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 ~~network dynamic~~ models have to be identified using the collected microarray data. The identification strategy is to ~~identify~~outline the cellular protein interaction network ~~one~~ protein by ~~one~~ protein, i.e. for p~~=~~ = 1, ..., N. The regulatory parameters are identified by solving a least square parameter estimation problem [63].

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

$$y_p[t+1] = \left[y_p[t] y_1[t] \quad \cdots \quad y_p[t] y_{Q_p}[t] \quad x_p[t] \quad y_p[t] \right] \cdot \begin{bmatrix} b_{p1} \\ \vdots \\ b_{pQ_p} \\ \alpha_p \\ (1-\beta_p) \end{bmatrix} + \omega_p[t] \quad \text{(S1)}$$

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

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 [22-23] 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~~ can be presented as ~~the following equation.~~follows:

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

~~Where~~where $Y_p = \left[y_p[2] \quad y_p[3] \quad \cdots \quad y_p[L] \right]^T$, $\Phi_p = \left[\phi_p[1] \quad \phi_p[2] \quad \cdots \quad \phi_p[L-1] \right]$, and L is the number of ~~data points of~~ microarray data points after cubic spline interpolation. The parameter estimation problem can then be formulated as the following least square minimization problem:

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

Since ~~it is still lack of~~ large-scale ~~measurement~~measurements of protein activities are lacking, mRNA expression profiles are used ~~to substitute for the protein activity levels~~ instead when identifying the interaction parameters. Although the mRNA expression ~~level~~levels can't be completely representative of ~~protein expression level, but there are partially positive correlation between mRNA and~~ the corresponding protein expression levels, they are at least partially and positively correlated [64-65]. The least square minimization problem in equation (S3) can be solved for the optimal estimate

θ_P^{T} by quadratic programming [63]. We can estimate the regulatory interaction parameters b_{pq} ~~one~~ protein by ~~one~~ 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 ~~the~~a potential PPI network.

Determination of ~~the~~ cross-correlation values

Keeping the ~~pair~~pairwise relationship of these N-k pairs to maintain ~~the~~ dependence between (x_i, y_{i+k}), z_i ~~are~~is sampled with N-k ~~times of~~ replacements to form a ~~bootstrap~~bootstrapped sample~~,~~ $Z^* = \{ z_i^* : i = 1, ..., N\text{-}k$ and z_i^* belongs to $Z^*\}$. The correlation coefficient from the ~~bootstrap~~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 ~~bootstrap~~bootstrapped correlation coefficients are ~~sorted to be~~derived as $-1 \le c_{(1)}^*(k) \le c_{(2)}^*(k) \le \ldots \le c_{(B)}^*(k) \le 1$. ~~Then~~In this case, the two-sided percentile interval of $(1-\alpha)$ is given by $\left[c_{(B\alpha/2)}^*(k), c_{(B(1-\alpha/2))}^*(k)\right]$ ~~in this case~~ [31]. If this percentile interval does not contain 0, then the null hypothesis is rejected at ~~the~~ significance level ~~of~~ α. Otherwise, the data fails to reject the null hypothesis, again at ~~the~~ significance level ~~of~~ 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:

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

where $I\{\cdot\}$ is the indicator function whose value is one when the event is true and zero otherwise. The time-lagged correlation (TlC) of $\check{x}$ and $\check{y}$ is defined as $c(j)$ ~~that has~~having the smallest p-value (i.e. $TlC(\check{x}, \check{y}) = c(j)$ if $p(j) \le p(k)$ $\forall k \ne j$).

註解 [Editor2]:

CHECK: Most of the mathematical text in this section has incorrect symbols (e.g., £) or garbled symbols, making the text have no mathematical sense; please check and possibly revise this equation before submission to your target journal.