Online Supplementary Material

Supplementary Methods

Network parameter identification via time series microarray data

After constructing the dynamic models of the candidate cellular protein interaction network, the interaction parameters in the models have to be identified using the collected microarray data. The identification strategy is to outline the cellular protein interaction network protein by protein, i.e. for p=1, ..., N. The regulatory parameters are identified by solving a least square parameter estimation problem [1].

Equation (1) can be rewritten as the following regression form:

$$y_p[t+1] = \begin{bmatrix} y_p[t]y_1[t] & \cdots & y_p[t]y_{Q_P}[t] & x_p[t] & y_p[t] \end{bmatrix} \cdot \begin{bmatrix} b_{p1} \\ \vdots \\ b_{pQ_P} \\ \alpha_p \\ (1-\beta_p) \end{bmatrix} + \omega_p[t] \qquad \text{(S1)}$$

$$= \phi_p[t] \cdot \theta_p + \omega_p[t]$$

where $\phi_p[t]$ indicates the regression vector and θ_p is the parameter vector to be estimated. In order to avoid overfitting, the cubic spline method [2, 3] was used to interpolate extra time points for the gene expression data. For simplicity, at different time points i.e. t=1, ..., L, equation (S1) could be presented as follows:

$$Y_p = \Phi_p \cdot \theta_p + \Omega_p \qquad \text{(S2)}$$

where $Y_p = \begin{bmatrix} y_p[2] & y_p[3] & \cdots & y_p[L] \end{bmatrix}^T$, $\Phi_p = \begin{bmatrix} \phi_p[1] & \phi_p[2] & \cdots & \phi_p[L-1] \end{bmatrix}$, and L is the number of microarray data points after cubic spline interpolation. The parameter estimation problem can then be formulated as the following least square minimization problem

$$\min_{\theta_p} \frac{1}{2} \left\| Y_p - \Phi_p \cdot \theta_p \right\|_2^2 \qquad \text{(S3)}$$

The least square minimization problem in equation (S3) can be solved for the optimal estimate

θ_P^{T} by the quadratic programming [1]. We can estimate the regulatory interaction parameters b_{pq} protein by protein for the candidate PPI network via time series microarray data: i.e., through the estimated interaction abilities $\hat{b}_{pq},\ p=1,..,N,\ q=1,..,Q$, we can estimate the regulatory interaction parameters of a potential PPI network.

Determination of cross-correlation values

Keeping the pairwise relationship of these *N-k* pairs to maintain dependence between $(x_i,\ y_{i+k})$, z_i is sampled with N-k replacements to form a bootstrapped sample *Z** = { z_i^*: i = 1, ..., *N-k* and z_i^* belongs to *Z**}. The correlation coefficient from the bootstrapped sample *Z** is computed and denoted as $c^*(k),\ -1 \le c^*(k) \le 1$. Repeating the resampling procedure B times, we will observe $c_1^*(k),\ c_2^*(k),\ \ldots,\ c_B^*(k)$. These bootstrapped correlation coefficients are derived as $-1 \le c_{(1)}^*(k) \le c_{(2)}^*(k) \le \ldots \le c_{(B)}^*(k) \le 1$. In this case, the two-sided percentile interval of $(1-\alpha)$ is given by $\left[c_{(B\times\alpha/2)}^*(k),\ c_{(B\times(1-\alpha/2))}^*(k)\right]$ [4]. If this percentile interval does not contain 0, then the null hypothesis is rejected at significance level α. Otherwise, the data fails to reject the null hypothesis, again at significance level α. Since the p-value is the smallest value of α for which the null hypothesis will be rejected based on the observation, the p-value for this test is estimated as follows:

$$\hat{p}(k) = \min\left\{\hat{p}_+(k),\ 1-\hat{p}_+(k)\right\},\quad \text{where } \hat{p}_+(k) = \sum_{i=1}^{B} I\left\{c_i^*(k) \ge 0\right\} / B \qquad \text{(S4)}$$

where I{·} is the indicator function whose value is one when the event is true and zero otherwise. The time-lagged correlation (*TlC*) of $\vec{x}$ and $\vec{y}$ is defined as $c(j)$ having the smallest *p*-value (i.e. $TlC(\vec{x},\vec{y}) = c(j) \text{ if } p(j) \le p(k)\ \forall k \ne j$). Note that $-1 \le TlC(\vec{x},\vec{y}) \le 1$. So that $TlC(\vec{x},\vec{y})$ is the cross-correlation between $\vec{x}$ and $\vec{y}$.

References

1. Coleman TF, Hulbert LA: A Direct Active Set Algorithm for Large Sparse Quadratic Programs with Simple Bounds. *Math Program* 1989, 45(3):373-406.
2. Bar-Joseph Z, Gerber G, Gifford DK, Jaakkola TS, Simon I: A new approach to analyzing gene expression time series data. In: *Proceedings of the sixth annual international conference on Computational biology; Washington, DC, USA*. 565202: ACM 2002: 39-48.
3. De Boor C: A practical guide to splines : with 32 figures, Rev. edn. New York: Springer; 2001.
4. Efron B, Tibshirani R: An introduction to the bootstrap. New York: Chapman & Hall; 1993.