Enumeration Algorithms and Statistical Significance

Koji Tsuda
AIST Computational Biology Research Center
JST ERATO Minato Project

Joint work with Aika Terada,
Mariko Okada-Hatakeyama, Jun Sese

Published on July 23, 2013

PNAS

Statistical significance of combinatorial regulations

Aika Terada[a,b,c], Mariko Okada-Hatakeyama[d], Koji Tsuda[c,e,1], and Jun Sese[a,b,1]

[a]Department of Computer Science and [b]Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; [c]Minato Discrete Structure Manipulation System Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Sapporo, Hokkaido 060-0814, Japan; [d]Laboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences (IMS-RCAI), Yokohama, Kanagawa 230-0045, Japan; and [e]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved July 3, 2013 (received for review February 4, 2013)

More than three transcription factors often work together to enable cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is not only computationally nontrivial but also extremely unlikely because of multiple testing correction. The exponential growth in the number of tests forces us to set a strict limit on the maximum arity. Here, we propose an efficient branch-and-bound algorithm called the "limitless arity multiple-testing procedure" (LAMP) to count the exact number of testable combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists significant combinations without any limit, whereas the family-wise error rate is rigorously controlled under the threshold. In the human breast cancer transcriptome, LAMP discovered statistically significant combinations of as many as eight binding motifs. This method may contribute to uncover pathways regulated in a coordinated fashion and find hidden associations in heterogeneous data.

Bonferroni correction | gene expression

deliberately excluding such tests. Here, we propose an efficient branch-and-bound algorithm, called the "limitless arity multiple-testing procedure" (LAMP). LAMP counts the exact number of "testable" motif combinations and derives a tighter bound of FWER, which allows the calibration of the Bonferroni factor as the FWER is controlled rigorously under the threshold.

In comparison with existing methods that can find only two-motif combinations, our testing procedure may contribute to finding larger fractions of regulatory pathways and TF complexes, thus providing more concrete evidence for further investigation. In legacy yeast expression data (29), a four-motif combination corresponding to a known pathway was found using LAMP, whereas only two motifs in the combination had been predicted using the existing method. When applied to human breast cancer transcriptome data (30), combinations of up to eight motifs were found to be statistically significant.

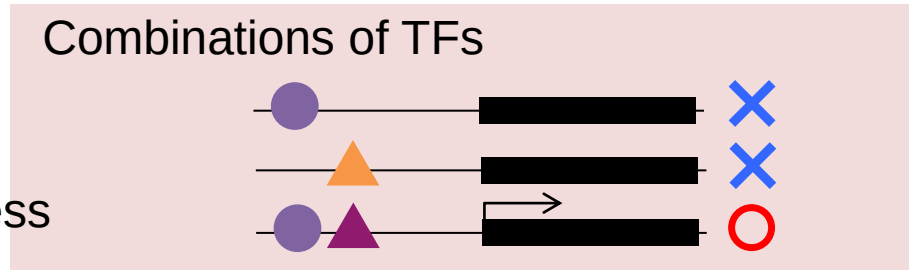
Results

Method Overview. To present our strategy for combinatorial regu-

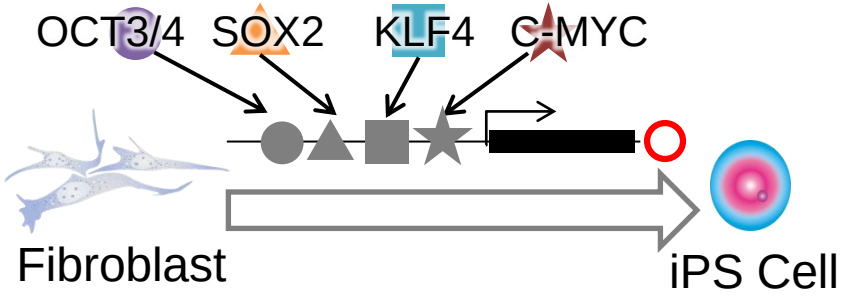
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Transcription factors (TFs) work in combination

