Detecting and modeling time shifts in microarray time series data applying Gaussian processes

Max Zwießele

Eberhard Karls Universität Tübingen

September 13, 2010

MAX-PLANCK-GESELLSCHAFT

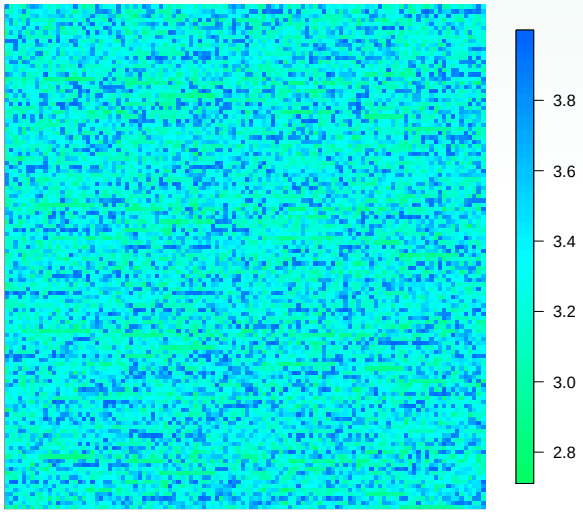
BIOLOGISCHE KYBERNETIK

Why Time Shifts in Microarray Time Series?

Why Microarrays?

Microarray Advantages:

- Huge capacity (thousands of gene expression levels at a time).
- Multiple experiments over time (and replicates):
  - → Time Series of experiment.
  - → Time Series with several replicates.
- Different conditions:
  - → Detect differentially expressed genes.



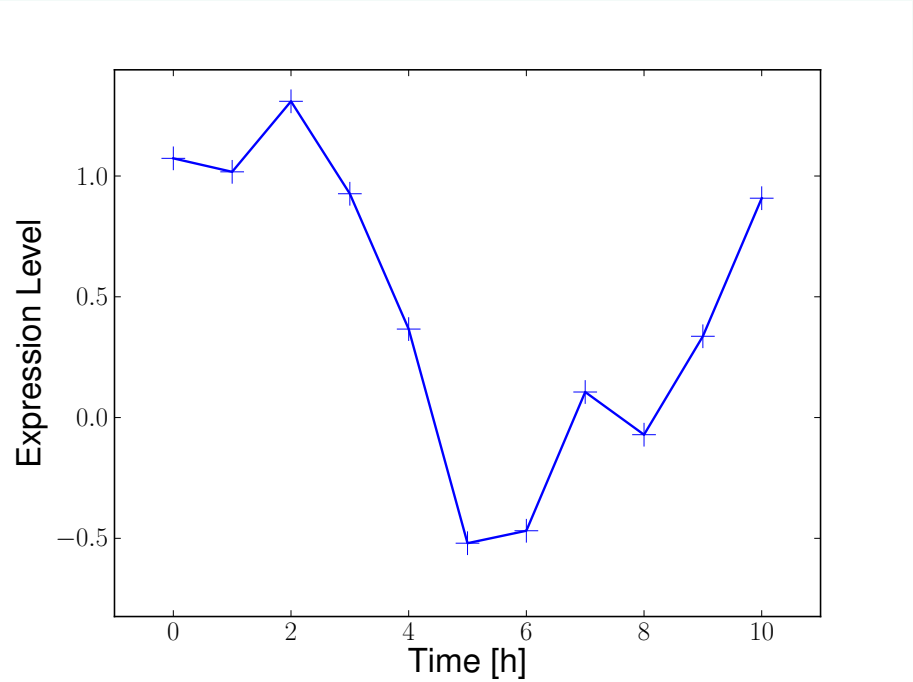
BIOLOGISCHE KYBERNETIK

Why Time Shifts in Microarray Time Series?

Why Microarrays?

Microarray Advantages:

- Huge capacity (thousands of gene expression levels at a time).
- Multiple experiments over time (and replicates):
  - → Time Series of experiment.
  - → Time Series with several replicates.
- Different conditions:
  - → Detect differentially expressed genes.



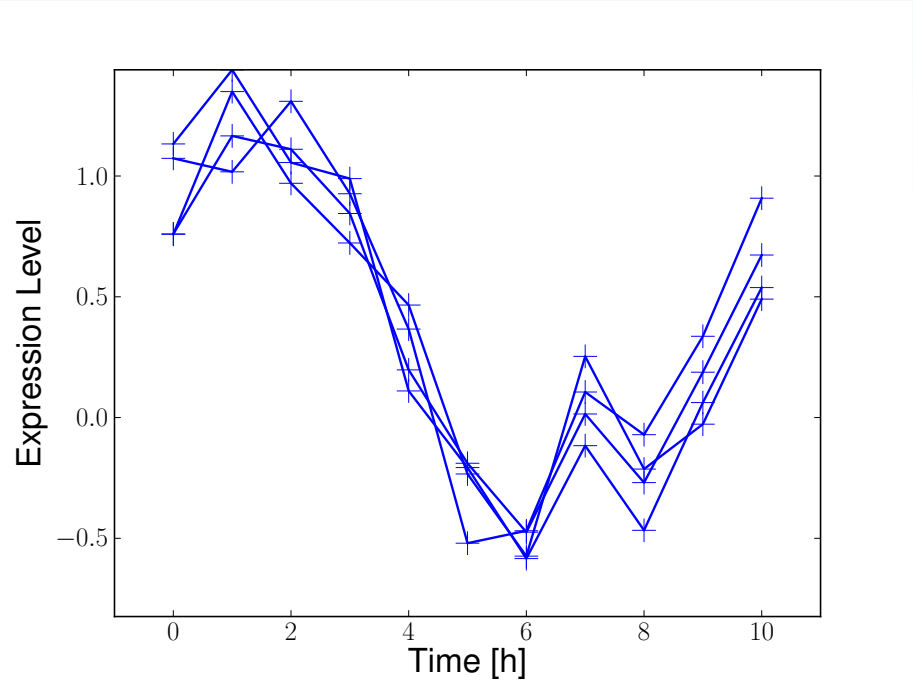
BIOLOGISCHE KYBERNETIK

Why Time Shifts in Microarray Time Series?

Why Microarrays?

Microarray Advantages:

- Huge capacity (thousands of gene expression levels at a time).
- Multiple experiments over time (and replicates):
  - → Time Series of experiment.
  - → Time Series with several replicates.
- Different conditions:
  - → Detect differentially expressed genes.



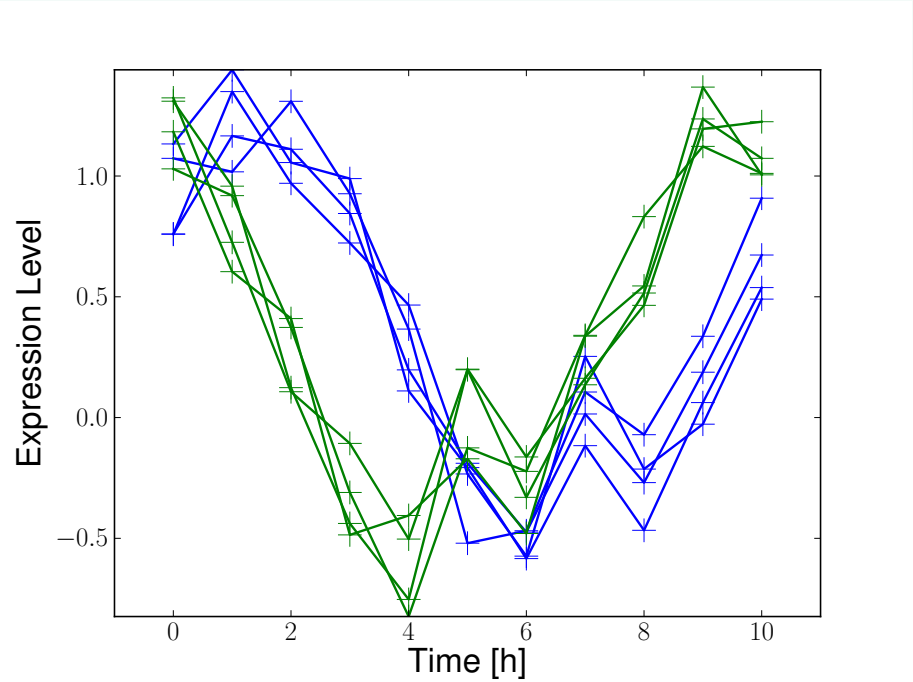
BIOLOGISCHE KYBERNETIK

Why Time Shifts in Microarray Time Series?

