Hypothesis generation in bioinformatics

Meelis Kull
BIIT-group
University of Tartu

Estonian CS Theory Days
Nelijärve, Feb 6, 2011

Bioinformatics

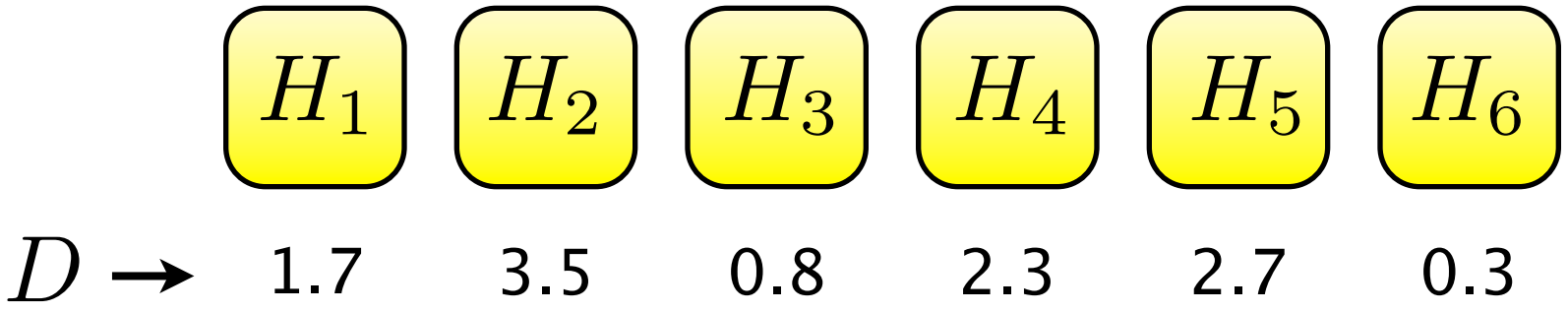
- Applying computationally intensive techniques to increase the understanding of biological processes.
- **Input: biological data**
  - manually curated or directly from experiments
- **Output: biological hypotheses**
  - one or more
  - possibly scored and ranked

A study in bioinformatics resulting with multiple hypotheses

- Collecting data
  - reading literature
  - browsing databases
  - making experiments
- Hypotheses generation
  - may be manual
- Hypotheses evaluation
  - scoring
  - ranking
  - filtering

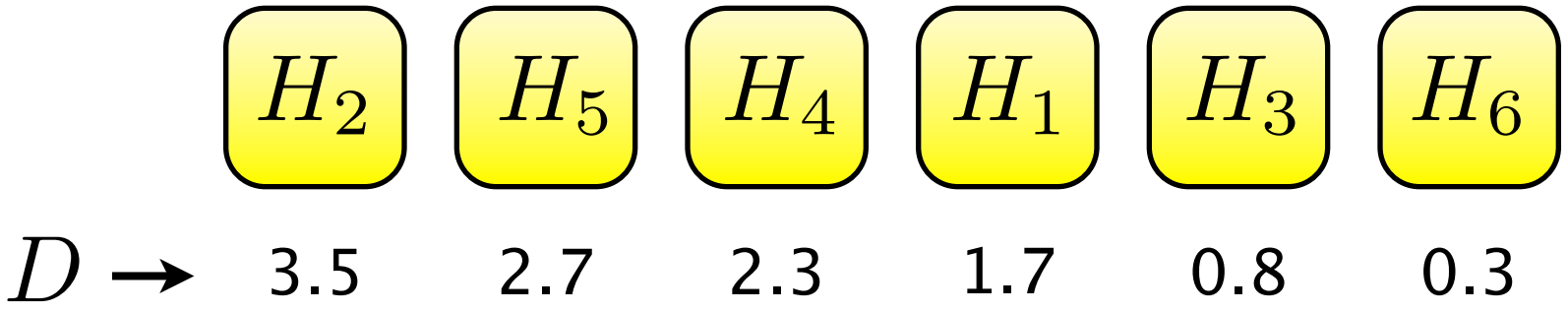
Hypothesis evaluation

- Each hypothesis must have a statistic
  - a function for calculating a real-valued score based on data
  - higher score means better hypothesis



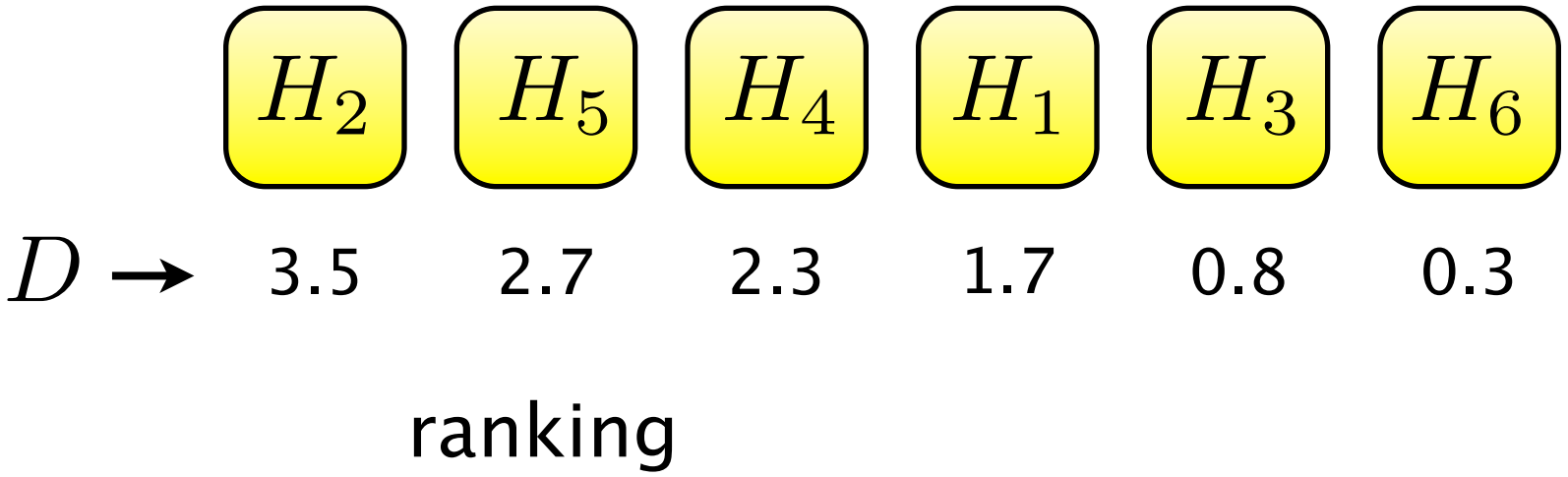
Hypothesis evaluation

- Each hypothesis must have a statistic
  - a function for calculating a real-valued score based on data
  - higher score means better hypothesis



