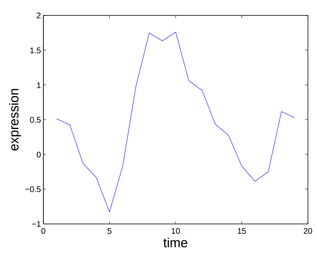
Prediction of protein-to-protein interactions

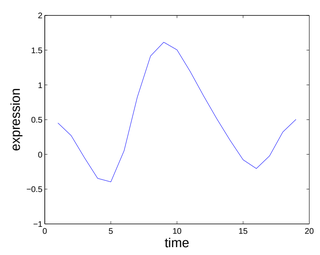
C. Gilissen[a], **P. Groot**[b], **P. Lucas**[b],
J. Veltman[a], **A. Geurts van Kessel**[a], **M. Egmont-Petersen**[a]

a) Department of Human Genetics - UMC St.Radboud Nijmegen
b) Faculty of computing sciences - Radboud University, Nijmegen

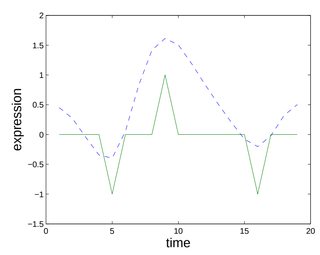
c.gilissen@antrg.umcn.nl



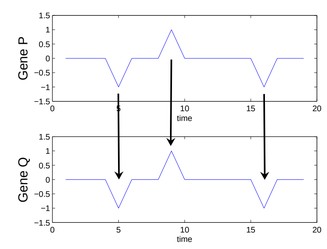
(a)



(b)



(c)



(d)

Data set 1	Data set 2
YCR042C,YHL025W	YCR040W,YIL115C
YCR042C,YHR202W	YCR042C,YKR051W
YCR042W,YIL049W	YCR042C,YML107C
YDL002C,YBR092C	YCR042C,YHR202W
YDL002C,YCR042C	YCR091W,YDR540C
YDL002C,YEL052W	***YDL006W,YER042W***
YDL006W,YER042W	YDL26W,YNL159C
YDL034W,YBR054W	YDL034W,YBR054W
YDL054C,YAL019W	YDL036C,YIL065W
YDL058W,YFL048C	YDL058W,YJL037W
YDL058W,YHR066W	YDL058W,YHR004C
YDL076C,YCR066W	YDL058W,YLR241W

(e)

YDL006W, YER042W 0.54%
YDL058W,YHR004C 0.37%
YCR042C,YHR202W 0.23%
YDL034W,YER054W 0.11%

(f)

Figure 1: *An overview of the applied method. a) original expression data; b) expression data after smoothing; c) data discretised based on local extrema; d) comparison of discrete patterns; e) removal of chance findings by comparing predictions of two data sets; f) ranking of predictions according to confidences of the local extrema.*

Abstract

We use local extrema in microarray time series data as the basis for discretisation. By comparing discretised gene expressions using similarity functions, we discover putative protein-to-protein interactions. We validate the results by use of public true positive and true negative databases for protein-to-protein interactions and demonstrate the high predictiveness of these local extrema as a time series feature.

1. Method

1.1 Smoothing

We create a smoothed version of our original signal h, which removes noise and effectively captures the properties of the signal at a higher scale (Figure 1b). We obtain such a smoother signal by convolving the original signal h with the Gaussian function:

$$h(t) * k(t) = \int_{-\infty}^{\infty} h(t-\tau)k(t)\,d\tau \qquad (1)$$

where $k(t)$ is the Gaussian function:

$$k(t;m,s) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(t-m)^2}{2s^2}\right). \qquad (2)$$

By varying s, we can vary the scale level (and thus the smoothing) of the scale-space representation.

