Supplementary Text S1 for "A model-based method for transcription factor target identification with limited data"

Antti Honkela et al.

1 Derivation of the Gaussian process model

The linear system of ordinary differential equations underlying our model is

$$\frac{dp(t)}{dt} = f(t) - \delta p(t) \tag{1}$$

$$\frac{dm_j(t)}{dt} = B_j + S_j p(t) - D_j m_j(t), \tag{2}$$

where $f(t)$ denotes the TF mRNA, $p(t)$ the TF protein and $m_j(t)$ the target gene mRNA concentration. Assuming steady-state initial conditions $p(0) = 0$ and $m_j(0) = B_j/D_j$, the solution of the system is

$$p(t) = \exp(-\delta t) \int_0^t f(v) \exp(\delta v)\, dv \tag{3}$$

$$m_j(t) = \frac{B_j}{D_j} + S_j \exp(-D_j t) \int_0^t \exp(D_j u) \exp(-\delta u) \int_0^u f(v) \exp(\delta v)\, dv\, du. \tag{4}$$

Both of these are linear operators of $f(t)$. Hence, placing a Gaussian process prior on $f(t)$ implies a joint Gaussian process model over all $(f(t), p(t), m_j(t))$ [**?**, **?**]. This Gaussian process is completely characterised by its mean and covariance functions. Assuming $\mathrm{E}[f(t)] = 0$, the above solutions (**??**)-(**??**) imply $\mathrm{E}[p(t)] = 0$, $\mathrm{E}[m_j(t)] = B_j/D_j$.

What remains is to determine the covariance functions. These can be evaluated as expectations

$$k_{xy}(t, t') = \mathrm{E}[(x(t) - \mathrm{E}[x(t)])(y(t') - \mathrm{E}[y(t')])], \tag{5}$$

where $x, y \in \{f, p, m_j\}$. Assuming the squared exponential covariance for $f(t)$,

$$k_{ff}(t, t') = a \exp\left(-\frac{(t - t')^2}{l^2}\right), \tag{6}$$

all the required covariance functions can be derived in closed form by repeated application of the identity

$$\int_0^t \exp(Du) \operatorname{erf}(u/l + E)\, du = \frac{1}{D}\Bigg[\exp(Dt) \operatorname{erf}(E + t/l) - \operatorname{erf}(E) + \exp\left(\left(\frac{Dl}{2}\right)^2 - EDl\right) [\operatorname{erf}(E - Dl/2) - \operatorname{erf}(E - Dl/2 + t/l)]\Bigg]. \tag{7}$$

1.1 Covariance function k_{fp}

The covariance of the TF mRNA and TF protein k_{fp} is the same as the cross-covariance derived in [**?**]. In our model, this is only needed for inference of the protein concentration (such as in Figs. 1 and 2). The covariance is

$$\begin{aligned} k_{fp}(t,t') &= \exp(-\delta t') \int_0^{t'} \exp(\delta u) k_{ff}(u,t) du \\ &= \frac{\sqrt{\pi} a l}{2} \exp\left(\left(\frac{\delta l}{2}\right)^2 + \delta(t-t')\right) [\mathrm{erf}(\delta l/2 + t/l) - \mathrm{erf}(\delta l/2 + (t-t')/l)]. \end{aligned} \tag{8}$$

1.2 Covariance function k_{fm_j}

The covariance of the TF mRNA and target mRNA k_{fm_j} is

$$\begin{aligned} k_{fm_j}(t,t') = S_j \exp(-D_j t') \int_0^{t'} \exp((D_j - \delta)u) \int_0^u \exp(\delta v) k_{ff}(t,v)\, dv\, du \\ = S_j \frac{\sqrt{\pi} a l}{2(\delta - D_j)} \exp(-(D_j + \delta)t')) \\ \left(\exp\left(\left(\frac{D_j l}{2}\right)^2 + D_j t + \delta t'\right) [\mathrm{erf}(D_j l/2 + t/l) - \mathrm{erf}(D_j l/2 + (t-t')/l)] \right. \\ \left. - \exp\left(\left(\frac{\delta l}{2}\right)^2 + \delta t + D_j t'\right) [\mathrm{erf}(\delta l/2 + t/l) - \mathrm{erf}(\delta l/2 + (t-t')/l)]\right). \end{aligned} \tag{9}$$

1.3 Covariance function k_{pp}

Again following [**?**], the covariance of the TF protein k_{pp} is

$$\begin{aligned} k_{pp}(t,t') &= \exp(-\delta(t+t')) \int_0^t \exp(\delta u) \int_0^{t'} \exp(\delta u') k_{ff}(u,u') du' du \\ &= \frac{\sqrt{\pi} a l}{4\delta} \exp\left(\left(\frac{\delta l}{2}\right)^2 - \delta(t+t')\right) [h(t',t) + h(t,t')], \end{aligned}$$

where

$$\begin{aligned} h(t',t) = \Bigg\{ \exp[2\delta t] \Bigg[\mathrm{erf}\left(\left(\frac{\delta l}{2}\right) + \frac{t}{l}\right) - \mathrm{erf}\left(\left(\frac{\delta l}{2}\right) + \frac{t-t'}{l}\right) \Bigg] \\ + \Bigg[\mathrm{erf}\left(\left(\frac{\delta l}{2}\right) - \frac{t'}{l}\right) - \mathrm{erf}\left(\left(\frac{\delta l}{2}\right)\right) \Bigg] \Bigg\}. \end{aligned} \tag{10}$$

This is only needed for inference of the protein concentrations.