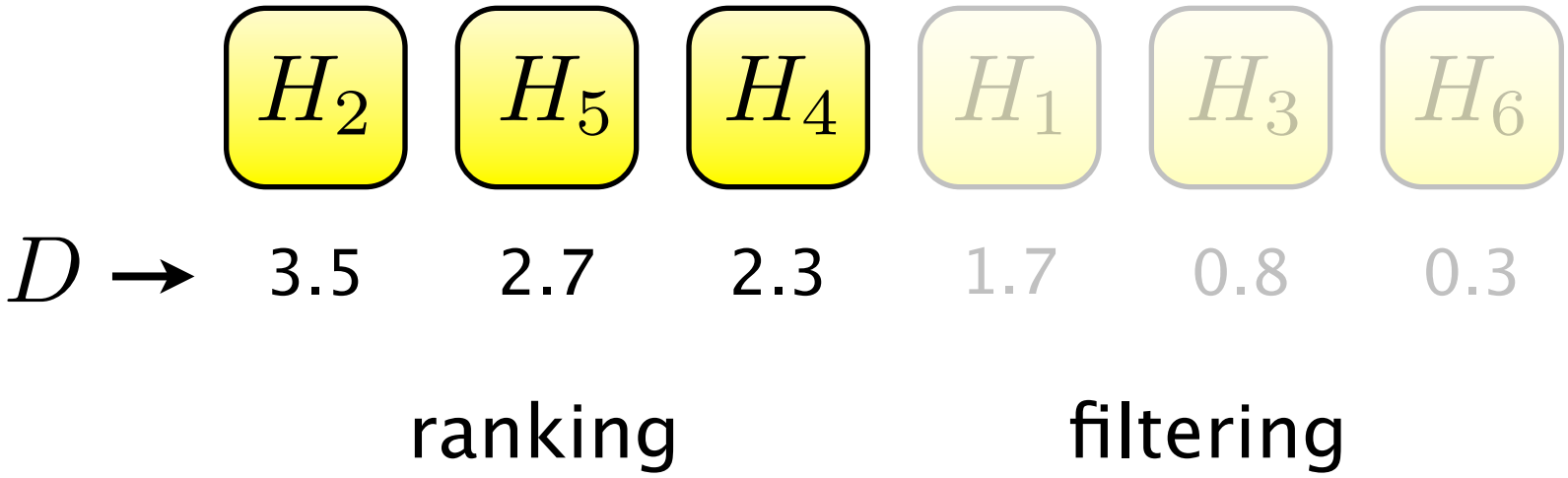
Hypothesis evaluation

- Each hypothesis must have a statistic
  - a function for calculating a real-valued score based on data
  - higher score means better hypothesis



Hypothesis evaluation

- Each hypothesis must have a statistic
  - a function for calculating a real-valued score based on data
  - higher score means better hypothesis



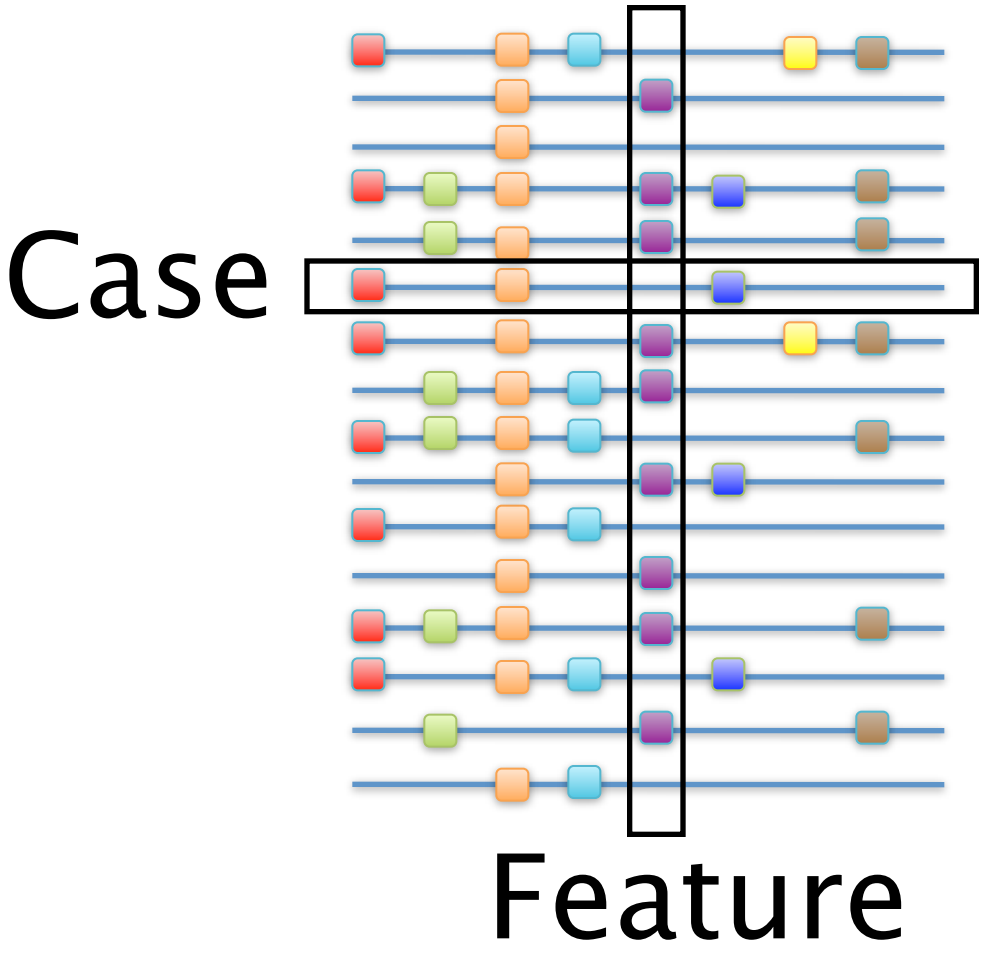
Hypothesis evaluation

- Input:
  - The data
  - The hypothesis statistics
- Scoring
- Ranking
- Filtering
- Output:
  - Filtered ranked scored hypotheses

P-value based hypothesis evaluation

- Input:
  - A data generating model based on our current understanding of the system
  - The data
  - The hypothesis statistics
- Scoring
- P-value calculation:
  - how probable it is to get so high score according to the model
- Ranking
- Filtering
- Output:
  - Filtered ranked p-valued hypotheses

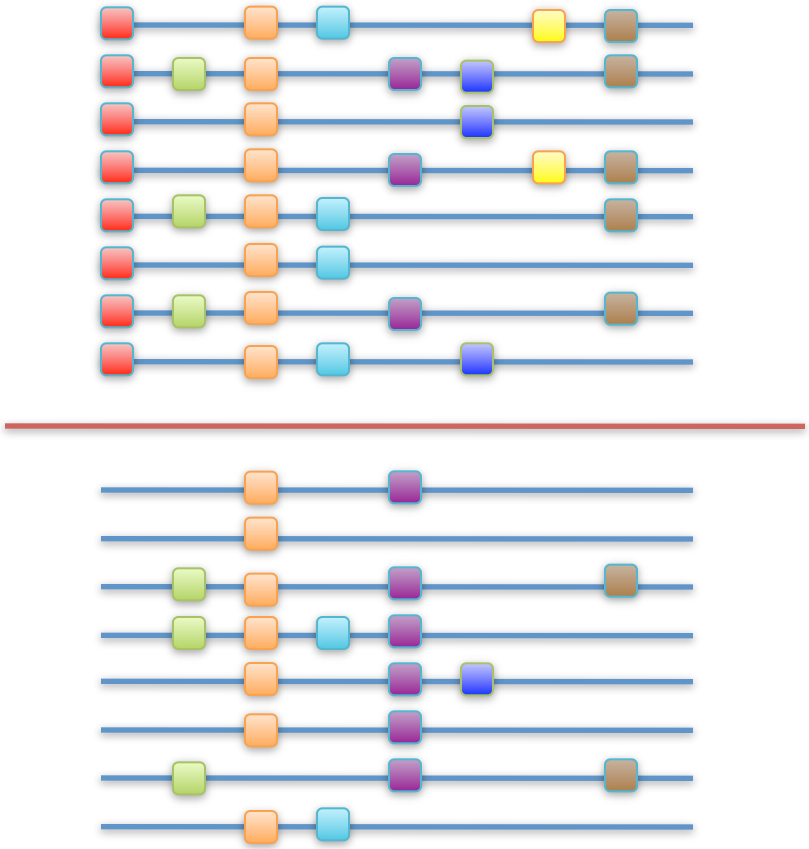
Sample data: Cases with features



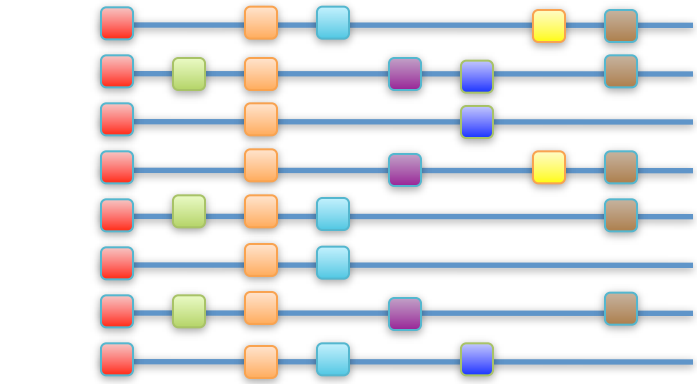
Task:
Find features that are functionally related to the red feature

