Identification of associated transcription factors in promoters and their related enhancer regions

Cornelia Meckbach

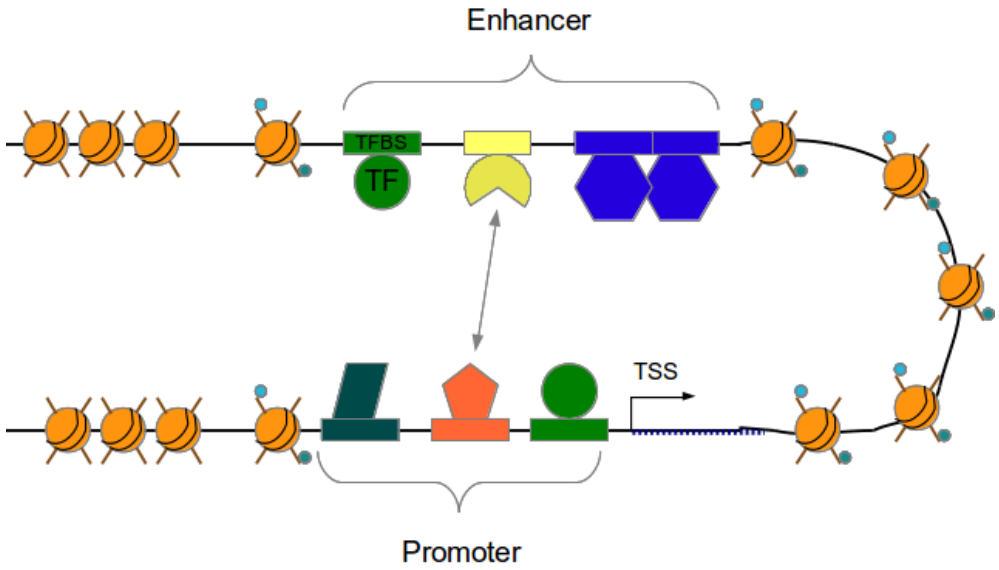
Institute of Bioinformatics

March 8, 2018

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

UNIVERSITÄTSMEDIZIN
GÖTTINGEN UMG

Motivation

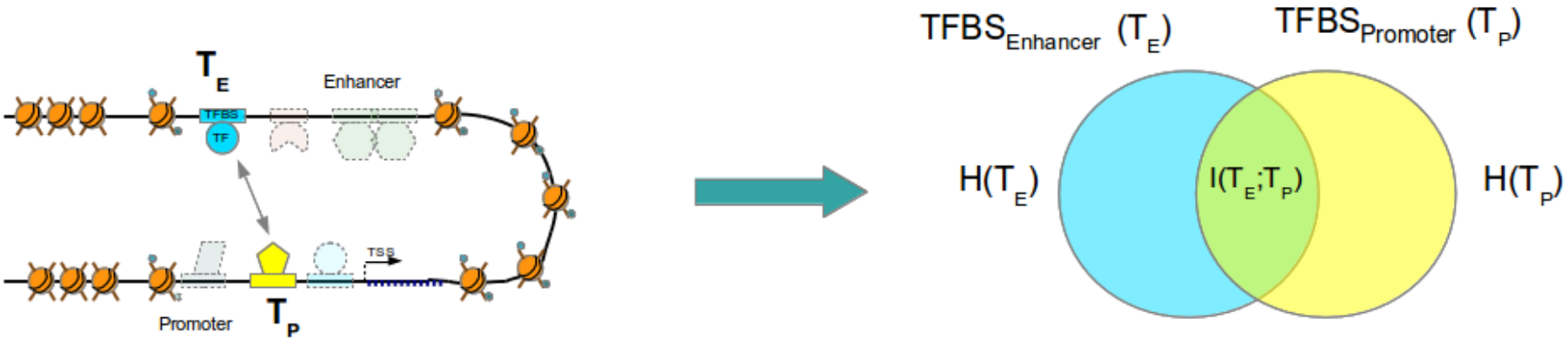


- Transcription factors (TFs) on paired enhancer and promoter regions are associated if they are involved in the pairing process.