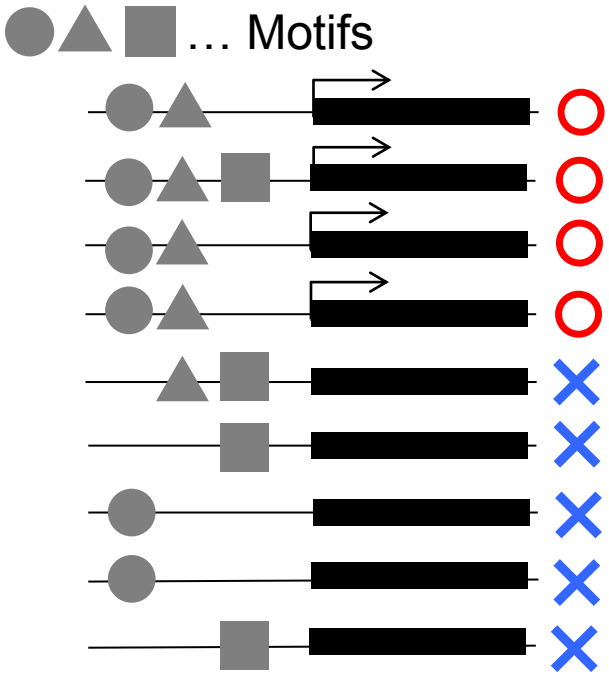
- Often several TFs are necessary to induce the expression of downstream genes



Example: Yamanaka Factor（K. Okita *et al.*, Nature, 2007）



Find statistically significant combinations of TF binding motifs



Contingency table for ●▲

	Up-regulated	No-regulated
With Motif Combination	4	0
Without	0	5

P-value by Fisher exact test
0.0079

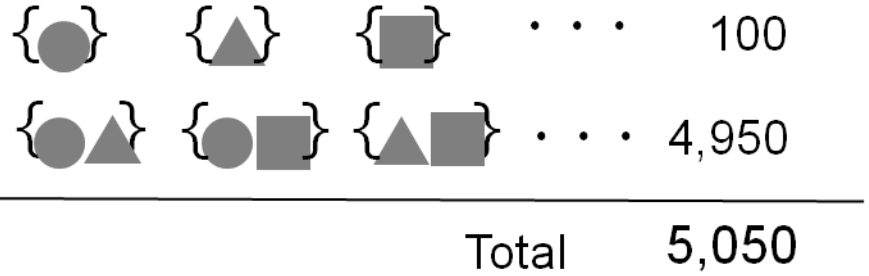
Significant?
No – You have to apply multiple testing procedure

Bonferroni Correction

- Family-wise error rate(FWER)
  - At least one false discovery occurs
- P-value threshold δ is determined such that FWER is below α
- For m tests,

$$\delta = \frac{\alpha}{m}$$

- 100 motifs in total
- Number of tests



- Corrected threshold

  $\delta = 0.05/5050$

  $= 9.9 \times 10^{-6}$
- Bonferroni is too conservative!

New Proposal: Limitless Arity Multiple testing Procedure (LAMP)

- Count the exact number of “testable” combinations
  - Infrequent combinations do not affect family-wise error rate
  - Stepwise procedure involving itemset mining
- Calibrate the correction factor to the smallest possible value
- Discovered statistically significant motif combinations in yeast and breast cancer expression data

Raw p-value

	Up regulated	No regulated
With Motif Combination	a	b
Without	c	d

- Null Hypothesis H
  - Two variables are independent
- P-value: p(a,b,c,d)
  - Probability of observing stronger table than observed
  - If smaller than α, reject H (discovery!)
- Type-I error: reject H when it is true
- Probability of type-I error must satisfy

$$P(p < \alpha \mid H) \leq \alpha$$

Multiple Tests

- m null hypotheses $H_1,\ldots,H_m$
- V: Number of rejections in m tests
- Probability that more than one type-I error occurs: Family-wise error rate (FWER)