Why Microarrays?

Microarray Advantages:

- Huge capacity (thousands of gene expression levels at a time).
- Multiple experiments over time (and replicates):
  - → Time Series of experiment.
  - → Time Series with several replicates.
- Different conditions:
  - → Detect differentially expressed genes.



BIOLOGISCHE KYBERNETIK

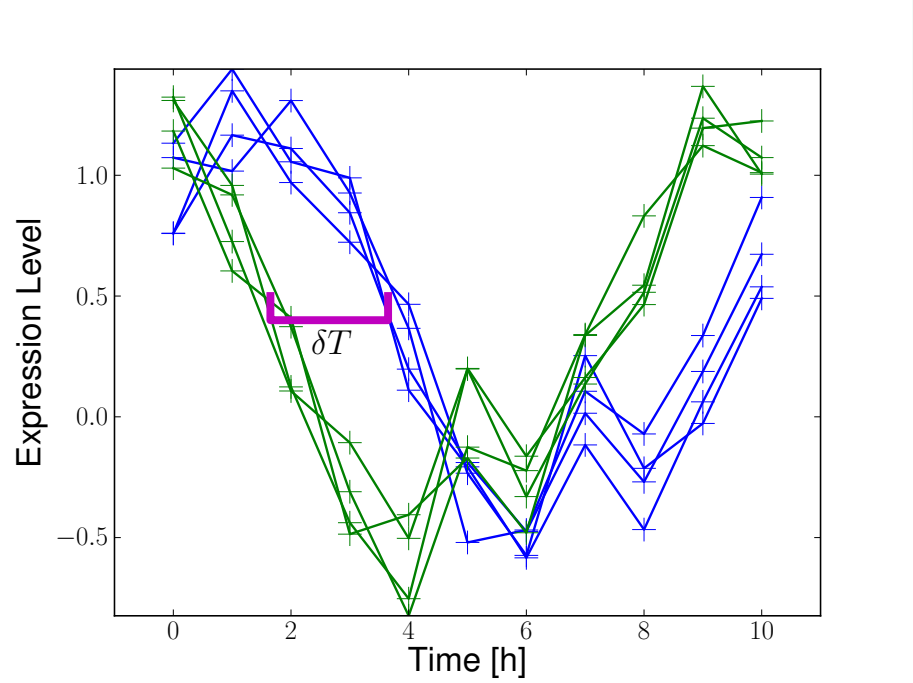
Why Time Shifts in Microarray Time Series?

Why Time Shifts?

1. *Technical Problems:* [3]
   - Varying efficiencies of intermediary steps:
   - → Imaging scanner (Laser).
   - → Amount of oligos per cell.
2. *Biological Problems:*
   - Dependency to environmental conditions.
   - Synchronization of replicates is hard. (Cell-Cycle)

Assumption:

Time shifts in data



BIOLOGISCHE KYBERNETIK

Outline

BIOLOGISCHE KYBERNETIK

Outline

BIOLOGISCHE KYBERNETIK

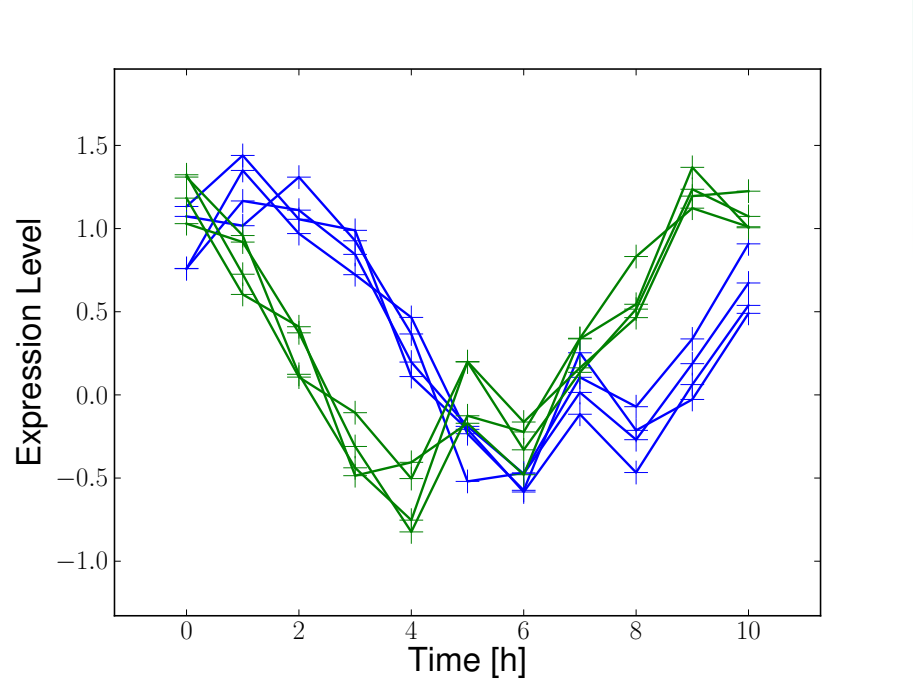
Dealing with Time Shifts in Data

Stegle *et al*'s Approach

- Used Gaussian Processes (GPs) to search functions

$$f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{y} \quad . \qquad (1)$$

- For differential expression detection, they compared:
  - individual model against
  - shared model.



- Providing the Bayes factor:

$$BF = \log \frac{p(A|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(B|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(\hat{\theta}_{\mathcal{I}})}{p(A, B|\mathcal{H}, \hat{\theta}_{\mathcal{S}})p(\hat{\theta}_{\mathcal{S}})} \qquad (2)$$

BIOLOGISCHE KYBERNETIK

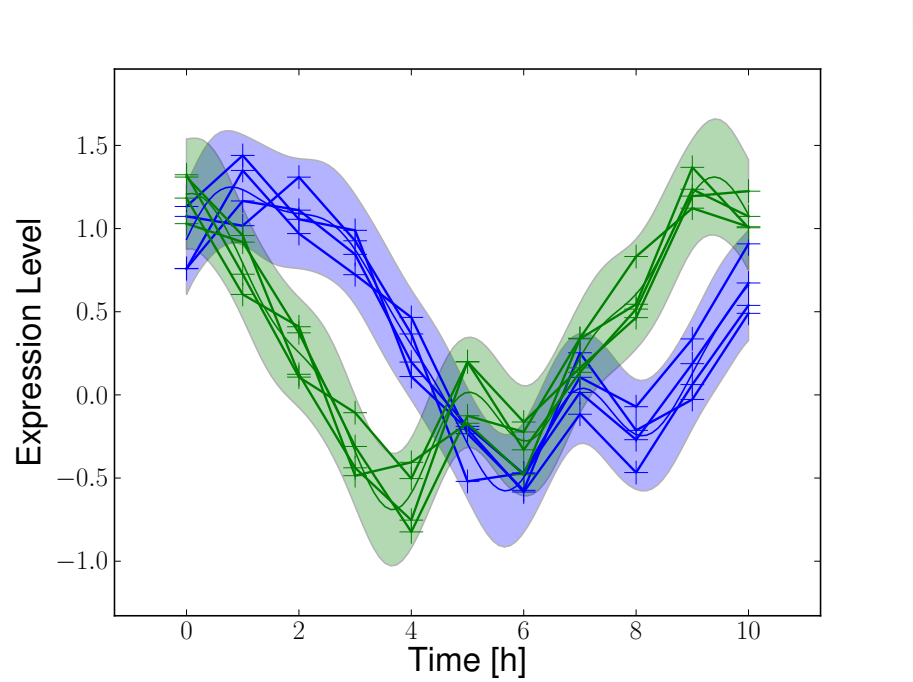
Dealing with Time Shifts in Data

Stegle *et al*'s Approach

- Used Gaussian Processes (GPs) to search functions

$$f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{y} \quad . \qquad (1)$$

- For differential expression detection, they compared:
  - individual model against
  - shared model.



