Identification of potentially collaborating transcription factors using pointwise mutual information

Dr. Mehmet Gültas

Department of Breeding Informatics,
University of Göttingen

Outline

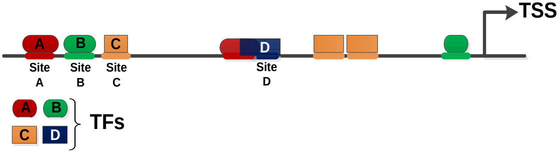
1. Cooperative TFs
2. PC-TraFF Algorithm
3. Functionality of PC-TraFF

Outline

1. Cooperative TFs
2. PC-TraFF Algorithm
3. Functionality of PC-TraFF

Motivation

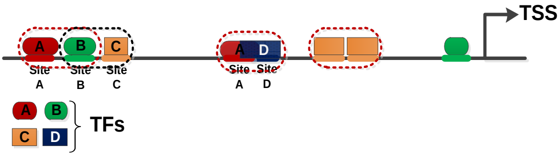
Transcription factors (TFs) ⇒ a special class of gene regulatory proteins



- TFs bind to specific DNA motifs in the genome
- Single TF motifs are not sufficient to analyze regulatory networks

Motivation

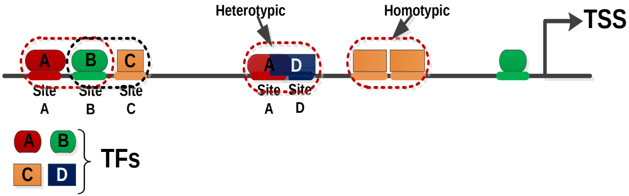
Transcription factors (TFs) ⇒ a special class of gene regulatory proteins



- TFs bind to specific DNA motifs in the genome
- Single TF motifs are not sufficient to analyze regulatory networks
- TFs bind to the promoter regions in a cooperative manner
  - Partner choose of TFs is not random
  - It depends on the evolution of the protein family

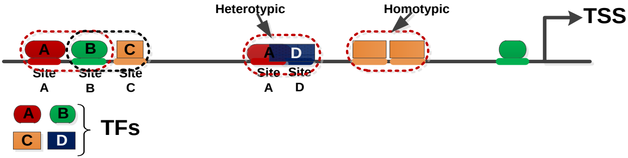
Motivation

Transcription factors (TFs) ⇒ a special class of gene regulatory proteins



- Cooperations:
  - synergistic or antagonistic interactions
  - between homotypic or heterotypic TFs

Aim



- **Aim of this study:** Identification of potential collaborating TF-pairs
- **Metric:** Pointwise Mutual Information (PMI)

$$PMI(x;y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

- $p(x,y)$ is the joint probability of x and y
- $p(x)$ and $p(y)$ are the marginal probabilities of x and y

Pointwise Mutual Information

PMI is a powerful metric for document summarization processes as well as for the detection of word collocations in linguistics.

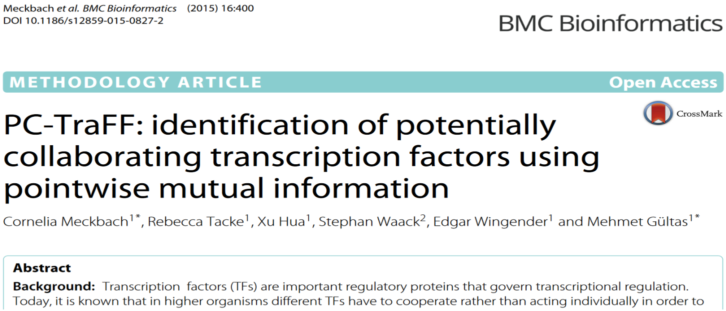
PMI from linguistics to bioinformatics

PMI in bioinformatics		**PMI in linguistics**
Genome	$\Rightarrow$	Document
Sequences	$\Rightarrow$	Sentences
TFBSs	$\Rightarrow$	Words

Outline

1. Cooperative TFs
2. PC-TraFF Algorithm
3. Functionality of PC-TraFF

PC-TraFF Algorithm



Meckbach *et al. BMC Bioinformatics* (2015) 16:400
DOI 10.1186/s12859-015-0827-2

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information

Cornelia Meckbach[1*], Rebecca Tacke[1], Xu Hua[1], Stephan Waack[2], Edgar Wingender[1] and Mehmet Gültas[1*]