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i)$$

- Multiple testing procedures aim to control FWER under α

Bonferroni Correction

- Given threshold δ、FWER is bounded as

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i) \leq \sum_{i=1}^{m} P(p_i \leq \delta \mid H_i) \quad \text{Union bound}$$

$$\leq m\delta \quad \text{Definition of p-value}$$

- Thus, setting δ=α/m calibrate FWER bound to α

	Up-regulated	Not regulated	
With Motif Combination	a	b	x
Without	c	d	N-x
	n_u	N-n_u	N

Occurrence Frequency

- P-value by Fisher exact test cannot be smaller than

$$f(x) = \binom{n_u}{x} / \binom{N}{x}$$

- No chance of false discovery, if $f(x) \geq \delta$

$$P(p < \delta \mid H) = 0$$

Tarone Correction (Biometrics, 1990)

- Considering minimum p-value, FWER is bounded as follows

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i) \leq \sum_{i=1}^{m} P(p_i \leq \delta \mid H_i)$$ Union bound

$$= \sum_{\{i \mid f(x_i) \geq \delta\}} P(p_i \leq \delta \mid H_i)$$ Use minimum p-value to remove hypotheses

$$\leq \mid \{i \mid f(x_i) \geq \delta\} \mid \delta$$ Definition of p-value

- Take maximum δ that keeps FWER bound below α

- FWER is represented as

$$g(\delta) = |\{i \mid f(x_i) \geq \delta\}| \, \delta$$

- Identify all motif combinations that satisfy

$$f(x) \geq \delta$$

- Inverse function

$$f^{-1}(\delta) = \lambda \ s.t. \ f(\lambda) \leq \delta \leq f(\lambda - 1)$$

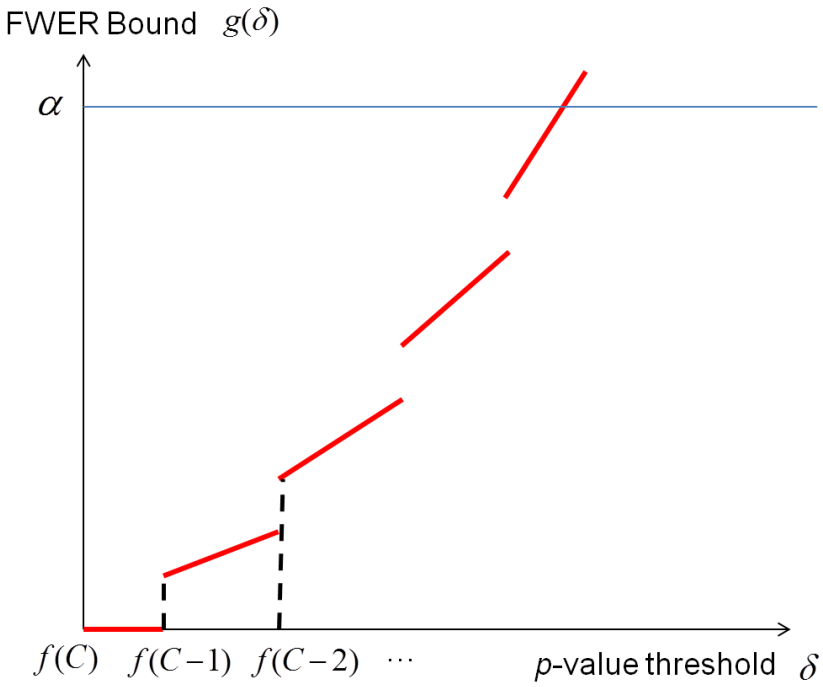
- Find all combinations whose frequency is λ or more by itemset mining
- FWER bound is computed as