$\Rightarrow$ Identification of associated TFs on enhancer and related promoter regions based on their transcription factor binding sites (TFBSs).

Inspired by:
Wong, KC (2017). **MotifHyades: expectation maximization for de novo DNA motif pair discovery on paired sequences**. Bioinformatics,

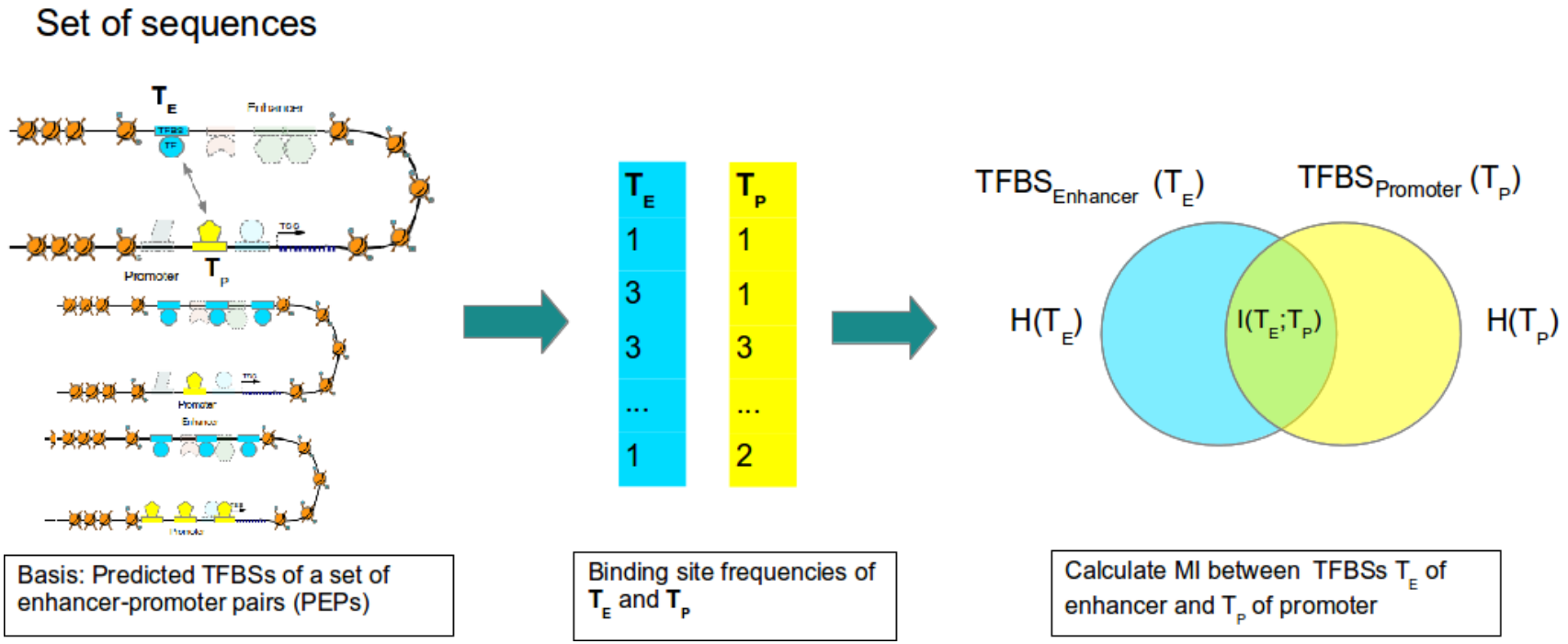
Mutual information



- Identification of associated TFs of promoter-enhancer pairings (PEPs) using mutual information (MI)

