Zhao et al. Proteome Science 2012, 10(Suppl 1):S10
http://www.proteomesci.com/content/10/S1/S10

PROCEEDINGS Open Access

# Computational synchronization of microarray data with application to *Plasmodium falciparum*

Wei Zhao[1,2], Justin Dauwels[1,2*], Jacquin C Niles[1,3], Jianshu Cao[1,4*]

*From* IEEE International Conference on Bioinformatics and Biomedicine 2011
Atlanta, GA, USA. 12-15 November 2011

## Abstract

**Background:** Microarrays are widely used to investigate the blood stage of *Plasmodium falciparum* infection. Starting with synchronized cells, gene expression levels are continually measured over the 48-hour intra-erythrocytic cycle (IDC). However, the cell population gradually loses synchrony during the experiment. As a result, the microarray measurements are blurred. In this paper, we propose a generalized deconvolution approach to reconstruct the intrinsic expression pattern, and apply it to *P. falciparum* IDC microarray data.

**Methods:** We develop a statistical model for the decay of synchrony among cells, and reconstruct the expression pattern through statistical inference. The proposed method can handle microarray measurements with noise and missing data. The original gene expression patterns become more apparent in the reconstructed profiles, making it easier to analyze and interpret the data. We hypothesize that reconstructed gene expression patterns represent better temporally resolved expression profiles that can be probabilistically modeled to match changes in expression level to IDC transitions. In particular, we identify transcriptionally regulated protein kinases putatively involved in regulating the *P. falciparum* IDC.

**Results:** By analyzing publicly available microarray data sets for the *P. falciparum* IDC, protein kinases are ranked in terms of their likelihood to be involved in regulating transitions between the ring, trophozoite and schizont developmental stages of the *P. falciparum* IDC. In our theoretical framework, a few protein kinases have high probability rankings, and could potentially be involved in regulating these developmental transitions.

**Conclusions:** This study proposes a new methodology for extracting intrinsic expression patterns from microarray data. By applying this method to *P. falciparum* microarray data, several protein kinases are predicted to play a significant role in the *P. falciparum* IDC. Earlier experiments have indeed confirmed that several of these kinases are involved in this process. Overall, these results indicate that further functional analysis of these additional putative protein kinases may reveal new insights into how the *P. falciparum* IDC is regulated.

## Introduction

Approximately 40% of the global population is at risk for contracting malaria, and an estimated 780,000 people die annually from this disease [1]. Human malaria is caused by five *Plasmodium* species, of which *P. falciparum* is responsible for the majority of human fatalities. The disease is transmitted when an infected mosquito bites a person, and injects sporozoites that migrate to and develop in the liver before merozoites are released into the bloodstream and invade red blood cells (RBCs) [2]. Within the RBC, *P. falciparum* undergoes a well-defined developmental cycle (IDC) during a 48-hour period that is characterized by three main stages, namely: rings, trophozoites and schizonts [2]. Schizont-infected RBCs rupture at the end of this cycle to release merozoites that can invade RBCs and reestablish a new IDC. In addition to the morphologic changes that characterize parasite development during the IDC, changes in gene expression accompany [3,4] and most likely drive this developmental program.

* Correspondence: jdauwels@ntu.edu.sg; jianshu@mit.edu
[1]Singapore-MIT Alliance for Research and Technology, Centre for Life Sciences, 28 Medical Drive, Singapore 117456
Full list of author information is available at the end of the article

Gene expression during the 48-hour IDC has been densely profiled at 1-hour intervals using microarray technology in an effort to understand how overall gene expression patterns help shape blood stage parasite biology [3,4]. These studies revealed that the levels of many transcripts are reproducibly high or low at characteristic times within the IDC. Genes involved in key biological processes most relevant to a given IDC stage are generally coordinately up- and down- regulated. In fact, a 'just-in-time' model to describe transcriptionally regulated gene expression in *P. falciparum* has been proposed [4,5]. Here, a transcript's level is proposed to peak just prior to when its encoded protein product is most critically needed. Regulation of general biological processes such as metabolism, DNA synthesis, protein turnover and red blood cell invasion, for example, is well-described by this model [3,4]. While the full complement of parasite proteins controlling IDC progression has not been identified, the just-in-time principle could be useful for identifying key, transcriptionally regulated proteins that play important roles in regulating this process.

Protein kinases represent one such protein class, and have previously been implicated in regulating various aspects of the cell cycle and development in *P. falciparum* [6-8]. The genome encodes a predicted 85 [9] or 99 [10] protein kinases, which include an expanded and divergent FIKK family, of which there are 20 members [9,11]. In *Plasmodium spp.*, several protein kinases have been shown to be important in regulating blood stage biology, including parasite egress from infected RBCs [6-8]. Recently, a large-scale knockout screening effort identified 36 out of the 65 protein kinases evaluated (no FIKKs included) as likely essential to the development of *P. falciparum* blood stage parasites [12]. Overall, these studies highlight the important contribution of the protein kinases to regulating critical aspects of *P. falciparum* biology during the IDC.

Therefore, we have been interested in addressing whether computationally applying the just-in-time principle to publicly available microarray data for *P. falciparum* is a reasonably efficient strategy for identifying protein kinases that are important to transition through the IDC. Such an approach, if successful, could prioritize protein candidates for more exhaustive experimental analysis. Two fundamental assumptions underlie our analytical framework, namely: (1) an important subset of protein kinases regulating transition through the IDC is transcriptionally regulated, and this is captured in the publicly available microarray data; and (2) increases in protein kinase transcript levels predictably precede peak protein synthesis, the latter defining the time at which a given protein kinase plays its critical role in IDC progression. Our approach requires explicitly addressing the confounding factor of decaying synchrony of parasite cultures in order to improve the reliability of predictions. For microarray experiments, parasite cultures are initially synchronized. However, these cultures gradually lose synchrony over the experimental course [3]. Consequently, gene expression levels at discrete time points for individual parasites are not directly inferred from the observed microarray data, which reflect an ensemble of increasingly asynchronous parasite transcription profiles. To address this issue, we introduce a deconvolution approach for reconstructing the intrinsic gene expression pattern of individual parasites from microarray data. This work was partly presented in our previous paper [8]. We identify and account for three main factors driving asynchrony in gene expression profiles during a microarray experiment, namely: (1) diversity of infection time; (2) diversity of growth rate; and (3) the emergence of early stage parasites intermixed with later stage parasites. By including these considerations into our computational framework, we are able to more accurately reconstruct single cell, global gene expression profiles, from which the temporal expression profile for individual protein kinases are determined.

## Methods

A publicly available microarray data set on three *P. falciparum* strains (HB3, 3D7 and Dd2) is used in this work [4,13]. In the HB3 data set, the expression profiles of 4345 oligonucleotide sequences have been measured at 48 time points with 1 hour interval. The $23^{rd}$ and $29^{th}$ data points are missing for all oligonucleotide sequences. The oligonucleotide sequences associated with protein kinases involved in the *P. falciparum* life cycle are retrieved from the PlasmoDB database. Some protein kinases have several unique oligonucleotide sequences. For those protein kinases, an average trace is calculated from the curves associated with each oligonucleotide sequence. In this fashion, the gene expression profiles of 65 protein kinases are collected from the data set of HB3. Along the same lines, 52 protein kinases and 51 protein kinases are collected from the data set 3D7 and Dd2 respectively. In the rest of this section, we present a computational method to extract the intrinsic gene expression pattern from the microarray data.

Gene expression levels obtained in microarray experiments are aggregates across many individual iRBCs. First, we will derive a set of linear equations (15) that relates the microarray data to the intrinsic expression pattern of individual iRBC. Next, by solving the corresponding linear inverse problem (16), we reconstruct the expression pattern. All symbols used in this paper are explained in the Table 1.

### Analysis of decaying synchrony

We assume that three main factors drive the iRBCs out of synchronization in the microarray experiment,

**Table 1 Explanations of symbols**

| Symbols | Explanations |
|---|---|
| $M$ | the total amount of cells in the media, it consists of RBCs and iRBCs |
| $S(t)$ | the number of schizonts which infect RBCs at time $t$ |
| $R(t)$ | the number of fresh RBCs infected by schizont at time $t$ |
| $a_{\text{in}}$ | the average number of RBCs infected by one schizont during the infection period |
| $a_{\text{af}}$ | the average number of RBCs infected by one schizont after the infection period |
| $\widetilde{L}$ | the normalized life span of individual iRBCs |
| $p_{\widetilde{L}}(l)$ | the probability density function of normalized life span $\widetilde{L}$ |
| $L$ | the average life span of iRBCs |
| $L'$ | the individual life span of iRBCs |
| $\ell$ | the cell age of iRBC in hours |
| $\{f_i(\ell),\ \ell \in [0, L]\}$ | the intrinsic gene expression pattern of protein $i$ |
| $\ell_{\text{re}}$ | the rescaled cell age according to its normalized life span $\widetilde{L}$ |
| $\{f_i(\ell_{\text{re}}),\ \ell_{\text{re}} = \widetilde{L}\ell\ ,\ \ell \in [0, L']\}$ | the gene expression profile of individual iRBC on protein $i$ |
| $S_f(t)$ | the number of fast-growing iRBCs which infect RBCs at time $t$ |
| $R_f(t)$ | the number of RBCs infected by fast-growing iRBCs at time $t$ |
| $N(t)$ | the total number of iRBCs that have been infected at time $t$ |
| $N(t, \ell_{\text{re}})$ | the number of iRBCs which reach rescaled cell age $\ell_{\text{re}}$ at time $t$ |
| $e_i(t)$ | the observed expression level of protein $i$ at time $t$ |
| $\widetilde{f}_i(\ell)$ | the normalized gene expression pattern on protein $i$ |
| $s$ | the transitions between three stages of iRBC: ring, trophozoite, and schizont |
| $L_i(s)$ | the likelihood that protein kinases $i$ is involved in regulating the stage transition $s$ |
| $T_s$ | the time point when stage transition $s$ occur |
| $n$ | the average number of merozoites released by one schizont |
| $p$ | the probability that one merozoites does not infect any RBC |
| $V$ | the volume of merozoites can travel after it is released from schizont |
| $V_{\text{total}}$ | the volume of whole media |
| $m$ | the total number of RBCs in the media |

namely: (A) diversity of infection time, (B) diversity of growth rate, and (C) the emergence of early stage parasites intermixed with later stage parasites. These are discussed below.

***Diversity of infection time***

As illustrated in Figure 1, the invasion of RBCs does not occur simultaneously. In the microarray experiment [4], late-stage schizonts are synchronized by six sorbitol treatments on three generations. Prior to the first microarray time point, fresh RBCs are infected by late-stage schizonts within two hours, raising the parasitemia from 5% to 16%. After the invasion period, 80% of parasites are in the ring stage. Let $M$ stand for the total number of cells in the media. In other words, 5% × $M$ of schizonts infect 80% × 16% × $M$ of ring during two-hour invasion, remaining 20% × 16% × $M$ of schizonts are still alive after the invasion period. Although the concentration of RBCs is reduced from 14% to 3.3% immediately after the invasion period, RBCs can still be infected as long as schizonts remain. Therefore, a large amount of RBCs will be infected after the two-hour invasion.