Hypotheses:
Brown
Blue

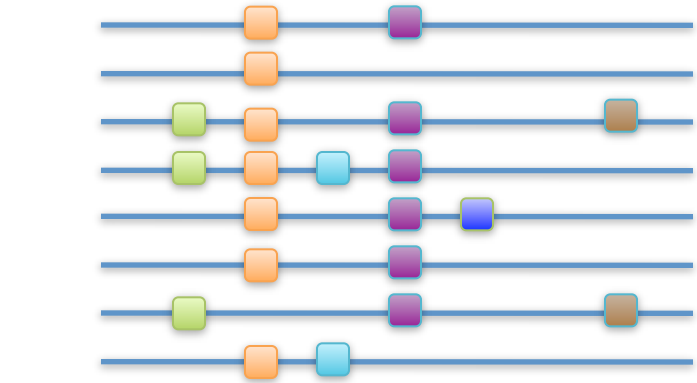
Sample data: Cases with features



Sample data: Cases with features



Foreground



Background

D

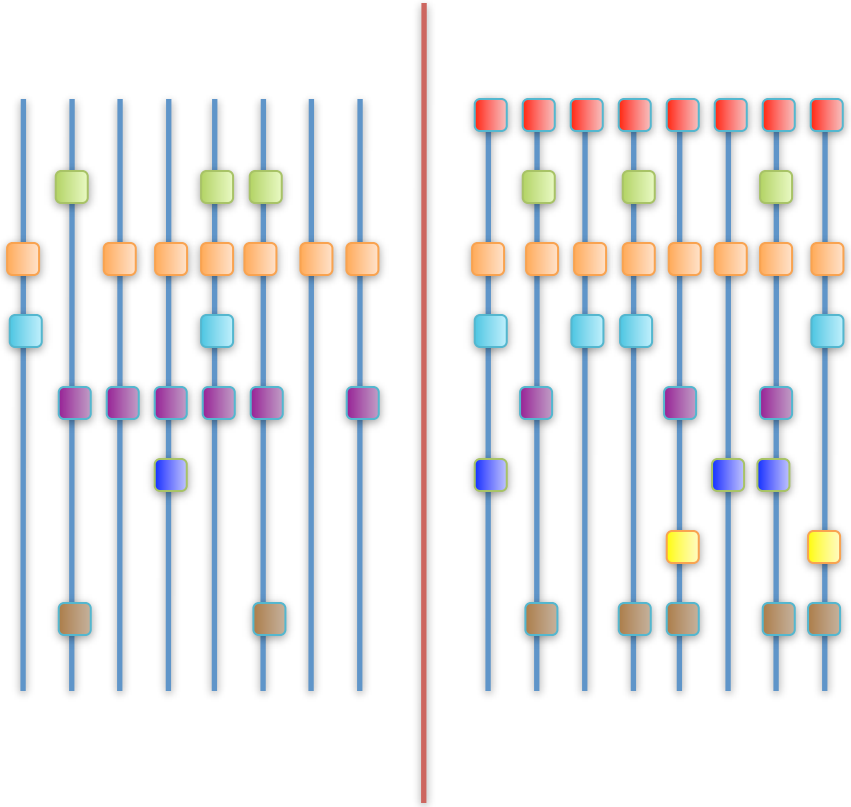
$D_1^{\mathsf{F}}\ D_2^{\mathsf{F}}\ \cdots$

$D_1^{\mathsf{B}}\ D_2^{\mathsf{B}}\ \cdots$

3 8 4 3 3 2 5

3 7 2 6 1 0 2

Sample data: Cases with features



Car is moving
Car is green
Engine is running
Car looks clean
Gas pedal is pressed
Car has eight seats
Car is yellow
Gear is engaged

BG	FG
3	3
7	8
2	4
6	3
1	3
0	2
2	5

- A hypothesis – feature i is functionally related to car moving
- We have 7 features, therefore 7 hypotheses
- QUESTION:
  - What statistics to use to score the hypotheses?

Foreground count statistic

$$h_i^{\mathsf{FG}}(D) = D_i^{\mathsf{F}}$$

	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathsf{FG}}(D)$	3	8	4	3	3	2	5

Bias statistic

$$h_i^{\mathsf{BIAS}}(D) = \frac{D_i^{\mathsf{F}}}{D_i^{\mathsf{F}} + D_i^{\mathsf{B}}}$$

	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathsf{BIAS}}(D)$	0.50	0.53	0.67	0.33	0.75	1.00	0.71

Hypergeometric p-value statistic

$$h_i^{\mathrm{HYPER}}(D) = \sum_{k=D_i^{\mathrm{F}}}^{D_i^{\mathrm{F}}+D_i^{\mathrm{B}}} \frac{\binom{m}{k}\binom{m}{D_i^{\mathrm{F}}+D_i^{\mathrm{B}}-k}}{\binom{2m}{D_i^{\mathrm{F}}+D_i^{\mathrm{B}}}}$$

	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathrm{HYPER}}(D)$	0.70	0.50	0.30	0.98	0.28	0.23	0.16