- Providing the Bayes factor:

$$BF = \log \frac{p(A|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(B|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(\hat{\theta}_{\mathcal{I}})}{p(A, B|\mathcal{H}, \hat{\theta}_{\mathcal{S}})p(\hat{\theta}_{\mathcal{S}})} \qquad (2)$$

BIOLOGISCHE KYBERNETIK

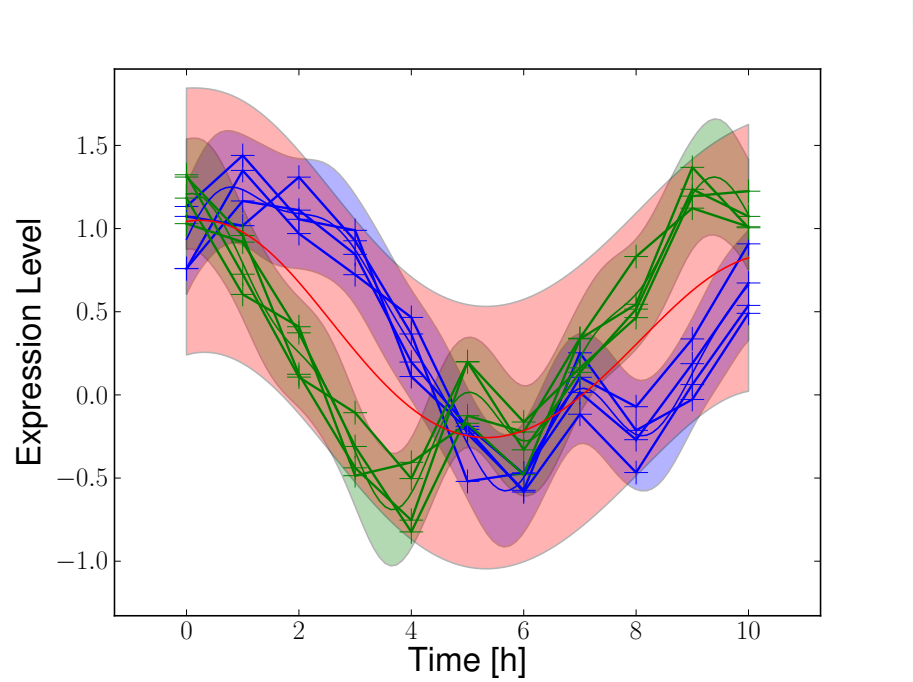
Dealing with Time Shifts in Data

Stegle *et al*'s Approach

- Used Gaussian Processes (GPs) to search functions

$$f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{y} \quad . \qquad (1)$$

- For differential expression detection, they compared:
  - individual model against
  - shared model.



- Providing the Bayes factor:

$$BF = \log \frac{p(A|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(B|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(\hat{\theta}_{\mathcal{I}})}{p(A, B|\mathcal{H}, \hat{\theta}_{\mathcal{S}})p(\hat{\theta}_{\mathcal{S}})} \qquad (2)$$

BIOLOGISCHE KYBERNETIK

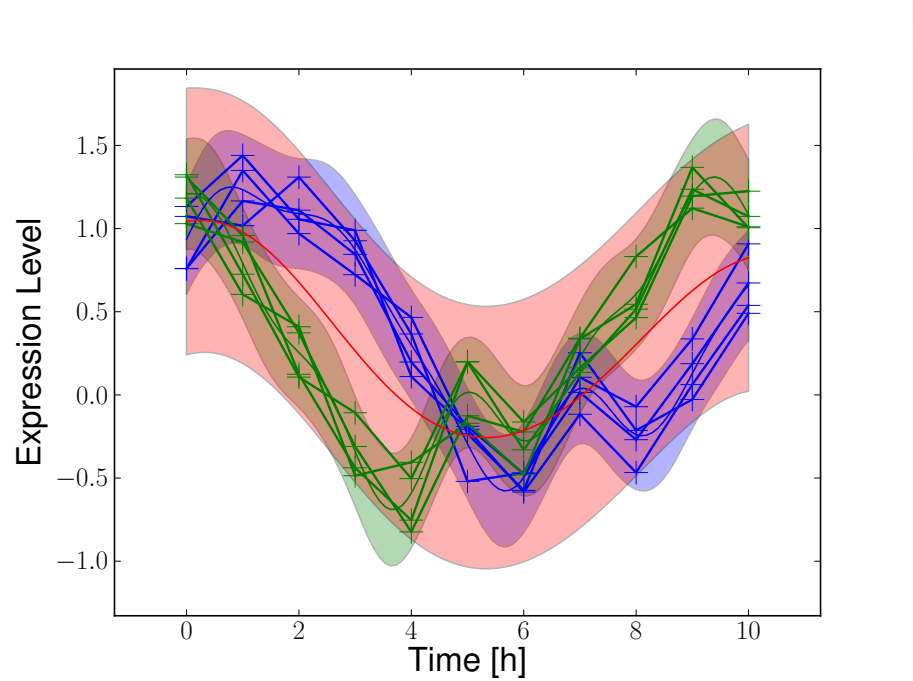
Dealing with Time Shifts in Data

Stegle *et al*'s Approach

- Used Gaussian Processes (GPs) to search functions

$$f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{y} \quad . \qquad (1)$$

- For differential expression detection, they compared:
  - individual model against
  - shared model.



- Providing the Bayes factor:

$$BF = \log \frac{p(A|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(B|\mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}})p(\hat{\theta}_{\mathcal{I}})}{p(A, B|\mathcal{H}, \hat{\theta}_{\mathcal{S}})p(\hat{\theta}_{\mathcal{S}})} \qquad (2)$$

BIOLOGISCHE KYBERNETIK

Dealing with Time Shifts in Data

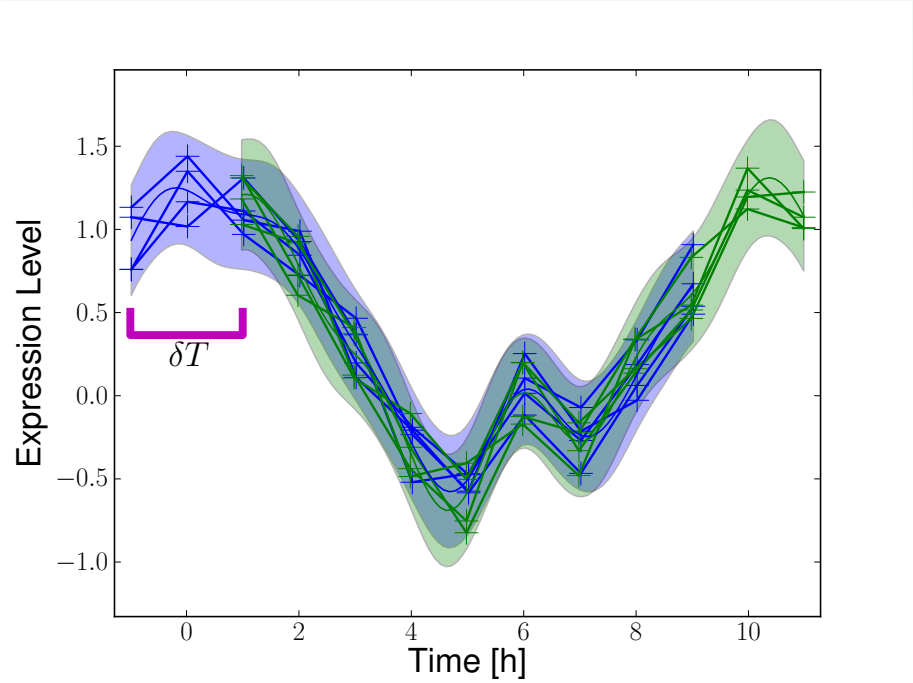
Regression w.r.t. Time Shifts

- We allow functions f to shift time line

$$f : \mathbf{x} \mapsto f(\mathbf{x} - \delta T) = \mathbf{y} \quad , \quad (3)$$

$\rightarrow$ Matching time series .