Let $R(t)$ denote the number of fresh RBCs infected by schizont at time $t$ (hours). In the perfectly synchronized case, $R(t)$ should be a Dirac delta function, which means all iRBCs are simultaneously infected at the same time. In microarray experiment, however, $R(t)$ has a high value during the invasion period, and it maintains positive as long as schizonts remain.

Let $S(t)$ stand for the number of schizonts which infect RBCs at time $t$. The parameters $a_{\text{in}}$ and $a_{\text{af}}$ denote the average number of RBCs infected by one schizont during and after the infection period respectively. Therefore, the expression of $R(t)$ can be written as:

$$R(t) = \begin{cases} a_{\text{in}}S(t), & \text{if t} \in \text{[infection period]}, \\ a_{\text{af}}S(t), & \text{if t} \in \text{[after infection period]}. \end{cases} \quad (1)$$

At the end of this section, we explain how to estimate the parameters $a_{\text{in}}$, $a_{\text{af}}$ and function $S(t)$.

***Diversity of growth rate***

The iRBCs grow at different rates [4]. Consequently, as illustrated in Figure 2, synchrony gradually decays, even if all iRBCs are simultaneously infected at the same

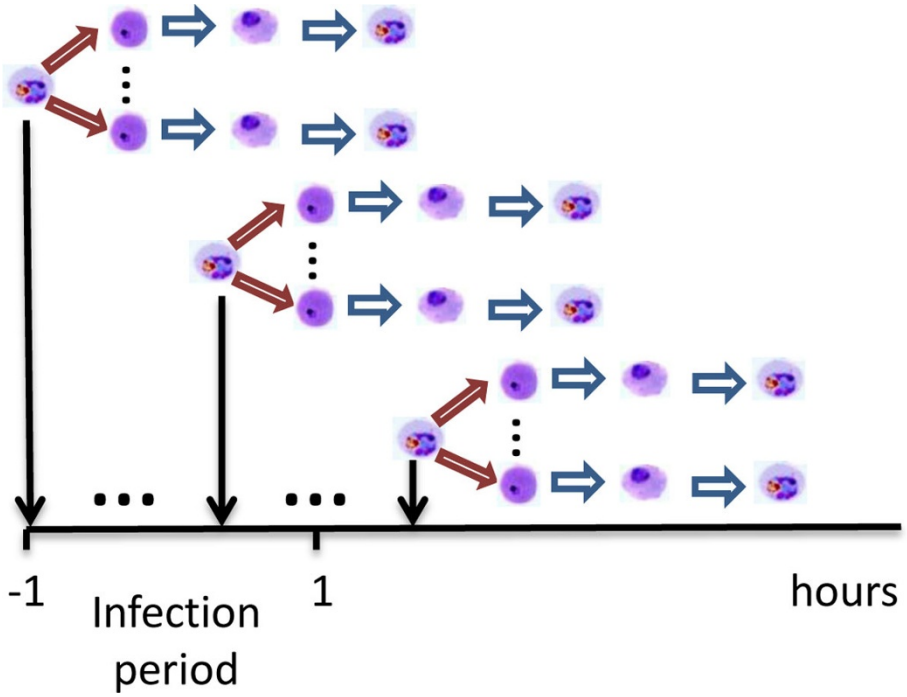


**Figure 1 Diversity of infection time among iRBCs**. The invasion of RBCs does not occur simultaneously. RBCs will be continuously infected as long as schizonts remain.

time. Let the random variable $\widetilde{L}$ indicate the normalized life span of iRBC, which is the quotient of individual life span $L'$ and the average life span $L$:

$$\widetilde{L} = \frac{L'}{L}. \tag{2}$$

$\tilde{L}$ is assumed to follow a normal distribution: $\widetilde{L} \sim N(1, \sigma^2)$. Hence the probability density function $p_{\widetilde{L}}(l)$ of normalized life span $\widetilde{L}$ can be written as:

$$p_{\widetilde{L}}(l) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(l-1)^2}{2\sigma^2}}, \tag{3}$$

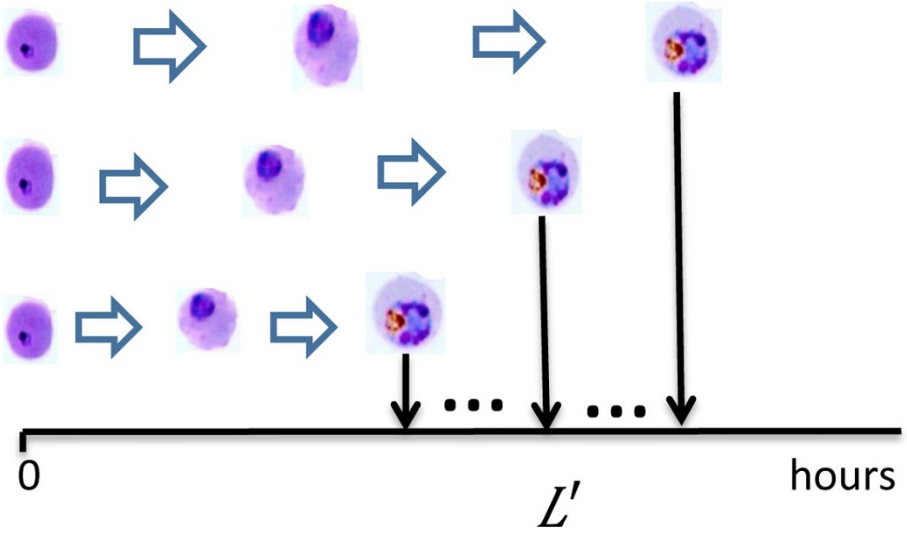


**Figure 2 Diversity of growth rate among iRBCs**. The iRBCs grow at different rates.

where the value of $\sigma$ is adjusted to fit the experimental observations. The details will be discussed later in this paper.

Let $\{f_i(\ell), \ell \in [0, L]\}$ be the intrinsic gene expression pattern of protein $i$ on one complete life span, where $\ell$ denotes the cell age of iRBC (hours). Since $f_i(\ell)$ represents the common pattern shared by individual RBC, the expression profile of individual iRBC is assumed to be $\{f_i(\ell_{re}), \ell_{re} = \ell/\widetilde{L}, \ell \in [0, L']\}$, where $\ell_{re}$ denotes the rescaled cell age according to its normalized life span $\widetilde{L}$ For instance, one iRBC has been infected for $\ell$ hours, its current expression level on the protein $i$ is written as $f_i(\ell/\widetilde{L})$, where $\widetilde{L}$ is the normalized life span of the corresponding RBC.

### *The emergence of early stage parasites intermixed with later stage parasites*

Due to the diversity of growth rate, a few iRBCs can reach the late stage of schizont early. As a result, those fast-growing iRBCs can infect additional fresh RBCs, as illustrated in Figure 3. This phenomenon has been observed in experiments [4]. Let $S_f(t)$ denote the number of fast-growing iRBCs which reach end of their life span at time $t$. Let $R_f(t)$ be the number of RBCs which are infected by these iRBCs at time $t$. We assume that $R_f(t)$ is proportional to $S_f(t)$:

$$R_f(t) = a_{af} S_f(t). \tag{4}$$

The invasion factor $a_{af}$ stands for the average number of fresh RBCs that will be infected by one schizont after the invasion period.

As shown in (2), the normalized life span $\widetilde{L}$ stands for the quotient of individual life span $L'$ and the average life span $L$. The number of iRBCs that start and end their life span at time $t$ are denoted as $R(t)$ and $S_f(t)$ respectively. Given the probability density function $p_{\widetilde{L}}(l)$ of normalized life span $\widetilde{L}$, the number of iRBCs that have reached the end of their life span at time $t$ can be written as follow:

$$\begin{aligned}\int_{-\infty}^{t} S_f(t')\mathrm{d}t' &= \int_{-\infty}^{+\infty} R(t') P_{\widetilde{L}}\left(\widetilde{L} < \frac{t-t'}{L}\right)\mathrm{d}t' \\ &= \int_{-\infty}^{+\infty} R(t') \int_{-\infty}^{\frac{t-t'}{L}} p_{\widetilde{L}}(l)\mathrm{d}l\mathrm{d}t'.\end{aligned} \tag{5}$$

Therefore, the expression of $S_f(t)$ can be derived from (5) as:

$$\begin{aligned}S_f(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{t} S_f(t')\mathrm{d}t' \\ &= \frac{1}{L}\int_{-\infty}^{+\infty} R(t') p_{\widetilde{L}}\left(\frac{t-t'}{L}\right)\mathrm{d}t'.\end{aligned} \tag{6}$$

Therefore, we have the expression of $R_f(t)$ by substituting (6) into (4):

$$R_f(t) = \frac{a_{af}}{L}\int_{+\infty}^{+\infty} R(t') p_{\widetilde{L}}\left(\frac{t-t'}{L}\right)\mathrm{d}t'. \tag{7}$$

### Simulation of iRBCs population distribution

Let $N(t)$ denote the total number of iRBCs at time $t$. $N(t)$ consists of 3 parts: the late-schizonts that infect fresh RBCs during the infection period and have not yet burst at time $t$, the first generation of iRBCs (infected by late-schizonts around infection period) that have not yet reached the end of their life span at time $t$, and second generation of iRBCs (infected by fast-growing iRBCs) that have not yet reached the end of their life span at

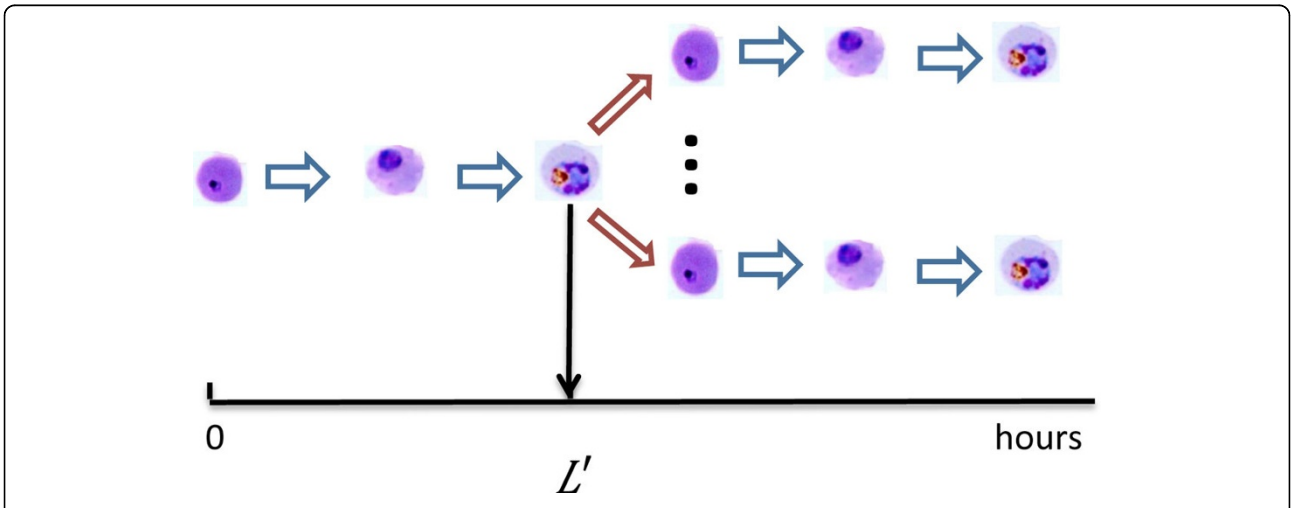


**Figure 3 The emergence of early stage parasites intermixed with later stage parasites**. Additional fresh RBCs will be infected by fast-growing iRBCs which reach the late stage of schizont early.