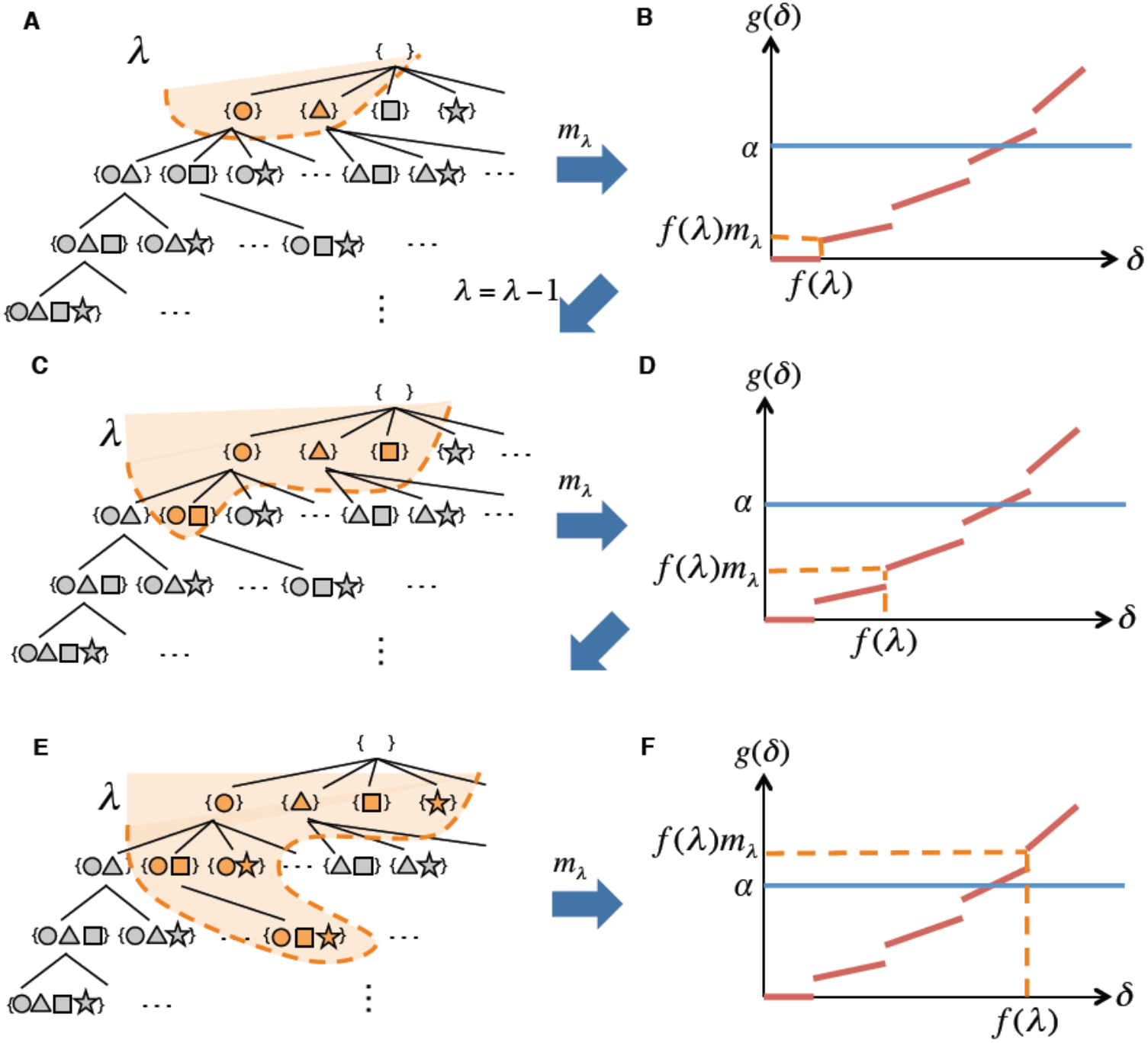
$$g(\delta) = m' \delta$$

m': Number of motif combintions whose frequency is λ or more

Finding optimal δ that calibrates FWER bound to α

- FWER bound is piecewise linear
- Repeat itemset mining with decrementing the frequency parameter
- A line segment drawn by a mining call
- Finish if line segment reaches α




A
B
C
D
E
F
λ
{ }
{○} {△} {□} {☆}
{○△} {○□} {○☆} {△□} {△☆}
{○△□} {○△☆} {○□☆}
{○△□☆}
m_λ
$\lambda = \lambda - 1$
$g(\delta)$
α
$f(\lambda)m_\lambda$
$f(\lambda)$
δ