1.4 Covariance function k_{pm_j}

The covariance of the TF protein and target mRNA k_{pm_j} is

$$
\begin{aligned}
\frac{k_{pm_j}(t,t')}{S_j \exp(-\delta t - D_j t')} &= \int_0^{t'} \exp((D_j - \delta)u') \int_0^{t} \exp(\delta v) \int_0^{u'} \exp(\delta v') k_{ff}(v,v')\, dv'\, dv\, du' \\
&= \frac{\sqrt{\pi} a l}{4\delta} \exp\left(\left(\frac{\delta l}{2}\right)^2\right) \Bigg(\frac{2\delta \exp(-D_j t' - \delta t)}{\delta^2 - D_j^2} [\mathrm{erf}(\delta l/2 - t/l) - \mathrm{erf}(\delta l/2)] \\
&\quad + \frac{\exp(-\delta(t+t'))}{\delta - D_j} [2\,\mathrm{erf}(\delta l/2) - \mathrm{erf}(\delta l/2 - t'/l) - \mathrm{erf}(\delta l/2 - t/l)] \\
&\quad + \frac{\exp(\delta(t'-t))}{\delta + D_j} [\mathrm{erf}(\delta l/2 + t'/l) - \mathrm{erf}(\delta l/2 - (t-t')/l)] \\
&\quad + \frac{\exp(\delta(t-t'))}{\delta - D_j} [\mathrm{erf}(\delta l/2 + (t-t')/l) - \mathrm{erf}(\delta l/2 + t/l)] \Bigg) \\
&\quad + \frac{\sqrt{\pi} l}{2(\delta^2 - D_j^2)} \exp\left(\left(\frac{D_j l}{2}\right)^2 - D_j t' - \delta t\right) \Bigg(\mathrm{erf}(D_j l/2 - t'/l) - \mathrm{erf}(D_j l/2) \\
&\quad + \exp((D_j + \delta)t)[\mathrm{erf}(D_j l/2 + t/l) - \mathrm{erf}(D_j l/2 + (t-t')/l)] \Bigg).
\end{aligned}
\tag{11}
$$

This is only needed for inference of the protein concentrations.

1.5 Covariance function $k_{m_j m_k}$

The final covariance between target genes $k_{m_j m_k}$ is

$$
\begin{aligned}
k_{m_j m_k}(t,t') &= S_j S_k \exp(-D_j t - D_k t') \int_0^{t} \exp((D_j - \delta)u) \int_0^{t'} \exp((D_k - \delta)u') \\
&\qquad \int_0^{u} \exp(\delta v) \int_0^{u'} \exp(\delta v') k_{ff}(v,v')\, dv'\, dv\, du'\, du \\
&= \frac{\sqrt{\pi} a l S_j S_k}{2} \Big(h_{jk}(t,t',\delta) + h_{kj}(t',t,\delta) - h_{jk}(t,t',D_j) - h_{kj}(t',t,D_k) \Big)
\end{aligned}
\tag{12}
$$

where

$$
\begin{aligned}
h_{jk}(t,t',D_x) &= \exp\left(\left(\frac{D_x l}{2}\right)^2\right) \frac{\exp(-D_x t - D_k t')}{(D_x + \delta)(D_j - \delta)} \Bigg\{ \\
&\quad \left(\frac{\exp((D_k - \delta)t') - 1}{D_k - \delta} + \frac{1}{D_k + D_x} \right) [\mathrm{erf}(D_x l/2 - t/l) - \mathrm{erf}(D_x l/2)] \\
&\quad + \frac{\exp((D_k + D_x)t')}{D_k + D_x} [\mathrm{erf}(D_x l/2 + t'/l) - \mathrm{erf}(D_x l/2 - (t-t')/l)] \Bigg\}.
\end{aligned}
\tag{13}
$$

2 Gaussian process inference

Denoting all the observations of replicate r by $\boldsymbol{y}_r$ and a diagonal matrix with their measurement variance parameters by $\Sigma_r = \mathrm{diag}(\sigma_{1f}^2, \dots, \sigma_{nf}^2, \{\sigma_{1jm}^2, \dots, \sigma_{njm}^2\})$, the full kernel is $K_r = K + \Sigma_r$, where K can be evaluated using the above formulae.

Based on standard Gaussian process regression [?], the posterior distribution of a vector $\boldsymbol{x}$ consisting of values of f, p and m_j, not necessarily at times of observations, is Gaussian with

$$\boldsymbol{x}|\boldsymbol{y}_r \sim \mathcal{N}(\boldsymbol{\mu}_x + K_{\boldsymbol{xy}} K_r^{-1}(\boldsymbol{y}_r - \boldsymbol{\mu} y), K_{\boldsymbol{xx}} - K_{\boldsymbol{xy}} K_r^{-1} K_{\boldsymbol{yx}}), \tag{14}$$

where $K_{\boldsymbol{xy}}$ and $K_{\boldsymbol{xx}}$ can again be evaluated using the above formulae.

References

[1] Rasmussen, CE, Williams, CKI (2006) *Gaussian Processes for Machine Learning* (MIT Press).

[2] Lawrence, ND, Sanguinetti, G, Rattray, M (2007) in *Advances in Neural Information Processing Systems*, eds Schölkopf, B, Platt, JC, Hofmann, T (MIT Press, Cambridge, MA) Vol. 19, pp 785–792.