Binomial p-value statistic

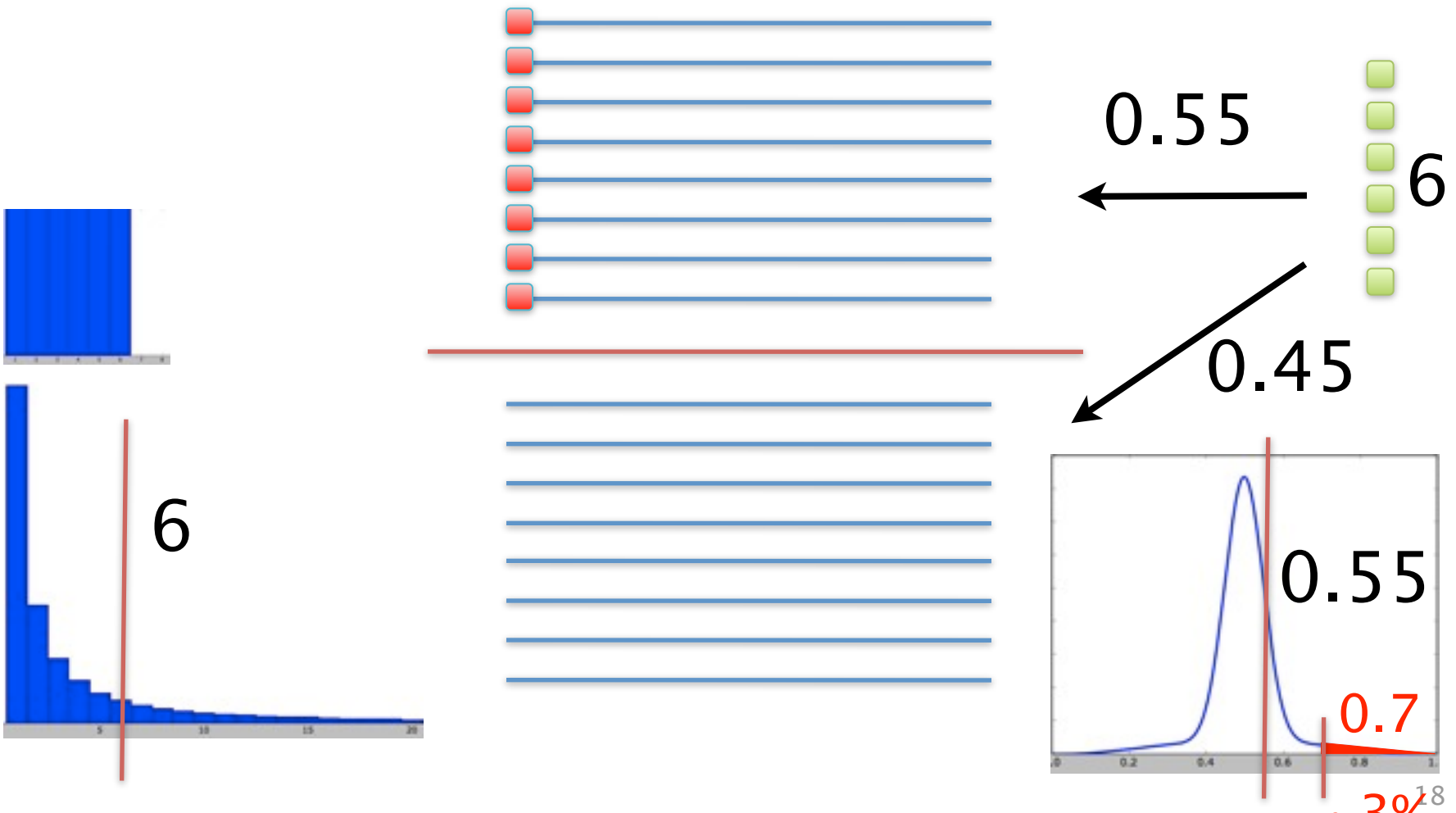
$$h_i^{\mathrm{BINOM}}(D) = \sum_{k=D_i^{\mathrm{F}}}^{D_i^{\mathrm{F}}+D_i^{\mathrm{B}}} \binom{D_i^{\mathrm{F}} + D_i^{\mathrm{B}}}{D_i^{\mathrm{F}}} 0.5^k 0.5^{D_i^{\mathrm{F}}+D_i^{\mathrm{B}}-k}$$

	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathrm{BINOM}}(D)$	0.66	0.50	0.34	0.91	0.31	0.25	0.23

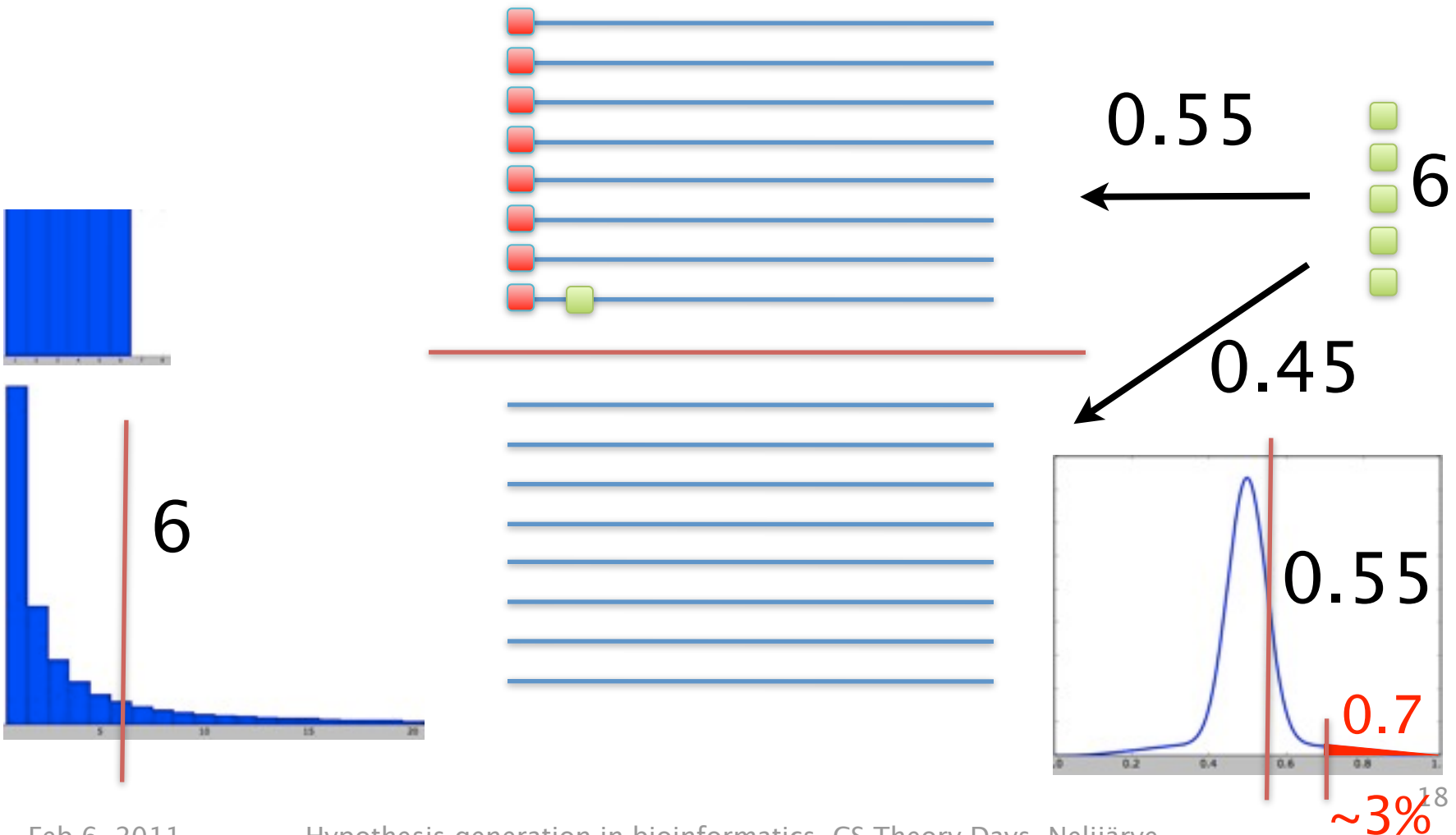
Comparison

	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathsf{FG}}(D)$	3	8	4	3	3	2	5
$h_i^{\mathsf{BIAS}}(D)$	0.50	0.53	0.67	0.33	0.75	1.00	0.71
$h_i^{\mathsf{HYPER}}(D)$	0.70	0.50	0.30	0.98	0.28	0.23	0.16
$h_i^{\mathsf{BINOM}}(D)$	0.66	0.50	0.34	0.91	0.31	0.25	0.23