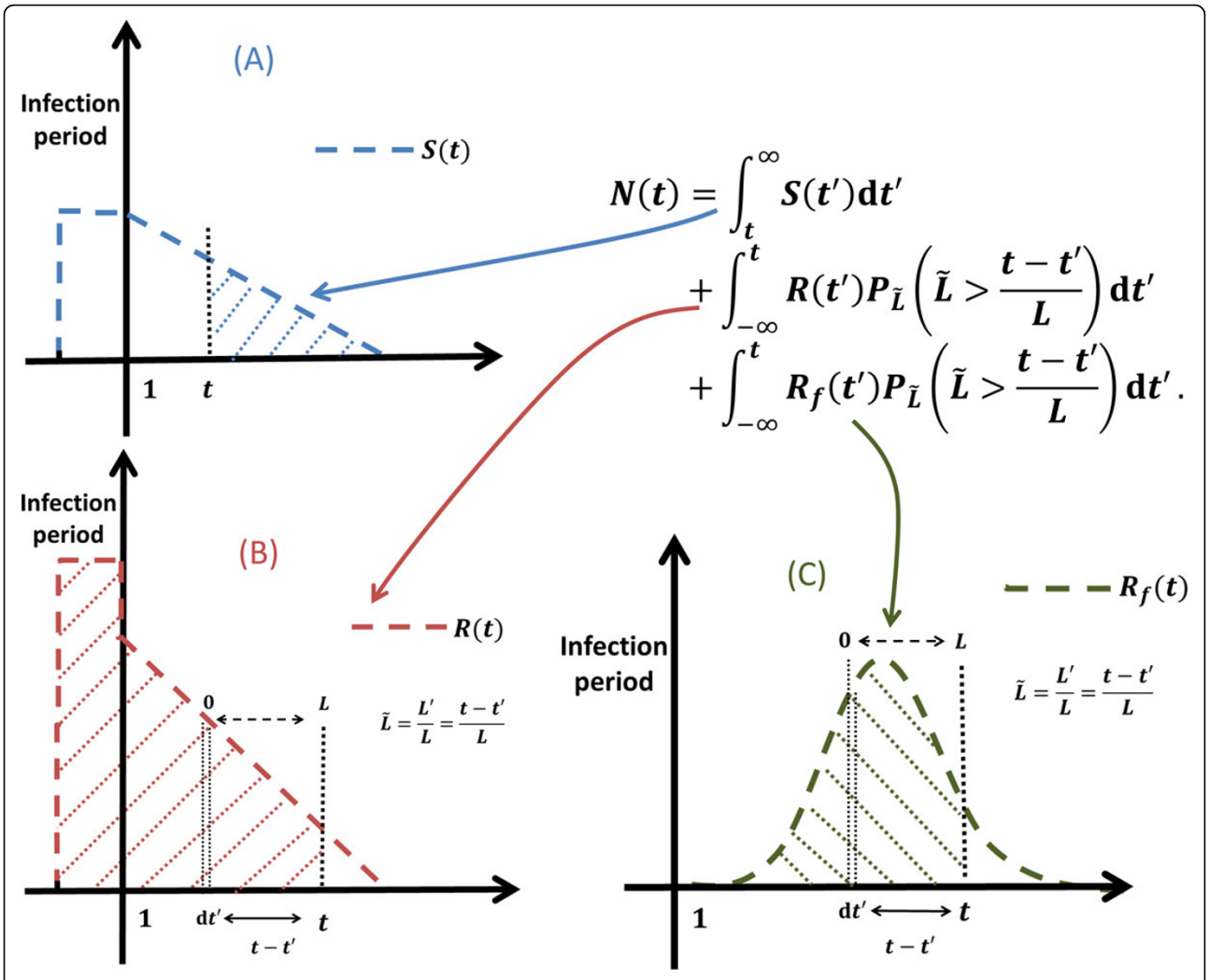


**Figure 4 The total number of iRBCs $N(t)$ at time $t$.** (A) The late-schizonts that infect fresh RBCs during infection period and have not yet burst at time $t$. (B) The first generation of iRBCs (infected by late-schizonts around infection period) that have not yet reached the end of their life span at time $t$. (C) Second generation of iRBCs (infected by fast-growing iRBCs) that have not yet reached the end of their life span at time $t$.

time $t$. Therefore, as illustrated in Figure 4, we can decompose $N(t)$ as:

$$N(t) = \int_t^{+\infty} S(t')\mathrm{d}t' + \int_{+\infty}^{t} R(t')p_{\tilde{L}}\left(\tilde{L} > \frac{t-t'}{L}\right)\mathrm{d}t' + \int_{+\infty}^{t} R_f(t')p_{\tilde{L}}\left(\tilde{L} > \frac{t-t'}{L}\right)\mathrm{d}t', \quad (8)$$

where $S(t)$ stands for the number of late-schizonts bursts at time $t$, $R(t)$ denotes the number of RBCs infected by late-schizonts at time $t$, and $R_f(t)$ is the number of RBCs which are infected by fast-growing iRBCs at time $t$. The expressions of $R(t)$ and $R_f(t)$ are given by (1) and (7) respectively. The expression of $S(t)$ will be derived in (25).

Let $N(t, \ell_{re})$ stand for the number of iRBCs that reach the rescaled cell age $\ell_{re}$ at time $t$. In other words, given the time t, $N(t, \ell_{re})$ indicates the distribution of iRBCs over a complete life span of iRBC (see Figure 5). Therefore, $N(t, \ell_{re})$ and $N(t)$ satisfy following relation:

$$N(t) = \int_{-\infty}^{L} N(t, \ell'_{re})\mathrm{d}\ell'_{re}. \quad (9)$$

Hence, as illustrated in Figure 6, the number of iRBCs that have not yet reached the rescaled cell age $\ell_{re}$ time $t$ can be expanded from (8) as follows:

$$\begin{aligned}\int_{-\infty}^{\ell_{re}} N(t, \ell'_{re})\mathrm{d}\ell'_{re} &= \int_t^{+\infty} S(t')P_{\tilde{L}}\left(\tilde{L} < \frac{t'-t}{L-\ell_{re}}\right)\mathrm{d}t' + \int_{-\infty}^{t} R(t')P_{\tilde{L}}\left(\tilde{L} > \frac{t-t'}{\ell_{re}}\right)\mathrm{d}t' \\ &\quad + \int_{-\infty}^{t} R_f(t')P_{\tilde{L}}\left(\tilde{L} > \frac{t-t'}{\ell_{re}}\right)\mathrm{d}t' \\ &= \int_t^{+\infty} S(t')\int_{-\infty}^{\frac{t'-t}{L-\ell_{re}}} p_{\tilde{L}}(\ell)\mathrm{d}l\mathrm{d}t' + \int_{-\infty}^{t} R(t')\int_{\frac{t-t'}{\ell_{re}}}^{\infty} p_{\tilde{L}}(l)\mathrm{d}l\mathrm{d}t' \\ &\quad + \int_{-\infty}^{t} R_f(t')\int_{\frac{t-t'}{\ell_{re}}}^{+\infty} p_{\tilde{L}}(l)\mathrm{d}l\mathrm{d}t'.\end{aligned} \quad (10)$$

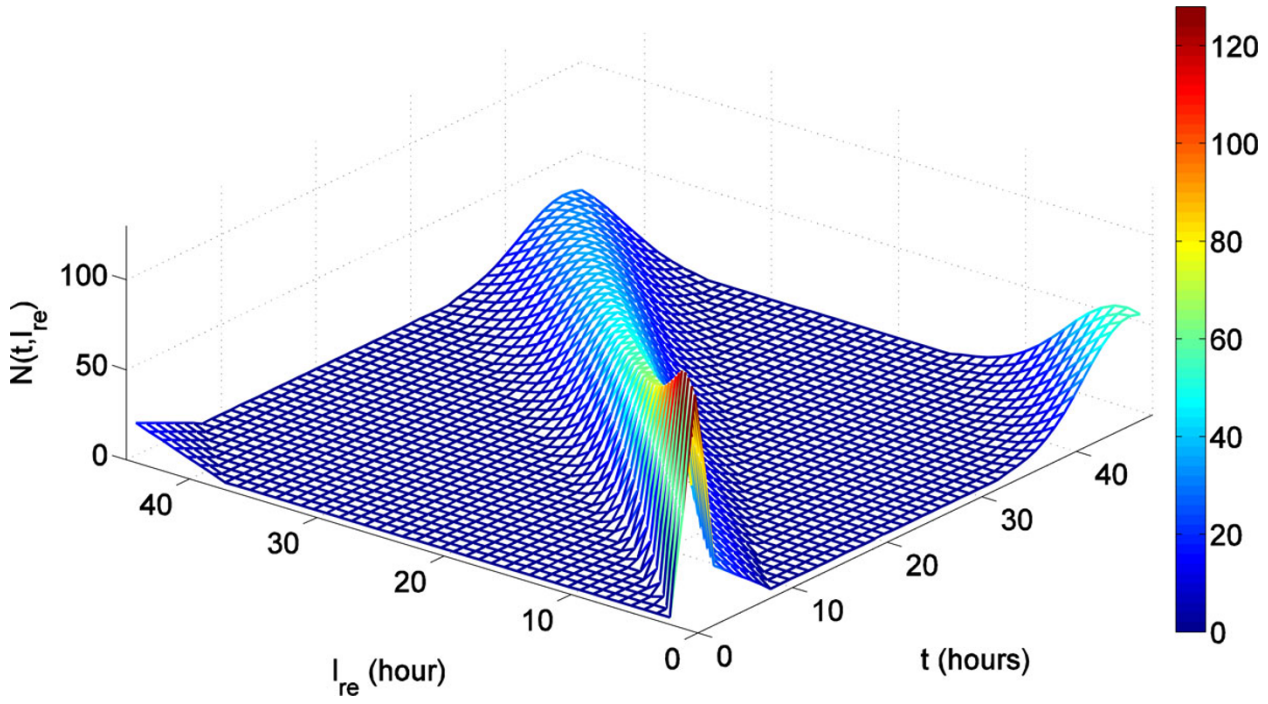


**Figure 5 iRBCs population distributions change over time**. N($t$, $\ell_{re}$) with $\sigma$ = 0.1. It indicates the number of iRBCs that reach the rescaled cell age $\ell_{re}$ at time $t$.

