BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Modeling miRNA-mRNA interactions: fitting chemical kinetics equations to microarray data

Zijun Luo[1,2*], Robert Azencott[1] and Yi Zhao[2*]

Abstract

Background: The miRNAs are small non-coding RNAs of roughly 22 nucleotides in length, which can bind with and inhibit protein coding mRNAs through complementary base pairing. By degrading mRNAs and repressing proteins, miRNAs regulate the cell signaling and cell functions. This paper focuses on innovative mathematical techniques to model gene interactions by algorithmic analysis of microarray data. Our goal was to elucidate which mRNAs were actually degraded or had their translation inhibited by miRNAs belonging to a very large pool of potential miRNAs.

Results: We proposed two chemical kinetics equations (CKEs) to model the interactions between miRNAs, mRNAs and the associated proteins. In order to reduce computational cost, we used a non linear profile clustering method named minimal net clustering and efficiently condensed the large set of expression profiles observed in our microarray data sets. We determined unknown parameters of the CKE models by minimizing the discrepancy between model prediction and data, using our own fast non linear optimization algorithm. We then retained only the CKE models for which the optimized fit to microarray data is of high quality and validated multiple miRNA-mRNA pairs.

Conclusion: The implementation of CKE modeling and minimal net clustering reduces drastically the potential set of miRNA-mRNA pairs, with a high gain for further experimental validations. The minimal net clustering also provides good miRNA candidates that have similar regulatory roles.

Keywords: miRNA, Chemical kinetics modeling, Minimal net clustering

Background

Transcriptional and translational processes are fundamental cell mechanisms, involving three main molecular species: messenger RNA (mRNA) and their associated proteins, as well as microRNAs (miRNAs).

The miRNAs are small non-coding RNAs of roughly 22 nucleotides in length, which can bind with and inhibit protein coding mRNAs through complementary base pairing. A given miRNA can potentially bind and silence hundreds of mRNAs across a number of signaling pathways. These repressive miRNA-mRNA interactions occur in multiple cellular processes [1-3], and involve two distinct modalities: they may directly degrade their target mRNAs, or more often inhibit their translation [4-9].

The best characterized features determining the targets of a specific miRNA are the conserved Watson-Crick pairing to the 5' region (positions 2–7) of the miRNA, which are the so-called "seed pairing rules" [3,10-13]. Since seed pairing rules are neither sufficient nor necessary for miRNA-target functions [4,14], they have usually been combined with microarray or proteomic analysis to find potential miRNA-target pairs[15-17]. Classical microarray data analysis relies mostly on massive application of correlation analysis and linear statistical techniques to simultaneously acquired gene expression profiles. Combined with profile thresholding and heat map displays, these techniques provide commonly used clues for qualitative inference.

In [18], Principal Component Analysis and linear correlation had been linked with comparative sequence analysis to study microarray data recorded during mouse stem cells differentiation, and to broadly predict potential miRNA-mRNA interactions.

To go beyond the results of the linear microarray analysis applied on time-course microarray data in [18], we have formalized two basic architectures for repressive

*Correspondence: boluomiduo1@gmail.com; zhao.yi@hitsz.edu.cn
[1]Department of Mathematics, University of Houston, 4800 Calhoun, Houston, TX, USA
[2]School of Natural Sciences and Humanities, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, Guangdong, China

miRNA-mRNA interactions: "Transcription Degradation" (TD) and "Translation Inhibition" (TI).

Traditional chemical kinetics equations had been proposed to model the transcriptional and translational processes without involving the interaction of miRNAs [19,20].

We have derived Chemical Kinetics Equations (CKEs) to model the dynamics of TD and TI motifs, in the spirit of [20-30]. The equations are algebraically invariant under affine transformation and allow data condensation to reduce computational cost. We have implemented "minimal net" clustering method, which can control the maximum diameter of the clusters, to condense large data sets of gene expression.

Modeling by nonlinear CKEs involves complicated parameter estimation problems to fit the very large set of expression profiles recorded by microarrays. We have developed innovative fast algorithms dedicated to CKE parameter estimation, by optimization of the quality of fit between model and data.

We validate only the parameterized motifs having a high quality of fit to data. To reach robust conclusions we apply a "parameter parsimony" principle, favoring the models having the smallest number of parameters. And we have also evaluated the robustness of our parameter estimation algorithm by algorithm by direct testing on simulated data. These tests did validate our parameter estimation method is quite robust. This nonlinear approach goes further than well-established analysis based on correlation techniques combined with heat map displays.

Methods

Basic interaction architectures

To validate if an miRNA M does indeed repress a given gene G, we model the chemical interactions of M and G within a small network containing the pair (M, G). We now sketch two basic interaction architectures and their CKE models.

Transcription Degradation Motifs (TD-motifs)

We call "TD-motif" any interaction architecture, as sketched in Figure 1, involving a single miRNA-mRNA pair (M,G) where G is in Target(M), and M degrades the transcription of G. The TD motif includes also two sets of proteins rep(G) and act(G), namely the transcriptional "repressors" and "activators" of G, denoted by

$$rep(G) = \{R_1, \ldots R_k\} \quad \text{and} \quad act(G) = \{A_1, A_2, \ldots, A_q\}$$

Let $g(t), p(t), m(t), r_i(t), a_j(t)$, be the expression levels at time t for the chemical species G, P, M, R_i, A_j. We model the transcription process by a CKE similar to CKEs proposed in [20,22,25,29,30], but with a complementary term encoding the repressive impact of miRNA M on its target mRNA G:

$$\frac{dg(t)}{dt} = -\beta g(t) - \nu g(t) m(t) + \kappa REP(t)[1 - ACT(t)] \tag{1}$$

where $\beta > 0$ is the degradation rate of G, $\nu > 0$ is the reaction rate between G and M, $\kappa > 0$ is the product of the transcription rate by the concentration c of DNA templates, concentration which we assume to be some constant not depending on time (see [20]).

The percentage $0 \leq F(t) = REP(t)[1 - ACT(t)] \leq 1\%$ is the fraction of existing DNA templates which are committed at time t to transcription of the mRNA gene G. Here the percentages $REP(t)$ and $ACT(t)$ are modeled by the following products,

$$\begin{aligned} REP(t) &= REP_1(t) \times \ldots \times REP_q(t) \\ ACT(t) &= ACT_1(t) \times \ldots \times ACT_k(t) \end{aligned} \tag{2}$$

where

$$\begin{aligned} REP_i(t) &= \frac{1}{(1 + u_i r_i(t))^{SR_i}} \\ ACT_j(t) &= \frac{1}{(1 + w_j a_j(t))^{SA_j}} \end{aligned} \tag{3}$$

The parameters $SR_i > 0, u_i > 0$ and $SA_j > 0, w_j > 0$ are the number of binding sites and the affinity constant for the transcriptional factors R_i and A_j.

Note that the transcription repressors R_i combine multiplicatively their individual impacts REP_i in $REP(t)$, and that the REP_i are analogous to Hill function (see [19,25]);

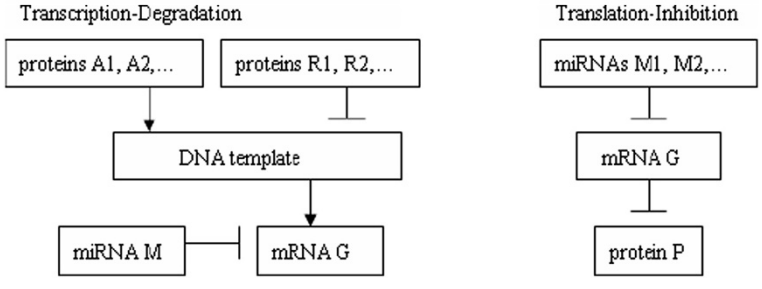

Figure 1 TD-motif and TI-motif.

similar remarks apply to the transcription activators. The multiplicative expressions of $REP(t)$ and $ACT(t)$ are typical of a so-called "cis-regulatory" function and have been derived by J. Goutsias [20].

The term $\kappa REP(t)[1-ACT(t)]\,dt$ is the concentration of new G molecules synthesized by transcription during the small time interval $[t, t+dt]$, while the repressive interactions of M and G eliminates $vg(t)m(t)dt$ molecules of G, and natural decay destructs $\beta g(t)dt$ molecules of G.

Translation-Inhibition Motifs (TI-motifs)

We call "TI-motif" any interaction architecture, as sketched in Figure 1, involving a set $M_1, \ldots, M_r$ of miRNAs inhibiting the translation of mRNA gene G, by repressing the expression of the protein P generated by G. Let $(p(t), m_i(t), g(t))$ be the concentrations at time t of protein P, miRNA M_i, and mRNA G. In the spirit of [20,21], we model the translation inhibition dynamics by the CKE

$$\frac{dp}{dt} = -\gamma p(t) + \lambda g(t) H(t) \tag{4}$$

where $\gamma > 0$ and $\lambda > 0$ are the degradation rate and the translation rate for protein P and where $H(t)$ is the percentage of G molecules committed at time t to the translation of G. Thus $H(t)$ encodes the inhibiting impact of the miRNAs $M_1, \ldots, M_k$ on the translation of G, and is modeled as a product of terms similar to $REP(t)$:

$$\begin{aligned} &H(t) = H_1(t) \times \ldots \times H_k(t) \quad \text{where} \\ &H_i(t) = \frac{1}{(1+u_i m_i(t))^{SM_i}} \end{aligned} \tag{5}$$

The parameters $SM_i > 0$ and $u_i > 0$ are resp. the number of binding sites and the affinity constant controlling the inhibiting impact of miRNA M_i on the translation of gene G. Note that $H(t)$ decreases when the miRNA concentrations $m_i(t)$ increase.