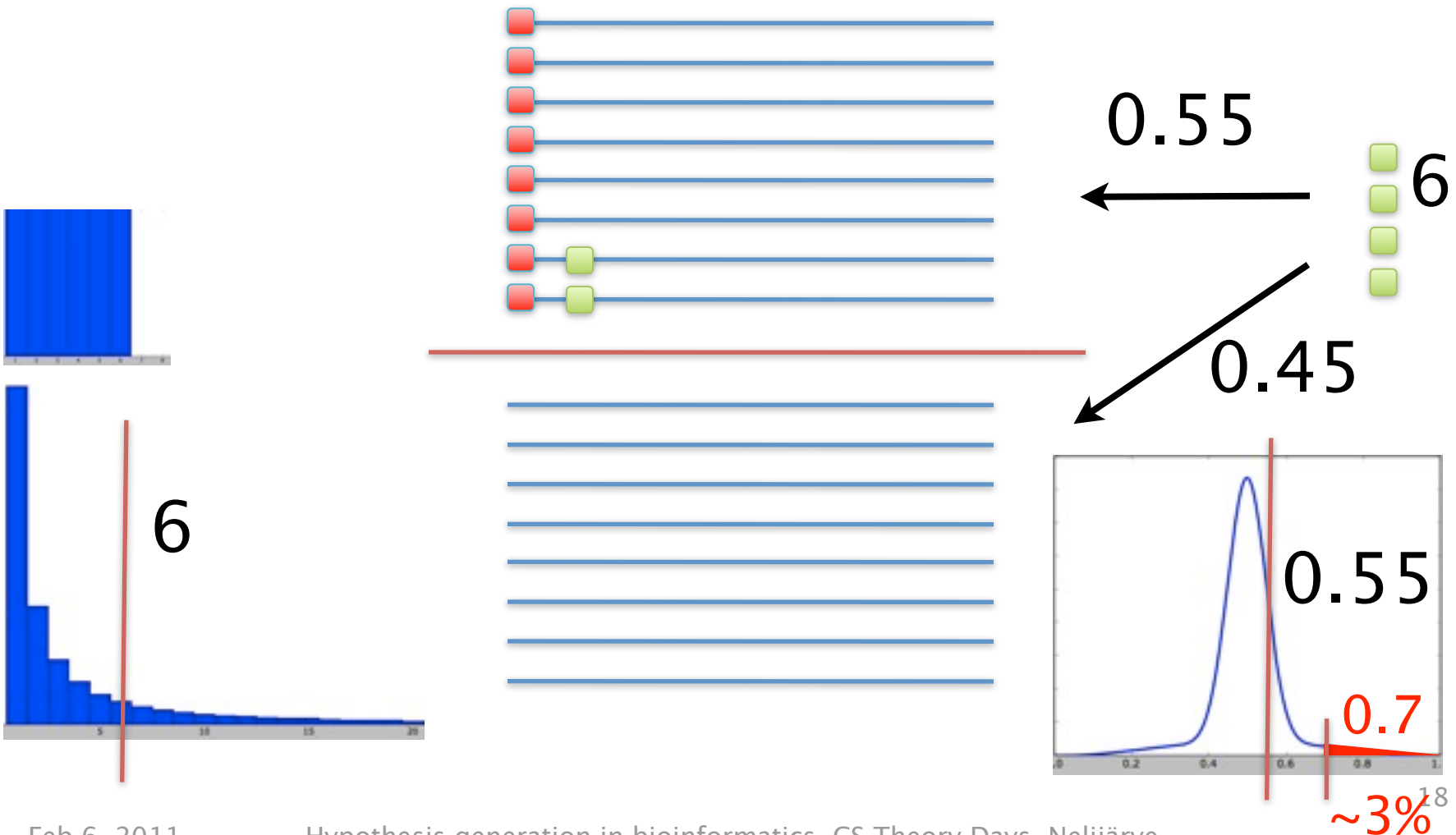
Artificial data: 10000 cases, 10000 features



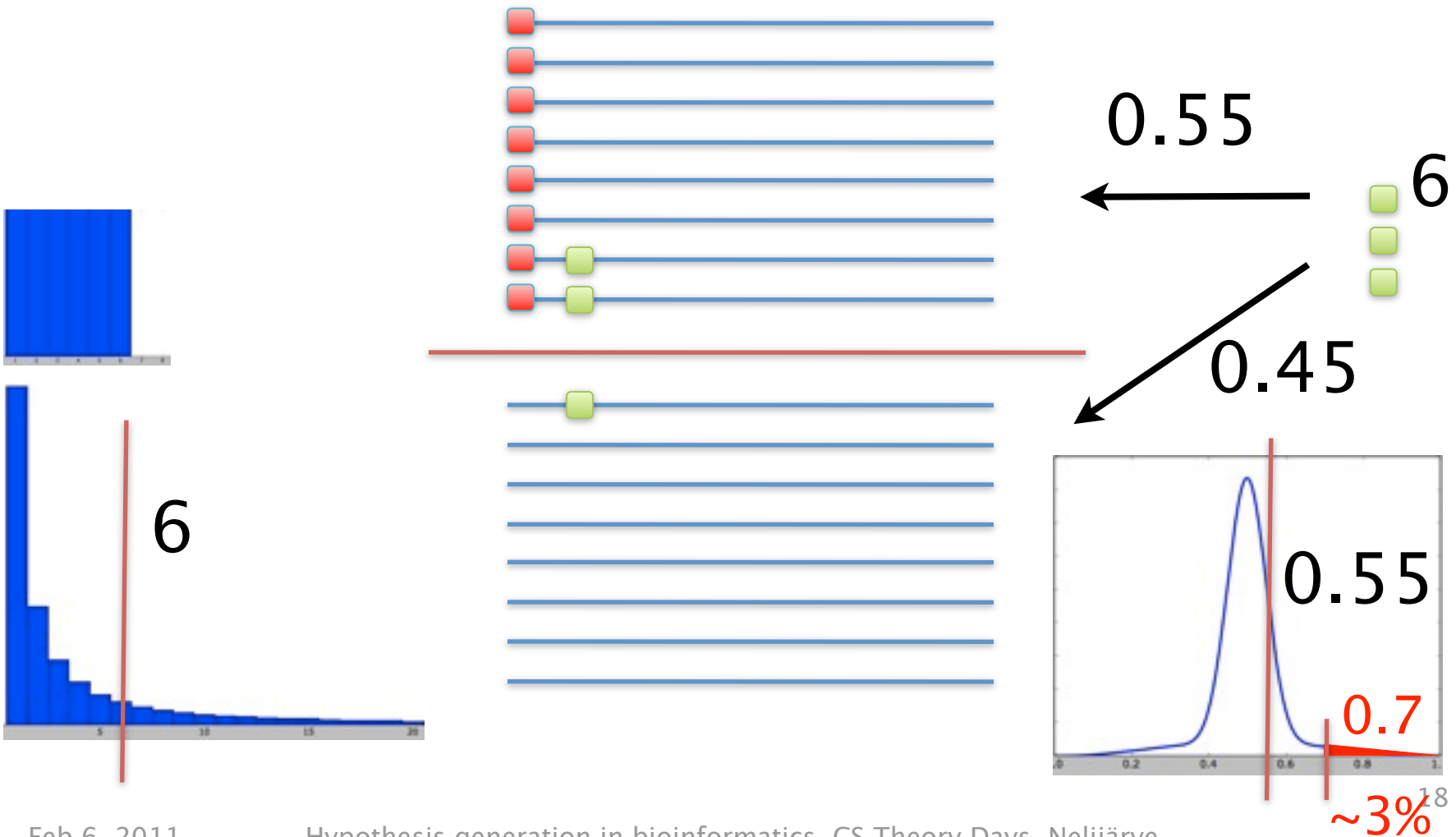
Artificial data: 10000 cases, 10000 features