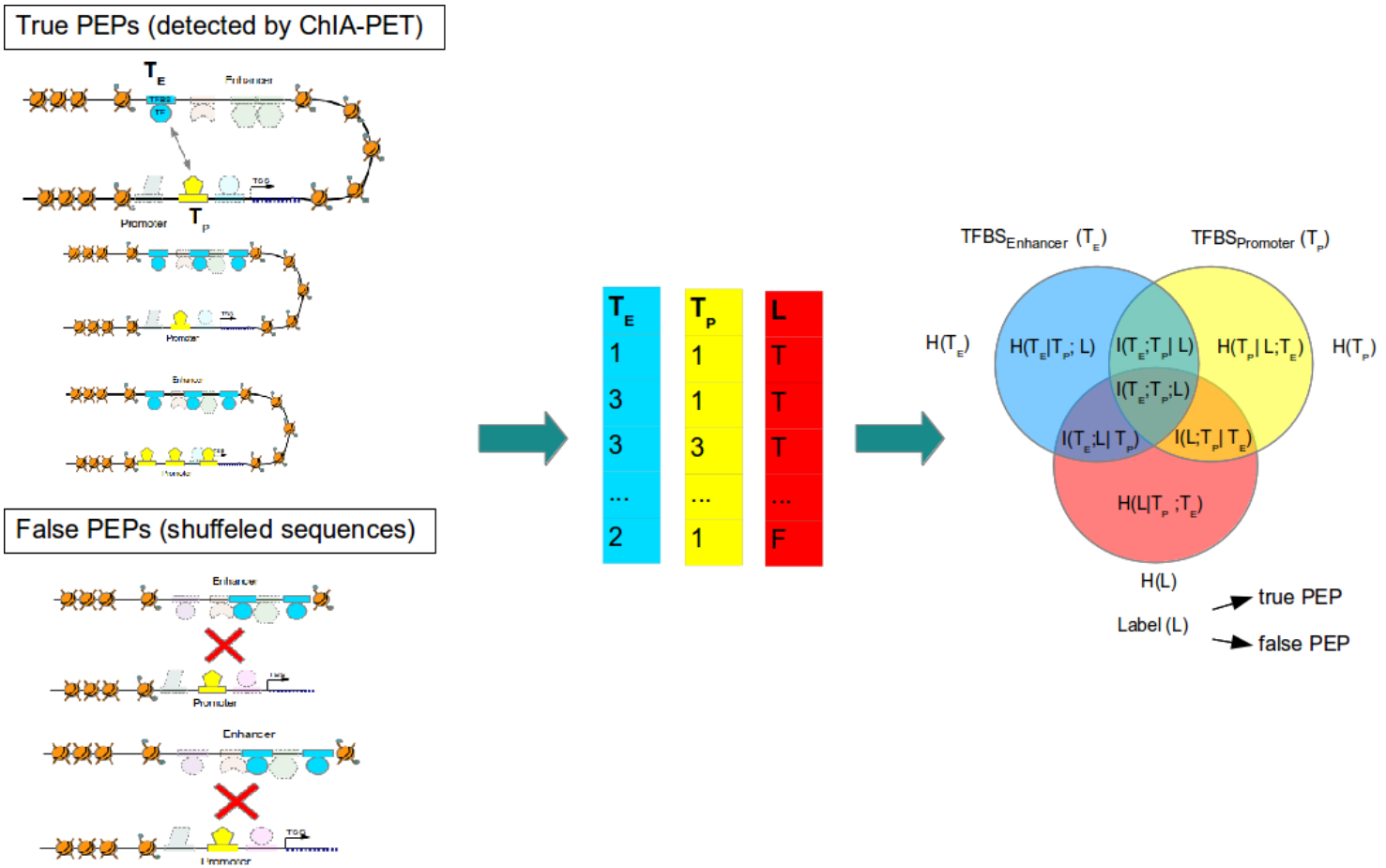
→ Two TFs are associated with each other if their binding behavior is in dependence of each other.