- Now compare
  - individual model against
  - shared model.



BIOLOGISCHE KYBERNETIK

Dealing with Time Shifts in Data

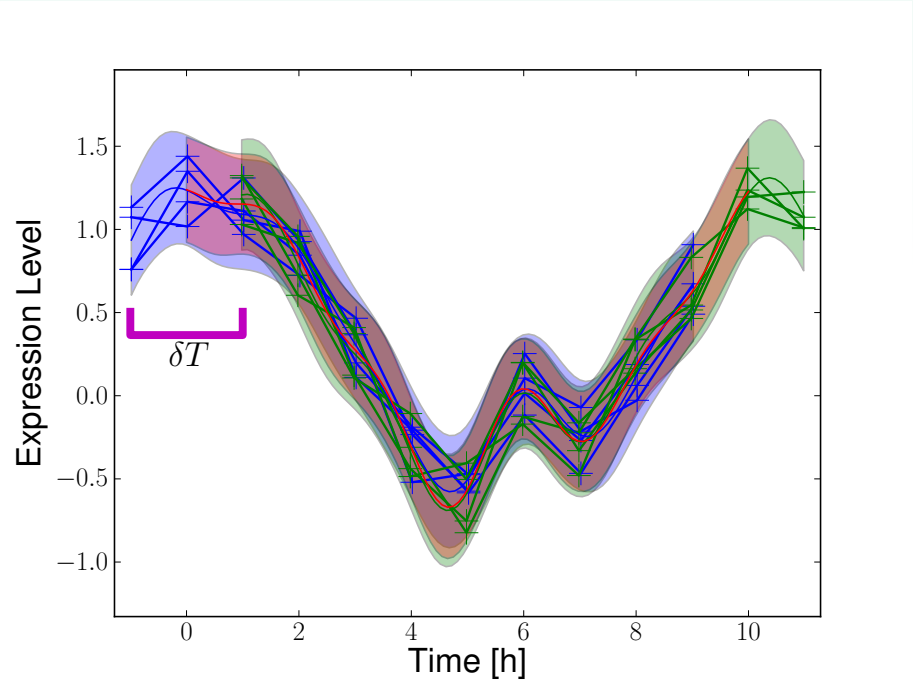
Regression w.r.t. Time Shifts

- We allow functions f to shift time line

$$f : \mathbf{x} \mapsto f(\mathbf{x} - \delta T) = \mathbf{y} \quad , \quad (3)$$

$\rightarrow$ Matching time series .

- Now compare
  - ⊟, ⊟ individual model against
  - ⊟ shared model.



BIOLOGISCHE KYBERNETIK

Gaussian Process Regression w.r.t. Time Shifts

Probability for unseen data

Two main results of GP regression [1] of interest:

1. Probability for previous unseen outputs $\mathbf{y}^*$:

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{D}, \theta) = \mathcal{N}\big(\mathbf{y}^*|\underline{\mathbf{y}}^*, \mathrm{cov}(\mathbf{y}^*)\big), \quad \text{where} \tag{4}$$

$$\underline{\mathbf{y}}^* = K_{\mathrm{x}^*,\mathrm{x}}[K_{\mathrm{x}} + \sigma^2\mathbf{E}]^{-1}\mathbf{y} \ , \tag{5}$$

$$\mathrm{cov}(\mathbf{y}^*) = K_{\mathrm{x}^*} - K_{\mathrm{x}^*,\mathrm{x}}[K_{\mathrm{x}} + \sigma^2\mathbf{E}]^{-1}K_{\mathrm{x},\mathrm{x}^*} \ , \tag{6}$$

and $\mathbf{E}$ denotes the identity matrix.

2. Probability for hyperparameters to describe the training data:

$$\log p(\mathbf{y}|\mathbf{x}, \theta) = -\frac{1}{2}\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \ , \tag{7}$$

$$\mathbf{K} = K_{\mathrm{x}} + \sigma^2\mathbf{E} \ . \tag{8}$$

Gaussian Process Regression w.r.t. Time Shifts

Novel Covariance Function with Time Shift Parameter

Requirements:

- Covariance matrix $K_{\mathrm{x},\mathrm{x}^*} = [k(x_r, x'_{r'}) : \forall x_r, x'_{r'} \in \mathbf{x}, \mathbf{x}^*]$.
- Time shift $T_r \in \mathbf{T}$ per replicate $(\mathbf{x}, \mathbf{y})_r$.

New covariance function with time shift per replicate:

$$k(x_r, x'_{r'}) := A^2 \cdot \exp\left(-\frac{d^2}{2L^2}\right) + \sigma \quad , \tag{9}$$

$$d = \left\|(x_r - T_r) - (x'_{r'} - T_{r'})\right\| \quad . \tag{10}$$

Important: The most probable set of hyperparameters $\hat{\theta} = (A, L, \mathbf{T}, \sigma)$ are learned from data! (Appendix 1)

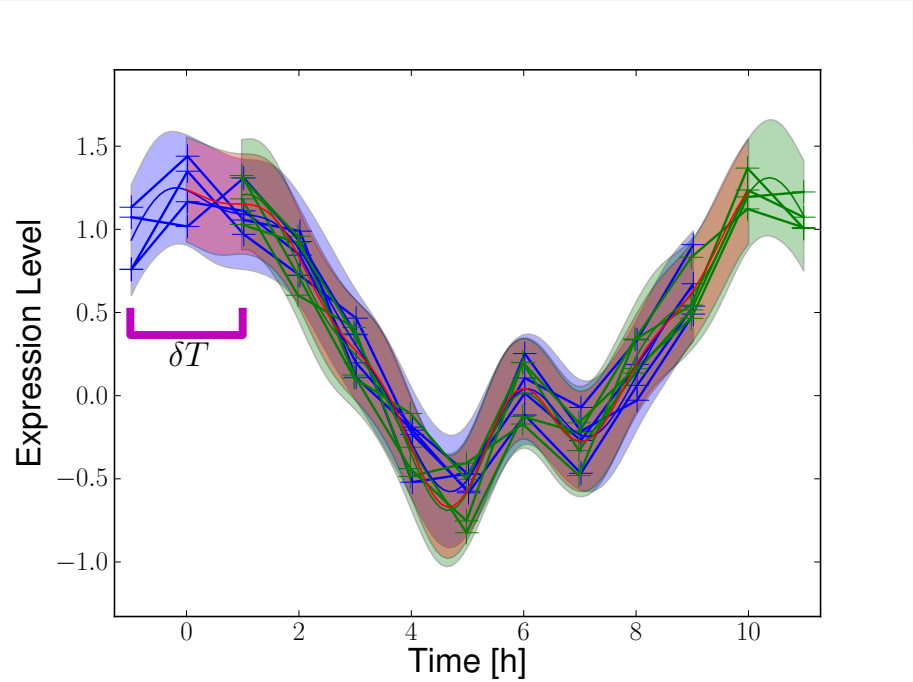
BIOLOGISCHE KYBERNETIK

Gaussian Processes Regression w.r.t. Time Shifts

Bayes Factors [2]

Three Bayes factors for hypothesis testing:

1. Individual vs. shared regression, included $\mathbf{T}$:
2. Individual vs. shared regression, where for shared $\mathbf{T} = (0, \ldots, 0)$:
3. Shared with, vs. shared without time shift:



$$\mathcal{S}^{\mathcal{I}}_{\delta T} := \log \frac{p(\mathbf{C}_k|\mathcal{H}_{GP}, \hat{\theta}^{\mathbf{T}}_{\mathcal{I}}) \cdot p(\mathbf{C}_{k'}|\mathcal{H}_{GP}, \hat{\theta}^{\mathbf{T}}_{\mathcal{I}}) \cdot p(\hat{\theta}^{\mathbf{T}}_{\mathcal{I}})}{p(\mathbf{C}_k \cup \mathbf{C}_{k'}|\mathcal{H}_{GP}, \hat{\theta}^{\mathbf{T}}_{\mathcal{S}}) \cdot p(\hat{\theta}^{\mathbf{T}}_{\mathcal{S}})} = 2.225 \quad (11)$$

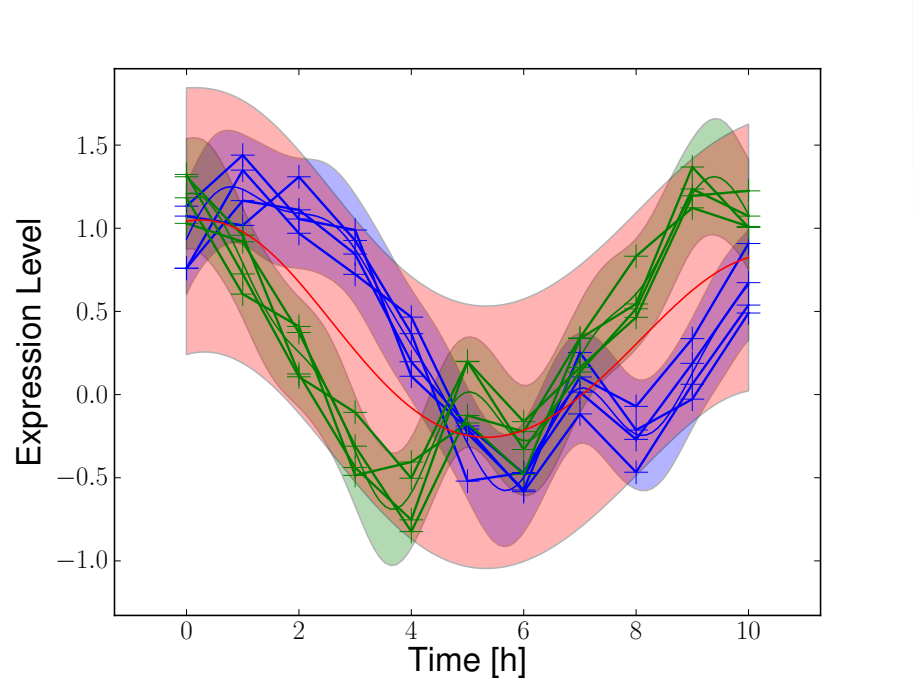
BIOLOGISCHE KYBERNETIK

Gaussian Processes Regression w.r.t. Time Shifts

Bayes Factors [2]

Three Bayes factors for hypothesis testing:

1. Individual vs. shared regression, included $\mathbf{T}$:
2. Individual vs. shared regression, where for shared $\mathbf{T} = (0, \ldots, 0)$:
3. Shared with, vs. shared without time shift:



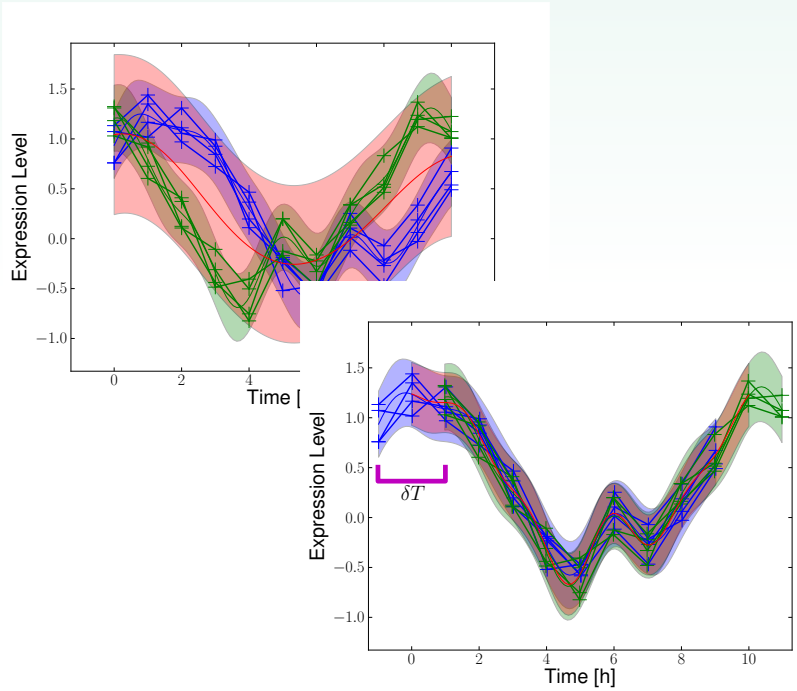
$$\mathcal{S}_0^{\mathcal{I}} := \log \frac{p(\mathbf{C}_k | \mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}}^{\mathsf{T}}) \cdot p(\mathbf{C}_{k'} | \mathcal{H}_{GP}, \hat{\theta}_{\mathcal{I}}^{\mathsf{T}}) \cdot p(\hat{\theta}_{\mathcal{I}}^{\mathsf{T}})}{p(\mathbf{C}_k \cup \mathbf{C}_{k'} | \mathcal{H}_{GP}, \hat{\theta}_{\mathcal{S}}^{(0,\ldots,0)}) \cdot p(\hat{\theta}_{\mathcal{S}}^{(0,\ldots,0)})} = 82.780 \quad (12)$$

Gaussian Processes Regression w.r.t. Time Shifts

Bayes Factors [2]

Three Bayes factors for hypothesis testing:

1. Individual vs. shared regression, included $\mathbf{T}$:
2. Individual vs. shared regression, where for shared $\mathbf{T} = (0, \ldots, 0)$:
3. Shared with, vs. shared without time shift:



$$\mathcal{S}^{\mathcal{S}} := \log \frac{p(\mathbf{C}_{\mathrm{k}} \cup \mathbf{C}_{\mathrm{k}'} | \mathcal{H}_{GP}, \hat{\theta}_{\mathcal{S}}^{\mathrm{T}}) \cdot p(\hat{\theta}_{\mathcal{S}}^{\mathrm{T}})}{p(\mathbf{C}_{\mathrm{k}} \cup \mathbf{C}_{\mathrm{k}'} | \mathcal{H}_{GP}, \hat{\theta}_{\mathcal{S}}^{(0,\ldots,0)}) \cdot p(\hat{\theta}_{\mathcal{S}}^{(0,\ldots,0)})} = 80.556 \quad (13)$$

BIOLOGISCHE KYBERNETIK

Outline

BIOLOGISCHE KYBERNETIK

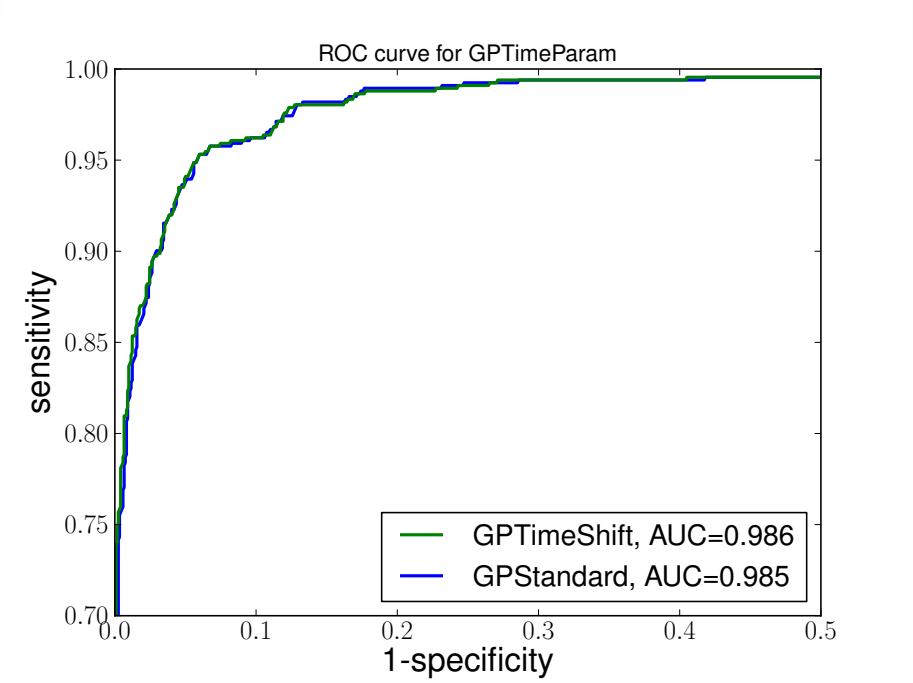
Detecting Differentially Expressed Genes