Therefore, $N(t, \ell_{re})$ can be derived from (10) as follows:

$$\begin{aligned} N(t,\ell_{re}) &= \frac{\mathrm{d}}{\mathrm{d}\ell_{re}}\int_{-\infty}^{\ell_{re}} N(t,\ell'_{re})\mathrm{d}\ell'_{re} \\ &= \int_t^{+\infty} S(t')p_{\widetilde{L}}\left(\frac{t'-t}{L-\ell_{re}}\right)\frac{t'-t}{(L-\ell_{re})^2}\mathrm{d}t' + \int_{-\infty}^{t} R(t')p_{\widetilde{L}}\left(\frac{t-t'}{\ell_{re}}\right)\frac{t-t'}{\ell_{re}^{\,2}}\mathrm{d}t' \\ &+ \int_{-\infty}^{t} R_f(t')p_{\widetilde{L}}\left(\frac{t-t'}{\ell_{re}}\right)\frac{t-t'}{\ell_{re}^{\,2}}\mathrm{d}t'. \end{aligned} \quad (11)$$

By substituting the expression of $R_f(t)$ (7) into (11), $N(t, \ell_{re})$ can be written as:

$$\begin{aligned} N(t,\ell_{re}) &= \int_t^{+\infty} S(t')p_{\widetilde{L}}\left(\frac{t'-t}{L-\ell_{re}}\right)\frac{t'-t}{(L-\ell_{re})^2}\mathrm{d}t' + \int_{-\infty}^{t} R(t')p_{\widetilde{L}}\left(\frac{t-t'}{\ell_{re}}\right)\frac{t-t'}{\ell_{re}^{\,2}}\mathrm{d}t' \\ &+ \frac{a_{af}}{L}\int_{-\infty}^{t} p_{\widetilde{L}}\left(\frac{t-t'}{\ell_{re}}\right)\frac{t-t'}{\ell_{re}^{\,2}}\mathrm{d}t'\int_{-\infty}^{+\infty} R(t'')p_{\widetilde{L}}\left(\frac{t'-t''}{L}\right)\mathrm{d}t''. \end{aligned} \quad (12)$$

### Modeling of gene expression level

The gene expression levels obtained in microarray experiments are aggregates across many individual iRBCs. This superposition across iRBCs is modeled by means of a linear system. Let $e_i(t)$ denote the observed expression level of protein $i$ at time $t$. As discussed earlier, $N(t, \ell_{re})$ denotes the distribution of iRBCs on rescaled cell age $\ell_{re}$ and $\{f_i(\ell_{re}), \ell_{re} = \ell/\widetilde{L}, \ell \in [0.L']\}$ stands for the gene expression level of individual iRBC according to its rescaled cell age $\ell_{re}$. Therefore, the observed expression level $e_i(t)$ can be written as a integral over one complete life span of iRBCs as follows:

$$e_i(t) = \int_0^L N(t,\ell_{re})f_i(\ell_{re})\mathrm{d}\ell_{re}. \quad (13)$$

We represent the continuous function $f_i(s)$ and $N(t, \ell_{re})$ as a series of discrete points $\{f_i(1), f_i(2),..., f_i(T)\}$, and $\{N(t,1), N(t,2),...,N(t,T)\}$ respectively. Therefore, $e_i(t)$ can be approximated as:

$$e_i(t) \approx \sum_{\ell_{re}=1}^{L} N(t,\ell_{re})f_i(\ell_{re})\Delta\ell_{re}. \quad (14)$$

In microarray experiments, gene expression levels are measured at a series of discrete time points. Hence, the resulting observed expression level $e_i(t)$ is a series of discrete value. In the dataset HB3, for example, the expression levels are measured every hour over a period of 48 hours. Therefore, a linear system can be derived based on (14):

$$\underbrace{\begin{pmatrix} N(1,1) & N(1,2) & \dots & N(1,L) \\ N(2,1) & N(2,2) & \dots & N(2,L) \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}}_{\mathrm{A}} \underbrace{\begin{pmatrix} f_i(1) \\ f_i(2) \\ \vdots \\ f_i(L) \end{pmatrix}}_{\mathrm{x}} = \underbrace{\begin{pmatrix} e_i(1) \\ e_i(2) \\ \vdots \end{pmatrix}}_{\mathrm{b}}. \quad (15)$$

The observation matrix $A$ can be calculated by means of the equation (12). The element of matrix $A$ at row $t$ and column $\ell_{re}$ denotes the number of iRBCs that reach

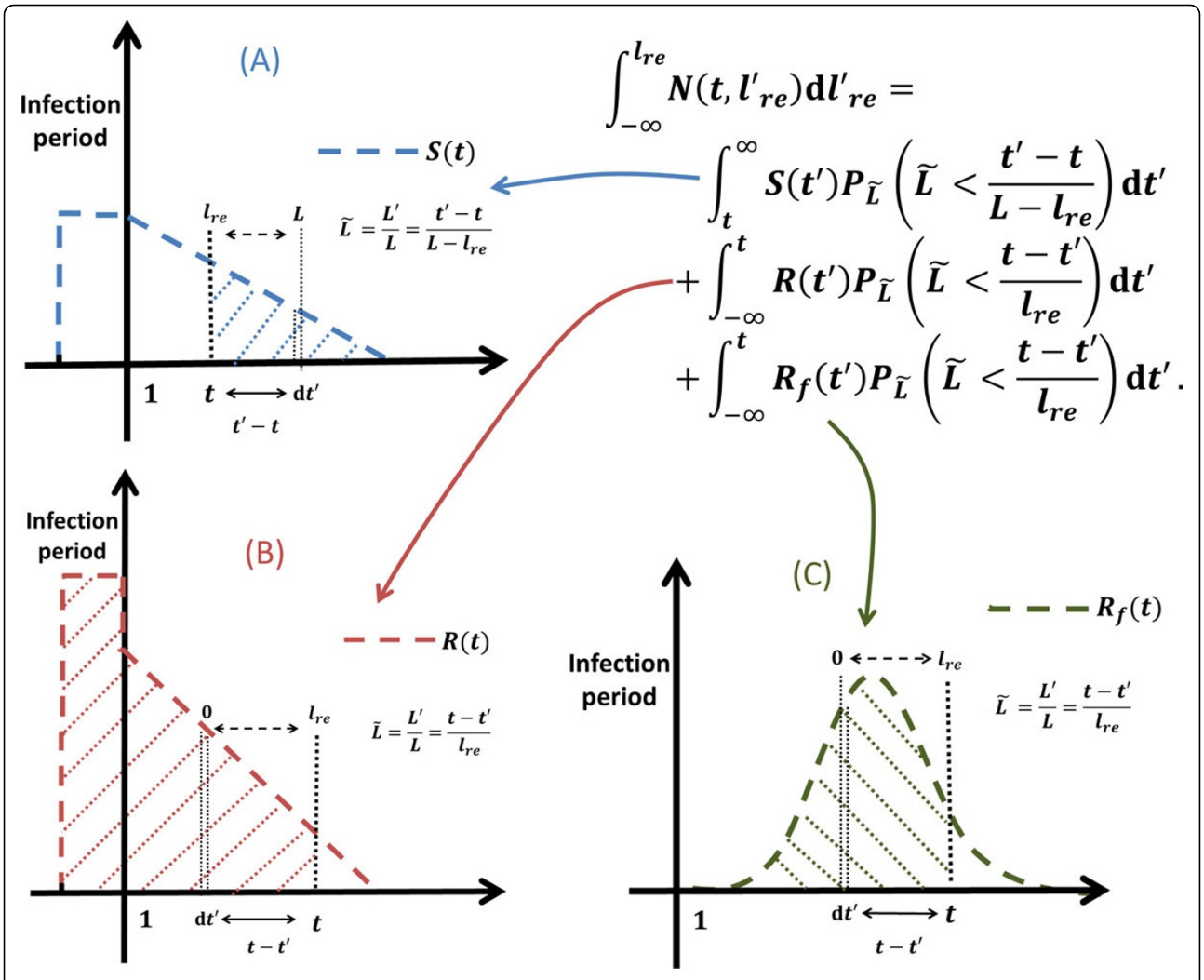


**Figure 6 The number of iRBCs that have not reached the rescaled cell age $\ell_{re}$ at time t**. (A) The late-schizonts that infect fresh RBCs around infection period and have not reached the rescaled cell age $\ell_{re}$ at time t. (B) The first generation of iRBCs that have not reached the rescaled cell age $\ell_{re}$ at time t. (C) Second generation of iRBCs that have not reached the rescaled cell age $\ell_{re}$ at time *t*.

the rescaled cell age $l_{re}$ at time point $t$. The constant vector $b$ stands for the gene expression levels observed in microarray experiment. The unknown variable vector $x$ is the intrinsic gene expression pattern of individual cell. We can find $x$ by solving the discrete linear inverse problem (15).

### Reconstruction of synchronized expression profile

For each protein $i$, the observed expression level $e_i$ is modeled as the superposition of the expression level $f_i$ of individual iRBCs, as described in (15). To solve the described linear inverse problem (15), we minimize an objective function. The objective function contains the squared error $(Ax - b)(Ax - b)^{\top}$. The intrinsic gene expression pattern of iRBCs is assumed to be a smooth curve. Therefore, we also include square difference as $\sum_{k=1}^{L-1}(x_k - x_{k+1})^2 + (x_L - x_1)^2$ in the objection function. We also need to impose the constraint $x \geq 0$ because expression levels are positive. In summary, we compute the intrinsic gene expression pattern $x$ as follows:

$$x = \arg\min_{x \geq 0} \left\{ \underbrace{(Ax-b)(Ax-b)^{\top}}_{\text{square error}} + c \underbrace{\left[ \sum_{k=1}^{L-1} (x_k - x_{k+1})^2 + (x_L - x_1)^2 \right]}_{\text{gradients}} \right\}. \quad (16)$$

We solve (16) numerically by means of the function *fmincon* in the Optimization Toolbox of Matlab (MATLAB 7.9, The MathWorks Inc.)

### Prediction of protein kinases

As we discussed earlier, an increased gene expression level before a cell stage transition is regarded as a sign

that the corresponding protein kinase is involved in that stage transition. We denote the likelihood that protein kinases $i$ is involved in regulating the stage transition $s$ by $L_i(s)$. Let $T_s$ stand for the time point when stage transition $s$ occur. The expression of likelihood $L_i(s)$ is derived in following.