Mutual information



- Consider a set of experimentally validated PEPs for a cell line (e.g. by ChIA-PET)
  - Predict all TFBSs of the underlying sequences
  - Calculate MI for a TFBS pair T_E and T_P

Multivariate mutual information

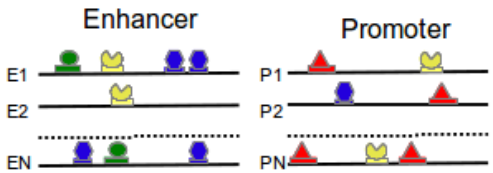


- How much information contains TFBS T_E about T_P by considering the interaction type (label).

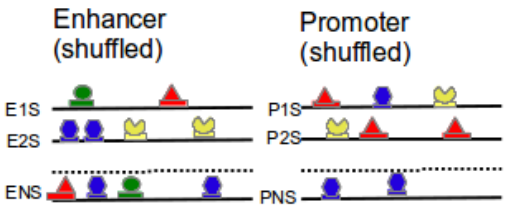
Workflow

Prediction of TFBSs using Match™

True PEPs: enhancer and promoter pairs (ChIA-PET)



False PEPs: shuffled enhancer and promoter sequences of positive samples



Creation of joint TFBS-sequence matrix with label

	TFBSs of Enhancer			TFBSs of Promoter			
	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEP 1	$f_{1,1}$						l_1
PEP 2							
...							
PEP N	$f_{N,1}$						l_N

$f_{i,j}$ is the number of TFBS j in sequence i
l_i is the label of PEP i (true/false pairing)

Normalization and interval building of TFBS count values

$I(T_E1;T_P1;L)$

	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEI 1	$i_{1,1}$						l_1
PEI 2							
...							
PEI N	$i_{N,1}$						l_N

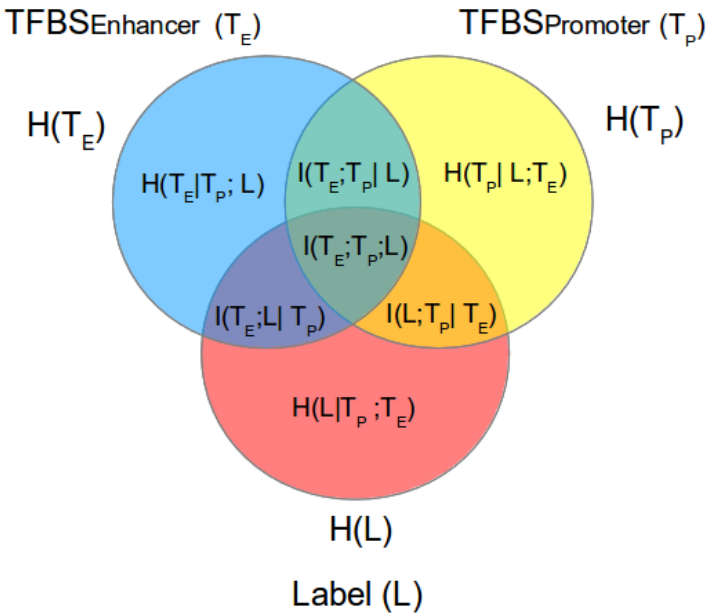
$i_{i,j}$ is the interval of count value of TFBS j in sequence i
l_i is the label of PEP i (true/false pairing)

Result

TFBS Enhancer	TFBS Promoter	MI
T_E1	T_P1	0.01
T_E1	T_P2	0.001
...		