Applications to Yeast Transcriptome

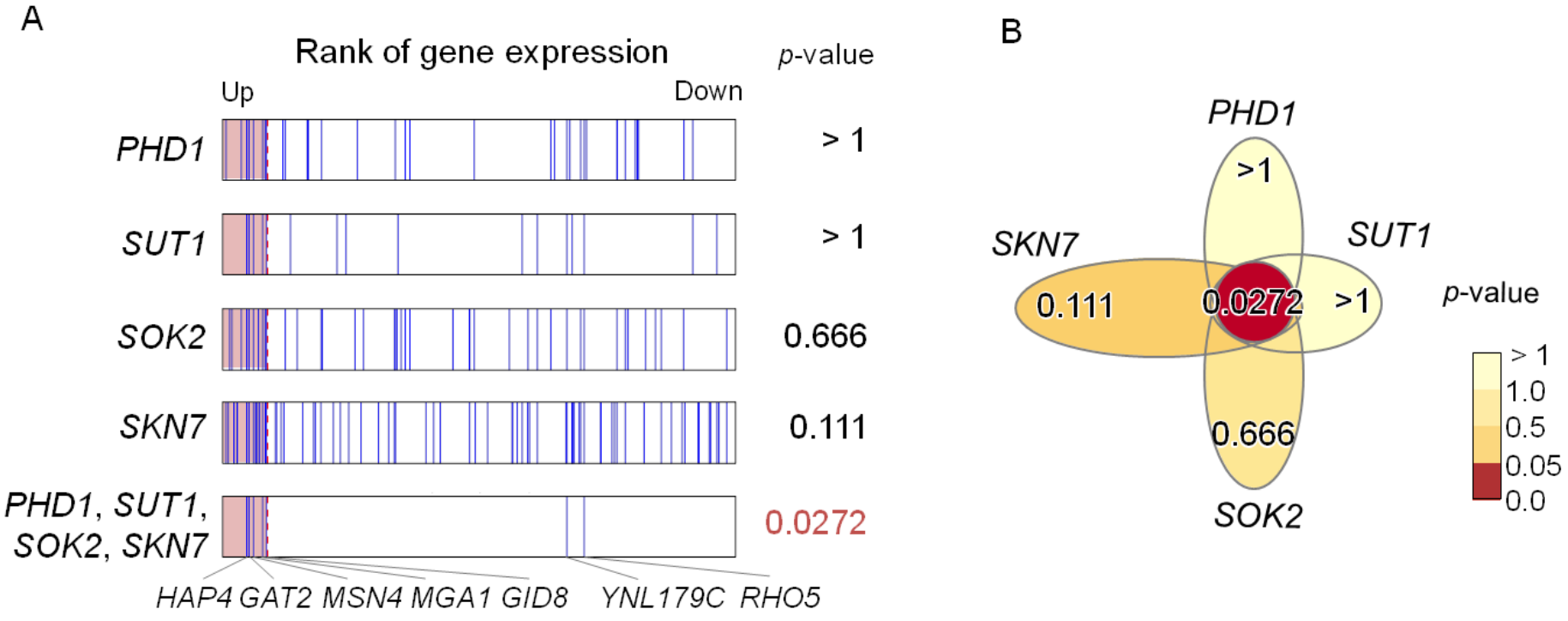
- Microarray data by Gasch et al
- Binding motif data by SGD Database
- 102 motifs, each binding to 30.1 genes on average
- Expressions of about 6000 genes measured on 173 different conditions