Our model for TI motif has been inspired by J. Goutsias [20], and we present below the hypotheses and arguments justifying the expression of $H(t)$. There is a key difference between the CKEs modeling TD and TI motifs. For TI motifs, the concentration of G-molecules committed to translation is $g(t)H(t)$, where $g(t)$ is the concentration of G-molecules and $H(t) < 100\%$.

For TD motifs, the concentration of G-molecules synthesized by transcription is $\kappa F(t) = \alpha c F(t)$ with $F(t) < 100\%$, where α and c are respectively the transcription rate and the concentration of DNA templates, assumed to be constant in time.

Derivation of chemical kinetics equations

Our derivation of the regulation equation for TI motif is quite similar to presentation given for TD motifs in [20], but has several changes in assumptions and formulations. To derive the CKEs (1) and (4), we propose a few hypotheses.

- Hyp. 1: (TD-motifs) The molecules of the miRNA repressor of gene G can strongly bind only at one unique specific site of G- molecules, and once a single such strong bind occurs, the corresponding G molecule degrades extremely fast.
- Hyp. 2: (TI-motifs) Each miRNA M_j in the set $rep(G)$ of translation inhibitors of G can weakly bind with G but only at specific binding sites constituting a set $BIND_j$ of size S_j. The sets $BIND_1, BIND_2, \ldots$ are pairwise disjoint. Once a G molecule thus binds with one inhibitor M_j, then this G molecule will fail to translate.
- Hyp. 3: For any given G-molecule, call X_j the random number of sites in $BIND_j$ which actually bind with one "M_j"-molecule. We assume that the random variables $X_1, X_2, \ldots.$ are independent.

Hyp.1 is based on the fact that only a small fraction of all messenger RNAs have more than a single miRNA binding site and miRNA bound with an mRNA gene G has a limited effect on the mRNA G, but affects more substantially the protein P generated by G ([4,31].

Consider first any TI motif involving an mRNA gene G generating the protein P , as well as a set $rep(G) = \{M_1, M_2, \ldots M_k\}$ of k translation inhibitors. We now derive the CKE (4) for this TI motif. Call $p(t), g(t), m_j(t)$ the concentrations of P, G, M_j and let $H(t)$ be at time t the percentage of existing G molecules committed to the translation of G into P. The basic CKE driving translation of G into P is then, as seen in [20],

$$\frac{dp}{dt} = -\gamma p(t) + \lambda g(t) H(t) \tag{6}$$

where γ, λ, are the degradation and translation rates of P.

The main point is to compute $H(t)$.

For $k = 0$, there are no translation inhibitors, and hence $H(t) = 100\%$. For $k = 1$, let $Q[S]$ be the set of G molecules with exactly S binding sites bound to M_1 molecules, where $0 \le S \le S_1$. Let $q(S,t)$ be the concentration of $Q[S]$ molecules. We have the forward and backward reactions

$$Q[S] + M_1 \rightharpoonup Q[S+1] \quad \text{and } Q[S+1] \rightharpoonup Q[S] + M_1 \tag{7}$$

By molecular collision theory [23], the concentration $\rho(t)$ of free M_1 molecules which bind at time t on $Q[S]$-molecules to produce $Q[S+1]$-molecules by forward reaction 7, is proportional to the product of $m_1(t)q(S,t)$ by the number $(S_1 - S)$ of vacant binding sites. Hence, for some constant c,

$$\rho(t) = c(S_1 - S) m_1(t) q(S,t) \tag{8}$$

Similarly the concentration $\tau(t)$ of M_1 molecules freed by backward reaction 7 is given by

$$\tau(t) = \tilde{c}(S+1)q(S+1,t) \tag{9}$$

for some constant $\tilde{c}$. At chemical equilibrium, we have $\tau(t) = \rho(t)$, and hence

$$q(S+1,t) = um_1(t)\frac{S_1 - S}{S+1}q(S,t)$$

where u is a new constant. By recurrence on S, this implies

$$q(S,t) = C_{S_1}^S [um_1(t)]^S q(0,t) \quad \text{where } C_{S_1}^S = \frac{(S_1)!}{S!\,(S_1 - S)!} \tag{10}$$

This entails

$$\begin{aligned} g(t) &= \sum_{S=0}^{S_1} q(S,t) = q(0,t)\sum_{S=0}^{S_1} C_{S_1}^S [um(t)]^S \\ &= [1 + um_1(t)]^{S_1} q(0,t) \end{aligned} \tag{11}$$

The set of G-molecules committed to translation into P at time t is identical to $q(0,t)$. Thus $H(t) = q(0,t)/g(t)$, and (11) yields $H(t) = 1/[1 + um_1(t)]^{S_1}$.

Using hypothesis Hyp 3, a recurrence on k extends the argument just given for $k = 1$, to prove that for $k \geq 1$, the percentage $H(t)$ of G molecules committed to translation into P is given by:

$$H(t) = \frac{1}{[1 + u_1 m_1(t)]^{S_1}} \times \ldots \times \frac{1}{[1 + u_k m_k(t)]^{S_k}} \tag{12}$$

which proves CKE (4) for TI motifs.

For TD motifs, the cis-regulation function $F(t)$ in CKE (1) has been derived in [20]. By using hypothesis 1 and molecular collision theory, we directly deduce that the concentration the concentration of mRNA G degraded by miRNA M is proportional to the product of both concentrations of G and M. Thus we justify the miRNA degradation term $-vg(t)m(t)$.

Invariance by affine profile transformations

Affine profile transformations

Call $\Gamma \subset R^q$ the set of all possible expression level profiles $r(t)$ indexed by time $t_1, \ldots, t_q$. To any pair $T = (a,b)$ of real numbers, we associate an *affine profile transformation*

$$r \to Tr \quad \text{defined by } r(t) \to (ar(t) + b)$$

for all time dates t. Let $\mathcal{T}$ be the set of all such affine profile transformations.

In microarray data sets, expression levels of genes are recorded via optical analysis of fluorescence intensities, and hence depend strongly on experimental acquisition modalities. So relative expression levels between pairs of recorded chemical species are more meaningful quantities, and graphic displays of microarray data by "heat maps" often involve logarithms of raw data.

Our microarray data record expression levels which as a first approximation can be viewed as unknown affine transformations of concentrations. Since our CKEs (1), (4) were derived for concentrations, we need to check how these CKEs and their parameters change under generic affine profile transformations.

Affine invariance of CKE models

Consider first a TD-motif involving an mRNA gene G, a protein P, an miRNA M, transcription repressive proteins $R_1, R_2, \ldots$, and transcription activating proteins $A_1, A_2, \ldots$. The CKE model (1) links the concentrations g, p, m, r_i, a_j of G, M, P, R_i, A_j. Let $\hat{g}, \hat{p}, \hat{m}, \hat{r}_i, \hat{a}_j$ be the corresponding recorded expression profiles. Assume that each such recorded profiles $\hat{r}$ is linked to the concentration profile 4 by some *unknown* affine profile transformation T_r.

From CKE (1), one directly deduces that $\hat{g}, \hat{p}, \hat{m}, \hat{r}_i, \hat{a}_j$ verify a new CKE having an algebraic form completely similar to CKE (1), but where the original parameters $\beta, \nu, \kappa, u_i, w_j$ are replaced by new parameters $\hat{\beta}, \hat{\nu}, \hat{\kappa}, \hat{u}_i, \hat{w}_j$, and where the integers $SR_1, SR_2, \ldots$, and $SA_1, SA_2, \ldots$, remain unchanged.

The new parameters are easily expressed in terms of the original ones and of the coefficients of the affine transformations, but this is irrelevant practically since we will use microarray data to directly compute the new parameters for a CKE of type (1) linking recorded expression levels.

Similar computations for TI-motifs show that this affine invariance property also holds for CKE (4). Hence to model a TD or a TI motif, we can fit a CKE model of type (1) or (4) to recorded expression profiles, even though the theoretical model justification involved true concentrations, which are not directly measured by microarrays.

The key assumption is that, for each chemical species C, the expression levels $\hat{c}(t)$ of C recorded by microarray are approximately linked to the concentration $c(t)$ of C by some affine relation $\hat{c}(t) = a(C)c(t) + b(C)$, where the unknown coefficients $a(C)$ and $b(C)$ may depend on the species C.

The preceding algebraic model invariance under multiple affine profile transformations strongly suggests that *adequate distances between dynamic profiles of recorded expression levels should be invariant under affine profiles transformations*, as developed in the next section.

Condensation of expression levels profiles

Microarray data typically record several tens of thousands gene expression profiles. So computational costs to fit microarray data to all potential TD-motifs and TI-motifs would of course be prohibitive. A natural option to reduce combinatorial explosion is to cluster the observed profiles.