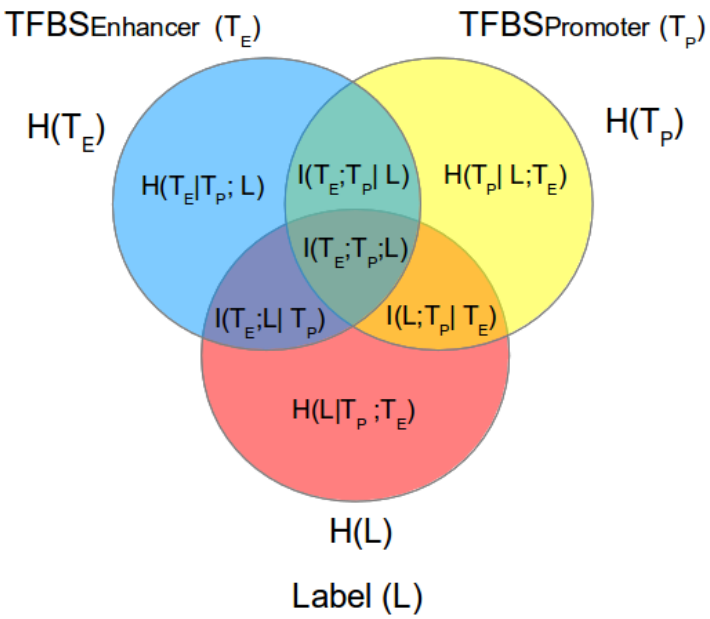
Normalize MI to avoid side effects

Information theoretic measures



Multivariate mutual information
$I(T_E;T_P;L)$

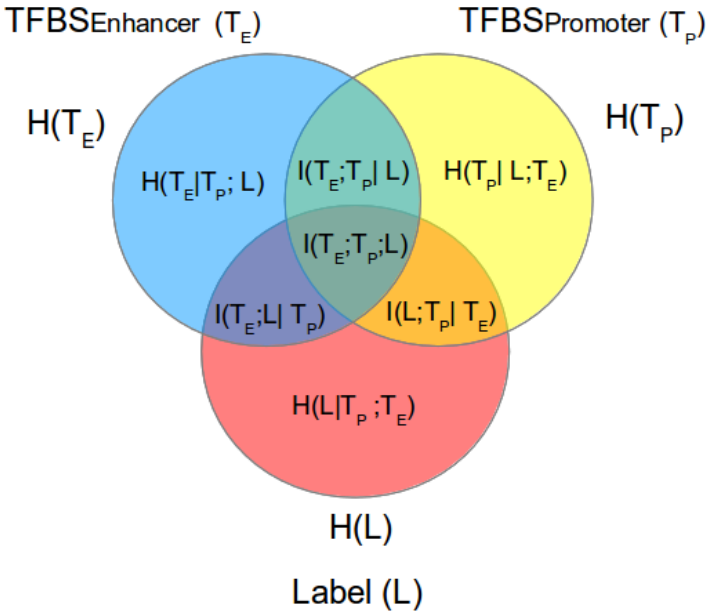
Information theoretic measures



Multivariate mutual information
I(T_E;T_P;L)

Mutual information of joint $T_E T_P$ with L
I(T_E, T_P;L)