Let $\tilde{f}_i(\ell)$ be the normalized expression pattern such that its integral over one iRBC life span is equal to 1:

$$\tilde{f}_i(\ell) = \frac{f_i(\ell)}{\sum_{\ell=1}^{L} f_i(\ell)\Delta l}, \tag{17}$$

where $\ell = 1,..., L$. Since protein kinases are often regulated at translational and post-translational levels [5], $L_i(s)$ is estimated as the maximum sum of $W$ consecutive points of $\tilde{f}_i(\ell)$ that appear $H$ hours prior to $T_s$.

$$L_i(s) = \max_{T_s - H \le \ell_1 \le T_s - W + 1} \left\{ \sum_{\ell=\ell_1}^{\ell_1+W-1} \tilde{f}_i(\ell) \right\}. \tag{18}$$

Consequently, protein kinase $i$ can be prioritized by its likelihood $L_i(s)$ of being involved in a stage transition.

### Parameter estimation

The proposed method for reconstructing expression patterns consists of two main steps. First, the linear system described in (15) is built by calculating the cell age distribution $N(t, \ell_{re})$ (12) for each element of the observation matrix $A$. Second, the expression pattern is reconstructed by solving the discrete linear inverse problem (16). Therefore, the performance of our method mainly depends on the calculation of $N(t, \ell_{re})$. As shown in equation (12), $N(t, \ell_{re})$ is dominated by three sets of parameters: the infection factors $a_{in}$ and $a_{af}$, the burst rate of schizonts $S(t)$, and the standard deviation $\sigma$ of normalized life span $\tilde{L}$.

The infection factors $a_{in}$, $a_{af}$, and burst rate of schizonts $S(t)$ can be accurately estimated from parasitemia and percent representation of iRBC observed at each time point. As shown in Figure 7, the percent representation of iRBC is available at each time point [4]. However, the parasitemia is given only at two time points, one during and one after the infection period. In this section, we explain how we estimate all parameters from the specification of the microarray experiments [4].

#### *Infection factors*

In the proposed model, the infection factors, $a_{in}$ and $a_{af}$ denote the number of RBC that can be infected by one bursted schizont during the infection period and after the period respectively. As discussed earlier, 5% × $M$ schizonts infect 80% ×16% × $M$ of ring cell during invasion period, and the remaining 20% × 16% × $M$ schizonts are still alive after the invasion period. Therefore, the value of $a_{in}$ can be deduced as follows:

$$a_{in} = \frac{16\% \times 80\% \times M}{5\% \times M - 16\% \times 20\% \times M} \approx 7.11. \tag{19}$$

After the invasion period, the cell concentration (both fresh and infected) is reduced from 14% to 3.3%. To estimate the value of $a_{af}$, we propose a simple model to describe how infection factor is influenced by the cell concentration.

At the end of the life span, the cell membrane of schizont bursts, and merozoites are released to infect other RBCs [14]. We denote by $p$ the probability that a merozoite does not infect any RBC, and let $n$ represent the average number of merozoites released by one schizont. The value of $n$ is estimated to be 14, because one schizont usually contains 12 to 16 merozoites [14]. Therefore, the average number of RBCs infected by one bursted schizont is given by $n(1 - p)$.

In the following, we derive an expression for the probability $p$. First, we assume that each merozoite can travel a certain space after it is released from schizont, and let $V$ be the volume of this space. Second, we also assume that if a RBC appears in the space indicated by $V$, it will be immediately infected by the corresponding merozoite. Let $V_{total}$ denote the volume of whole media. Let $m$ stand for the total number of RBCs remaining in the media. Hence, the value of $p$ can be estimated as the probability that none of the $m$ RBCs appear in the space $V$. If the RBCs are uniformly distributed in $V_{total}$, it follows:

$$p = \left( \frac{V_{total} - V}{V_{total}} \right)^m. \tag{20}$$

The concentration of RBCs is reduced by adjusting the volume of culture $V_{total}$ from 1000 milliliter to 4500 milliliter after the infection period [4]. Therefore, the expression of $a_{in}$ and $a_{af}$ can be obtained by substituting (20) into $n(1 - p)$ as follows:

$$\begin{cases} a_{in} = 14 \left( 1 - \left( \frac{1000 - V}{1000} \right)^m \right) \\ a_{af} = 14 \left( 1 - \left( \frac{4500 - V}{4500} \right)^{m'} \right). \end{cases} \tag{21}$$

At the beginning of the experiment, the culture is initialized by 115.0 milliliter of purified RBC [4]. Human blood has 4 to 6 million RBC per microliter (cubic millimeter), and the corresponding hematocrit is about 45% [15]. Hence, the number of RBC $m$ before the infection period can be estimated as:

$$m = \frac{5,000,000 \times 115,000}{45\%} \approx 1.28 \times 10^{12}. \tag{22}$$

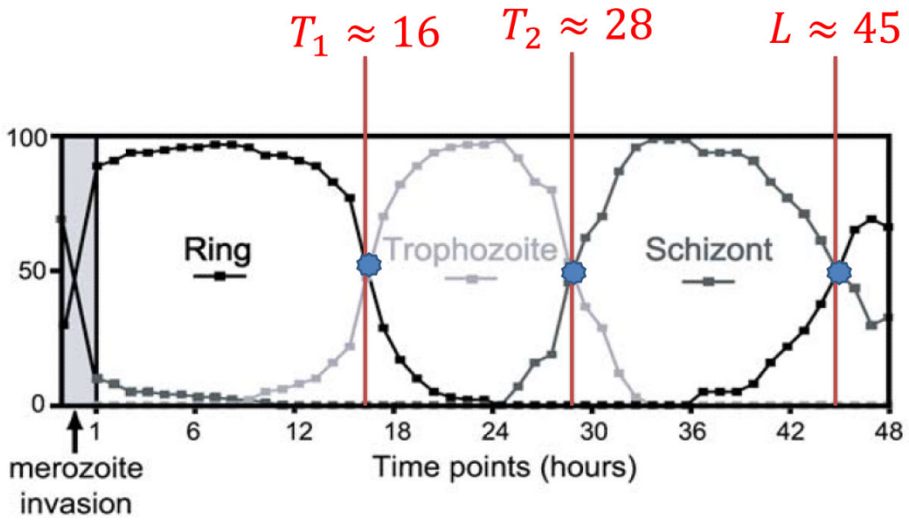


**Figure 7 Percent representation of iRBCs observed in experiment**. The boundaries ($T_1$, $T_2$ and $L$) between the three stages (ring, trophozoite, and schizont) are approximated from the percent representation of iRBCs observed at every time point in experiment [4].

After the infection period, the parasitemia is raised from 5% to 16%. Therefore, the number of RBC before the infection period $m$ and after the infection period $m'$ satisfy following relation:

$$\frac{m}{100\% - 5\%} = \frac{m'}{100\% - 16\%}. \tag{23}$$

Consequently, the number of RBC after the infection period $m'$ can be estimated. By substituting the value of $a_{\text{in}}$, $m$ and $m'$ into (21), the value of $V$ and $a_{\text{af}}$ can be deduced as follows:

$$\begin{cases} V \approx 5.55 \times 10^{-10} \text{millilitre}, \\ a_{\text{af}} \approx 1.82. \end{cases} \tag{24}$$

The derivation of $V$ is related to the concept of mean free paths in physics, roughly the diffusion length for the first binding event or the lifetime [16].

***Burst rate of schizonts***

The number of schizonts infecting RBCs at time $t$ is denoted by $S(t)$. As discussed earlier, 5% × $M$ - 16% × 20% × $M$ schizonts burst in the two-hour infection period, leaving 16% × 20% × $M$ schizonts alive till around 7 hours after the invasion period (8th microarray data point). In other words, 36% of schizonts burst in the first two hours followed by the remaining 64% of schizonts which burst within the next 7 hours. Therefore, $S(t)$ is approximated as a piecewise linear function whose integral on these two periods is 36 and 64 respectively. The value of $S(t)$ in the first two hours is also assumed to be a constant. As shown in Figure 8, consequently, the expression of $S(t)$ can be written as follows:

$$s(t) = \begin{cases} 18, & \text{if } -1 \le t < 1, \\ -2.54t + 20.54, & \text{if } 1 \le t \le 8.1, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

***Standard deviation of normalized life span***

Given $a_{\text{in}}$, $a_{\text{af}}$, and $S(t)$, the iRBCs population distributions $N(t, \ell_{\text{re}})$ (12) depend on the probability density function $p_{\widetilde{L}}(l)$ of normalized life span $\widetilde{L}$ (3). The normalized life span $\widetilde{L}$ has been assumed to follow a normal distribution. By definition, the mean of $\widetilde{L}$ is equal to 1. The standard deviation $\sigma$ is estimated such that the resulting iRBCs population distributions $N(t, \ell_{\text{re}})$ are in agreement with the experimental data.

Over one complete life cycle, iRBCs go through three life stages, i.e., ring, trophozoite and schizont, as follows:

$$\text{iRBC stage} = \begin{cases} \text{ring}, & \text{if } 0 \le \ell_{\text{re}} < T_1, \\ \text{trophozoite}, & \text{if } T_1 \le \ell_{\text{re}} < T_2, \\ \text{schizont}, & \text{if } T_2 \le \ell_{\text{re}} \le L. \end{cases} \tag{26}$$

As illustrated in Figure 7, the value of $T_1$, $T_2$ and $L$ are approximated from the percent representation of iRBCs. Given the boundaries between the three stages, iRBCs population distributions $N(t, \ell_{\text{re}})$ can represent the percent representation of iRBCs at each time point as follows. By definition, $N(t, \ell_{\text{re}})$ stands for the number of iRBCs that reach the rescaled cell age $l_{re}$ at time $t$. Hence, the number of iRBC at a specific stage (ring, trophozoite, and schizont) can be estimated as a

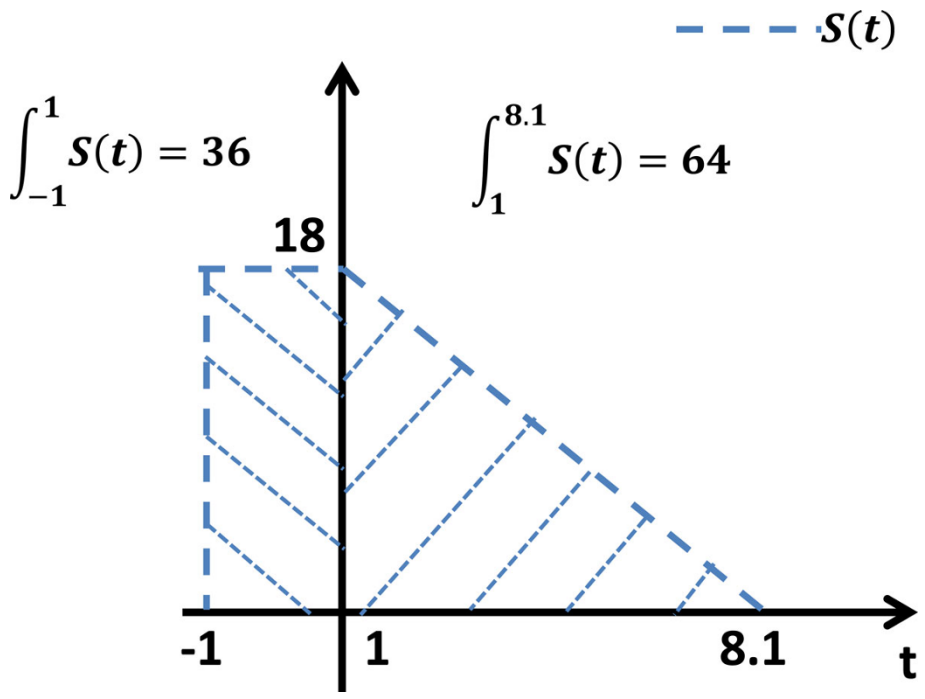


**Figure 8 Burst rate of schizonts**. The number of schizonts $S(t)$ infecting RBCs at time $t$ is approximated as a piecewise linear function according to experimental observations.