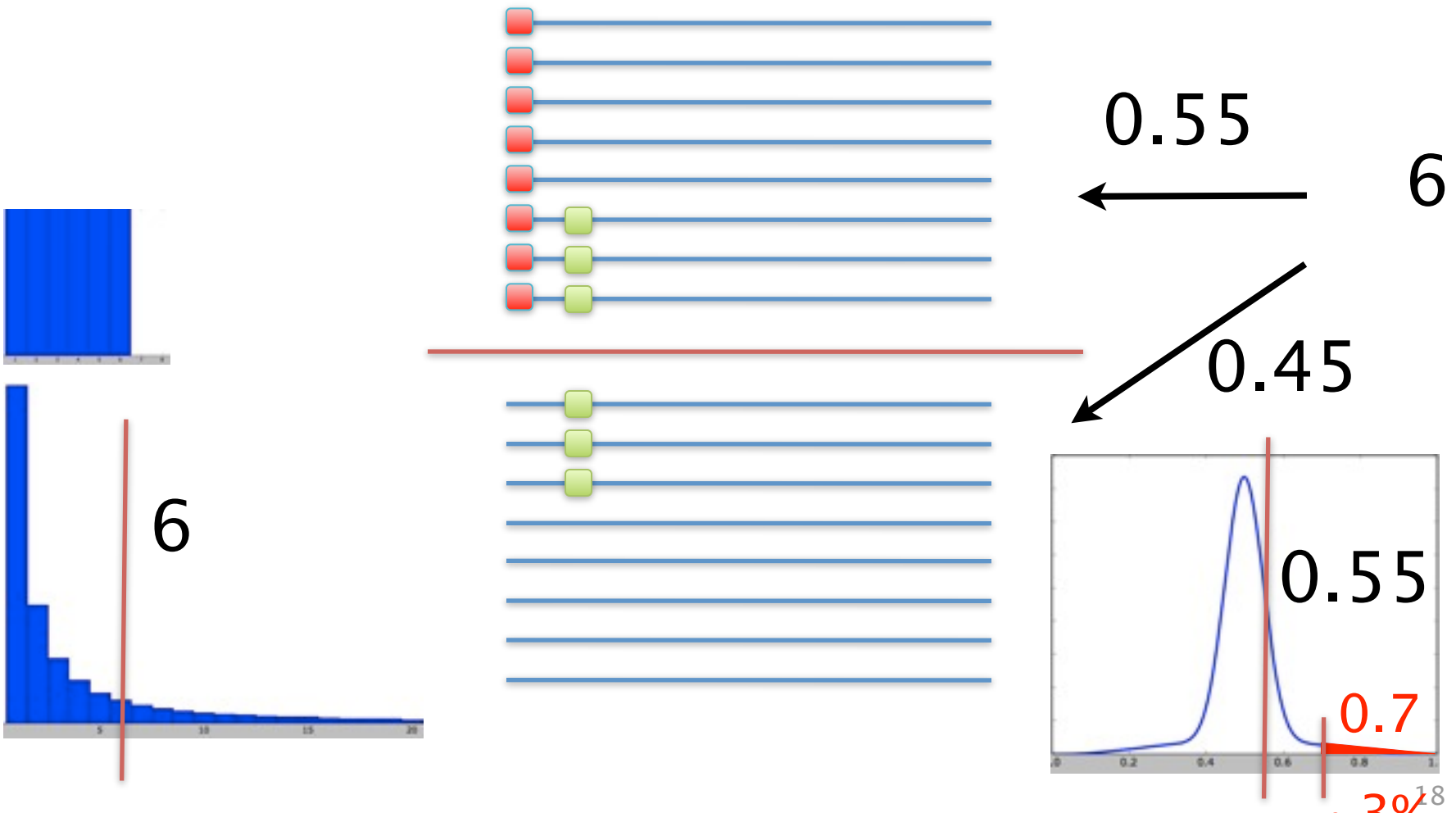
Artificial data: 10000 cases, 10000 features



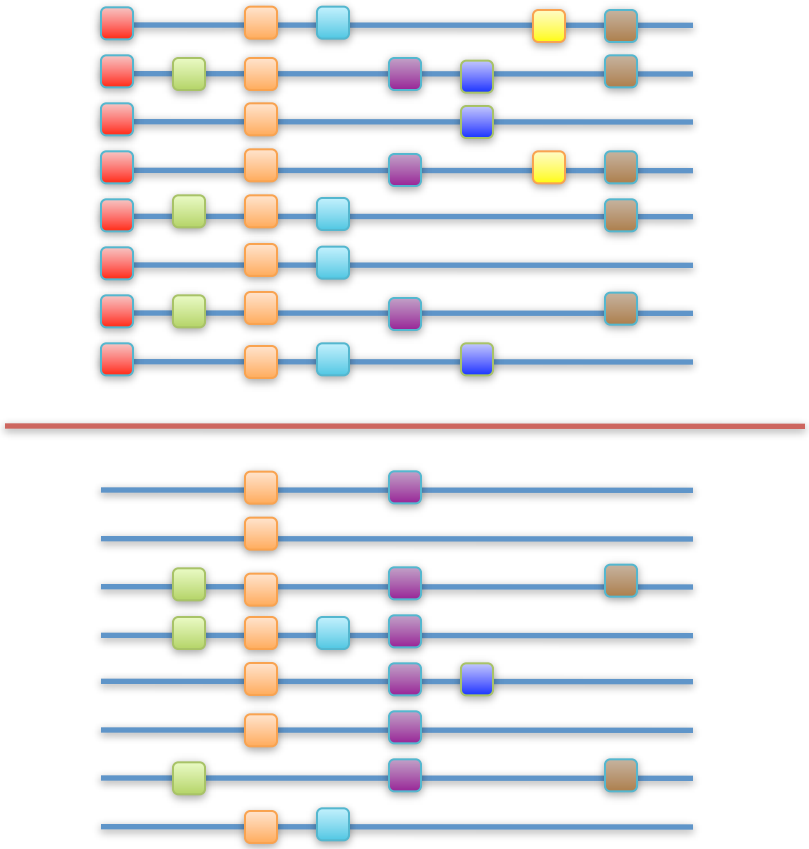
Artificial data: 10000 cases, 10000 features