Statistically significant TF combinations under a heat shock condition

Corrected p-value（p-value*K）

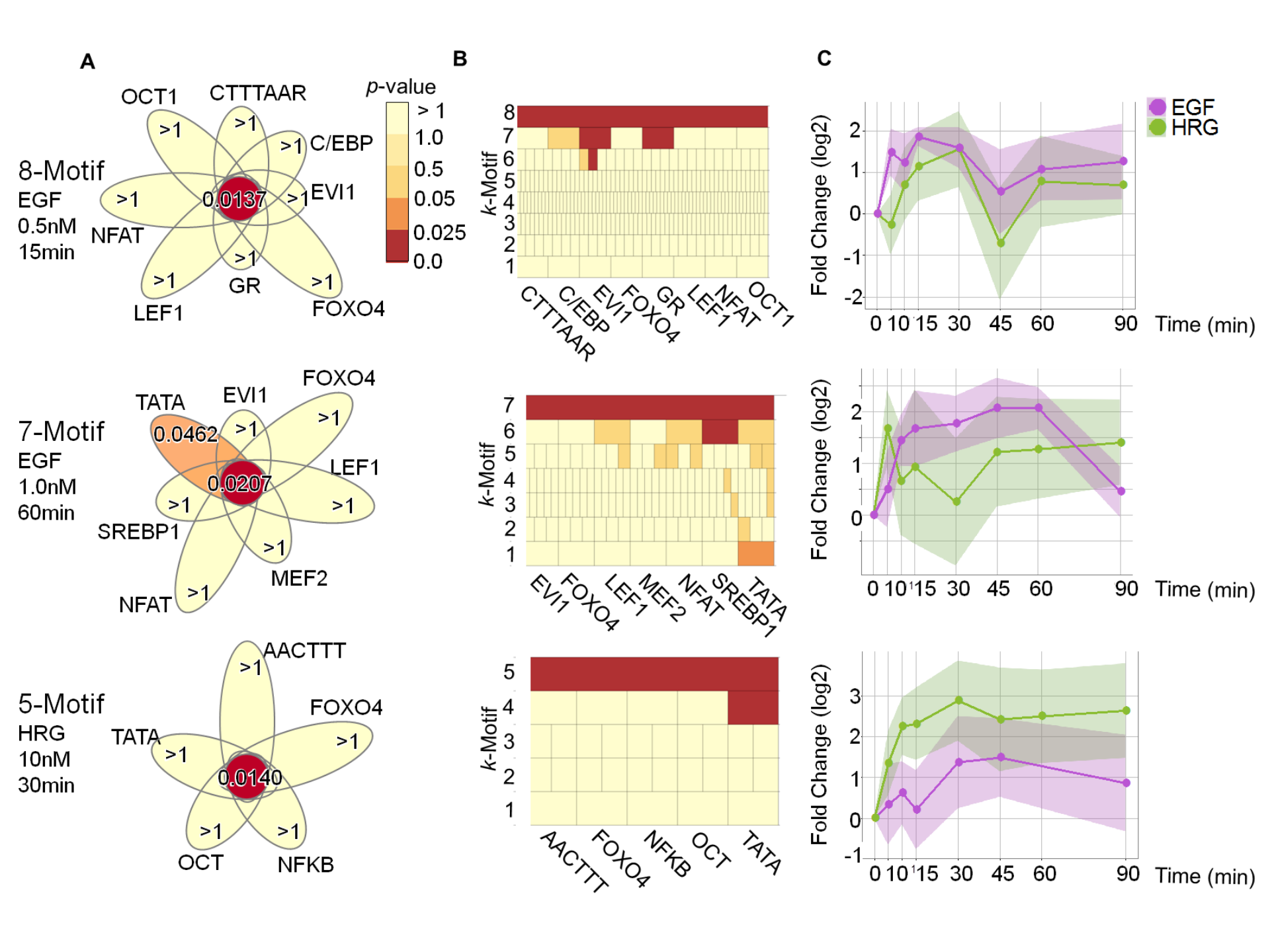
Combination	LAMP (≦102)	Bonferroni (≦4)
	K = 303	K = 4,426,528
HSF1	4.41E-24	6.44E-20
MSN2	3.73E-11	5.45E-07
MSN4	0.00053	> 1
SKO1	0.00839	> 1
SNT2	0.0192	> 1
PHD1, SUT1, SOK2, SKN7	0.0272	> 1

Red：significant


A
Rank of gene expression
p-value
Up
Down
PHD1
> 1
SUT1
> 1
SOK2
0.666
SKN7
0.111
PHD1, SUT1,
SOK2, SKN7
0.0272
HAP4 GAT2 MSN4 MGA1 GID8 YNL179C RHO5
B
PHD1
>1
SKN7
SUT1
0.111
0.0272
>1
0.666
SOK2
p-value
> 1
1.0
0.5
0.05
0.0