integral of $N(t, \ell_{re})$ on the rescaled cell age $l_{re}$. Therefore, the percent of ring cell at time $t$ can be calculated as follows:

$$\text{percent of ring at time } t = \frac{\overbrace{\int_0^{T_1} N(t,\ell_{re})d\ell_{re}}^{\text{ring}}}{\underbrace{\int_0^{T_1} N(t,\ell_{re})d\ell_{re}}_{\text{ring}} + \underbrace{\int_{T_1}^{T_2} N(t,\ell_{re})d\ell_{re}}_{\text{trophozoite}} + \underbrace{\int_{T_2}^{L} N(t,\ell_{re})d\ell_{re}}_{\text{schizont}}}. \quad (27)$$

The percent of trophozoite and schizont can be calculated similarly.

The distribution $N(t, \ell_{re})$ for $\sigma = 0.1$ is shown in Figure 5. The percent of ring, trophozoite, and schizont calculated from $N(t, \ell_{re})$ is illustrated in Figure 9. The corresponding calculated percent representation of iRBCs depends on the value of $\sigma$. Consequently, we tune the value of $\sigma$ such that percent representation observed in experiment, as illustrated in Figure 7, coincides with the simulated percent representation, as illustrated in Figure 9.

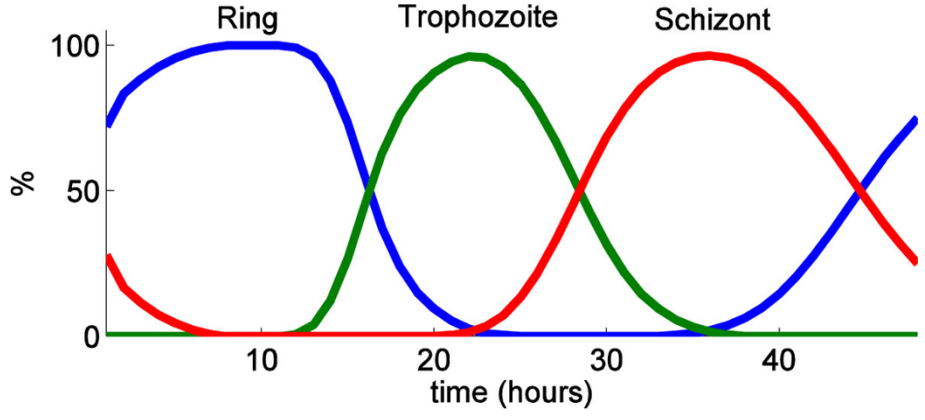


**Figure 9 Percent representation of iRBCs calculated from iRBCs population distributions**. The percent of iRBCs (ring trophozoite, and schizont) calculated from iRBCs population $N(t, \ell_{re})$ with $\sigma = 0.1$.

## Results on synthetic data

In this section, the proposed method is evaluated on synthetic microarray data.

### Generate synthetic microarray data

The data obtained in microarray experiments are aggregates across many cells. This superposition across cells is modeled by means of a linear system (15). The matrix $A$ in that linear system depends on the parameters ($a_{in}$, $a_{af}$, $\sigma$, and $S(t)$), estimated from microarray experiments [4]. Synthetic microarray data is generated by substituting the expression pattern of individual iRBC $x$ into the linear system. As shown in Figure 10, we generate synthetic microarray data for four expression patterns (A, B, C and D). Each of them simulates a protein kinase that serves a specific function in cell stage transition, and hence has high expression level in a short time period before a cell stage transition.

The microarray experiments measure the superposition across cells. Due to the decay of synchrony among cells, the intrinsic expression pattern is blurred in microarray data. By comparing the intrinsic expression patterns and synthetic microarray data shown in Figure 10, we demonstrate how decaying synchrony can blur the expression pattern.

### Tolerance to signal noise and missing data points

Due to the assay complexity in microarray experiments, signal noise [17] and missing time points [18] are common phenomena observed in real microarray data. To evaluate the performance of our method under similar conditions, expression patterns are reconstructed from synthetic microarray data contaminated by Gaussian white noise and/or missing data points. The resulting expression patterns are compared to the original expression patterns (A, B, C and D).

As shown in Figure 11, the synthetic microarray data of expression pattern (A) are contaminated by Gaussian white noise. The signal to noise ratios (SNR) of the resulting microarray data are 10, 15 and 20 dB respectively. To calculate matrix $A$, we choose the same parameter values ($a_{in}$, $a_{af}$, $\sigma$, and $S(t)$) as used to generate synthetic microarray data. In the next section, we will investigate how the results vary with other choice of those parameter values. The resulting matrix $A$ is substituted into the linear system described in equation (15), and the expression pattern is reconstructed by solving the corresponding linear inverse problem (16). The resulting expression patterns shown in Figure 11 are quite similar to the original expression pattern (A) given in Figure 10.

Along the same lines, the tolerance to missing data points is evaluated using synthetic microarray data with missing data points. We randomly remove data points (10, 20 and 30) from the microarray data of expression pattern respectively. Then expression pattern are reconstructed from contaminated microarray data. As shown in Figure 12, the proposed method generates a reliable

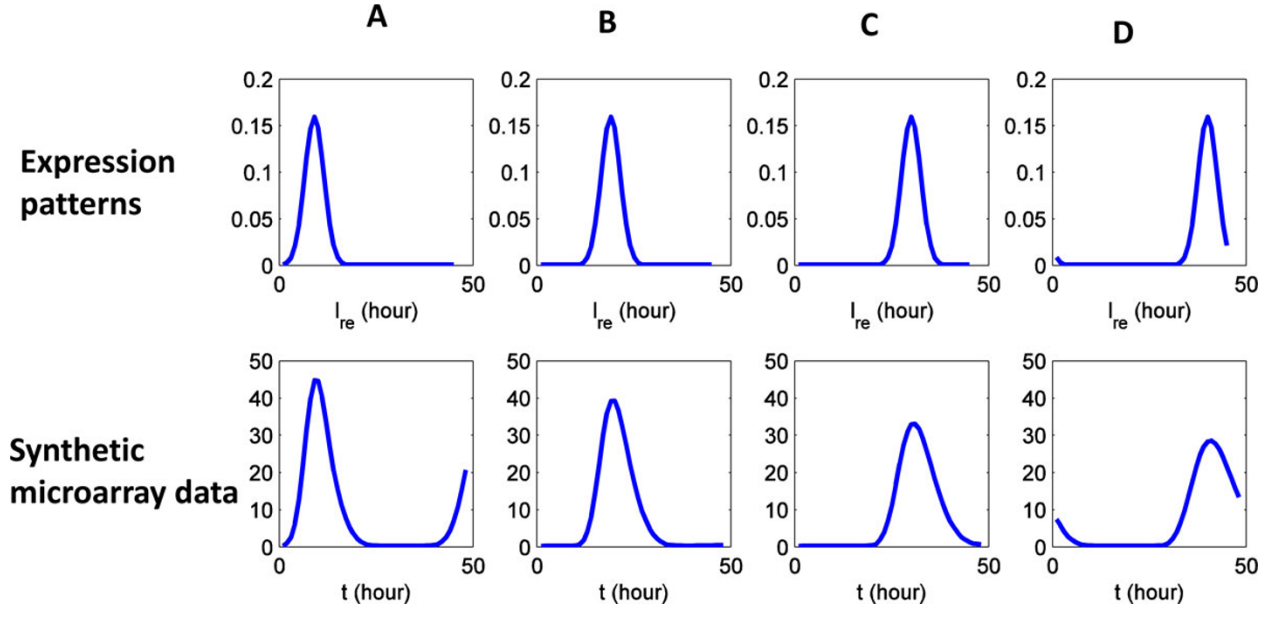


**Figure 10 Synthetic microarray data generated from known expression patterns**. Synthetic microarray data are generated for four expression patterns (A, B, C and D).

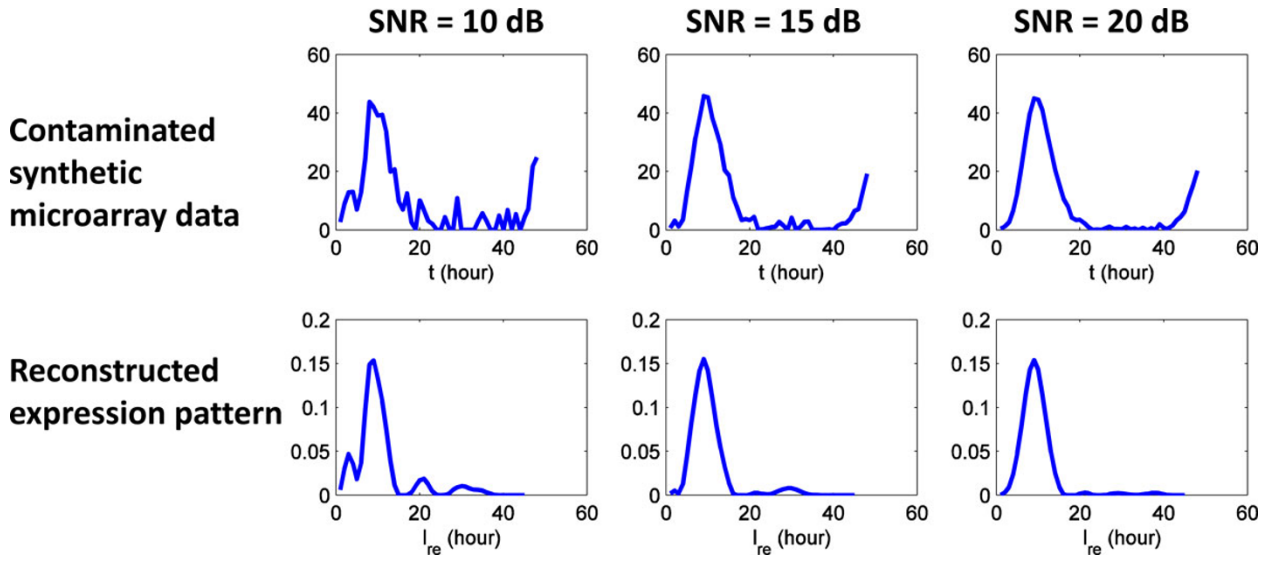


**Figure 11 Microarray data contaminated by Gaussian white noise**. Gaussian white noise is used to contaminate the synthetic microarray data of expression pattern (A). The reconstructed patterns are quite similar to the original expression pattern (A).

estimate for the original expression pattern (A), even when 30 of 48 time points are missing in the microarray data.