Artificial data: 10000 cases, 10000 features



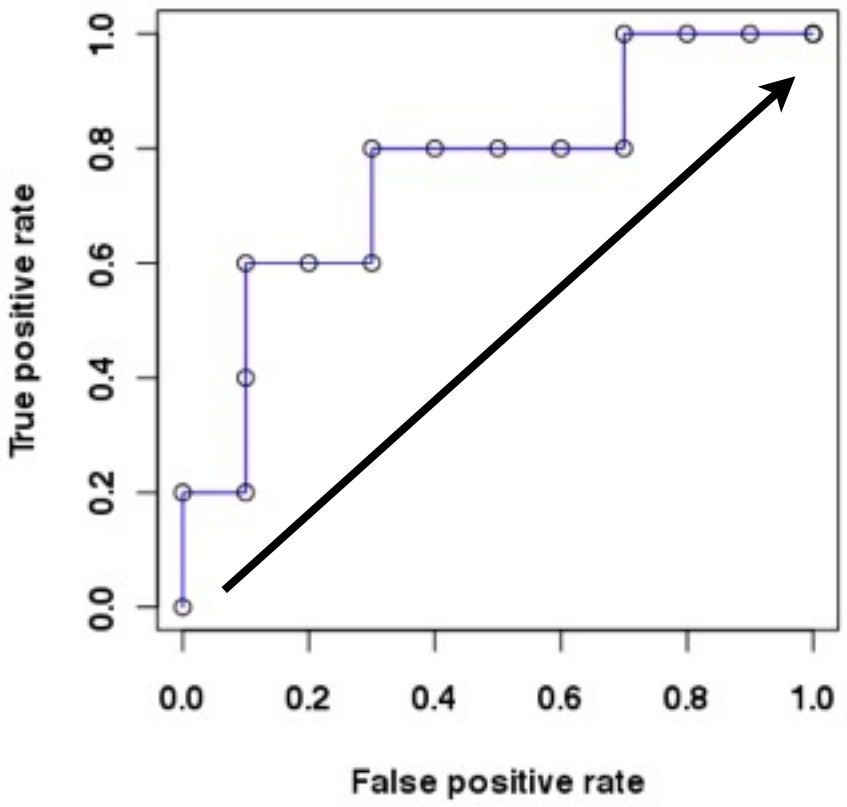
Artificial data



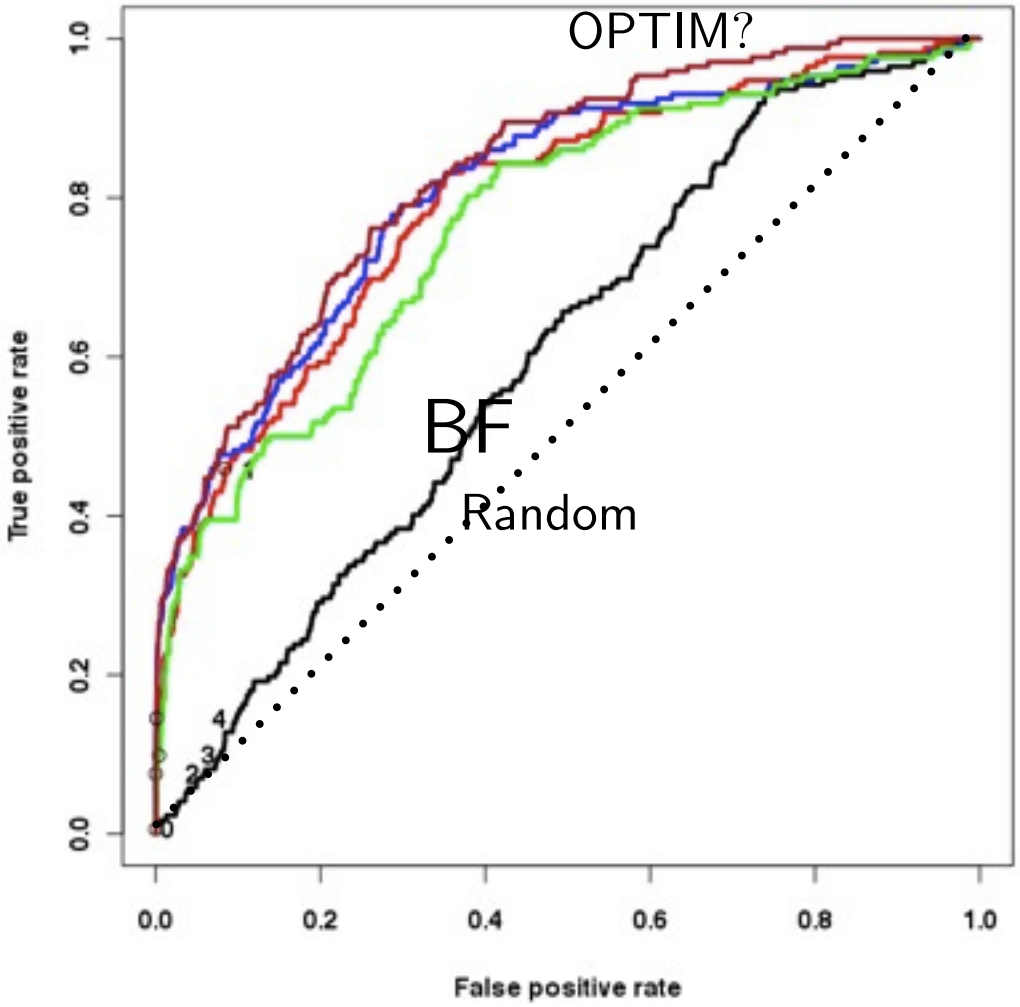
	i=1	i=2	i=3	i=4	i=5	i=6	i=7
D_i^{F}	3	8	4	3	3	2	5
D_i^{B}	3	7	2	6	1	0	2
$h_i^{\mathsf{FG}}(D)$	3	8	4	3	3	2	5
$h_i^{\mathsf{BIAS}}(D)$	0.50	0.53	0.67	0.33	0.75	1.00	0.71
$h_i^{\mathsf{HYPER}}(D)$	0.70	0.50	0.30	0.98	0.28	0.23	0.16
$h_i^{\mathsf{BINOM}}(D)$	0.66	0.50	0.34	0.91	0.31	0.25	0.23
True bias	**0.46**	**0.53**	**0.52**	**0.44**	**0.80**	**0.55**	**0.74**
True bias ≥ 0.7	–	–	–	–	+	–	+