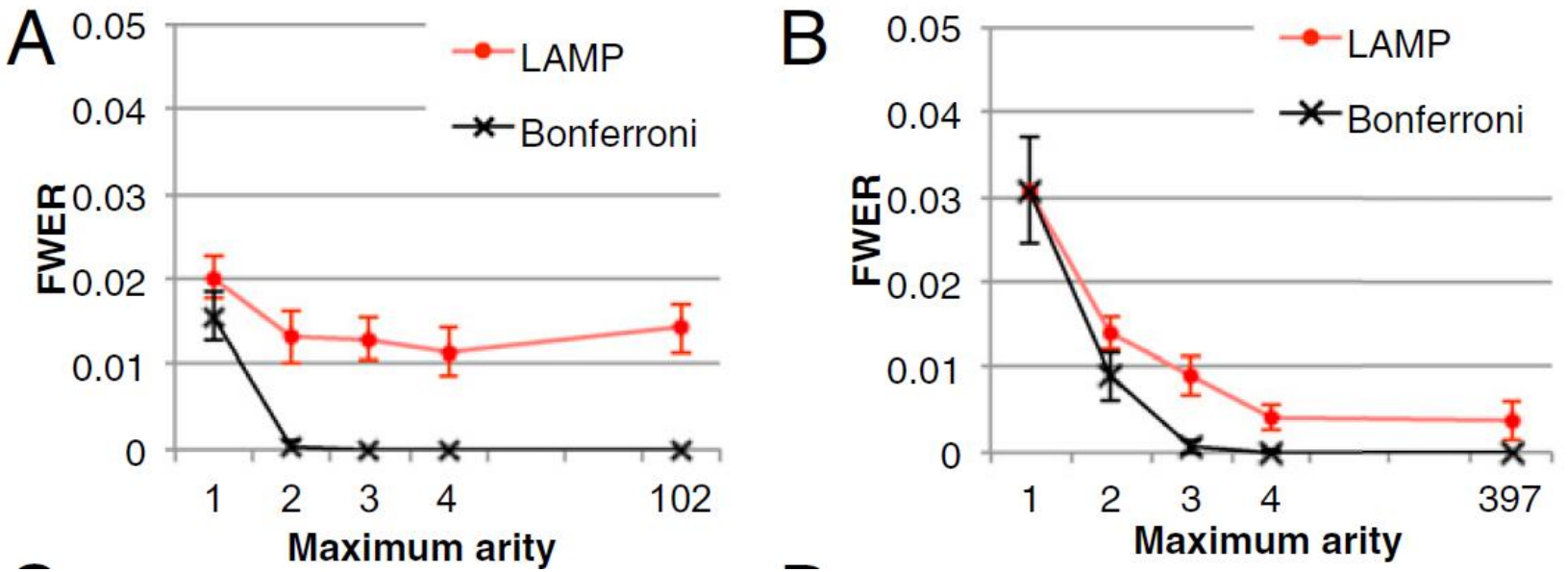
Application to MCF7 human breast cancer cells (GSE6462)

- Treated with epidermal growth factor (EGF) or heregulin (HRG)
  - 0.1, 0.5, 1, 10 nM
- Expression measured 5, 10, 15, 30, 45, 60 mins after
- Motifs taken from MSigDB
- 397 motifs, Approx. 12000 genes
- LAMP K= 1,174,108 ~ 3,750,336
- Bonferroni K=1.4 x 10^{16} (maximum arity =8)

