Multiple Datasets Tabular Combination Model : Supporting Material for Main Paper

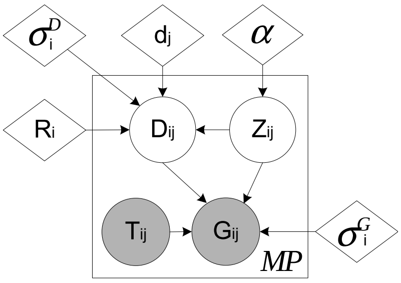
1 Model Description

Multiple Datasets Tabular Combination (MDTC) Model is designed for pairwise correspondence inference from multiple time series datasets. To simplify our discussion, we use gene regulatory correspondence as a working example, and present how to apply the MDTC Model to infer pairwise regulatory correspondence between transcription factors (TFs) and genes from multiple time series Microarray expression datasets collected under a variety of experimental conditions.

Figure 1(a) is the tabular representation of MDTC Model. Each column represents a particular TF-gene pair. The first row represents global parameters for each of the pairs. Each of the other rows represents an expression experiment.

Time scale factor	Experimental Conditions	TF-gene pair 1	...	TF-gene pair P
1	Canonical condition	d_1	...	d_P
R_1	Condition 1	$V_{1,1}$	...	$V_{1,P}$
R_2	Condition 2	$V_{2,1}$	...	$V_{2,P}$
...	...	...	...	...
R_M	Condition M	$V_{M,1}$	...	$V_{M,P}$

(a)



(b)

Figure 1: Two representations of Multiple Datasets Tabular Combination Model. (a) The tabular representation. $V_{ij} = \{T_{ij}, G_{ij}, D_{ij}, Z_{ij}\}$, where T_{ij} and G_{ij} are the normalized expression profiles of the TF and gene of TF-gene pair j under condition i, respectively. D_{ij} is the actual lag between TF-gene pair j under condition i. Z_{ij} is a binary variable which indicates whether there is a regulatory correspondence between TF-gene pair j under condition i. R_i is the time scale factor for condition i. d_j is the canonical lag for pair j. (b) The graphical model representation. Observed variables are shaded. For an experiment in which an correspondence exists between the TF and the gene (that is, if $Z_{ij} = 1$), the actual lag D_{ij} depends on the canonical lag, d_j, and the specific time scale factor, R_i. And the observed expression profile for the gene, G_{ij}, is constrained by the TF's profile, T_{ij}, and the actual lag D_{ij}.

The MDTC Model associates four variables with each cell in the table: (a) the observed expression profiles of the TF and gene in this cell (denoted by T_{ij} and G_{ij}, respectively), which have been normalized to have zero mean and unit standard deviation, (b) whether there exists a regulatory correspondence for this TF-gene pair in this dataset (denoted by the binary variable Z_{ij}), and if so, (c) the actual lag (denoted by D_{ij}) for this correspondence. Both of Z_{ij} and D_{ij} are unobserved. In addition, the MDTC Model associates one parameter with each column: the canonical lag, d_j, for the pair. This parameter allows us to associate different lags with different pairs. Also, the MDTC Model associates one parameter with each row: the time scale factor, R_i. R_i represents the linear transformation required to transform the time unit of one experiment to another. It has been shown that this linear transformation provides a good fit for mapping one time scale to another in expression experiments [1]. Thus, while one can use a more complicated transformation, with the proposed model we restrict our attention to linear transformation.

Given the column and row parameters, the expected lag, $Expected\_Lag_{ij}$, for each cell can be computed as the product of the canonical lag, d_j, for the TF-gene pair, and the time scale factor, R_i, for this experiment. The modelling assumption here is that the expected lag between the expression profiles of the TF and the gene is determined by both the typical lag for this pair and the experimental condition of the particular dataset, namely:

$$Expected\_Lag_{ij} = R_i \times d_j \tag{1}$$

Figure 1(b) uses a graphical model to illustrate the dependencies among variables in the MDTC Model. Let $T_{ij}(t)$ and $G_{ij}(t)$ denote the values of TF and gene's profiles of TF-gene pair j under condition i evaluated at time point t, respectively, and let L_i denote the length of the i^{th} experiment, θ denote the model parameters $\{R_i, d_j, (\sigma_i^D)^2, (\sigma_i^G)^2\}$ $(i = 1, .., M; j = 1, .., P)$. The conditional probabilities in this graphical model are defined as:

$$\begin{aligned}
P(G_{ij}(t)|T_{ij}, D_{ij}, Z_{ij} = 1, \theta) &\sim \begin{cases} \mathcal{N}(0, (\sigma_i^G)^2) & t \in [0, D_{ij}] \\ \mathcal{N}(T_{ij}(t - D_{ij}), (\sigma_i^G)^2) & t \in (D_{ij}, L_i] \end{cases} && (2) \\
P(G_{ij}(t)|T_{ij}, D_{ij}, Z_{ij} = 0, \theta) &\sim \mathcal{N}(0, 1); && (3) \\
P(G_{ij}|T_{ij}, D_{ij}, Z_{ij}, \theta) &= \exp\,(\int_{t=0}^{L_i} \log P(G_{ij}(t)|T_{ij}, D_{ij}, Z_{ij})\,dt); && (4) \\
P(T_{ij}(t)|\theta) &\sim \mathcal{N}(0, 1); && (5) \\
P(T_{ij}|\theta) &= \exp(\int_{t=0}^{L_i} \log(P(T_{ij}(t)))\,dt); && (6) \\
P(D_{ij}|Z_{ij} = 1, \theta) &\sim \mathcal{N}(Expected\_Lag_{ij}, (\sigma_i^D)^2); && (7) \\
P(D_{ij}|Z_{ij} = 0, \theta) &\sim Uniform(0, L_i); && (8) \\
P(Z_{ij} = 1|\theta) &= \alpha; && (9)
\end{aligned}$$

Equations 2, 3 and 4 reflect the key assumption in our model. When $Z_{ij} = 1$, the gene's profile is a lagged noisy repeat of the TF's profile (Equation 2). The distance of this lag is D_{ij}, which accounts for the translation, accumulation and transportation time of the TF. A Gaussian noise with variance, $(\sigma_i^G)^2$, is added to this repeat. This noise represents the biological and experimental noise that might lead to slight difference from the expected expression level. Prior to its activation by the TF (between time point 0 and D_{ij}), the gene's profile is modelled as Gaussians with zero mean and the similar variance. When $Z_{ij} = 0$, the gene might be either regulated by another TF or not activated. To reflect this uncertainty, each point in the profile is modelled as a Gaussian with zero mean and unit standard deviation (Equation 3, we normalize profiles to zero mean and unit standard deviation). The overall dependency of G_{ij} on its parents are derived from Equation 2 and 3, and expressed in Equation 4.

Since we do not try to explain the profile of the TF in the pair, Equation 5 and 6 assign equal probability to any TF profile.

In Equation 7 and 8, when $Z_{ij} = 1$, the actual lag, D_{ij}, is assumed to follow a Gaussian distribution[1] whose mean is $Expected\_Lag_{ij}$ which is equal to $R_i \times d_j$, and variance is $(\sigma_i^D)^2$. $(\sigma_i^D)^2$ represents biological and experimental noise which may lead to lags that are slightly different from the expected lag. When $Z_{ij} = 0$, no correspondence exists for this pair. Thus, the lag D_{ij} is not meaningful and value in its range is equally probable.

Finally, in Equation 9, we assign the same prior probability, α, to every Z_{ij}. This prior can be determined by domain knowledge about the expected number of interactions.

To summarize, we list the model variables and parameters as follows:

[1] Since the actual lag should be between zero and the length of the profile, the distribution is in fact a truncated Gaussian distribution. In practice, we did sample D_{ij} according to this truncated distribution. However, since the normalization term for this truncated Gaussian is very close to 1, we ignored its effect when updating the parameters in M-step.

1 Hidden variable : Z_{ij} and D_{ij};

2 Observed variable : T_{ij} and G_{ij};

3 parameters : $\theta = (R_i, d_j, (\sigma_i^D)^2, (\sigma_i^G)^2)$

where $i = 1, ..., M$ and $j = 1, ..., P$.

We are now ready to define the likelihood of our model. From the variable dependencies described by the graphical model in Figure 1(b), we can write the likelihood of each cell in Figure 1(a). Using the cell likelihood, the complete likelihood for our model is the product of the likelihood of all cells. The complete log-likelihood can be written as:

$$\begin{aligned} LL &= \sum_{i=1}^{M}\sum_{j=1}^{P} \log P(T_{ij}, G_{ij}, D_{ij}, Z_{ij}|\theta) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{P} \log P(G_{ij}|T_{ij}, D_{ij}, Z_{ij}, \theta) + \log P(T_{ij}|\theta) + \log P(D_{ij}|Z_{ij}, \theta) + \log P(Z_{ij}|\theta) \end{aligned} \tag{10}$$

where M is the number of experimental conditions (rows), and P is the number of pairs (columns).

