Electronic Supplementary Material (ESI) for Molecular BioSystems

Elucidating phosphorylation dynamics of the ERK MAP kinase

Supplementary material

Tina Toni, Yu-ichi Ozaki, Paul Kirk, Shinya Kuroda, Michael P. H. Stumpf

1 Data preprocessing

The experimental data consist of hundreds of measurements at 24 time points. We average over all values at each time point and obtain a single time series for each of the molecular species, MAPKK and Mpp. We call these "the averaged data". The MAPKK time series data will be used as an input to the models, and the Mpp data for comparison with the simulation output of the models. Because MAPKK and Mpp are present in the cell as monomers as well as in complex with other molecular species, their measurements give the sum of fluorescent intensities as follows: the MAPKK dataset measures $MAPKK_{tot}$ at each time point, and the Mpp dataset measures the sum of Mpp and $Mpp \cdot MKP$ intensities. There are no data available for any of the other molecular species: M, Mp, or any complex of the complexes $M \cdot MAPKK$, $Mp \cdot MAPKK$, $Mp \cdot MKP$, and therefore the model selection algorithm needs to be able to cope with missing data. This is one of the advantages of the ABC-based algorithms; the user can define the distance function that suit the available data best.

Before using the data, we first need to determine the scaling constants. The data consist of quantitative measurements of the fluorescent intensity, which is determined by the amount of fluorescent particles attached with antibodies to MAPKK and Mpp. This fluorescent intensity is related to the *concentration* of MAPKK and Mpp and before doing the model selection, we translate the fluorescent intensities into concentrations (Figure 1). Firstly, our experimental collaborators have demonstrated a linear relationship between the fluorescent intensities and the concentrations obtained by western blotting [1], and have determined the background intensity (this is, whether linear functions in Figure 1 pass through the origin or not). Secondly, we use the information from Yamada *et al.* [2], which states that the peak concentrations that both MAPKK and Mpp reach are 1500 nM, and fit a linear function through the data to determine the quantitative relationship between the concentration data and the fluorescent intensity for both molecular species:

$$Mpp_{conc} = 0.0047 \; Mpp_{fl.intensity} \tag{1a}$$

$$MAPKK_{conc} = 0.00554 \; MAPKK_{fl.intensity} + 60, \tag{1b}$$

where subscript *conc* represents the concentration (in nM units) and subscript $fl.intensity$ the fluorescent intensity.

Another nontrivial issue is how to deal with the MAPKK time series data as an input to the models. Firstly, we want to exclude the experimental noise (which is reflected in the MAPKK data) in the input to the models, and secondly, we need the MAPKK input on the whole time interval rather than only at discrete time points. In order to smooth the noise, we use Gaussian process regression (Figure 2), which has gained popularity in systems biology in recent years [3]. Many different uses of Gaussian processes are known, and here we use them as a regression tool [4–6]. This gives us a continuous trajectory through the noisy data points, which we use as a better alternative to the linear interpolation between

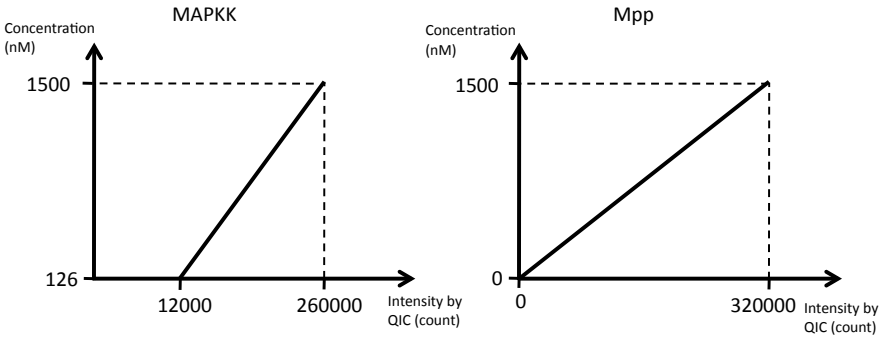

Figure 1. Linear functions that determine the relationship between the fluorescent intensity and concentration of MAPKK and Mpp data. Through experiments and information from the literature we determine the maximum fluorescent intensity and maximum concentration, and points of crossing the x-axis. We then fit a linear function $y = kx + n$ through these points. The resulting linear functions (1) determine the relationship between fluorescent intensity and concentration.

the experimental time points.