Experiment

- GPTimeShift applied on *Arabidopsis thaliana* data set
  - Plant leafs inoculated by fungal pathogen *Botrytis cinerea*.
  - Harvested every 2h up to 48h post-inoculation.
  - Contains four replicates for each condition of $n = 30,336$ genes.
- Goal: Detect differentially expressed genes.

Detecting Differentially Expressed Genes

Result



Test set of $n = 1,890$ genes, manually classified into either 'differentially expressed' or '**not** differentially expressed'.

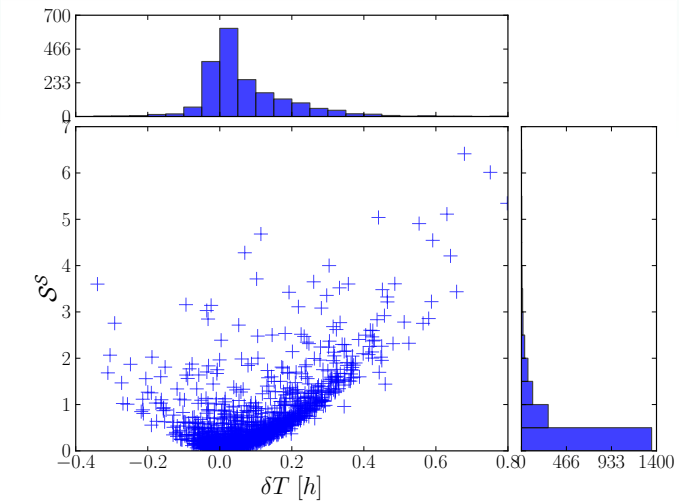
BIOLOGISCHE KYBERNETIK

Detecting Differentially Expressed Genes

Conclusion

No significant improvement by including time shifts:

- Microarray problems solved by preprocessing and normalization.
- No significant time shifts in data.



Correlation of Transcription Factors and Time Shifts

Experiment

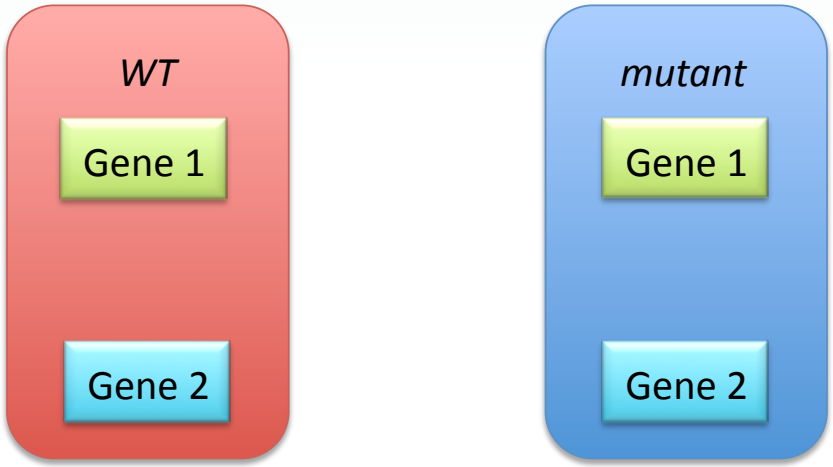
- GPTimeShift applied on *yeast* data set:
  - Knocked out *mei4* (coding for a TF).
  - Harvested every 1h up to 8h post-perturbation.
  - Only one replicate per condition of $n = 4,410$ genes.
- Goal: Detect TFs and corresponding targets by time shift correlations.

BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

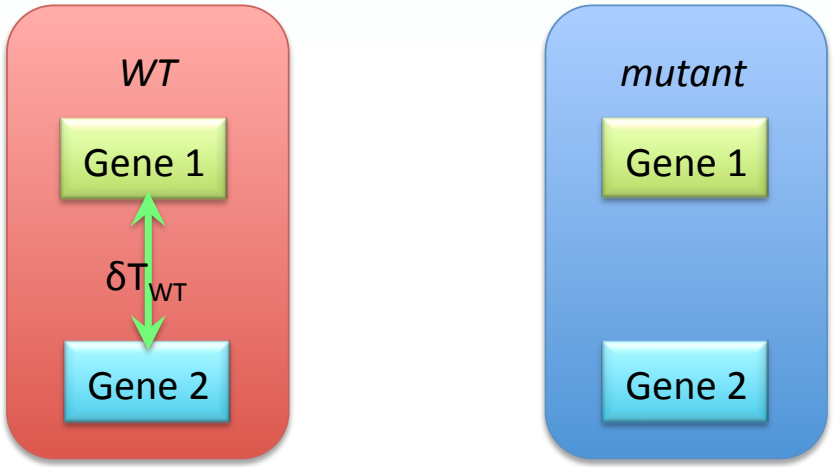


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

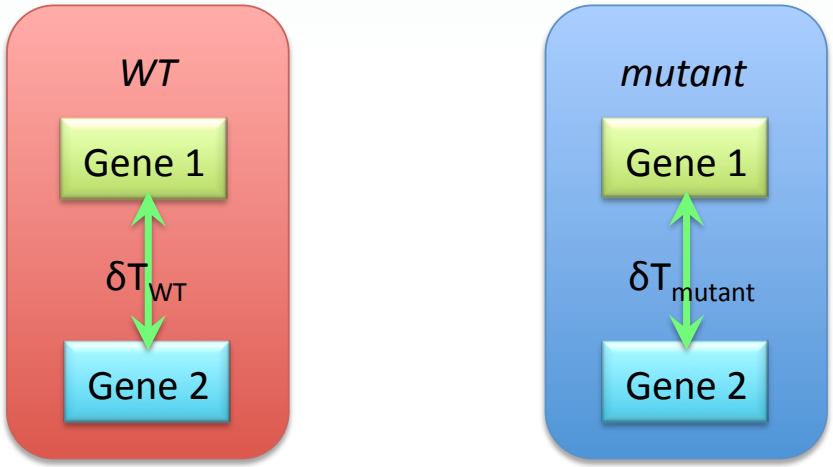


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

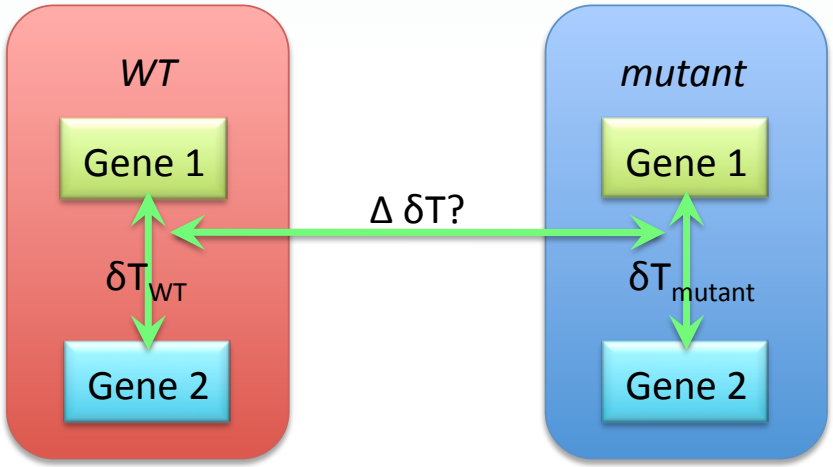


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

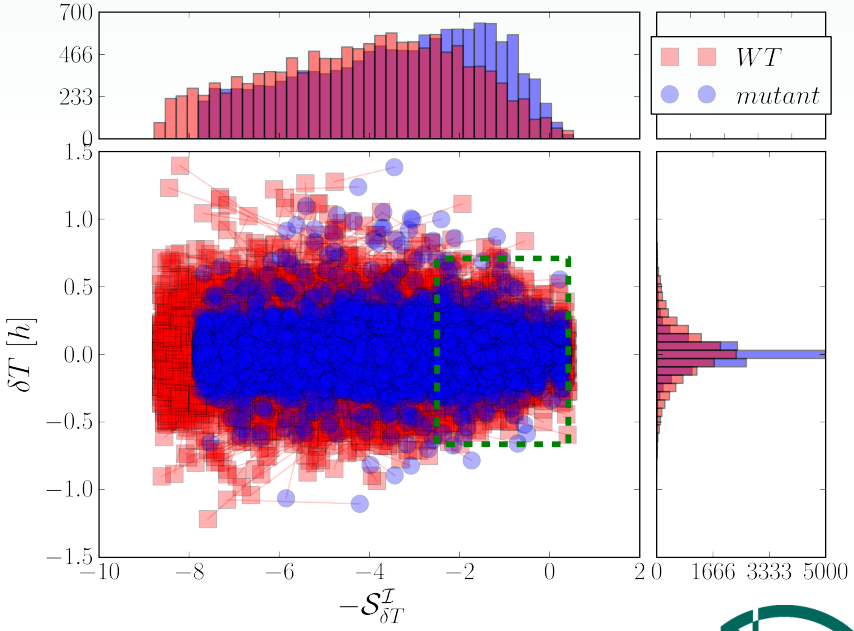


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

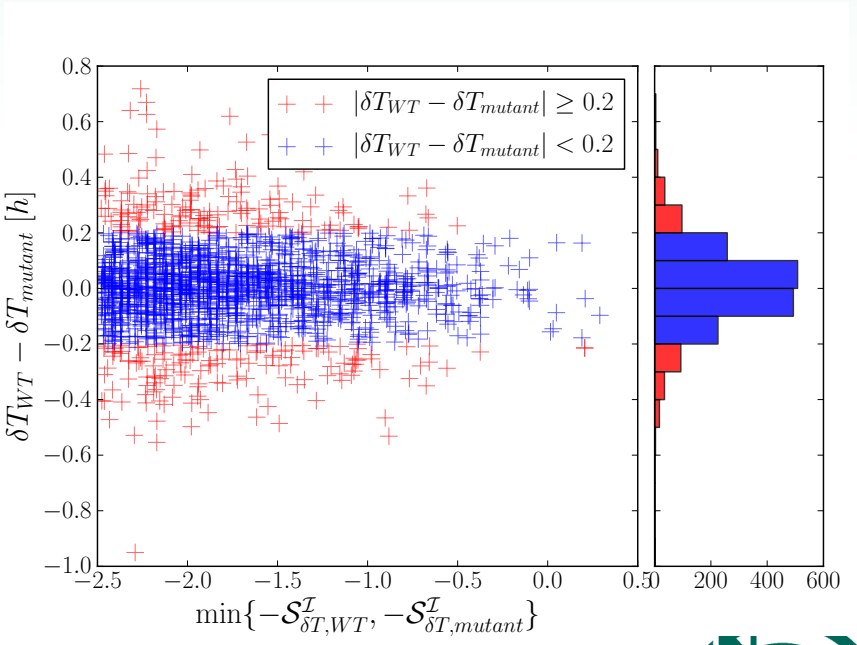


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Results

- Compare time shifts within conditions (*WT* and *mutant*) of gene pairs against each other.
- Get correlated gene pairs ($-\mathcal{S}^{\mathcal{I}}_{\delta T} > -2.5$), due to comparing time series of different genes.
- Select only tails of time shift difference distribution.

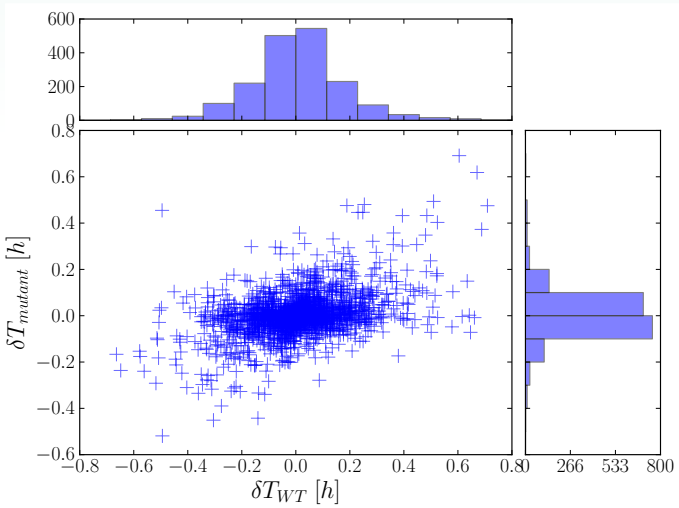


BIOLOGISCHE KYBERNETIK

Correlation of Transcription Factors and Time Shifts

Conclusion

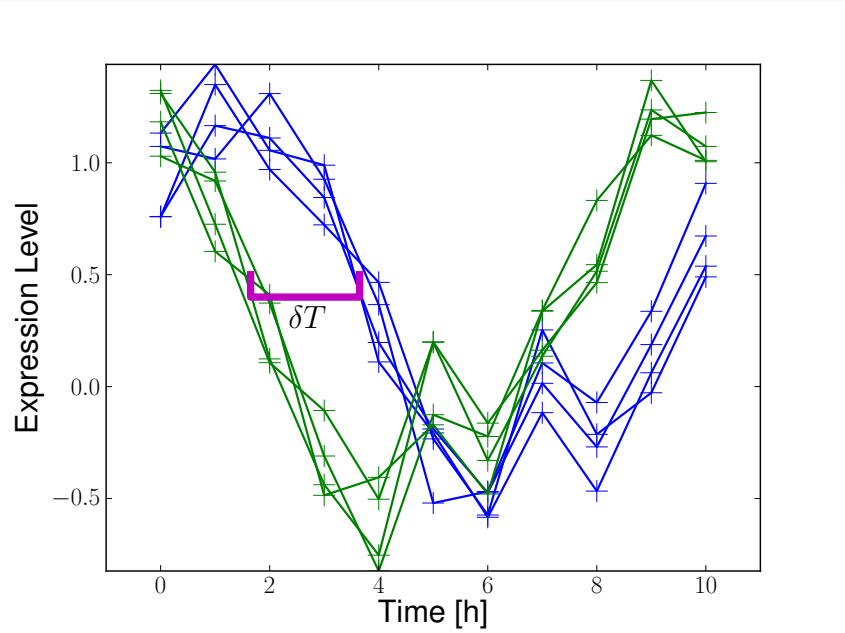
- Unfortunately only a subset of predicted pairs are TFs:
  - Algorithm for searching TFs and targets has to be adapted.
- Proof of principle:
  - TFs lead to time shifts in data.
  - Knocking out a TF leads to less time shifts.