Information theoretic measures

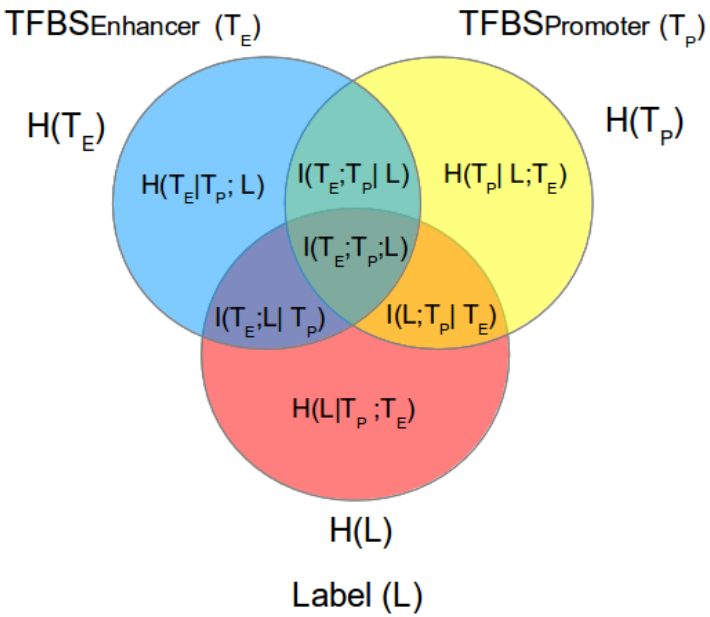


Multivariate mutual information
$I(T_E;T_P;L)$

Mutual information of joint $T_E T_P$ with L
$I(T_E, T_P;L)$

Conditional mutual information
$I(T_E;T_P|L)$

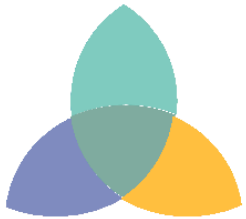
Information theoretic measures



Multivariate mutual information
$I(T_E;T_P;L)$

Mutual information of joint $T_E T_P$ with L
$I(T_E,T_P;L)$

Conditional mutual information
$I(T_E;T_P|L)$



Dual total correlation
$DTC(T_E;T_P;L)$

Synthetic example I

	TFBSs of Enhancer			TFBSs of Promoter			
	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEP1	1	10	1	2	1	1	T
PEP2	1	10	2	2	1	2	T
PEP3	1	10	3	2	1	3	T
PEP4	1	10	4	2	1	4	T
PEP5	4	9	5	9	2	1	F
PEP6	4	1	6	9	3	2	F
PEP7	4	0	7	9	4	3	F
PEP8	4	2	8	9	5	4	F

- Synthetic TFBS-sequence matrix: An entry f_{ij} in the matrix is the frequency of TFBS T_j in sequence i.
- One row corresponds to a PEP
- The label column indicates the pairing type (true/false pair)

Synthetic example I

	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEP1	1	10	1	2	1	1	T
PEP2	1	10	2	2	1	2	T
PEP3	1	10	3	2	1	3	T
PEP4	1	10	4	2	1	4	T
PEP5	4	9	5	9	2	1	F
PEP6	4	1	6	9	3	2	F
PEP7	4	0	7	9	4	3	F
PEP8	4	2	8	9	5	4	F

① Perfect associated TFBS pair

Synthetic example I

	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEP1	1	10	1	2	1	1	T
PEP2	1	10	2	2	1	2	T
PEP3	1	10	3	2	1	3	T
PEP4	1	10	4	2	1	4	T
PEP5	4	9	5	9	2	1	F
PEP6	4	1	6	9	3	2	F
PEP7	4	0	7	9	4	3	F
PEP8	4	2	8	9	5	4	F