Since $\log P(T_{ij}|\theta)$ is a constant, maximizing this likelihood is equivalent to maximizing the other three terms, which are (i) $P(G_{ij}|T_{ij}, D_{ij}, Z_{ij}, \theta)$, how well the time series expression profiles for the TF and gene align under the actual lag, D_{ij}, (ii) $P(D_{ij}|Z_{ij}, \theta)$, how much D_{ij} deviates from the expected lag given by Equation 1, and (iii) $P(Z_{ij}|\theta)$, the prior of Z_{ij}. This reflects the essence of our model, choosing an actual lag that balances the desire to find the best match between the profiles of TF and gene (measured by (i)) against the desire to choose an actual lag near the expected lag (measured by (ii)) for this cell.

1.1 Estimating the model parameters

We now describe an EM algorithm to estimate the model parameters, θ, by seeking to maximize the expected likelihood.

In E-step, we calculate the expectation of the complete log-likelihood. The expectation is under the distribution of the hidden variables given the observed variables and the parameters in last iteration. Namely,

$$\begin{aligned} E(LL) &= \sum_{i=1}^{M}\sum_{j=1}^{P} E(\log P(T_{ij}, G_{ij}, D_{ij}, Z_{ij}|\theta)) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{P} E(\log P(G_{ij}|T_{ij}, D_{ij}, Z_{ij}|\theta)) \\ &\quad + E(\log P(T_{ij}|\theta)) + E(\log P(D_{ij}|Z_{ij}, \theta)) + E(\log P(Z_{ij}|\theta)) \end{aligned} \tag{11}$$

This expectation is intractable in that it contains the integral over joint distribution of Z_{ij} and D_{ij}. We used Gibbs sampling to approximate this expectation.

Firstly, we can sample D_{ij} from the distribution :

$$P(D_{ij}|T_{ij}, G_{ij}, Z_{ij}, \theta) \propto P(G_{ij}|T_{ij}, D_{ij}, Z_{ij}, \theta) \times P(D_{ij}|Z_{ij}, \theta) \tag{12}$$

where $P(D_{ij}|Z_{ij},\theta)$ is a Gaussian when $Z_{ij}=1$, and uniform when $Z_{ij}=0$, and

$$\begin{aligned}
P(G_{ij}|T_{ij},D_{ij},Z_{ij}=1,\theta) &= \exp(\int_{t=0}^{D_{ij}} \log(\frac{1}{\sqrt{2\pi}\sigma_i^G}\exp(-\frac{(G_{ij}(t))^2}{2(\sigma_i^G)^2}))\,dt \\
&\quad + \int_{t=D_{ij}}^{L_i} \log(\frac{1}{\sqrt{2\pi}\sigma_i^G}\exp(-\frac{(G_{ij}(t)-T_{ij}(t-D_{ij}))^2}{2(\sigma_i^G)^2}))\,dt) \\
&= \exp(\int_{t=0}^{D_{ij}} (-\frac{(G_{ij}(t))^2}{2(\sigma_i^G)^2})\,dt + \int_{t=D_{ij}}^{L_i} (-\frac{(G_{ij}(t)-T_{ij}(t-D_{ij}))^2}{2(\sigma_i^G)^2})\,dt \\
&\quad + L_i \times \log(\frac{1}{\sqrt{2\pi}\sigma_i^G})) \\
&= \exp(-\frac{1}{2(\sigma_i^G)^2}(\int_{t=0}^{D_{ij}} (G_{ij}(t))^2\,dt + \int_{t=D_{ij}}^{L_i} (G_{ij}(t)-T_{ij}(t-D_{ij}))^2\,dt) \\
&\quad + L_i \times \log(\frac{1}{\sqrt{2\pi}\sigma_i^G})) \\
&= \exp(-\frac{1}{2(\sigma_i^G)^2} \times SE_{ij}^1(D_{ij}) + L_i \times \log(\frac{1}{\sqrt{2\pi}\sigma_i^G})) \qquad (13)
\end{aligned}$$

where $SE_{ij}^1(D_{ij})$ is a function of D_{ij}, representing the squared error of the gene's profile compared to its expected mean when $Z_{i,j}=1$. This mean is defined by Equation 2. Similarly,

$$P(G_{ij}|T_{ij},D_{ij},Z_{ij}=0,\theta) = \exp(-\frac{1}{2} \times SE_{ij}^0 + L_i \times \log(\frac{1}{\sqrt{2\pi}})) \qquad (14)$$

where SE_{ij}^0 is the squared error of the gene's profile compared to zero, which is the mean of the gene's profile (defined by Equation 3) when $Z_{ij}=0$. In practice, we approximate the squared error by uniformly sampling a set of points on the domain of the profile and calculate the weighted sum of squared difference between two curves evaluated on these points.

