} = 1\}$ is obtained around $\hat{\mu} = 40$. The pattern of this row vector implies that the cell asynchrony has stronger effect on the intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ with high expression level around late life stage rather than around early life stage. This is a common observation for all row vectors in the 2-D plot of $D(\hat{\mu},\hat{\sigma})$.

Similarly, the column vector $\{D(\hat{\mu},\hat{\sigma})|\hat{\mu} = 30, 0.5 < \hat{\sigma} < 5\}$ stands for the change of $D(\hat{\mu},\hat{\sigma})$ on intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ with a fixed position of the peak $\hat{\mu} = 30$ but variant shapes ($0.5 < \hat{\sigma} < 5$). As shown in the Figure 1(d), the value in the column vector $\{D(\hat{\mu},\hat{\sigma})|\hat{\mu} = 30, 0.5 < \hat{\sigma} < 5\}$ continuously decreases when $\hat{\sigma}$ changes from $0.5$ to $5$. This suggests that the cell asynchrony has continuously decreasing effects on intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ if its bell-shaped expression curve become more disperse. This is also the common conclusion can be drawn from all column vectors of the $D(\hat{\mu},\hat{\sigma})$.

Summarizing, the 2-D plot of $D(\hat{\mu},\hat{\sigma})$ presented in Figure 1(e) suggests that the intrinsic patterns with high expression (smaller value of $\hat{\sigma}$) around late life stages (larger value of $\hat{\mu}$) are more likely to be affected by the cell asynchrony [4]. In the rest of this section, we will further investigate how the effects of cell asynchrony depend on the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ and hence the experimental conditions.

As depicted in Figure 2, we calculate $D(\hat{\mu},\hat{\sigma})$ with different parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$. The parameters are first initialized as $\{36, 7.11, 1.82, 0.1\}$. In each plot, one of the four parameter is selected and changed to either half or twice of its initial value. For example, Figure 2(a) and 2(b) show $D(\hat{\mu},\hat{\sigma})$ with parameter $r$ changed to 18 and 72 respectively. By comparing these two figures, we can observe how the $D(\hat{\mu},\hat{\sigma})$ is influenced by the value of $r$. When the parameter $r$ decreases from 72 to 18, $D(\hat{\mu},\hat{\sigma})$ increases considerably on all intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$. The same phenomenon is also observed when the parameter $\sigma$ increases from $0.05$ to $0.2$, as shown in Figure 2(g) and 2(h) respectively. By contrast, as presented in Figure 2(c) and 2(d), $D(\hat{\mu},\hat{\sigma})$ only slightly increases when $a_{\text{in}}$ decreases from $14.22$ to $3.56$ respectively. Similarly, from Figure 2(e) and 2(h), it can be seen that $D(\hat{\mu},\hat{\sigma})$ does not vary much when $a_{\text{af}}$ increases from $0.91$ to $3.64$ except for small values of $\hat{\mu}$.

From this analysis, it becomes clear that the parameters $r$ and $\sigma$ have a strong effect on cell asynchrony, whereas the parameters $a_{\text{in}}$ and $a_{\text{af}}$ have far less influence on cell asynchrony. Therefore, when measuring expression levels experimentally, it is crucial to properly measure the parameters $r$ and $\sigma$, either directly or indirectly.

## V. Conclusions

In this paper, we investigated the effects of cell asynchrony on time-series gene expression data of *P. falciparum*. By analyzing a linear model of cell asynchrony, we demonstrated how the cell asynchrony has varying effects on different intrinsic expression patterns, and how these effects are influenced by the experimental conditions (model parameters). The presented analysis may help to gain better understanding of the effect of cell asynchrony on expression data, and underlines the importance of measuring certain variables during experimental measurement of expression levels.

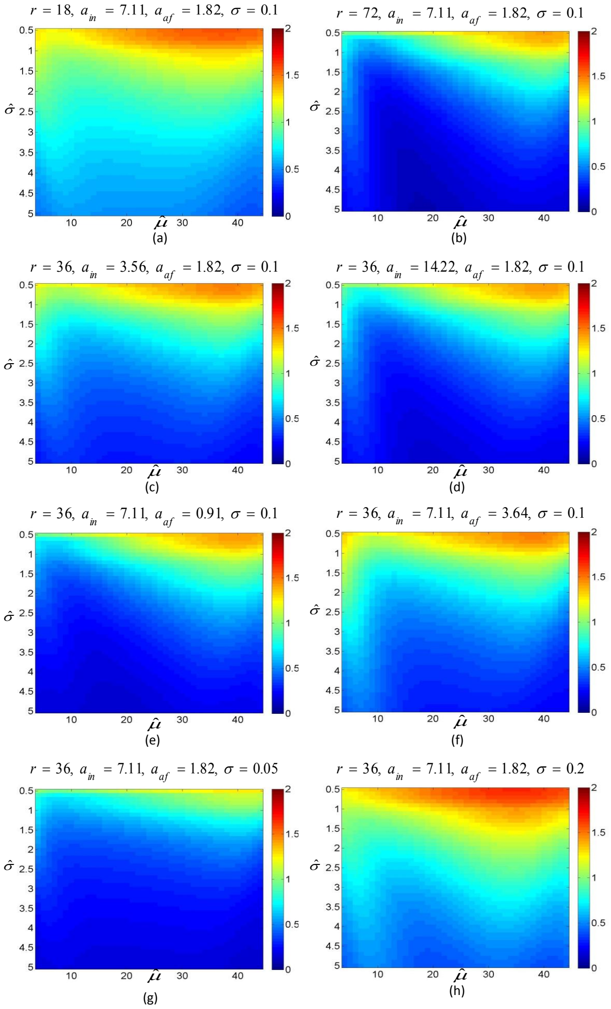

Fig. 2. The 2-D plots of $D(\hat{\mu},\hat{\sigma})$ calculated with different parameters $\{r, a_{\text{in}}, a_{\text{af}}, \sigma\}$. The value of $r$ is 18 in (a) and 72 in (b). Similarly, $a_{\text{in}}$ is equal to 3.56 in (c) and 14.22 in (d); $a_{\text{af}}$ is equal to 0.91 in (e) and 3.64 in (f); $\sigma$ is modified as 0.05 (g) and 0.2 (h).

## References

[1] *World Malaria Report*. Geneva, Switzerland: World Health Organization, 2013. [Online]. Available: http://www.who.int

[2] Z. Bozdech, M. Llinás, B. Lee, E. D. Wong, J. Zhu, and J. L. DeRisi, "The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum," *PLoS Biol*, vol. 1, no. 1, pp. e5+, Aug. 2003.

[3] M. Llinás, Z. Bozdech, E. D. Wong, A. T. Adai, and J. L. DeRisi, "Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains." *Nucleic acids research*, vol. 34, no. 4, pp. 1166–1173, 2006.

[4] W. Zhao, J. Dauwels, J. Niles, and J. Cao, "Computational synchronization of microarray data with application to Plasmodium falciparum," *Proteome Science*, vol. 10, no. Suppl 1, pp. S10+, 2012.

[5] Z. Bar-Joseph, S. Farkash, D. K. Gifford, I. Simon, and R. Rosenfeld, "Deconvolving cell cycle expression data with complementary information." *Bioinformatics (Oxford, England)*, vol. 20 Suppl 1, Aug. 2004.

[6] D. Siegal-Gaskins, J. N. Ash, and S. Crosson, "Model-Based Deconvolution of Cell Cycle Time-Series Data Reveals Gene Expression Details at High Resolution," *PLoS Comput Biol*, vol. 5, no. 8, pp. e1 000 460+, Aug. 2009.