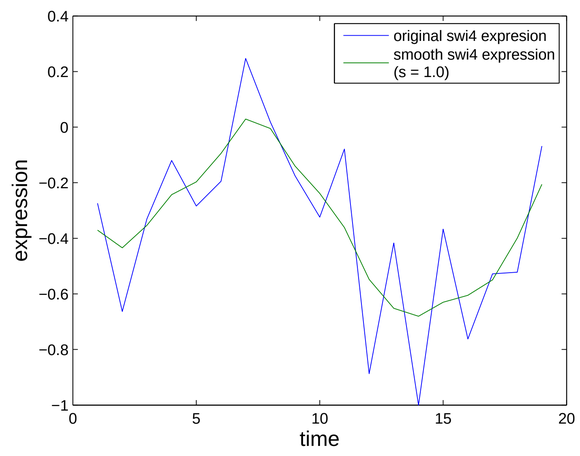


Figure 2: *Smoothing of swi4 expression, across 19 time samples. The original signal in blue; the smooth signal in green.*

1.2 Discretisation

We compute the regularized (smoothed) derivative of the gene expression time signal by convolution with the first-order derivative of the Gaussian function as:

$$\mathcal{D}(h * k)(t) = (h * \mathcal{D}k)(t) = (\mathcal{D}h * k)(t) \qquad (3)$$

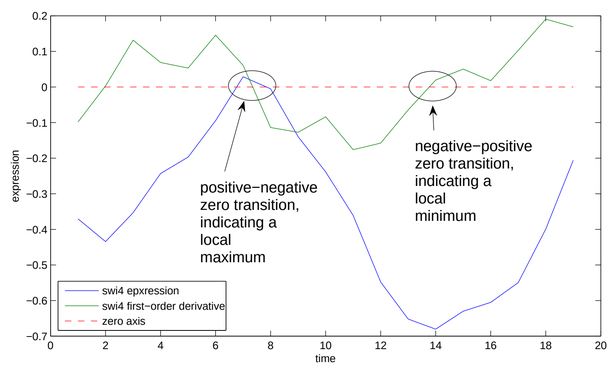


Figure 3: *Expression of swi4 (blue), and first-order derivative of swi4 (green), where zero transitions indicate local extrema.*

Discretisation is now straightforward,

- a positive-negative transition becomes a *maximum*
- a negative-positive transition becomes a *minimum*
- remaining samples are labeled *change*.

This results in a discrete time-series per gene where each sample can be either a maximum, minimum, or change (Figure 1c).

1.3 Similarity function

We distinguish four kinds of possible relations and the associated relation between the discrete time patterns, as shown in Figure 5. We find putative relations by comparing discrete gene patterns using these similarity functions (Figure 1d).

1.4 Integration of data sets

Many patterns are similar by mere chance. We remove these chance findings by combining the results of two independent data sets using a logical AND (Figure 1e).

1.5 Confidence intervals

For each detected local extrema t_l in a gene pattern we calculate the probability of a type I error, $P(t_l)$. The total confidence for a gene Q with n extrema is then calculated by the geometric mean:

$$C_Q = \left(\prod_{l=1}^{n}(1 - P(t_l))\right)^{\frac{1}{n}} \qquad (4)$$

We rank predictions according to their confidence (Figure 1f). For any relation consisting of genes Q and R the confidence measure becomes:

$$C_{QR} = C_Q \cdot C_R \qquad (5)$$

2. Results

We applied the method to two yeast datasets of Spellman et. al. and validated the resulting predictions by using a public Gold standard protein-to-protein interaction database. We compared the results to a similar approach where discretisation was based on per gene thresholding as shown in Figure 4.