Since our goal is to model expression profiles by ODEs of type (1) and (4), we need to control the diameters of all

clusters of expression profiles. This led us to reject hierarchical clustering as well as K-means clustering, and to implement in the space of expression profiles a "minimal-net" clustering technique, inspired by an innovative technique for automatic generation of prototypes in shape spaces (see [32]). In view of the preceding section, the diameters of clusters in the space of profiles should be measured by a distance invariant under affine profiles transformations.

Affine invariant distance between profiles

Recall that $\Gamma \subset R^q$ is the set of all profiles $r(t)$ indexed by time dates $t_1, \ldots, t_q$. The mean and variance of a profile r are denoted by

$$\overline{r} = \frac{1}{q}[r(t_1) + \ldots + r(t_q)] \; ;$$

$$var(r) = \frac{1}{q}[(r(t_1) - \overline{r})^2 + \ldots + (r(t_q) - \overline{r})^2]$$

Define normalized profiles $nor(r)$ and profiles correlations $corr(r_1, r_2)$ by

$$nor(r)(t) = \frac{r(t) - \overline{r}}{\sqrt{var(r)}} \; ;$$

$$corr(r_1, r_2) = < nor(r_1), nor(r_2) >$$

where $< \cdot, \cdot >$ is the usual scalar product in R^p.

We then define a *distance* $D(r_1, r_2)$ between profiles r_1 and r_2 by

$$D(r_1, r_2) = \sqrt{2 - 2\, corr(r_1, r_2))}$$

For any affine profile transformation T in $\mathcal{T}$, one has $nor(r) = nor(Tr)$, and hence $D(r_1, r_2) = D(nor(r_1), nor(r_2)$. Thus for any affine profile transformations T_1, T_2, and any profiles r_1, r_2 one has

$$D(T_1 r_1, T_2 r_2) = D(r_1, r_2)$$

So the profile distance D *is invariant by affine profile transformations*

Minimal net clustering

The set Γ of all profiles is now endowed with a distance D invariant by affine profile transformations. Call *mPR* the large set of miRNA profiles recorded at time $t_1, \ldots, t_q$ by our microarrays. We fix a maximum radius ε for profiles clusters. We seek to partition *mPR* into disjoint clusters $CL_1, \ldots, CL_r$ such that each CL_j has diameter inferior to 2ε, and we also want the number NN of clusters to be as small as possible. We have implemented an iterative algorithm to generate this type of minimal net clustering, in the spirit of [32].

Define a distance function $D(x, y)$, where x, y represent the observations of the data. Denote by $D(x, Y)$ the distance between an observation x and a set Y, where $D(x, Y) = \min_{y \in Y} D(x, y)$. Let Let $[x]$ denote the cluster containing single point x, and Γ be the set of all observations, the minimal net algorithm is as follows:

1. step 1, let $(x_1, y_1) = argmax_{x,y} D(x, y)$. Then let x_1 and y_1 become the representatives of two initial clusters. Let $CL_1 = [x_1]$, $CL_2 = [x_2]$, $C_1 = \{CL_1, CL_2\}$ and $R_1 = \Gamma \setminus C_1$ representing the remaining points in Γ excluding the 2 clusters.
2. After step $n - 1$, we obtain $C_{n-1} = \{CL_1, CL_2, \ldots CL_n\}$ and $R_{n-1} = \Gamma \setminus C_{n-1}$, where CL_j is a cluster that has single point, $j = 1, \ldots, n + 1$. In step n, let $HD = \max_{x \in R_{n-1}} D(x, C_{n-1})$, representing the maximum distance between observation x in R_{n-1} and set C_{n-1}. If $HD > \varepsilon$, find $x_{HD} = argmax_{x \in R_{n-1}} D(x, C_{n-1})$. Let $C_n = \{C_{n-1}, x_{HD}\}$, $R_n = \Gamma \setminus C_n$. Repeat until $HD \leq \varepsilon$.
3. Assume the loop stop at step NN, we have $C_{NN} = \{CL_1, CL_2, \ldots CL_{NN+1}\}$. For an observation x, find the point y_{CL_k}, belonging to cluster CL_k in C_{NN}, that is closest to x, i.e. $y_{CL_k} = argmin_{y \in C_{NN}}(D(x, C_{NN}))$, then assign x to the cluster CL_k.

We apply this minimal net clustering algorithm to the set of all miRNA profiles recorded by our microarrays. We define the cluster diameter $diam(CL)$ of cluster CL as the maximum distance of any two observations in the cluster, i.e. $diam(CL) = \max_{x,y \in CL} D(x, y)$, where. Compared with the commonly used clustering method such as K-means and Hierarchical clustering, the minimal net algorithm allows us to control the diameter of all clusters by setting the threshold ε representing the supremum radius of the cluster, while the K-means and hierarchical clustering is often used to determine the number of of clusters but not their diameters. If we select a very small threshold ε, the expression levels of genes in the same cluster can be considered almost identical.

Parameter estimation for the CKE models

Parameters estimation strategy

A generic strategy for parameter estimation in CKEs systems is to minimize a cost function evaluating the discrepancy between model predictions and experimental data. Each one of our CKE models has a single output variable, namely the expression level $g(t)$ of mRNA gene G for a TD motif, and the expression level $p(t)$ of protein P for a TI motif. The output variable $g(t)$ or $p(t)$ of a CKE model can be estimated by a function $\hat{g}(t)$ or $\hat{p}(t)$ once one knows the profiles of all other molecular species involved in the model. The estimators $\hat{g}(t)$ or $\hat{p}(t)$ are respectively determined by CKE (1) or CKE (4).

Each CKE models is parameterized by a parameter vector $\mathbf{w}$ of dimension n_{par}. The quality of fit of this model with recorded profiles data is quantified by *the size* $ERR(\mathbf{w})$ *of the estimation error* defined as follows

$$ERR(\mathbf{w}) = \max_t |f(t) - \hat{f}(t)|$$

where the output variable $f(t)$ is equal to $g(t)$ for TD motifs and to $p(t)$ for TI motifs. The concrete goal is to find the best parameter vector $\mathbf{w}$ by minimization of the lack of fit $ERR(\mathbf{w})$ over all possible values of w.

There are no magic solutions for such non linear minimization problems. Moreover fast computing was essential here, since we usually have to solve a very large number of similar "quality of fit maximization" problems when dealing with large microarray datasets.

We have tested several generic cost minimization approaches (see [33-35]) such as "genetic algorithms", as well as "gradient descent" to minimize a sum of squared modeling residuals. These two techniques turned out to require far too much computing time and were often unreliable due to their high dependence on initialization values.

We hence developed our own fast CKE parametrization algorithms to optimize the quality of fit between CKE models and microarray profiles data. This optimized quality of fit, adequately balanced by a systematic emphasis on parsimoniously parameterized models, becomes an essential clue to decide which potential interactions one should validate between miRNAs, mRNAs, and associated proteins.

Parameters parsimony requirement

Robustness of CKE parametrization is the main motivation for our parameter parsimony requirements. Consider CKE (1) modeling a TD-motif *TDM* involving n_{act} activators and n_{rep} repressors for the transcription of mRNA gene G, and one miRNA M degrading the transcription of G. Then the number n_{par} of unknown parameters is $n_{par} = 3 + 2(n_{act} + n_{rep})$. Each profile $g(t), m(t), a_i(t), r_j(t)$ is recorded at $t = t_1, \ldots, t_p$. The CKE outputs for *TDM* are $g(t_1), \ldots, g(t_p)$. Hence each microarray data set provides $(p-1)$ equations linking the n_{par} unknown parameters, since the recorded $g(t)$ should be very close to the predicted values $\hat{g}(t)$ obtained by solving the ODE (1) with initial value $g(t_1)$.

Vapnik's results on model fitting (see [36]) show that robust accuracy of parameter estimates requires fairly high values of the ratio of $(p-1)/n_{par}$. So $n_{par} << (p-1)$ is a necessary constraint, and we will impose the *parameter parsimony requirement* $n_{par} \leq (p-1)/4$. Indeed when $n_{par} > (t_p - 1)$, CKE models are overfitted and parameters are poorly estimated.

CKE Parameter estimation

TI-motifs: plausible ranges for parameters

Consider a generic TI motif *TIM* involving an mRNA gene G, its associated protein P, and k translation inhibiting miRNAs $[M_1, \ldots, M_k]$. Let $p(t), g(t), m_i(t)$ be the expression levels of P, G, M_i. We want to model *TIM* by CKE (4).

Parametrized by the vector

$$\mathbf{w} = [\gamma, \lambda, u_1, S_1, \ldots, u_k, S_k]$$

which involves $(2k+2) \leq 8$ parameters.

The degradation and translation rates γ and λ of protein P were unknown for most proteins.

According to results from [31], only a small percentage 2% of our 30,000 mRNAs have more than 2 potential binding sites for miRNAs, and only 0.02% mRNAs have as many as 7 such binding sites. So in our parameter estimations it is reasonable to restrict the number of binding sites S_j for miRNA M_j to be at most 5.

TI-motifs: optimizing the quality of fit

The inputs of CKE (4) are the initial value $p(t_1)$ and the expression levels $g(t), m_1(t), \ldots, m_k(t)$ recorded at time dates $t_1, \ldots, t_q$.