Z_{ij} was sampled from the conditional distribution,

$$P(Z_{ij}|D_{ij},T_{ij},G_{ij},\theta) \propto P(G_{ij}|T_{ij},Z_{ij},D_{ij},\theta) \times P(D_{ij}|Z_{ij},\theta) \times P(Z_{ij}|\theta) \qquad (15)$$

where $P(G_{ij}|T_{ij},D_{ij},Z_{ij},\theta)$ can be calculated by Equation 13 and Equation 14.

After sampling, we approximate the expectation in Equation 11, by

$$\begin{aligned}
E_{D,Z|T,G,\theta^{old}}(LL) &\approx \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P}(\log P(G_{ij}|T_{ij},D_{ij,s},Z_{ij,s},\theta) \\
&\quad + \log P(T_{ij}|\theta) + \log P(D_{ij,s}|Z_{ij,s},\theta) + \log P(Z_{ij,s}|\theta)) \qquad (16)
\end{aligned}$$

where $D_{ij,s}$ is the s^{th} sampled actual lag ,$Z_{ij,s}$ is the s^{th} sampled existence indicator, and S is the number of samples.

In M-step, we search for the parameters, θ, in order to maximize the expected log-likelihood approximated in E-step.

Since searching for $(\sigma_i^G)^2$ is independent of searching for the other parameters. We first consider searching R_i, d_j and $(\sigma_i^D)^2$ which are only related with,

$$\begin{aligned}
& \sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P} \log P(D_{ij,s}|Z_{ij,s},\theta) \qquad (17) \\
= & \sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P}(I(Z_{ij,s}=1)\log(\frac{1}{\sqrt{2\pi(\sigma_i^D)^2}}\exp((-\frac{(D_{ij,s}-R_i \times d_j)^2}{2(\sigma_i^D)^2})^2)) \\
& + I(Z_{ij,s}=0)\log(\frac{1}{L_i})) \qquad (18)
\end{aligned}$$

where I($\cdot$) is a indicator function, whose value is 1 when the expression in parenthesis is true, and 0 otherwise.

The searching problem becomes looking for R_i, d_j and $(\sigma_i^D)^2$ to maximize,

$$\sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P} I(Z_{ij,s}=1)\log(\frac{1}{\sqrt{2\pi(\sigma_i^D)^2}}\exp((-\frac{(D_{ij,s}-R_i\times d_j)^2}{2(\sigma_i^D)^2})^2)) \tag{19}$$

We zero the first derivative of Equation 19 with respect to R_i, d_j and $(\sigma_i^D)^2$, which leads to the following equations:

$$(\sigma_i^D)^2 = \frac{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)(D_{ij,s}-R_i\times d_j)^2}{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)} \tag{20}$$

$$R_i = \frac{\sum_{s=1}^{S}\sum_{j=1}^{P} D_{ij,s}\times d_j\times I(Z_{ij,s}=1)}{\sum_{s=1}^{S}\sum_{j=1}^{P} d_j^2\times I(Z_{ij,s}=1)} \tag{21}$$

$$d_j = \frac{\sum_{s=1}^{S}\sum_{i=1}^{M} D_{ij,s}\times R_i\times I(Z_{ij,s}=1)}{\sum_{s=1}^{S}\sum_{i=1}^{M} R_i^2\times I(Z_{ij,s}=1)} \tag{22}$$