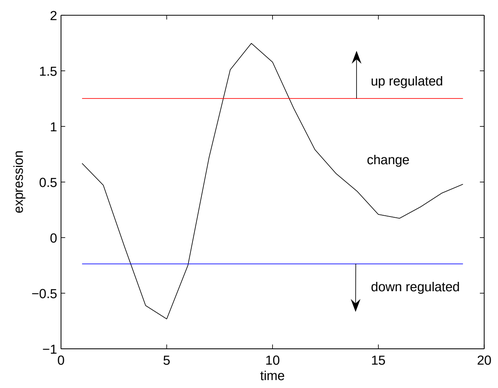


Figure 4: *Discretisation based on thresholding for RAD51. Thresholds are drawn at $c = 0.6$ of the distance between the mean and maximum / minimum.*

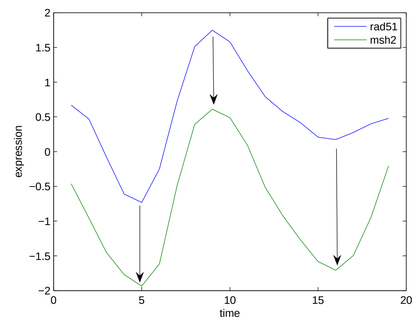
option	# pred.	% TP	% TN
no options	21	33%	14.29%
smoothing	43	11.63%	18.60%
lagfunction	146	8.90%	11.64%
no begin/end	118	13.56%	15.25%
$c = 0.99$	$1,968$	0.51%	17.47%
$c = 0.95$	$1,059$	0.76%	18.13%
$c = 0.90$	486	0.62%	16.87%
$c = 0.80$	98	0.00%	19.39%

Table 1: ***# pred.*** *is the number of predictions* ***% TP*** *is the percentage true-positives,* ***% TN*** *is the percentage true-negatives.*

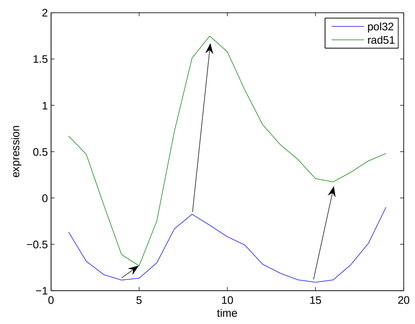
Results are shown in Table 1. Above: the results of the method with varying options: no options; increased smoothing factor; comparison using the lagfunction; ignoring extrema at the first and last samples of the series. Below: results from discretisation based on thresholding.

3. Conclusion

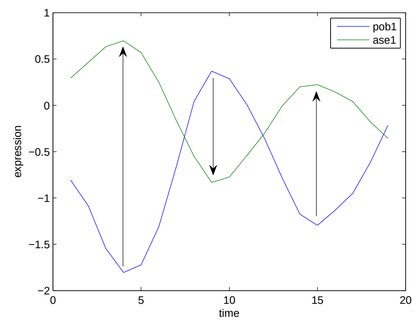
Local extrema are a feature in time-series gene expression data that can be used for finding biologically relevant interactions (e.g., protein-to-protein interactions). The applied method is invariant under scaling and shifting and can be adjusted for the amount of experimental noise.



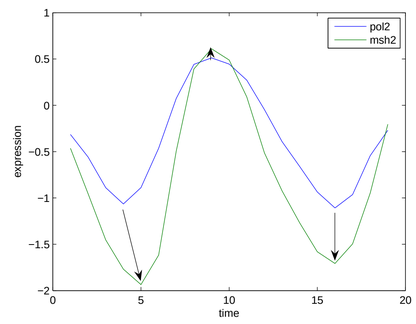
(a) protein-to-protein interactions, if $\forall_t : Q_t = R_t$ (i.e., the patterns are the same)



(b) controlled interactions, if $\forall_t : Q_t = R_{t+1}$ (i.e., the pattern of R is shifted one sample into the future with respect to the pattern of Q)



(c) inhibitory interactions, if $\forall_t : Q_t = -R_t$ (i.e., the patterns are opposite)



(d) Lagfunction, anticipating translocated extrema: $\forall_t : Q_t = R_t \vee Q_t = R_{t+l}$

Figure 5: *Discrete patterns corresponding to possible biological relations. Arrows indicate the similarity functions.*