Abstract

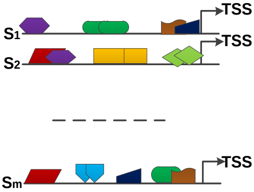
Background: Transcription factors (TFs) are important regulatory proteins that govern transcriptional regulation. Today, it is known that in higher organisms different TFs have to cooperate rather than acting individually in order to

Workflow of PC-TraFF

- **Input:**
  - Set of sequences
  - Position weight matrix (PWM) library

Workflow of PC-TraFF

- **Input:**
  - Set of sequences
  - Position weight matrix (PWM) library
- **Step1: Construction of the TFBS-sequence matrix**
  - Predict all TFBSs in the sequences by applying Match^{TM} program
  - Position weight matrix (PWM) library



$$\Longrightarrow \quad M_{m,n} = \begin{array}{c c} & \begin{array}{cccc} V\$_1 & V\$_2 & \cdots & V\$_n \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ \vdots \\ S_m \end{array} & \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \ddots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \end{array}$$

$\Rightarrow$ $x_{i,j}$ is the frequency of $TFBS_j$ in S_i

Step 2

- We identify important TFBSs in the sequences:
  - $PMI(s,t) \Leftrightarrow$ sequences and TFBSs

$$PMI(s;t) = \log_2 \frac{p(s_i,t_j)}{p(s_i)\cdot p(t_j)}$$

$$M_{m,n} = \begin{array}{c|ccccc|} & V\$_1 & V\$_2 & \cdots & \mathbf{V\$_j} & V\$_n \\ S_1 & x_{1,1} & x_{1,2} & \cdots & \mathbf{x_{1,j}} & x_{1,n} \\ S_2 & x_{2,1} & x_{2,2} & \cdots & \mathbf{x_{2,j}} & x_{2,n} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{S_i} & \mathbf{x_{i,1}} & \mathbf{x_{i,2}} & \cdots & \mathbf{x_{i,j}} & \mathbf{x_{i,n}} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ S_m & x_{m,1} & x_{m,2} & \cdots & \cdots & x_{m,n} \end{array}$$

$\rightsquigarrow$ $p(s_i)\cdot p(t_j) = p(s_i,t_j) \Rightarrow PMI(s;t) = 0$

$\rightsquigarrow$ $p(s_i)\cdot p(t_j) > p(s_i,t_j) \Rightarrow PMI(s;t) < 0$

Step 2

- We identify important TFBSs in the sequences:
  - $PMI(s,t) \Leftrightarrow$ sequences and TFBSs

$$PMI(s;t) = \log_2 \frac{p(s_i,t_j)}{p(s_i)\cdot p(t_j)}$$

$$M_{m,n} = \begin{array}{cccccc} & V\$_1 & V\$_2 & \cdots & \mathbf{V\$_j} & V\$_n \\ S_1 & x_{1,1} & x_{1,2} & \cdots & \mathbf{x_{1,j}} & x_{1,n} \\ S_2 & x_{2,1} & x_{2,2} & \cdots & \mathbf{x_{2,j}} & x_{2,n} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{S_i} & \mathbf{x_{i,1}} & \mathbf{x_{i,2}} & \cdots & \mathbf{x_{i,j}} & \mathbf{x_{i,n}} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ S_m & x_{m,1} & x_{m,2} & \cdots & \cdots & x_{m,n} \end{array}$$

⇝ $p(s_i)\cdot p(t_j) = p(s_i,t_j) \Rightarrow PMI(s;t) = 0$

⇝ $p(s_i)\cdot p(t_j) > p(s_i,t_j) \Rightarrow PMI(s;t) < 0$

✓ $\mathbf{p(s_i)\cdot p(t_j) < p(s_i,t_j) \Rightarrow PMI(s;t) > 0}$

Step 2

- We identify important TFBSs in the sequences:
    - $PMI(s,t) \Leftrightarrow$ sequences and TFBSs

$$PMI(s;t) = \log_2 \frac{p(s_i,t_j)}{p(s_i) \cdot p(t_j)}$$

$$M_{m,n} = \begin{array}{c|ccccc} & V\$_1 & V\$_2 & \cdots & \mathbf{V\$_j} & V\$_n \\ S_1 & x_{1,1} & x_{1,2} & \cdots & \mathbf{x_{1,j}} & x_{1,n} \\ S_2 & x_{2,1} & x_{2,2} & \cdots & \mathbf{x_{2,j}} & x_{2,n} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{S_i} & \mathbf{x_{i,1}} & \mathbf{x_{i,2}} & \cdots & \mathbf{x_{i,j}} & \mathbf{x_{i,n}} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ S_m & x_{m,1} & x_{m,2} & \cdots & \cdots & x_{m,n} \end{array}$$