① Perfect associated TFBS pair

② Associated TFBS pair in true PEPs

Synthetic example I

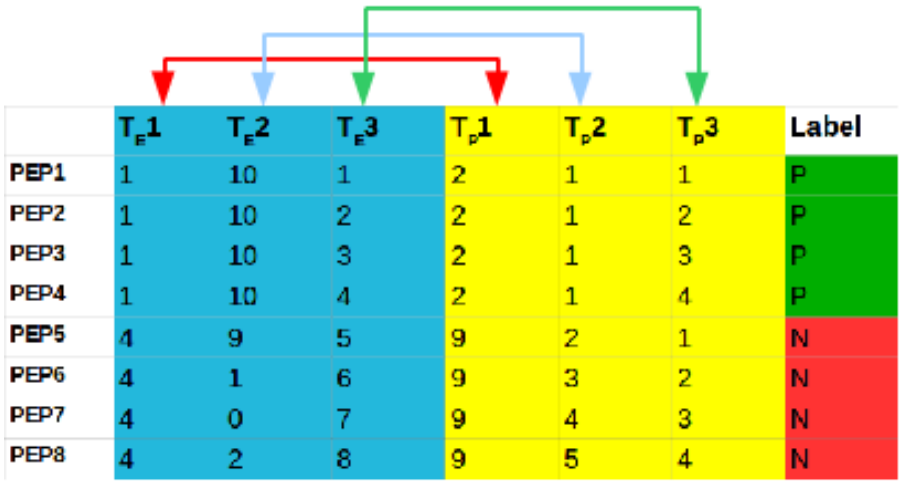
	T_E1	T_E2	T_E3	T_P1	T_P2	T_P3	Label
PEP1	1	10	1	2	1	1	T
PEP2	1	10	2	2	1	2	T
PEP3	1	10	3	2	1	3	T
PEP4	1	10	4	2	1	4	T
PEP5	4	9	5	9	2	1	F
PEP6	4	1	6	9	3	2	F
PEP7	4	0	7	9	4	3	F
PEP8	4	2	8	9	5	4	F

① Perfect associated TFBS pair

② Associated TFBS pair in true PEPs

③ Non-associated TFBS pair

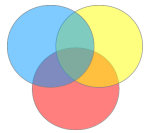
Synthetic example I :Result



	T_E1	T_E2	T_E3	T_p1	T_p2	T_p3	Label
PEP1	1	10	1	2	1	1	P
PEP2	1	10	2	2	1	2	P
PEP3	1	10	3	2	1	3	P
PEP4	1	10	4	2	1	4	P
PEP5	4	9	5	9	2	1	N
PEP6	4	1	6	9	3	2	N
PEP7	4	0	7	9	4	3	N
PEP8	4	2	8	9	5	4	N

1. Perfect associated TFBS pair
2. Associated TFBS pair in true PEPs
3. Non-associated TFBS pair

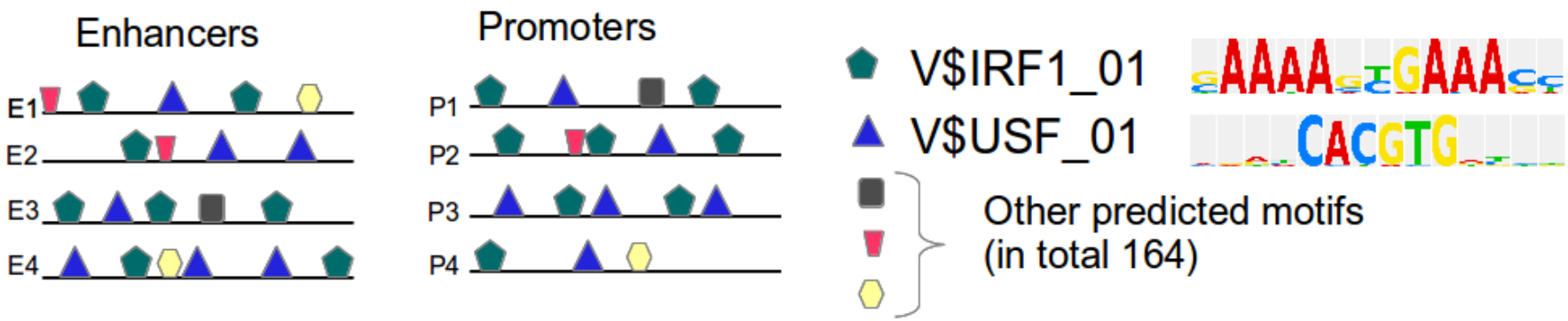
Table: Results for different measures for synthetic example I.