Real microarray data are usually contaminated by both signal noise and missing data points. Therefore, as shown in Figure 13, we conduct additional experiments on synthetic microarray data contaminated by both Gaussian white noise (SNR = 10, 20, and 30 dB) and 20 missing data points. The results of Figure 13 suggest that the proposed method can reliably reconstruct the expression pattern when the microarray data is simultaneously contaminated by both signal noise and missing data points.

### Sensitivity to model parameters

The performance of our method mainly depends on the calculation of cell age distribution $N(t, \ell_{re})$. As shown in equation (12), $N(t, \ell_{re})$ depends on three sets of

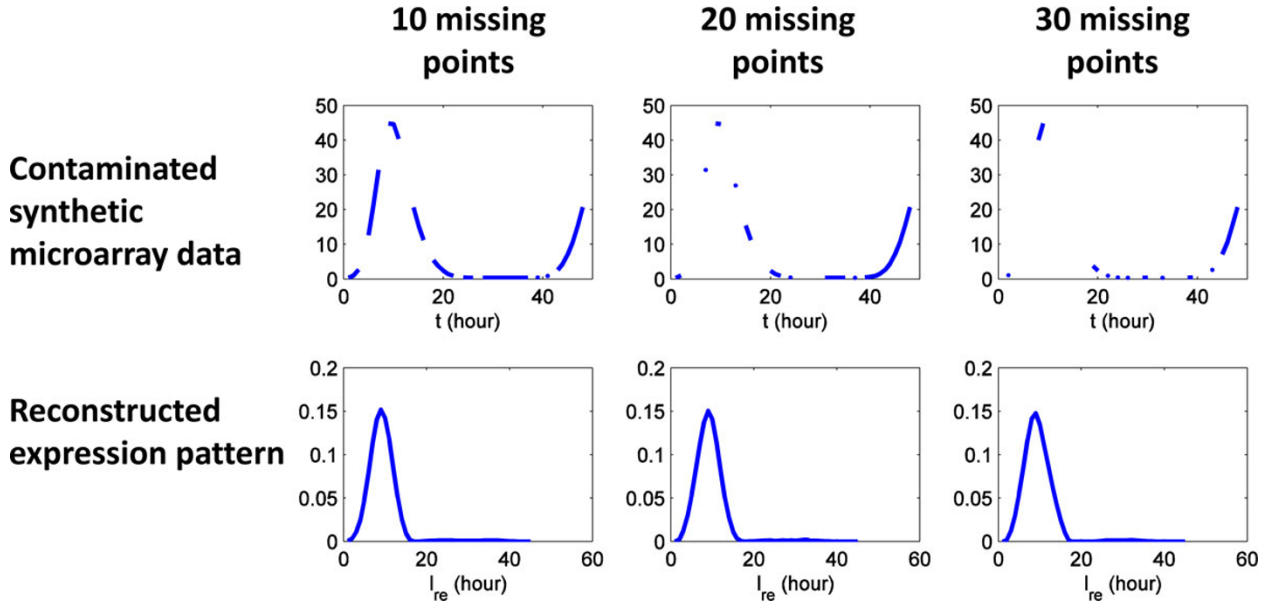


**Figure 12 Microarray data contaminated by random missing data points**. Data points (10, 20 and 30) are randomly remove from the microarray data of expression pattern. The expression pattern can be reliably reconstructed, even when 30 of 48 time points are missing in the microarray data.

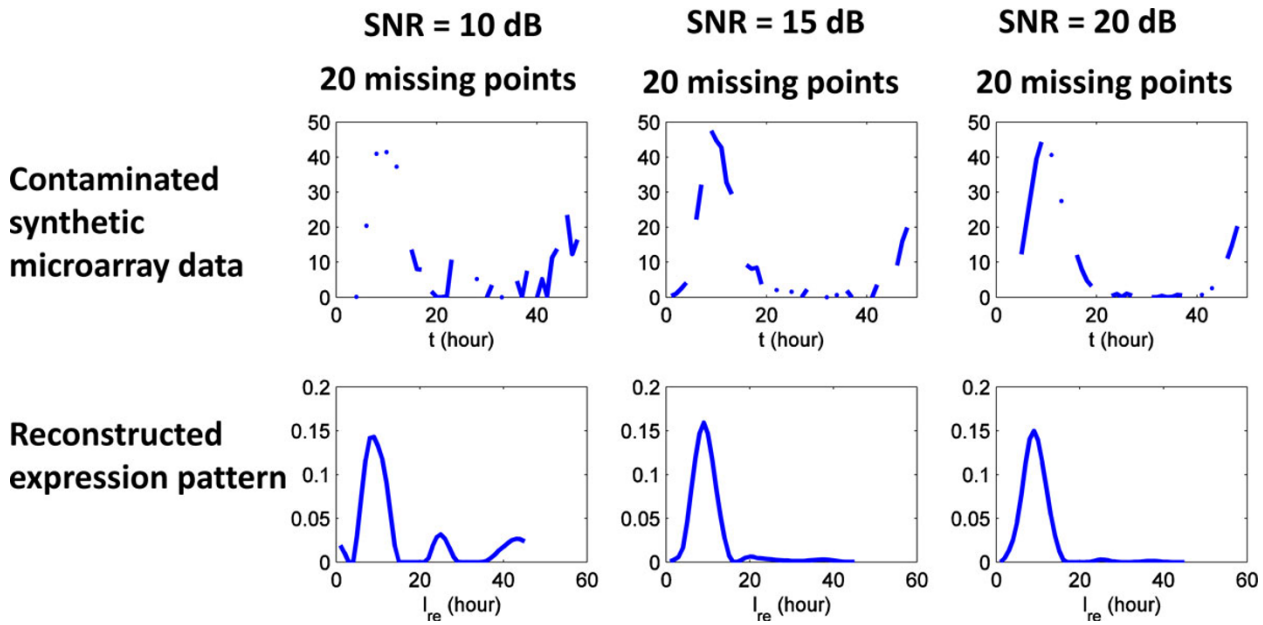


**Figure 13 Microarray data contaminated by both Gaussian white noise and random missing data points**. The proposed method can reliably reconstruct the expression pattern when the microarray data is simultaneously contaminated by both signal noise (SNR = 10, 20, and 30 dB) and missing data points (20).

parameters: the infection factors $a_{\text{in}}$ and $a_{\text{af}}$, the bufirst rate of schizonts $S(t)$, and the standard deviation $\sigma$ of normalized life span $\widetilde{L}$. In the experiments of the previous section, the same parameters are used to generate synthetic data and to reconstruct expression patterns. However, in real microarray data, these parameters are unknown, and need to be estimated from experimental observations. The infection factors $a_{\text{in}}$, $a_{\text{af}}$, and burst rate $S(t)$ of schizonts can be accurately calculated from parasitemia and percent representation of iRBC. It is difficult to precisely estimate the standard deviation $\sigma$ of normalized life span $\widetilde{L}$. In this section, we investigate how results vary with regards to the choice of $\sigma$.

Synthetic microarray data is generated from know expression pattern with $\sigma$ = 0.1. The synthetic microarray data is contaminated by Gaussian white noise (SNR = 10 dB) and 20 missing data points. The expression pattern is reconstructed with various values of $\sigma$ ($\sigma$=0.05, 0.1, or 0.15). As shown in Figure 14, the reconstructed expression patterns only slightly change as the value of $\sigma$ varies. Consequently, the reconstruction is robust to the choice of $\sigma$.

## Results on real data

Expression patterns are reconstructed for 68 protein kinases collected from the microarray data of *P. falciparum* (HB3, 3D7 and Dd2) [4,13]. Reconstructed expression patterns are substituted into the equation (18) to estimate the likelihood of each protein kinase being associated with a specific IDC transition, and hence, contributing to effecting either the stage transition itself or a process(es) needed in the subsequent stage. Data for the three broad transitions analyzed, namely ring-to-trophozoite, trophozoite-to-schizont and schizont-to-ring, are summarized in Additional Files 1, 2 and 3, respectively.

A primary motivation for developing this computational framework is to prioritize gene candidates with potentially important stage-dependent functions for detailed downstream experimental analysis of gene function. In the ring-to-trophozoite analysis, several members of the largely unstudied FIKK protein kinases emerge with relatively high probabilities of mediating important biology during this developmental transition. Several of these protein kinases are targeted to the infected RBC cytosol/membrane [19]. Two [MAL7P1.144, PFL0040c] have been previously studied using gene knockout approaches [20]. While non-essential to blood stage parasite growth and survival, these proteins help mediate the increased rigidity of infected RBCs observed in trophozoite stage parasites. Presumably, this requires modulation of the RBC cytoskeleton through a combination of RBC cytoskeletal and/or exported parasite protein phosphorylation and increased interactions between these [20]. The analysis here suggests that the other highly ranked family members (see Additional File 1) could also be mediating important yet unknown biology at this ring-trophozoite transition.