Discretizing the ODE (4) at time $t_1, \ldots, t_q$, we get

$$p(t_{j+1}) - p(t_j) = -\gamma p(t_j) + \lambda g(t_j) H(t_j) \tag{13}$$

where the percentage H(t) is as recalled above . By summation this implies, for $j = 2, 3, \ldots, q-1$, the relation

$$p(t_{j+1}) - p(t_1) = -\gamma Q(t_j) + \lambda L(t_j) \tag{14}$$

where

$$Q(t_j) = \sum_{n=1}^{j} g(t_n); \quad L(t_j) = sum_{n=1}^{j} g(t_n) H(t_n)$$

In view of equation (14), when all expression profiles involved have been recorded until time t_j, one can predict the still unknown value of $p(t_{j+1})$ by the following natural estimator $\hat{p}(t_{j+1})$,

$$\hat{p}(t_{j+1}) = p(t_1) - \gamma Q(t_j) + \lambda L(t_j) \tag{15}$$

The quality of fit of this CKE model with recorded profiles data will be quantified by *the size* $ERR(\mathbf{w})$ *of the estimation error* numerically computed as follows

$$ERR(\mathbf{w}) = \max_{j=1,\ldots,q} | p(t_j) - \hat{p}(t_j) | \tag{16}$$

which in view of (15) can be reformulated as

$$ERR(\mathbf{w}) = \max_{j=1,\ldots,T} | \pi_j + \mu_j \gamma - \nu_j \lambda | \tag{17}$$

where for $j = 1, \ldots, q$ we have set

$$\pi_j = p(t_j) - p(t_1)\mu_j = Q(t_j)\nu_j = L(t_j) \quad (18)$$

We seek a parameter vector $\mathbf{w}$ minimizing the cost function $ERR(\mathbf{w})$ in the parametric domain defined by

$$\gamma > 0\,;\ \lambda > 0\,;\ u_i > 0\,;\ 1 \le S_i \le 5$$

TI-motif: parameter estimation algorithm

For each $i = 1 \ldots k$, fix an arbitrary integer $1 \le S_i \le 5$. Call $\bar{m}_i$ and $\bar{H}_i$ the respective medians over time t of the functions $m_i(t)$ and $H_i(t)$. Since the function $H_i(t) = 1/[1 + u_i m_i(t)]^{S_i}$ is monotonous in $m_i(t)$, we have

$$\bar{H}_i = 1/[1 + u_i \bar{m}_i]^{S_i} \quad (19)$$

The assumption $0 < H_i(t) < 1$ yields $0 < \bar{H}_i(t) < 1$. We will discretize the possible values of $\bar{H}_i(t)$ by constraining them to belong to a grid $h_1, \ldots, h_s$ of $1 \le s \le 99$ percentage values equally spaced in $[0, 1]$. Since $\bar{m}_i$ is known, we invert equation (19) to compute a corresponding grid GRD_i of s potential values for $u_i = \frac{1}{\bar{m}_i}((\frac{1}{\bar{H}_i})^{1/S_i} - 1)$.

Now for $1 \le i \le k$, select and fix arbitrary values u_i in the grid GRD_i. After selecting as restrictive above the set of parameter vector $U = [u_1, S_1, \ldots u_k, S_k]$, the function $H(t)$ is then completely determined for all t.

Note that the set $\mathcal{U}$ of all possible such choices for U is of cardinal inferior to $N = (5s)^k$. Since we do explore each one of these possibilities separately, we need to keep the number N at a reasonable level, so we selected $s = 99$ for $k = 1$, $s = 60$ for $k = 2$, $s = 20$ for $k = 3$ etc.

Fix any U in $\mathcal{U}$. Since all recorded profiles involved in the TD motif are available at time $t_1, \ldots, t_q$, we can use the U value to directly compute all the values $H(t_j)$, and then all the numbers π_j, μ_j, ν_j defined by (18). We want to find values of the two last parameters $\gamma > 0$ and $\lambda > 0$ which will minimize

$$ERR(\gamma, \lambda) = \max_{j=1,\ldots,q} |\, \pi_j + \mu_j\gamma - \nu_j\lambda \,|$$

This problem is equivalent to minimizing the linear objective function:

$$\Psi(\gamma, \lambda, z) = z$$

under the $(2q + 3)$ linear inequality constraints

$$\gamma > 0\,; \quad \lambda > 0\,; \quad z > 0$$
$$z - (\pi_j + \mu_j\gamma - \nu_j\lambda) \ge 0 \quad \text{for } j = 1, \ldots, q$$
$$z + (\pi_j + \mu_j\gamma - \nu_j\lambda) \ge 0 \quad \text{for } j = 1, \ldots, q$$

This is a classical *constrained linear programming* problem, which can be solved by well known fast linear programming algorithms [37], to provide the optimal values γ^*, λ^*, z^*. Then $z^* = ERR(\gamma^*, \lambda^*)$ is the minimal value of $ERR(\gamma, \lambda)$.

These optimal values are functions $\gamma^*(U), \lambda^*(U), z^*(U)$ of the partial vector of parameters U in $\mathcal{U}$. We then select the optimal $\mathbf{U}^*$ in $\mathcal{U}$ as the value of U which minimizes $z^*(U)$ over $\mathcal{U}$. The optimal parametrization $\mathbf{w}^*$ of our TD motif is then given by $\mathbf{w}^* = [\gamma^*(U^*), \lambda^*(U^*), U^*]$. This new parametrization algorithm is fairly fast and has good accuracy. On a current laptop PC, our non-optimized MATLAB code implementation required less than 5 minutes of CPU time for the parametrization of a typical TI-model with 38 time points and 9 parameters [38]. After code optimization and a re-implementation in C, we expect this CPU time to be reduced to 2 minutes. The algorithm does not require any knowledge of the parameters ranges except for the number of binding sites S, which is an advantage for the range of reaction rates of of many molecules of interest are usually unknown.

TD-motifs: parameter estimation algorithm

The parametrization algorithms just presented also apply to TD models, as we now sketch. The output variable is now the expression level $g(t)$ of mRNA gene G.

Discretize the ODE (1) at time dates $t_1, \ldots, t_q$, to get

$$g(t_{j+1}) - g(t_j) = -\beta g(t_j) - \nu g(t_j) m(t_j) + \kappa F(t_j) \quad (20)$$

where F(t) is defined in (1). By summation this implies the relations

$$g(t_{j+1}) - g(t_1) = -\beta B(t_j) - \nu V(t_j) + \kappa K(t_j) \quad (21)$$

where

$$B(t_j) = \sum_{n=1}^{j} g(t_n)\,; \quad V(t_j) = \sum_{n=1}^{j} g(t_n) m(t_n)\,;$$
$$K(t_j) = sum_{n=1}^{j}\, F(t_n)$$

When all expression profiles involved have been recorded until time t_j, one can predict the unknown value of $g(t_{j+1})$ by the estimator $\hat{g}(t_{j+1})$,

$$\hat{g}(t_{j+1}) = g(t_1) - \beta B(t_j) - \nu V(t_j) + \kappa K(t_j) \quad (22)$$

The quality of fit of this CKE model with the recorded profiles data is quantified by *the size of the prediction error* which is a function $ERR(\mathbf{w})$ of the parameter vector $\mathbf{w}$

$$ERR(\mathbf{w}) = \max_{\mathbf{j}=1,\ldots,\mathbf{q}} |\, \mathrm{g}(\mathrm{t}_\mathbf{j}) - \hat{\mathrm{g}}(\mathrm{t}_\mathbf{j}) \,|$$

so that

$$ERR(\mathbf{w}) = \max_{\mathbf{j}=1,\ldots,\mathbf{q}} |\, \tau_\mathbf{j} + \rho_\mathbf{j}\beta + \eta_\mathbf{j}\nu - \theta_\mathbf{j}\kappa \,| \quad (23)$$

where we have set

$$\tau_j = g(t_j) - g(t_1) \quad \rho_j = B(t_j) \quad \eta_j = V(t_j) \quad \theta_j = K(t_j)$$

The TD model parameters $SR_i > 0, u_i > 0$ and $SA_j > 0, w_j > 0$ are the number of binding sites and the affinity constants for the transcriptional factors R_i and A_j. They

constitute a partial parameter vector U which we restrict as above by first imposing a moderate upper bound S_{max} on all the integers SR_i, SA_j. and by selecting, also as done above, adequate finite grids for the values of the u_i and the w_j. This constrains U to belong to a finite set $\mathcal{U}$. The cardinal N of $\mathcal{U}$ is forced to remain at most of order 10^5, by adequate constraints on S_{max} and on the coarseness of the u_i grids and the v_j grids.

To minimize $ERR(\mathbf{w})$, we fix as above an arbitrary U in $\mathcal{U}$. We can then compute all the numbers $\tau_j, \rho_j, \eta_j, \theta_j$. Since U is fixed, the error size $ERR(\mathbf{w})$ becomes a function $E(\beta, \nu, \kappa)$ of the last 3 positive parameters (β, ν, κ), still given by equation (23). As above the minimization of $E(\beta, \nu, \kappa)$ is equivalent to a constraint linear programming problem where we want to minimize the linear function $\Phi(\beta, \nu, \kappa, z) = z$ over the following set of $(2q + 4)$ linear inequalities

$$\beta > 0\,;\, \nu > 0\,;\, \kappa > 0\,;\, z > 0$$
$$z - (\tau_j + \rho_j\beta + \eta_j\nu - \theta_j\kappa) \geq 0 \quad \text{for } j = 1, \ldots, q$$
$$z + (\tau_j + \rho_j\beta + \eta_j\nu - \theta_j\kappa) \geq 0 \quad \text{for } j = 1, \ldots, q$$

Solving this constraint linear programming problem generates optimal parameters $(\beta^*, \nu^*, \kappa^*)$ and a minimal error z^*, which are all functions of U. One concludes as above by selecting an optimal U^* minimizing $z^*(U)$ over all U in $\mathcal{U}$.