⇝ $p(s_i) \cdot p(t_j) = p(s_i,t_j) \Rightarrow PMI(s;t) = 0$

⇝ $p(s_i) \cdot p(t_j) > p(s_i,t_j) \Rightarrow PMI(s;t) < 0$

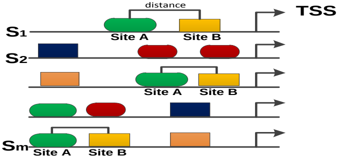
✓ $\mathbf{p(s_i) \cdot p(t_j) < p(s_i,t_j) \Rightarrow PMI(s;t) > 0}$

Step 3: potential collaborating TFs

PMI$(t_A; t_B)$ between two putative TFBSs t_a and t_b is calculated as follows:

$$PMI(t_A; t_B) = \log_2 \frac{p(t_A,t_B)}{p(t_A) \cdot p(t_B)}$$

- $p(t_A,t_B)$: joint probability of t_A and t_B
- $p(t_A)$ and $p(t_B)$: marginal probabilities

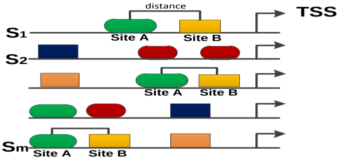


Step 3: potential collaborating TFs

PMI$(t_A; t_B)$ between two putative TFBSs t_a and t_b is calculated as follows:

$$PMI(t_A; t_B) = \log_2 \frac{p(t_A, t_B)}{p(t_A) \cdot p(t_B)}$$

- $p(t_A, t_B)$: joint probability of t_A and t_B
- $p(t_A)$ and $p(t_B)$: marginal probabilities



Cumulative PMI to determine cooperative TFs in the sequence set:

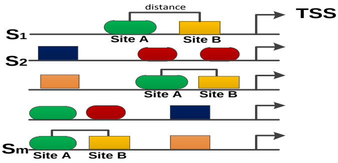
$$cPMI(t_A; t_B) = \sum_{s \in S} PMI(t_A; t_B)$$

Step 3: potential collaborating TFs

PMI$(t_A; t_B)$ between two putative TFBSs t_a and t_b is calculated as follows:

$$PMI(t_A; t_B) = \log_2 \frac{p(t_A, t_B)}{p(t_A) \cdot p(t_B)}$$

- $p(t_A, t_B)$: joint probability of t_A and t_B
- $p(t_A)$ and $p(t_B)$: marginal probabilities



Cumulative PMI to determine cooperative TFs in the sequence set:

$$cPMI(t_A; t_B) = \sum_{s \in S} PMI(t_A; t_B)$$

$\rightsquigarrow$ **z-score**$(cPMI(t_A; t_B)_{APC}) \geq 3$: **the pair is significant**

Outline

1. Cooperative TFs
2. PC-TraFF Algorithm
3. Functionality of PC-TraFF

Performance comparison

- Dataset analysis with PC-TraFF,
  - MatrixCatch, CPModule, and CrmMiner
- Sequence set analysis:

	The number of significant pairs			
Sequence sets	**PC-TraFF**	**MC**	**CPM**	**CrmM**
Genome-wide set	54	19	17	21
Breast cancer set	64	13	6	25

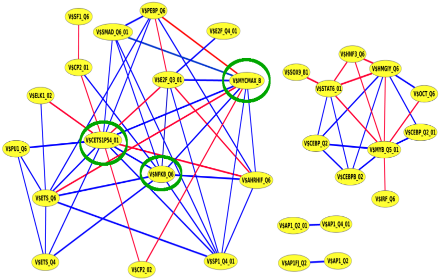
- Statistical comparison between PC-TraFF and previous methods:
  - **Positive pairs:** all experimentally validated pairs from interaction databases.
  - **Negative pairs:** all possible remaining pairs that can be detected using the PWM library.

	Sensitivity	**Specificity**	**MCC**
PC-TraFF	3.2%	99.3%	0.102
MatrixCatch	0.5%	99.9%	0.053
CPModule	0.5%	100%	0.06
CrmMiner	0.6%	99.6%	0.025

Sequence set analysis

Collaboration network

- Display the significant cooperations between TFs
- Nodes refer to related TFBSs
- Edges refer to a pairing between them



Summary

- Adopt an idea from the field of linguistics in the field of bioinformatics.
- Consider the genome as a document, the sequences as sentences, and TFBSs as words.
- PC-TraFF can identify known cooperative TF pairs as well as predict additional pairs.
- PC-TraFF algorithm has a tractable computational time and memory consumption.

Thank you for your attention!