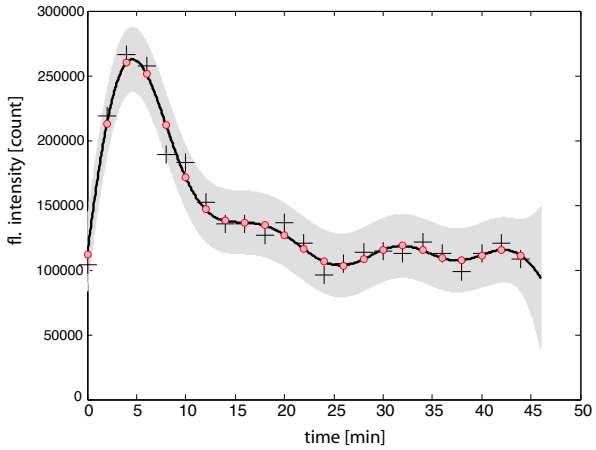

Figure 2. Gaussian process regression was used to smooth the MAPKK time series. The mean prediction (the black trajectory) is used as the smoothed MAPKK input to the models. Obtained by GPML Matlab code [5].

Moreover, we need to determine the total concentration of phosphatase MKP_{tot}, which we assume is constant. As we have found no references to it in the literature [2] we will (i) vary this number and do several model selection runs, (ii) perform the model selection with the phosphatase concentration treated as a parameter. We also need to determine the total MAPK concentration M_{tot}. We use the observation that the initial Mpp concentration before stimulation with growth factors equals about 5% of the total MAPK concentration.

2 Inputs to ABC SMC algorithm

Tolerance schedule: $\epsilon = \{5000, 3000, 1500, 1200, 1000, 900, 700, 600, 500\}$. Perturbation kernels: $KM_t(m|m^*) = 0.7$ if $m = m^*$ and 0.1 otherwise; $KP_t(\theta|\theta^*) = U(-\sigma, \sigma)$, $\sigma = 0.5(\max\{\theta\}_{t-1} - \min\{\theta\}_{t-1})$. Number of particles $N = 2000$. Distance function: square root of the sum of squared errors.

Prior distributions on the models: $P(M) = 0.25$, for $M \in \{DD, DP, PD, PP\}$.

Prior distributions on the parameters: $k1 \sim U(0, 0.1)$, $k_{-1} \sim U(0, 1)$, $k_2 \sim U(0, 10)$, $k_3 \sim U(0, 0.1)$, $k_{-3} \sim U(0, 1)$, $k_4 \sim U(0, 100)$, $h1 \sim U(0, 0.1)$, $h_{-1} \sim U(0, 1)$, $h_2 \sim U(0, 10)$, $h_3 \sim U(0, 0.1)$, $h_{-3} \sim U(0, 1)$, $h_4 \sim U(0, 100)$.

Initial conditions: $x_1 = M_{tot} - Mpp_0$, $x_3 = 0$, $x_4 = 0$, $x_5 = 0$, $x_6 = Mpp_0$, $x_8 = 0$, $x_9 = 0$, $Mpp_0 = 188$, x_2 is provided by the time series data, x_7 and M_{tot} as specified in the main text

3 Supplementary tables

$MKP_{tot} \backslash M_{tot}$	500	1000	2500	5000	8000	10000
200	3000	1200	500	500	500	500
500	3000	1200	400	500	500	500
1000	3000	1500	500	600	600	600
2000	3000	1500	500	500	500	500
3000	3000	1500	500	500	500	500
4000	3000	1500	500	500	500	500

Table 1. Distances reached in the ABC SMC inference runs for different combinations of constants MKP_{tot} and M_{tot}. Low numbers indicate that good fits to the data are obtained for high values of M_{tot}, while models do not fit to data as well for low values of M_{tot}.

References

1. Ozaki YI, Uda S, Saito TH, Chung J, Kubota H, et al. (2010) A quantitative image cytometry technique for time series or population analyses of signaling networks. PLoS ONE 5: e9955.

2. Yamada S, Taketomi T, Yoshimura A (2004) Model analysis of difference between EGF pathway and FGF pathway. Biochemical and Biophysical Research Communications 314: 1113–1120.

3. Gao P, Honkela A, Rattray M, Lawrence ND (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. Bioinformatics 24: i70–5.

4. MacKay DJC (2003) Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

5. Rasmussen C, Williams C (2006) Gaussian processes for machine learning. The MIT Press .

6. Kirk P, Stumpf M (2009) Gaussian process regression bootstrapping: Exploring the effects of uncertainty in time course data. Bioinformatics .