Quality of fit for CKE models

Consider any TD motif or TI motif $\mathcal{A}$. We have seen how to compute a parameter vector $\mathbf{w}^*$ optimizing the quality of fit between microarray data and our CKE model for $\mathcal{A}$. This was done by minimizing $ERR(\mathbf{w}) = \max_t |f(t) - \hat{f}(t)|$, where $f(t)$ is the main output variable of $\mathcal{A}$, and $\hat{f}(t)$ is the estimation of $f(t)$ based on the CKE parametrized by $\boldsymbol{w}$.

For the optimal CKE parametrization $\mathbf{w}^*$, the lack of fit to data can then be evaluated by $ERR(\mathbf{w}^*)$. However we have seen in section 'Invariance by affine profile transformations' that when comparing two expression profiles recorded by microarray, natural distances between profiles should be roughly invariant by changes of scale for these profiles. It would then be tempting to replace the absolute error of estimation $|f(t) - \hat{f}(t)|$ at time t by the relative error of estimation $\frac{|f(t)-\hat{f}(t)|}{f(t)}$. But relative errors become quite large whenever the output profile $f((t)$ is close to zero. To avoid such spuriously large error values, while still preserving scale invariance whenever $f(t)$ is not close to zero, we define the Smoothed Relative Error of estimation $SRE(t)$ at time t by the following formula, where $\bar{f}$ denotes the mean value of the profile f,

$$SRE(t) = \frac{|f(t) - \hat{f}(t)|}{f(t)} \quad \text{when } f(t) > 0.15\bar{f} \tag{24}$$

$$SRE(t) = \frac{|f(t) - \hat{f}(t)|}{\bar{f}} \quad \text{when } f(t) \leq 0.15\bar{f} \tag{25}$$

We finally quantify the *Modeling Error MODER* for the optimally parametrized CKE model of motif $\mathcal{A}$ by

$$MODER = \max_t SRE(t)$$

Results and discussion

Examples of application

We implemented our model of TI on microarray data of mouse stem cells undergoing RA-induced differentiation, as provided by LC Science Inc, and previously analyzed by classical techniques in [18]. We took the recorded expression profiles for proteins/mRNAs GCNF, Oct4, Nanog and Sox2 at time points (0, 1.5, 3, 6)/(0, 3, 6) and expression levels for 266 miRNAs on days 0, 1, 3, 6 from [38] during ES cell differentiation. These profile data were interpolated at 19 intermediary time points, by Piecewise Cubic Hermite Interpolation (PCHIP) and the number of parameters were limited to be 4, i.e. only 1 upstream miRNA was selected for the model, to satisfy the parameter parsimony requirement.

For miRNA M_i, the following linear transformation, which could be viewed as normalization, was done:

$$\hat{m}_i(t) = \frac{m_i(t) - \bar{m}_i(t)}{\sigma(m_i)} + 1$$

where $\sigma(m_i) = \sqrt{\sum_t (m_i(t) - \bar{m}_i(t))^2}$, $t = 0, 1/3, \ldots, 6$. Since $\| \frac{m_i(t)-\bar{m}_i(t)}{\sigma(m_i)} \| \leq 1$, $\hat{m}_i(t)$ is positive for $t = 0, 1/3, \ldots, 6$. Taking $\varepsilon = 0.15$, we applied Minimal Net clustering (the MATLAB code can be downloaded through Additional file 1) to the transformed data of miRNAs, $\hat{m}_i(t), i = 1, 2, \ldots, 266$, and obtained 107 clusters, in which the maximum cluster contains 14 miRNAs. We consider that the miRNAs belonging to the same cluster share the same normalized expression level within a negligibly small error. For each cluster $CL_j, j = 1, \ldots, 107$, we determine the miRNA MC_j which is the representative of the cluster CL_j for the distance D , and we let $mc_j(t)$ be the expression level of MC_j at time t. We call $mc_j(t)$ is the expression level of cluster CL_j.

We successively implemented the TI model with only one repressor miRNA (the MATLAB code can be downloaded through Additional file 2). After parametric modeling of our pre-selected 107 TI motifs, and evaluation of their quality of fit, we have validated only 3 clusters of miRNAs as translational inhibitors repressing protein Oct4. If an miRNA in CL_j is validated by modeling as a potential repressor, all other miRNAs belonging to CL_j are also potential repressors and can be validated numerically as well by the form invariability of the model under

affine transformation. Here we present one of these three validated miRNAs of Oct4 (see Figure 2), where the centroid of the cluster is miRNA mmu-miR-10a, while the other 6 miRNAs in the cluster are mmu-miR-203, 330, 342, 470 and 99b. We used TarBase [39] to search all the experimentally validated miRNAs targeting Oct4 (Pou5f1) in Mus musculus. By TarBase, only mmu-miR-470 has been experimentally validated and it is also numerically validated by our TI model.

For protein GCNF, only two miRNAs, mmu-let-7b and mmu-let-181a, have been experimentally validated by TarBase and both of them belong to the list of the 20 miRNAs numerically validated by our TI modeling (we presented the validated cluster containing mmu-let-7b in Figure 3). Our modeling approach did not validate any miRNA repressing both proteins Nanog and Sox2, while there are 3 miRNAs and 1 miRNA experimentally validated as separate repressors of Nanog and Sox2 respectively.

We here also present one example of TD motif for downstream factor mRNA Sox2 (Figure 4). With the assumption that the transcription factors are proteins Oct4 and Nanog [38], we validated cluster mmu-miR-134, 30a-3p, 30b, 335, 431, 433-3p, 434-3p, and 487b as degraders. Although (mmu-miR-134, Sox2) is also an experimentally tested pair, we will not discuss deep in detail the validation results of the TD motifs in this paper. The main reason is that the validation results of TD motifs depends much on our knowledge of the transcription factors. More discussion is in the subsection below.

Discussion

In [38], we pre-selected the potential miRNAs for each gene/protein by TargetScan 5.0 and miRanda before applying the two models. The pre-selection of miRNA candidates were not necessary though it greatly reduced the computational cost. However, the application of the CKE modeling was dependant on the target prediction algorithm, such as TargetScan or miRanda. Therefore, we introduced Minimal Net Clustering in this paper so that the data was condensed and the computational cost could be reduced by a purely numerical method without biological bias.

Since the results of TD model depend on information of transcription factors, the modeling validates not only miRNAs acting as mRNA degraders but also upstream transcription factors simultaneously. In [38] we numerically validated proteins GCNF, Oct4 and Nanog as transcription factors for mRNAs Oct4 and Nanog, while Marson et al. and Boyer et al. [40,41] claimed that Oct4 and Sox2 are bounded and act together with Nanog as transcription factors. Considering the impact of the transcription factors, the TD modeling may be less convincing unless the transcription factors are fixed as experimentally validated. In this paper we presented more details on the implementation and validation results of the TI model in

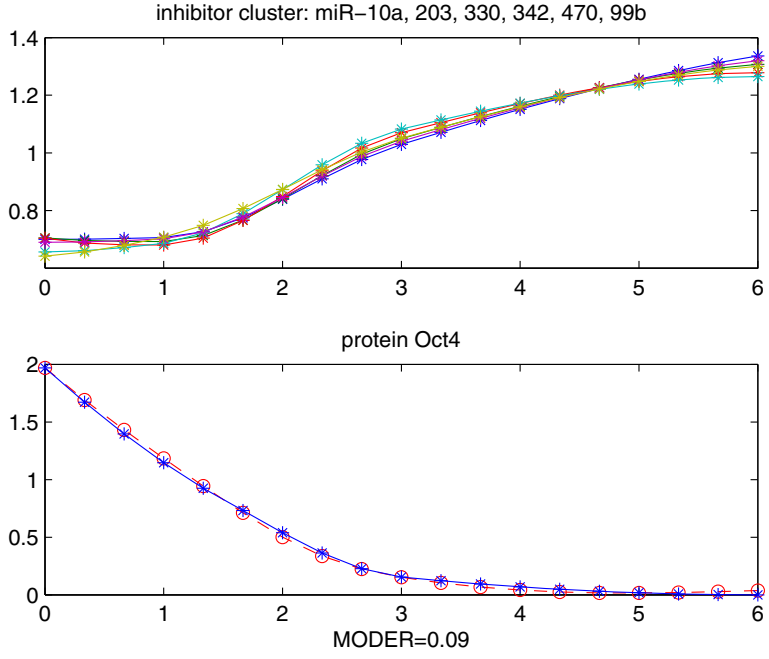

Figure 2 Example of validated TI motif. Example of TI motif repressing Oct4. All expression profiles are over days 0–6. Upper 6 profiles: normalized expression levels of numerically validated miRNAs in the same cluster: mmu-miR-10a, 203, 330, 342, 470 and 99b. Bottom 2 profiles: Blue line = recorded levels. Red dash line = predicted levels. "MODER" is the model global relative error of prediction.