BIOLOGISCHE KYBERNETIK

Summary

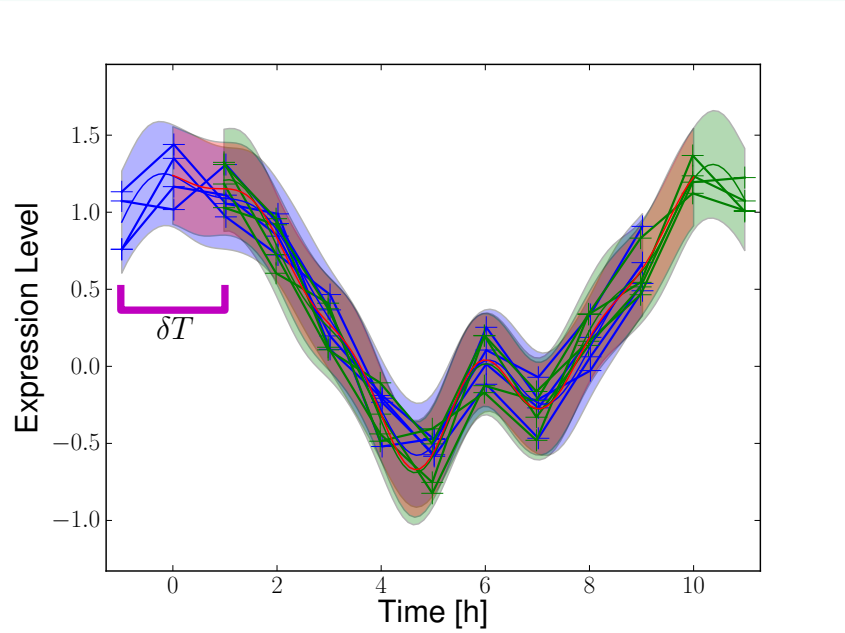
- Technical and Biological problems in microarray analyzes lead to time shifts.
- We used Gaussian Processes to model these time shifts.
- Modeling time shifts allowed us to
  - improve differential gene expression detection.
  - analyze TF(-target) and time shift correlations.



BIOLOGISCHE KYBERNETIK

Summary

- Technical and Biological problems in microarray analyzes lead to time shifts.
- We used Gaussian Processes to model these time shifts.
- Modeling time shifts allowed us to
  - improve differential gene expression detection.
  - analyze TF(-target) and time shift correlations.



BIOLOGISCHE KYBERNETIK

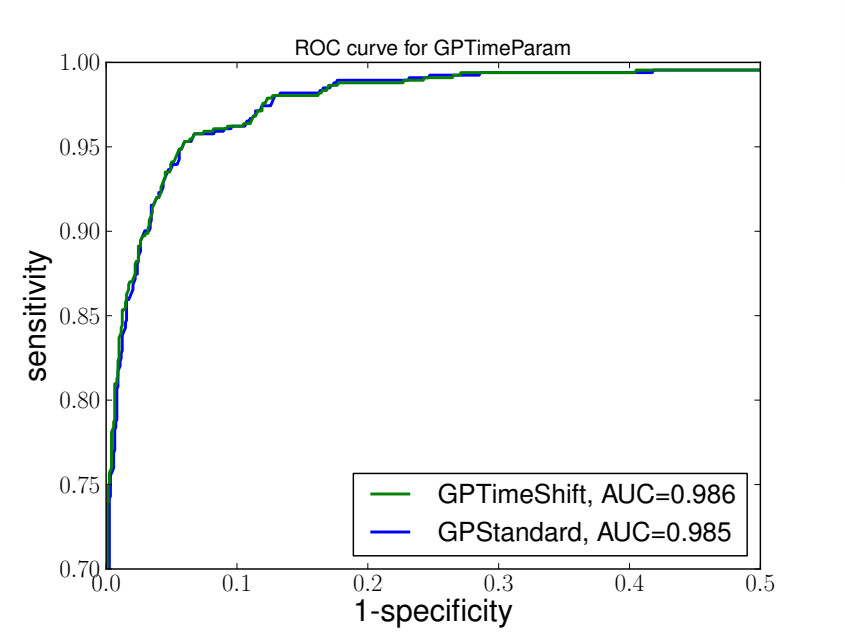
Summary

- Technical and Biological problems in microarray analyzes lead to time shifts.
- We used Gaussian Processes to model these time shifts.
- Modeling time shifts allowed us to
  - improve differential gene expression detection.
  - analyze TF(-target) and time shift correlations.

BIOLOGISCHE KYBERNETIK

Summary

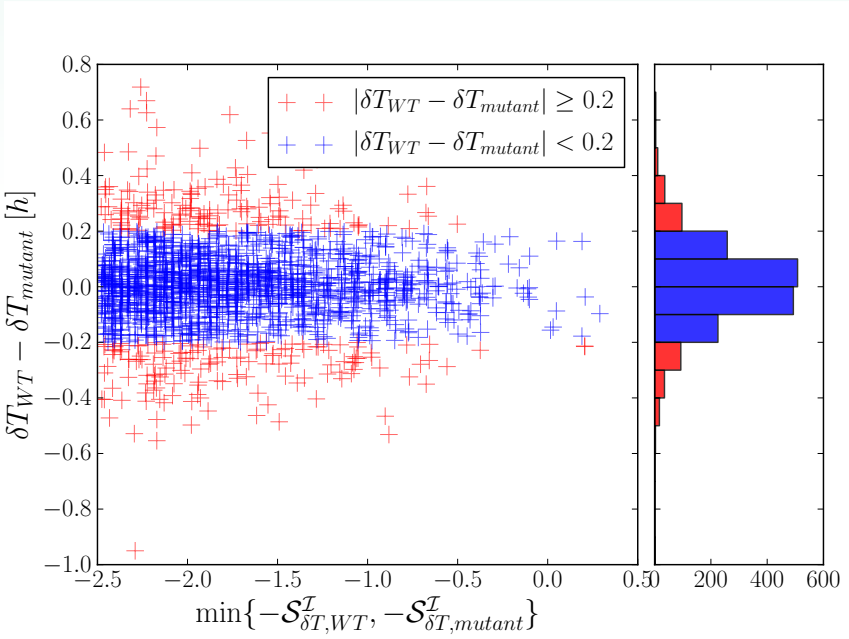
- Technical and Biological problems in microarray analyzes lead to time shifts.
- We used Gaussian Processes to model these time shifts.
- Modeling time shifts allowed us to
  - improve differential gene expression detection.
  - analyze TF(-target) and time shift correlations.



BIOLOGISCHE KYBERNETIK

Summary

- Technical and Biological problems in microarray analyzes lead to time shifts.
- We used Gaussian Processes to model these time shifts.
- Modeling time shifts allowed us to
  - improve differential gene expression detection.
  - analyze TF(-target) and time shift correlations.



BIOLOGISCHE KYBERNETIK

Discussion

References

Carl E. Rasmussen and Christopher K. I. Williams.
Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).
The MIT Press, December 2005.

O. Stegle, K. Denby, D.L. Wild, Z. Ghahramani, and K.M. Borgwardt.
A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series.
In *Research in Computational Molecular Biology: 13th Annual International Conference, Recomb 2009, Tucson, Arizona, USA, May 18-21, 2009, Proceedings*, page 201. Springer, 2009.

A. Zien, B. Schoelkopf, K. Tsuda, and JP Vert.
A primer on molecular biology.
Kernel methods in computational biology, page 3, 2004.

BIOLOGISCHE KYBERNETIK

Learning Hyperparameters [1]

With the probability for the outputs being described by the hyperparameters

$$\log p(\mathbf{y}|\mathbf{x},\theta) = -\frac{1}{2}\mathbf{y}^{\top}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \quad ,(7) \tag{15}$$

we can denote the most probable set of hyperparameters $\hat{\theta}$ with:

$$\hat{\theta} = \arg\max_{\theta}\{p(\theta|\mathbf{D})\} \tag{16}$$

$$= \arg\max_{\theta}\{p(\mathbf{y}|\mathbf{x},\theta)\cdot p(\theta)\} \tag{17}$$

$$= \arg\max_{\theta}\{\log\ p(\mathbf{y}|\mathbf{x},\theta) + \log\ p(\theta)\} \quad , \tag{18}$$

where $p(\theta)$ is the hyperprior probability for the hyperparameters θ.