TFBS of enhancer	**TFBS of promoter**	$I(T_E;T_P;L)$	$I(T_E,T_P;L)$	$I(T_E;T_P\|L)$	$DTC(T_E,T_P,L)$
T_E1	T_P1	**1.0**	**1.0**	0	**1.0**
T_E2	T_P2	0.43	0.43	0.43	0.86
T_E3	T_P3	0	0.33	0.66	**1.0**

Synthetic example II

- Use a given library of 166 PWMs to predict potential TFBSs in the sequences



- **True PEPs:** TFBSs V$IRF1_01 and V$USF_01 are randomly inserted 1 to 10 times in enhancer and promoter sequences
- **False PEPs:** Shuffled enhancer and promoter sequences

Synthetic example II: Results

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

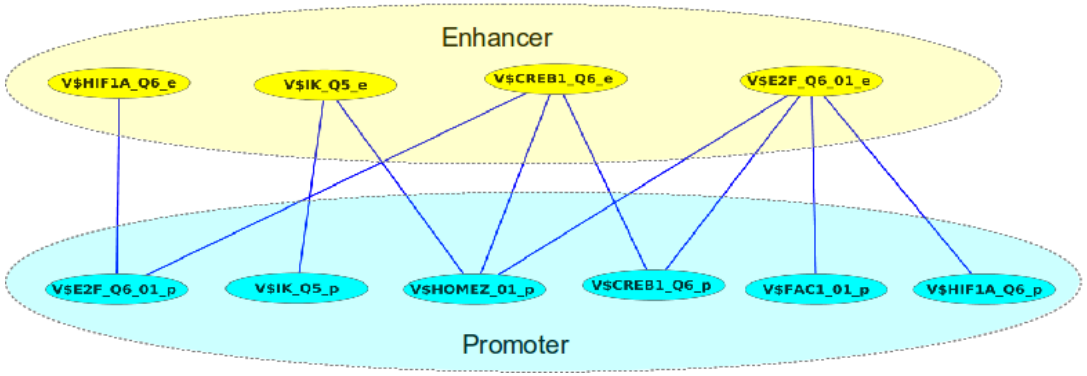
- In total there are 27556 TFBS pairs

→ Ranks of the inserted pairs?

Table: **Ranking position** of the inserted pairs.

TFBS of enhancer	**TFBS of promoter**	$I(T_E;T_P;L)$	$I(T_E,T_P;L)$	$I(T_E;T_P\|L)$	$DTC(T_E,T_P,L)$
V$IRF1_01	V$USF_01	1	160	5563	125
V$IRF1_01	V$IRF1_01	2	161	6848	215
V$USF_01	V$IRF1_01	3	408	4309	60
V$USF_01	V$USF_01	4	396	3524	23

Biolog. application: K562 cell line



TFBS enhancer (T_E)	Logoplot of enhancer motif	TFBS promoter (T_P)	Logoplot for promoter motif	I(T_E;T_P;L)
V$E2F_Q6_01		V$CREB1_Q6		0.0144
V$CREB1_Q6		V$CREB1_Q6_01		0.0122
V$CREB1_Q6		V$CREB1_Q6		0.0115
V$E2F_Q6_01		V$HOMEZ_01		0.0098
V$IK_Q5		V$IK_Q5		0.0095
V$IK_Q5		V$HOMEZ_01		0.0094
V$E2F_Q6_01		V$FAC1_01		0.0087
V$E2F_Q6_01		V$HIF1A_Q6		0.0084
V$CREB1_Q6		V$HOMEZ_01		0.0081
V$HIF1A_Q6		V$E2F_Q6_01		0.0075

Summary

- Workflow to detect associated TFs on enhancer and promoter regions based on their binding sites
- Compared four different information theoretic measures on synthetic data sets
- Multivariate mutual information I(T_E;T_P;L) performs best on both sets

Thanks

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

People

Edgar Wingender

Mehmet Gültas

Martin Haubrock

Sebastian Zeidler

Jürgen Dönitz

Rayan Daou

Darius Wlochowitz

Halima Alachram

Doris Waldmann

Torsten Schöps

Malte Sahrhage