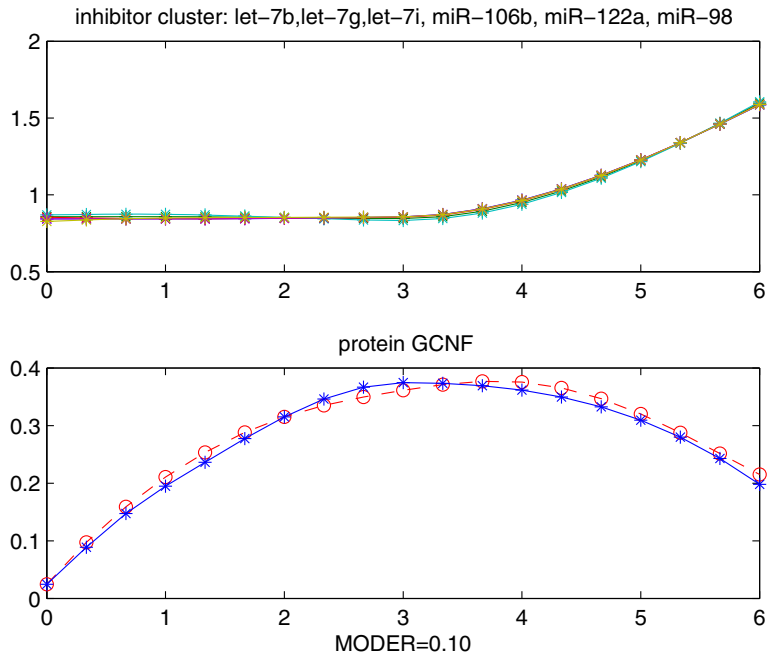

Figure 3 Example of validated TI motif. Example of TI motif repressing GCNF. All expression profiles are over days 0–6. Upper 6 profiles: normalized expression levels of numerically validated miRNAs in the same cluster: mmu-let-7b, let-7g, let-7i, miR-106b, miR-122a, miR-98. Bottom 2 profiles: Blue line = recorded levels. Red dash line = predicted levels. "MODER" is the model global relative error of prediction.

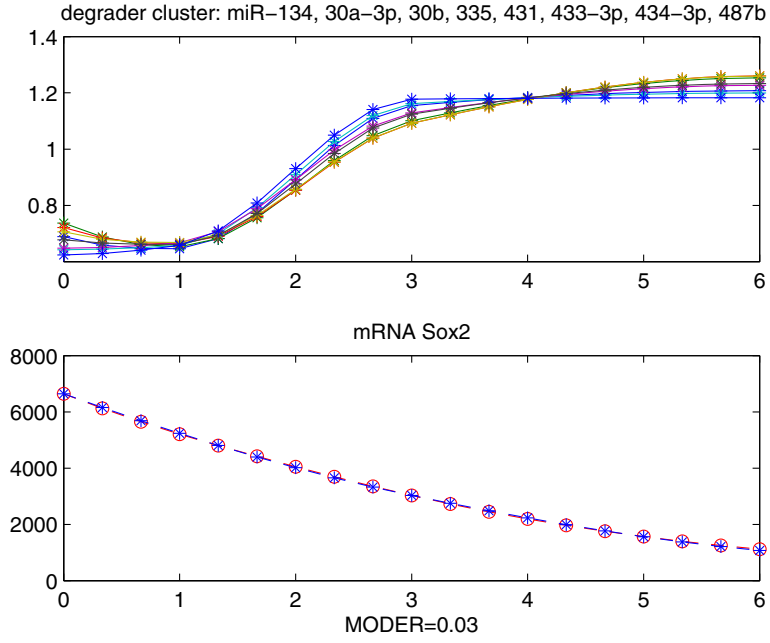

Figure 4 Example of validated TD motif. Example of TD motif repressing Sox2. All expression profiles are over days 0–6. Upper 8 profiles: normalized expression levels of numerically validated miRNAs in the same cluster: mmu-miR-134, 30a-3p, 30b, 335, 431, 433-3p, 434-3p, 487b. Bottom 2 profiles: Blue line = recorded levels. Red dash line = predicted levels. "MODER" is the model global relative error of prediction.

order to focus on the validation of miRNAs and avoid the influence of assumptions of transcription factors.

TarBase shows that, for protein GCNF the experimentally validated miRNAs are mmu-let-7b and mmu-miR-181a; for protein Oct4 the experimentally validated miRNA is mmu-miR-470; for protein Nanog the experimentally validated miRNAs are mmu-miR-134, mmu-miR-470, and mmu-miR-296; for protein Sox2 the experimentally validated miRNA is mmu-miR-134. In this paper, we have validated by TI modeling and minimal net clustering all the experimentally tested miRNA repressors of GCNF and Oct4. In our previous work [38], actually none of these miRNA repressors had yet been studied for modeling for the four proteins because of the restriction of pre-selection. And only the pair (mmu-miR-181a, GCNF) was validated by the classical correlation analysis done in Gu et al. [18] for the 4 proteins GCNF, Oct4, Nanog, Sox2. Therefore, the TI modeling combined with data condensation not only reduced computational cost but also clearly extended the set of miRNA inhibitors validated by model fitting to microarray data.

Since each numerically validated miRNA cluster may contain two or more miRNAs, the miRNAs in the same cluster could also be considered as potential candidate inhibitors for further experiments to validate. For instance, as Figure 2 shows, mmu-miR-203, 330, 342, 10a and 99b are potential candidates for protein Oct4 for they are in the same cluster as mmu-miR-470, which is validated by both the numerical modeling and experiments.

After the pair (mmu-miR-181a , GCNF) was well validated by our CKE model, we found that the cluster (see Figure 5, top left) containing mmu-miR-181a also includes mmu-miR-103 and mmu-miR-107, which are two known miRNAs that have the same roles in regulating insulin sensitivity and promoting metastasis of colorectal cancer [42,43]. We also checked that miR-103 and miR-107 have almost the same mature sequences:

mmu-miR-103: AGCAGCAUUGUACAGGGCUAUGA
mmu-miR-107: AGCAGCAUUGUACAGGGCUAUCA

After the pair (mmu-let-7b, GCNF) was well validated by our CKE model, we observed mmu-let-7g, and mmu-let-7i are in the same cluster as mmu-let-7b (see Figure 5, top right). It was claimed that let-7b and 7g reduce tumor growth in mouse models of lung cancer [44]. We then checked that indeed these three miRNAs have very similar mature sequences, namely

mmu-let-7b: UGAGGUAGUAGGUUGUGUGGUU
mmu-let-7g: UGAGGUAGUAGUUUGUACAGU
mmu-let-7i: UGAGGUAGUAGUUUGUGCUGU

To evaluate the correlation between mature sequence and expression profile of our set of 266 miRNAs, we systematically explored all the $266 \times 265/2 = 35,245$ pairs (mir_i, mir_j) of distinct miRNAs in this set, $i \neq j, i, j = 1, \ldots, 35245$. For each such pair we then computed

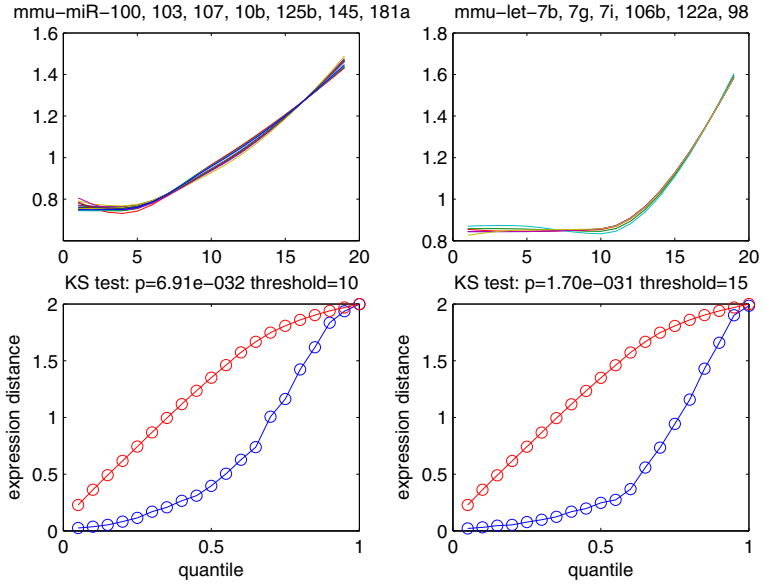

Figure 5 miRNA clusters and analysis of mature sequences and expression levels. Top left, expression profiles for a cluster of 7 miRNAs. Top right, expression profiles for a cluster of 6 miRNAs. Bottom left, the blue curve is the quantile curve of expression distances for miRNA pairs with high alignment scores ($NWA > 10$), the red curve is the analogous quantile curve for miRNA pairs with low alignment scores ($NWA \leq 10$). Note that the blue curve lies below the red curve indicates that when NWA increases, then expression distance tends to stochastically decrease. Bottom right, similar to bottom left but the NWA threshold changes to $NWA = 15$.

the Euclidean distance between the expression level profiles (expression distance in short) and the Needleman-Wunsch alignment score (NWA) between the mature sequences of each miRNA pair. We then divided the 35,245 miRNA pairs into two groups: GroupHigh includes the pairs with high alignment score ($NWA > 10$), i.e. GroupHigh includes miRNA pairs with similar mature sequences, and GroupLow includes the pairs with low alignment score ($NWA \leq 10$). We then compared the distribution of expression distances for all miRNA pairs in GroupHigh with the distribution of these distances for miRNA pairs in GroupLow. As seen in Figure 5 (bottom left), the quantiles of these distances in GroupLow are consistently larger than the corresponding quantiles in GroupHigh. This is fully confirmed by Kolmogorov-Smirnov test which yielded the very significant p-value 7×10^{-32}. The result still holds when we change the alignment score threshold $NWA = 10$ used to define GroupHigh and GroupLow (see Figure 5, bottom right, where the NWA threshold is now $NWA = 15$). We conclude that miRNAs having similar mature sequences tend, with high probability, to have similar expression levels. The above analysis and examples indicate that miRNAs belonging to the same cluster are good candidates to have similar mature sequence. Since the match between miRNA mature sequence and target sites is the main determinant for miRNA targets, miRNAs belonging to same cluster may hence also have similar regulatory roles.