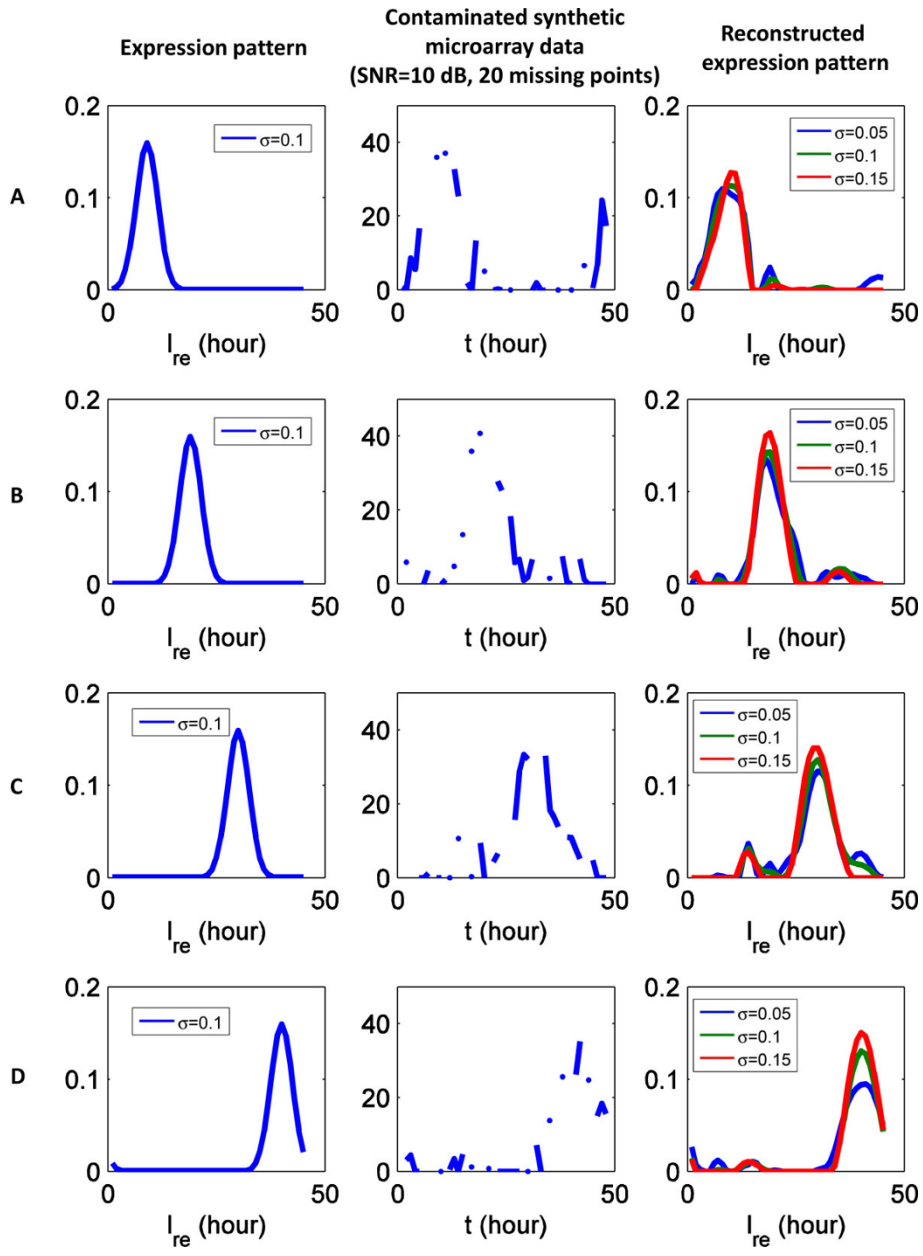


**Figure 14 Sensitivity to model parameter $\sigma$.** Synthetic microarray data is generated from know expression pattern with $\sigma$ = 0.1. The synthetic microarray data is contaminated by Gaussian white noise (SNR = 10 dB) and 20 missing data points. The reconstructed expression patterns only slightly change as the value of $\sigma$ varies ($\sigma$ =0.05, 0.1, or 0.15).

Interestingly, in the trophozoite-schizont and schizont-ring analyses, a number of protein kinases previously established to be essential and in some cases implicated in directly impacting the transition emerge with the highest probability rankings (Additional Files 2 and 3) [6-8,12]. In the schizont-ring analysis, for example, PFB0815 has previously been implicated in parasite motility/invasion/egress [6], and PF13_0211 in parasite egress from the RBC [7]. Furthermore, MAL13P1.278 is an essential Aurora kinase (Pfark1) that associates with spindle pole bodies during parasite schizogeny and is implicated in cell cycle regulation [8]. Overall, the top ranked genes in the trophozoite-schizont and schizont-ring analyses are highly enriched for protein kinases previously established to be essential to parasite survival. Therefore, it will be intriguing to experimentally examine the other highly ranked genes in both transition categories [PFI1415w, PF11_0079, PF10_0380, PF11_0510, PFI0095c, PFE0045c and PFC0945w] as these may also play critical roles in regulating parasite development

during these stages and could provide new opportunities for antimalarial drug development.

## Conclusions

This study proposes a new methodology to reconstruct intrinsic expression patterns from microarray data. We derive a linear system that relates the microarray data to the expression patterns. By solving the corresponding linear inverse problem, the expression patterns are reconstructed. The experiments conducted on synthetic data suggest that the proposed method can reliably reconstruct the expression pattern, even though both signal noise and missing data points contaminate the microarray data. By applying this method to *P. falciparum* microarray data, protein kinases are prioritized in terms of their likelihood of being involved in regulating some aspect(s) of the IDC. Indeed, the results of our analyses are supported by previously published experimental data confirming the involvement of several protein kinases in regulating parasite biology at or between developmental stage transitions in the IDC. Our results indicate that experimentally investigating the function of other putative protein kinases not previously studied but ranked highly in our analysis may provide new insights into *P. falciparum* biology.

## Additional material

**Additional file 1: Protein kinases are prioritized in terms of their likelihoods of being involved in the stage transition from ring to trophozoite**.

**Additional file 2: Protein kinases are prioritized in terms of their likelihoods of being involved in the stage transition from trophozoite to schizont**.

**Additional file 3: Protein kinases are prioritized in terms of their likelihoods of being involved in the stage transition from schizont to ring**.

**Acknowledgements**

This work was supported by the Singapore-MIT Alliance (SMA3) Graduate Fellowship granted to Z.W., and DOD ARO (Grant Number W911NF-09-1-0480).

This article has been published as part of *Proteome Science* Volume 10 Supplement 1, 2012: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2011: Proteome Science. The full contents of the supplement are available online at http://www.proteomesci.com/supplements/10/S1.

**Author details**

[1]Singapore-MIT Alliance for Research and Technology, Centre for Life Sciences, 28 Medical Drive, Singapore 117456. [2]School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. [3]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 56-341, Cambridge MA 02139, USA. [4]Department of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 6-237A, Cambridge MA 02139, USA.

**Authors' contributions**

Z.W. and J.D. prepared the first draft of the manuscript. J.N. proposed the initial idea, and helped with the data analysis and interpretation. Z.W., J.D., and J.C. conducted the theoretical analysis in this study. The numerical experiments were carried out by Z.W. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

Published: 21 June 2012

**References**

1. *World Malaria Report, 2011* Geneva, Switzerland: World Health Organization; 2011 [http://www.who.int/malaria/world_malaria_report_2011/en/].
2. Trampuz A, Jereb M, Muzlovic I, Prabhu RM: **Clinical review: Severe malaria.** *Crit Care* 2003, **7(4)**:315-323[http://dx.doi.org/10.1186/cc2183].
3. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De la Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301(5639)**:1503-1508[http://dx.doi.org/10.1126/science.1087025].
4. Bozdech Z, Llinás M, Lee B, Wong ED, Zhu J, DeRisi JL: **The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1**:e5+[http://dx.doi.org/10.1371/journal.pbio.0000005].
5. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR, Winzeler EA: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.** *Genome Research* 2004, **14(11)**:2308-2318[http://dx.doi.org/10.1101/gr.2523904].
6. Kato N, Sakata T, Breton G, Le Roch KG, Nagle A, Andersen C, Bursulaya B, Henson K, Johnson J, Kumar KA, Marr F, Mason D, McNamara C, Plouffe D, Ramachandran V, Spooner M, Tuntland T, Zhou Y, Peters EC, Chatterjee A, Schultz PG, Ward GE, Gray N, Harper J, Winzeler EA: **Gene expression signatures and small-molecule compounds link a protein kinase to Plasmodium falciparum motility.** *Nature Chemical Biology* 2008, **4(6)**:347-356[http://dx.doi.org/10.1038/nchembio.87].
7. Dvorin JD, Martyn DC, Patel SD, Grimley JS, Collins CR, Hopp CS, Bright AT, Westenberger S, Winzeler E, Blackman MJ, Baker DA, Wandless TJ, Duraisingh MT: **A Plant-Like Kinase in Plasmodium falciparum Regulates Parasite Egress from Erythrocytes.** *Science* 2010, **328(5980)**:910-912[http://dx.doi.org/10.1126/science.1188191].
8. Reininger L, Wilkes JM, Bourgade H, Miranda-Saavedra D, Doerig C: **An essential Aurora-related kinase transiently associates with spindle pole bodies during Plasmodium falciparum erythrocytic schizogony.** *Molecular microbiology* 2011, **79**:205-221[http://dx.doi.org/10.1111/j.1365-2958.2010.07442.x].
9. Ward P, Equinet L, Packer J, Doerig C: **Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote.** *BMC Genomics* 2004, **5**:79[http://dx.doi.org/10.1186/1471-2164-5-79].
10. Srinivasan N, Krupa A: **A genomic perspective of protein kinases in Plasmodium falciparum.** *Proteins* 2005, **58**:180-189[http://dx.doi.org/10.1002/prot.20278].
11. Doerig C, Billker O, Haystead T, Sharma P, Tobin AB, Waters NC: **Protein kinases of malaria parasites an update.** *Trends Parasitol* 2008, **24(12)**:570-577[http://dx.doi.org/10.1016/j.pt.2008.08.007].
12. Solyakov L, Halbert J, Alam MM, Semblat JP, Dorin-Semblat D, Reininger L, Bottrill AR, Mistry S, Abdi A, Fennell C, Holland Z, Demarta C, Bouza Y, Sicard A, Nivez MP, Eschenlauer S, Lama T, Thomas DC, Sharma P, Agarwal S, Kern S, Pradel G, Graciotti M, Tobin AB, Doerig C: **Global kinomic and phospho-proteomic analyses of the human malaria parasite Plasmodium falciparum.** *Nature Communications* 2011, **2**:565+[http://dx.doi.org/10.1038/ncomms1558].
13. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL: **Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains.** *Nucleic acids research* 2006, **34(4)**:1166-1173[http://dx.doi.org/10.1093/nar/gkj517].
14. Crewe W, Haddock DRW: *Parasites and human disease* Wiley; 1985 [http://www.worldcat.org/isbn/9780471010630].

15. Dean L: *Blood groups and red cell antigens* Bethesda, Md. : National Center for Biotechnology Information; 2005 [http://www.ncbi.nlm.nih.gov/books/NBK2261/].
16. Silbey RJ, Alberty RA, Bawendi MG: *Physical Chemistry* John Wiley & Sons; 2005 [http://www.mathematica-journal.com/issue/v9i3/newpublications/ISBN047121504X.html].
17. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(suppl 1)**: S96-S104[http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S96.abstract].
18. Liew AW, Law NF, Yan H: **Missing value imputation for gene expression data: computational techniques to recover missing data from available information.** *Briefings in bioinformatics* 2011, **12(5)**:498-513[http://dx.doi.org/10.1093/bib/bbq080].
19. Nunes MC, Goldring DP, Doerig C, Scherf A: **A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell.** *Molecular microbiology* 2007, **63(2)**:391-403[http://dx.doi.org/10.1111/j.1365-2958.2006.05521.x].
20. Nunes MC, Okada M, Scheidig-Benatar C, Cooke BM, Scherf A: **Plasmodium falciparum FIKK Kinase Members Target Distinct Components of the Erythrocyte Membrane.** *PLoS ONE* 2010, **5(7)**:e11747+[http://dx.doi.org/10.1371/journal.pone.0011747].

doi:10.1186/1477-5956-10-S1-S10
**Cite this article as:** Zhao *et al.*: Computational synchronization of microarray data with application to *Plasmodium falciparum*. *Proteome Science* 2012 **10**(Suppl 1):S10.