It is hard to find the close form for these equations. To overcome this problem, We used coordinate ascent to find approximate solutions to these equations. Specifically, we used the values of R_i, d_j and $(\sigma_i^D)^2$ in the t^{th} iteration (denoted by $R_{i,t}$, $d_{j,t}$ and $(\sigma_{i,t}^D)^2$, respectively) to calculate the values of R_i, d_j, and $(\sigma_i^D)^2$ in the next iteration, and got the following iterative solution for R_i, d_j and $(\sigma_i^D)^2$,

$$(\sigma_{i,t+1}^D)^2 = \frac{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)(D_{ij,s}-R_{i,t}\times d_{j,t})^2}{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)} \tag{23}$$

$$R_{i,t+1} = \frac{\sum_{s=1}^{S}\sum_{j=1}^{P} D_{ij,s}\times d_{j,t}\times I(Z_{ij,s}=1)}{\sum_{s=1}^{S}\sum_{j=1}^{P} d_{j,t}^2\times I(Z_{ij,s}=1)} \tag{24}$$

$$d_{j,t+1} = \frac{\sum_{s=1}^{S}\sum_{i=1}^{M} D_{ij,s}\times R_{i,t}\times I(Z_{ij,s}=1)}{\sum_{s=1}^{S}\sum_{i=1}^{M} R_{i,t}^2\times I(Z_{ij,s}=1)} \tag{25}$$

Finally, we search for $(\sigma_i^G)^2$ which is only related with

$$\sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P}\log P(G_{ij}|T_{ij}, D_{ij,s}, Z_{ij,s}, \theta) \tag{26}$$

Maximizing this probability is equivalent to maximizing:

$$\sum_{s=1}^{S}\sum_{i=1}^{M}\sum_{j=1}^{P}(I(Z_{ij,s}=1)\times(L_i\times\log(\frac{1}{\sigma_i^G})-\frac{1}{2(\sigma_i^G)^2}\times SE_{ij,s}^1(D_{ij,s}))) \tag{27}$$

Searching for optimal $(\sigma_i^G)^2$ that maximizes Equation 27 can be calculated by simple MLE techniques:

$$(\sigma_i^G)^2 = \frac{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)\times SE_{ij,s}^1(D_{ij,s})}{\sum_{s=1}^{S}\sum_{j=1}^{P} I(Z_{ij,s}=1)\times L_i} \tag{28}$$

1.2 inference

For inference, we define a confidence score, $conf_{ij}$, to be the posterior probability of Z_{ij} given the observed variables, $P(Z_{ij}|T_{ij}, G_{ij})$. This posterior can be approximated by our final samples in E-step of Z_{ij} and D_{ij}. We disregard samples of D_{ij}, and only use the samples of Z_{ij} to approximate this marginal posterior of Z_{ij}.

2 Predicting new pairs

The prediction algorithm makes predictions regarding which new TF-gene pairs are likely to exhibit regulatory correspondences under which experimental conditions. Given a new TF-gene pair we construct a new table with only one column, the column for the new TF-gene pair. The above iterative algorithm run on this table, holding the learned parameters ($R_i, (\sigma_i^D)^2, (\sigma_i^G)^2$) fixed. This algorithm estimates the canonical lag d for the new pair, as well as the confidence scores, $conf_i$, for each of the experimental conditions. In this way, we can predict for each condition how likely the new TF-gene pair is to exhibit a regulatory correspondence, according to the final confidence score assigned after the convergence of our prediction algorithm.

In order to compare the prediction results of our proposed algorithm with previous literature, we also classify each pair into two classes, pair with regulatory correspondence or pair without regulatory correspondence. This classification task is based on the final confidence scores. For a given TF-gene pair, our prediction algorithm arrives at a final vector of confidence scores. Each score in this vector corresponds to an experimental condition. Given this score vector, we apply a threshold, S. If the final confidence score for the i^{th} experimental condition is greater than S, we predict that there is a regulatory correspondence between this TF-gene pair under the i^{th} experimental condition. We then sum the number of conditions under which the pair is expected to have a regulatory correspondence. If this number is greater than some cutoff, C, this pair is predicted to be a pair with regulatory correspondence.

The entire gene regulatory network is often partly activated (only some edges are ”active” under a particular condition). Our threshold, S, is designed to discover this conditionally active regulatory correspondence. In addition, in order to pass our cutoff, C, a TF-gene pair must exhibit strong time lagged correspondence in multiple experiments. This may reduce our ability to identify all possible pairs (or our coverage). However, as shown in main paper this allows us to drastically reduce the set of false positives when compared to the analysis that is carried out using a single time series experiment.

References

[1] Z. Bar-Joseph, G. Gerber, T.S. Jaakkola, D.K. Gifford, and I. Simon. Continuous representations of time series gene expression data. *Journal of Computational Biology*, 3-4:341–356, 2003.