Robustness of parameter estimation

Estimation algorithms for nonlinear models may yield parameter estimates that are dependent on the particular set of data or on initial estimates of parameters. Since our parameter estimation algorithm is independent from initial estimates of parameters, we now focus on on the measurements errors affecting microarray data and on their impact for parameter estimation. Considering that

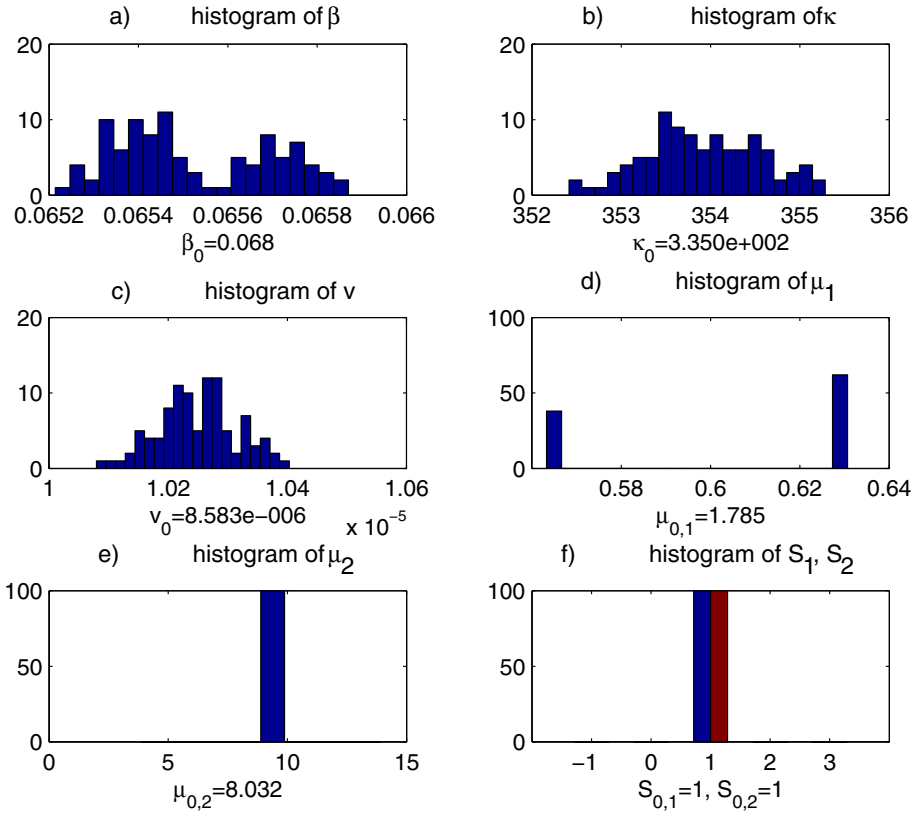

Figure 6 Histogram of re-estimated parameters for a TD model. a) the histogram of re-estimated parameter β, degradation rate of Sox2, the originally estimated β value of the model is below the graph. **b)** the histogram of re-estimated parameter κ, transcription rate of Sox2, the originally estimated κ value of the model is below the graph. **c)** the histogram of re-estimated parameter v, reaction rate of Sox2 and mmu-mir-21, the originally estimated v value of the model is below the graph. **d)** the histogram of re-estimated parameter μ_1, reaction constant of Sox2 and Oct4, the originally estimated μ_1 value of the model is below the graph. **e)** the histogram of re-estimated parameter μ_2, reaction constant of Sox2 and Nanog, the originally estimated μ_2 value of the model is below the graph. **f)** the histogram of re-estimated parameter S_1, S_2, number of binding sites of Oct4 and Nanog on Sox2 respectively, the originally estimated S_1 and S_2 value of the model is below the graph.

the noises of the microarray data are not negligible, we have analyzed the robustness of our parameters estimators when one perturbs randomly the observed expression levels of miRNAs. We have selected an arbitrary validated model *MD*, where the corresponding downstream factor is denoted as *D*, and has expression levels $d(t)$. Denote miRNAs $M_1, \ldots, M_j$ pertaining to model *MD* and call their expression levels $m_1(t), \ldots, m_j(t)$. Then we have perturbed the expression levels of the miRNAs by independent random noises having the recorded standard deviation $\sigma_1(t), \ldots, \sigma_j(t)$ and obtained simulated expression levels $sm_1(t), \ldots, sm_j(t)$. After injecting the perturbed expression levels $sm_1(t), \ldots, sm_j(t)$ into the model *MD* and get the predicted expression level $pd(t)$ of *D*. With $pd(t), m_1(t), \ldots, m_j(t)$ and expression levels of other upstream factors, we then applied our parameter estimation algorithm to re-estimate the model parameters. This procedure was repeated 100 times. Then for each model parameter, we plotted the histograms of those 100 re-estimated parameter values and compared them with the parameter values estimated from unperturbed data of model *MD*. This analysis showed that our parameter estimation algorithm is quite robust. Here we present the histograms of perturbed estimates of model parameters for one TD motif (Figure 6) and one TI motif (Figure 7).

Conclusion

We have separately modeled by chemical kinetics equations the 2 distinct modalities of the repressive actions of miRNAs on post-transcriptional processes of mRNA genes and the associated proteins. This was achieved by first defining the formal structure of two types of interaction architectures (Transcription Degradation motifs and Translation Inhibition motifs) linking miRNAs to subgroups of mRNA genes. The plausibility of each one of these potential TD motifs or TI-motifs was then evaluated by computerized parametric modeling, based on microarray data, of adequate formal chemical kinetics equations (CKEs).

We have sketched the formal derivation of 2 specific CKEs modeling by dynamic ODEs the interactions between concentrations of different species of molecules involved in each architecture. This led to a motif validation strategy based on the quantified quality of fit between our optimally parametrized models and the corresponding microarray data.

Our computerized parameter estimation is implemented by an innovative fast algorithm that does not require knowledge of range of molecular reaction rates. On a current standard laptop PC, our implementation of parameter estimation for a typical 9-parameters CKE model requires about 5 minutes of computing time.

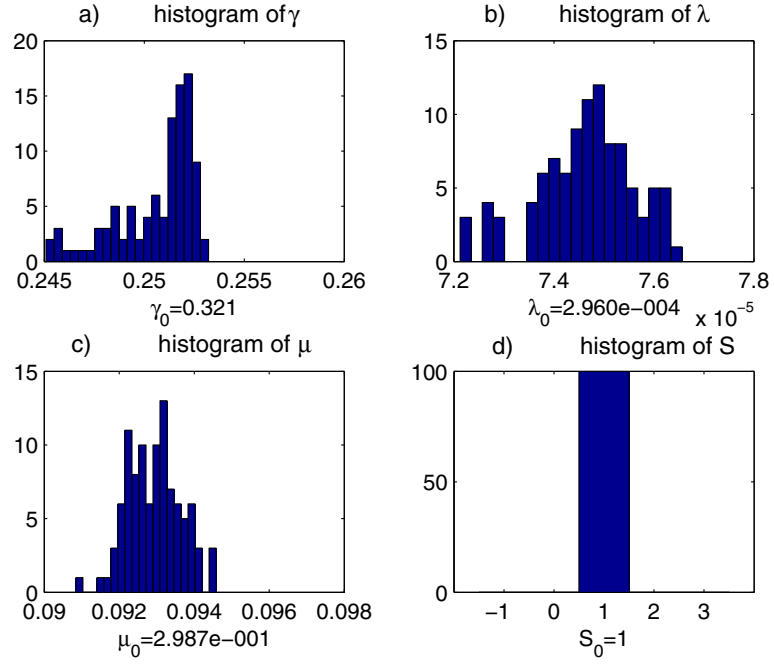

Figure 7 Histogram of re-estimated parameters for a TI model. a) the histogram of re-estimated parameter γ, degradation rate of GCNF, the originally estimated γ value of the model is below the graph. **b)** the histogram of re-estimated parameter λ, translation rate of GCNF, the originally estimated λ value of the model is below the graph. **c)** the histogram of re-estimated parameter μ, reaction constant of GCNF and mmu-let-7b, the originally estimated μ value of the model is below the graph. **d)** the histogram of re-estimated parameter *S*, number of binding sites of mmu-let-7b on GCNF respectively, the originally estimated *S* value of the model is below the graph.