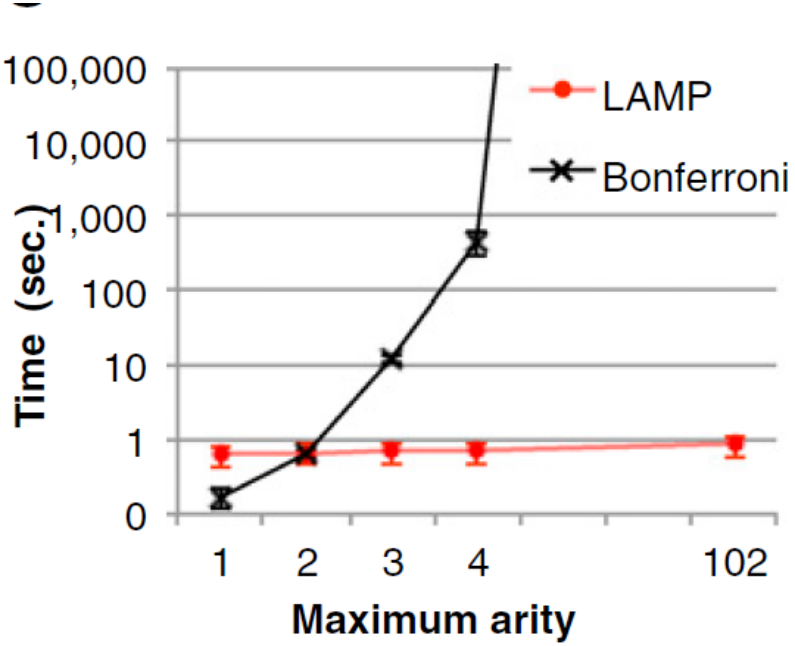

A
B
C
p-value
> 1
1.0
0.5
0.05
0.025
0.0
8-Motif
EGF
0.5nM
15min
OCT1
CTTTAAR
C/EBP
EVI1
NFAT
GR
LEF1
FOXO4
0.0137
>1
7-Motif
EGF
1.0nM
60min
TATA
EVI1
FOXO4
LEF1
SREBP1
MEF2
NFAT
0.0462
0.0207
5-Motif
HRG
10nM
30min
AACTTT
FOXO4
TATA
OCT
NFKB
0.0140
k-Motif
Fold Change (log2)
Time (min)
EGF
HRG
0
10
15
30
45
60
90

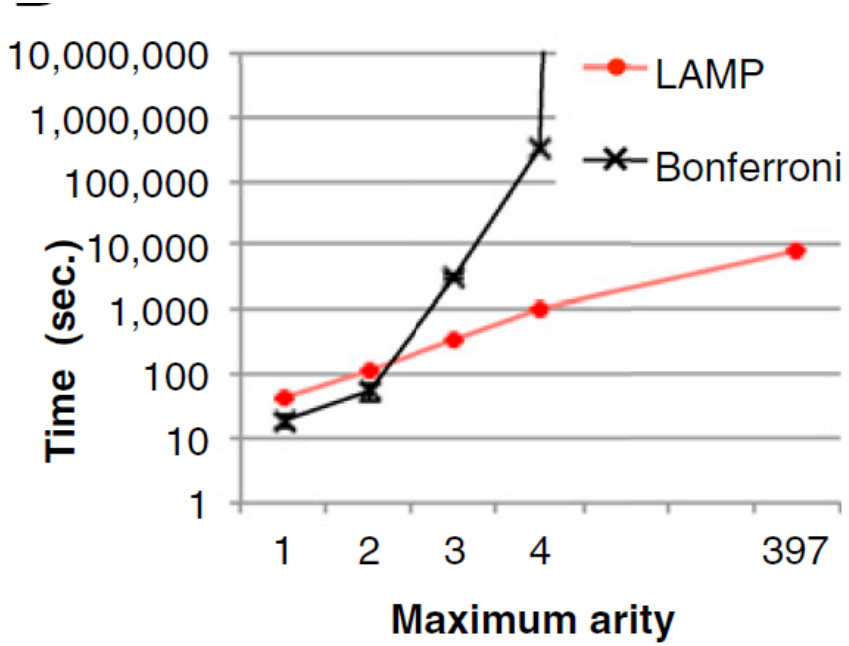
Empirical FWER

- LAMP's FWER is much closer to the designated value 0.05



Computational Time





Concluding Remarks

- LAMP is much more sensitive than Bonferroni, whereas FWER is strictly kept under threshold
- FDR version of LAMP may be possible
- Immediately applicable to sequences, trees and graphs
- Minimum p-value must be strictly positive
  - LAMP cannot be applied to t-test
  - Statistical tests with “robustness” can be combined with LAMP