Our parameter estimation algorithm also provides relatively high-quality optimization for the fit between model and microarray data, by integrating both global and local cost minimization techniques, in contexts where plausible ranges of values for most of the unknown parameters are not available in the literature. By perturbing the expression levels of miRNAs and re-estimating the parameters, we showed that our parameter algorithm has a satisfactory level of robustness. We believe that our parameter estimation technique with associated evaluation of quality of fit would be quite applicable as a generic algorithm to similar problems in chemical kinetics modeling of molecular interactions.

Modeling very large microarray data is computationally quite expensive. We have hence sketched clustering methods to condense large microarray data. This approach has of course been attempted before our work, but the main point is that we have carefully studied the mathematical compatibility of our CKE models with condensation of the profiles data. Since we have proved that the abstract form of our CKE models is invariant by arbitrary multiple affine transformations of profiles data, we have made sure to constrain the distance of two expression levels profiles to be invariant by these types of affine transformations.

We have implemented a Minimal Net Clustering algorithm based on this distance, which allows us to control the radius of the clusters. The number of CKEs to parameterize can be strongly reduced after condensation of the large data sets, and the affine invariance of our CKEs show that the condensed genes network can then still be modeled by similar CKEs.

By applying our TI modeling to multiple proteins such as GCNF, Oct4, Nanog, Sox2, we showed that 3 miRNA-target pairs experimentally validated can be also validated by the TI model.

Additional files

Additional file 1: Codes-minimal net clustering.

Additional file 2: Codes-parameter algorithm.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZL developed the chemical kinetics equations and the parameter estimation algorithm; RA proposed the transformation invariance of chemical kinetics equations and data condensation method; YZ did the computation work. ZL and RA wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Drs P. Gunaratne, A. J. Cooney, D. Lonard, X. Xu for multiple discussions concerning their experimental data and results on miRNA-mRNA interactions. These discussions led us to a joint publication.

Received: 10 March 2013 Accepted: 12 February 2014
Published: 18 February 2014

References

1. Ambros V: **The functions of animal microRNAs.** *Nature* 2004, **431**(7006):350–355.
2. Aravin A, Tuschl T: **Identification and characterization of small RNAs involved in RNA silencing.** *FEBS Lett* 2005, **579**(26):5830–5840.
3. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136**(2):215–233.
4. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I: **MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation.** *Nature* 2008, **455:**1124–1128.
5. Wang XJ, Reyes JL, Chua NH, Gaasterland T: **Prediction and identification of arabidopsis thaliana microRNAs and their mRNA targets.** *Genome Biol* 2004, **5**(9):R65.
6. Kawasaki H, Taira K: **MicroRNA-196 inhibits HOXB8 expression in myeloid differentiation of HL60 cells.** *Nucleic Acids Symp Ser* 2004, **48**(48):211–212.
7. Moxon S, Jing R, Szittya G, Schwach F, RPR L, Moulton V, Dalmay T: **Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening.** *Genome Res* 2008, **18**(10):1602–1609.
8. Williams AE: **Functional aspects of animal microRNAs.** *Cell Mol Life Sci* 2008, **65**(4):545–562.
9. Mazière P, Enright AJ: **Prediction of microRNA targets.** *Drug Discov Today* 2007, **12**(11–12):452–458.
10. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge C: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115:**787–798.
11. Lewis BP, Burge CB, Bartel D: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120:**15–20.
12. Brennecke J, Stark A, Russell RB, Cohen S: **Principles of microRNA-target recognition.** *PLoS Biol* 2005, **3:**e85.
13. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein E, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37:**495–500.
14. Chi SW, Hannon GJ, Darnell RB: **An alternative mode of microRNA target recognition.** *Nat Struct Mol Biol* 2012, **19**(3):321–327.
15. Baek D, Villé J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output.** *Nature* 2008, **455:**64–71.
16. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455:**58–63.
17. Mourelatos Z: **Small RNAs: The seeds of silence.** *Nature* 2008, **455:**44–45.
18. Gu P, Reid JG, Gao X, Shaw CA, Creighton C, Tran PL, Zhou X, Drabek RB, Steffen DL, Hoang DM, Weiss MK, Naghavi AO, El-daye J, Khan MF, Legge GB, Wheeler DA, Gibbs RA, Miller JN, Cooney AJ, Gunaratne PH: **Novel miRNA candidates and miRNA-mRNA pairs in ES cells.** *PLoS ONE* 2008, **3**(7):e2548.
19. de Jong H: **Modeling and simulation of genetic regulatory systems: a literature review.** *J Comput Biol* 2002, **9**(1):67–103.
20. Goutsias J, Kim S: **A nonlinear discrete dynamical model for transcriptional regulation: construction and properties.** *Biophys J* 2004, **86**(4):1922–1945.
21. Cornish-Bowden A: *Fundamentals of Enzyme Kinetics, 3rd Edition*. London: Portland Press; 2004.
22. Goutsias J, Lee NH: **Computational and experimental approaches for modeling gene regulatory networks.** *Curr Pharm Des* 2007, **13:**1415–1436.
23. Moore JW, Pearson RG: *Kinetics and Mechanism, 3rd Edition*. New York: Wiley; 1981.
24. Slonim DK, Yanai I: **Getting started in gene expression microarray analysis.** *PLoS Comput. Biol* 2009, **5**(10):e1000543.
25. Engl HW, Flamm C, Kügler P, Lu J, Müller S, Schuster P: **Inverse problems in systems biology.** *IOP Sci* 2009, **25:**123014.
26. Arkin A, Ross J, McAdams HH: **Stochastic kinetic analysis of developmental pathway bifurcation in phage l-infected escherichia coli cells.** *Genetics* 1998, **149:**1633–1648.
27. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay.** *Proc Natl Acad Sci* 2002, **99**(9):5860–5865.
28. Meir E, Munro EM, Odell GM, von Dassow G: **Ingeneue: a versatile tool for reconstituting genetic networks, with examples from the segment polarity network.** *J Exp Zool* 2002, **294:**216–251.

29. Müller S, Hofbauer J, Endler L, Flamm C, Widder S, Schuster P: **A generalized model of the repressilator.** *J Math Biol* 2006, **53:**905–937.
30. Widder S, Schicho J, Schuster P: **Dynamic patterns of gene regulation I: simple two-gene systems.** *J Theor Biol* 2007, **246:**395–419.
31. Hornstein E, Shomron N: **Canalization of development by microRNAs.** *Nat Genet* 2006, **38:**S20–S24.
32. Azencott R, Coldefy F, Younes L: **A distance for elastic matching in object recognition.** In *Pattern Recognition, Proceedings of 13th Int. Conf. ICPR 96, vol. 1*. IEEE; 1996:687–691.
33. Periwal V, Chow CC, Bergman RN, Ricks M, Vega GL, Sumner AE: **Evaluation of quantitative models of the effect of insulin on lipolysis and glucose disposal.** *Am J Physiol Regul Integr Comp Physiol* 2008, **295:**R1089–R1096.
34. Gregory PC: *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach With Mathematical Support*. London: Cambridge University Press; 2005.
35. Küegler P, Gaubitzer E, Müller S: **Parameter identification for chemical reaction systems using sparsity enforcing regularization: a case study for the Chlorite-Iodide Reaction.** *J Phys Chem A* 2009, **113:**2775–2785.
36. Vapnik V: *Statistical Learning Theory*. New York: Wiley; 1998.
37. Boyd S, Vandenberghe L: *Convex Optimization*. Cambridge: Cambridge University Press; 2004.
38. Luo Z, Xu X, Gu P, Lonard D, Gunaratne P, Cooney AJ, Azencott R: **Regulatory circuits of miRNAs in ES cell differentiation: a chemical kinetics modeling approach.** *PLoS One* 2011, **6**(10):e23263.
39. Vergoulis TI, Vlachos P, Alexiou G, Georgakilas M, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou A: **Tarbase 6.0: capturing the exponential growth of miRNA targets with experimental support.** *Nucl Acids Res* 2012, **40**(D1):D222–D229.
40. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, YR A: **Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells.** *Cell* 2008, **10:**1016.
41. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA: **Core transcriptional regulatory circuitry in human embryonic stem cells.** *Cell* 2005, **122**(6):947–956.
42. Trajkovski M, Hausser J, Soutschek J, Bhat B, Akin A, Zavolan M, Heim MH, Stoffel M: **MicroRNAs 103 and 107 regulate insulin sensitivity.** *Nature* 2011, **474**(7353):649–653.
43. Chen HY, Lin YM, Chung HC, Lang YD, Lin CJ, Huang J, Wang WC, Lin FM, Chen Z, Huang HD, Shyy JY, Liang JT, Chen RH: **miR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4.** *Cancer Res* 2012, **72**(14):3631–3641.
44. Esquela-Kerscher A, Trang P, Wiggins JF, Patrawala L, Cheng A, Ford L, Weidhaas JB, Brown D, Bader AG, Slack FJ: **The let-7 microRNA reduces tumor growth in mouse models of lung cancer.** *Cell Cycle* 2008, **7**(6):759–764.

doi:10.1186/1752-0509-8-19
Cite this article as: Luo *et al.*: Modeling miRNA-mRNA interactions: fitting chemical kinetics equations to microarray data. *BMC Systems Biology* 2014 **8**:19.