ROC-curve: Receiver Operating Characteristic

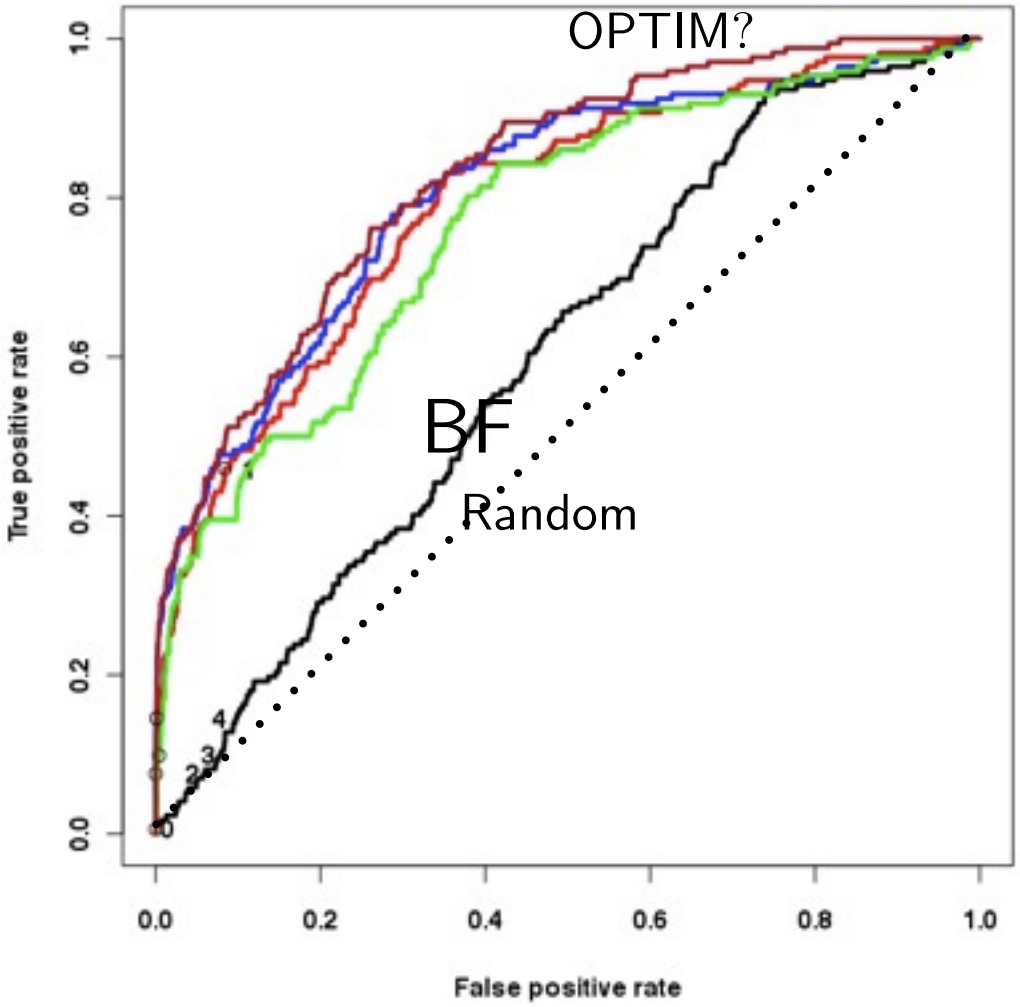
+ − + + − − + − − − − + − − −



Comparison

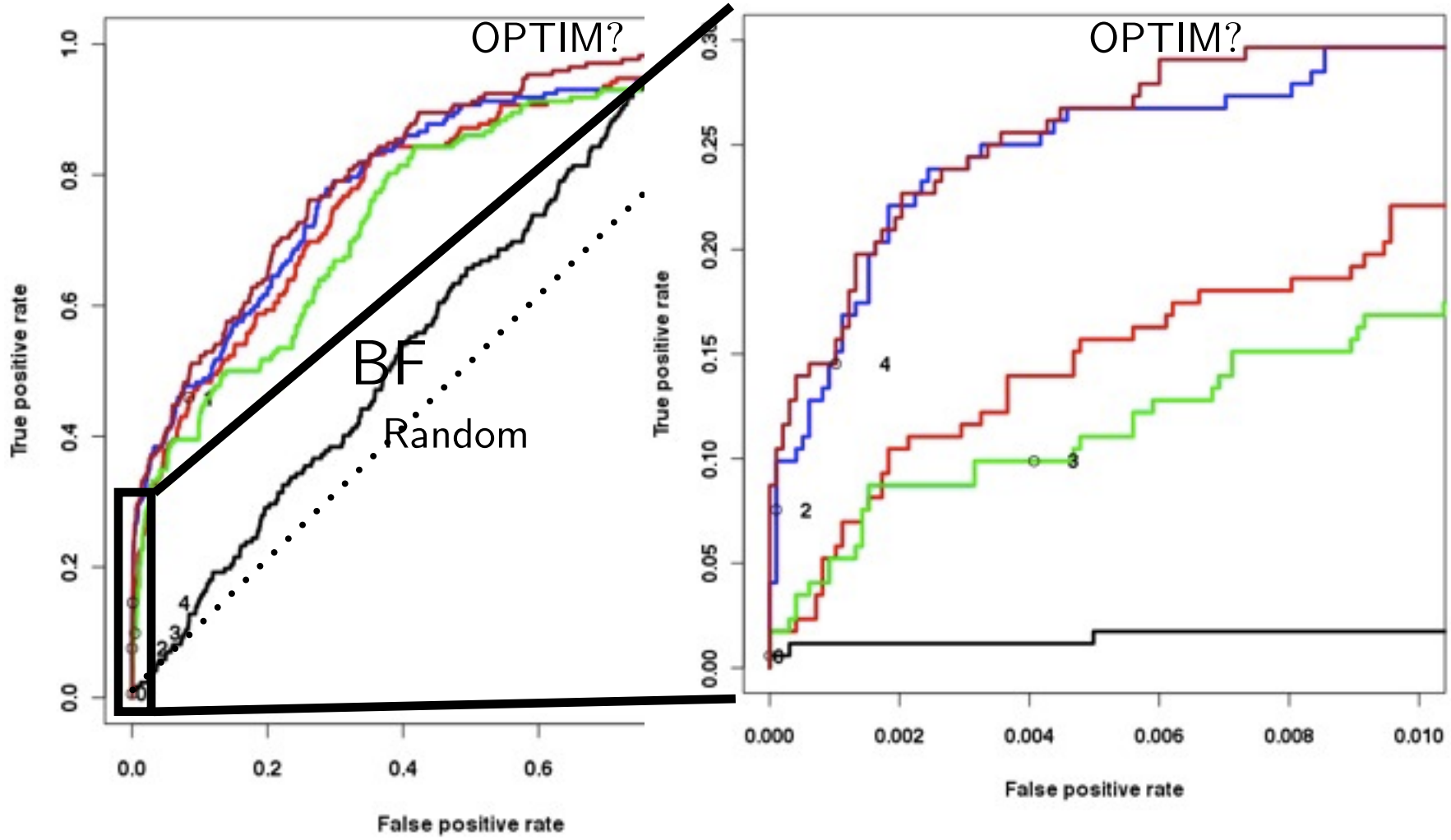

OPTIM?
BF
Random
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

Comparison

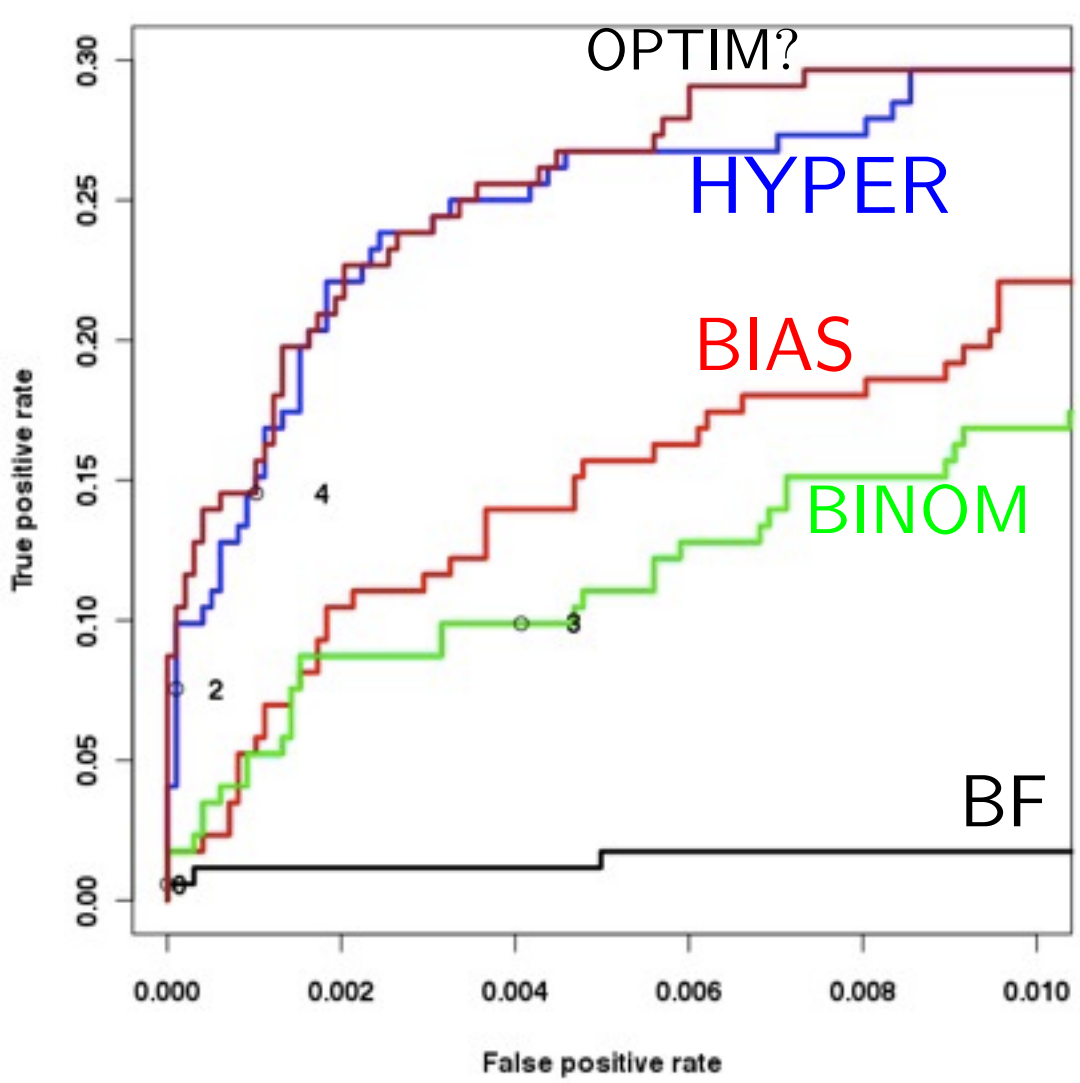


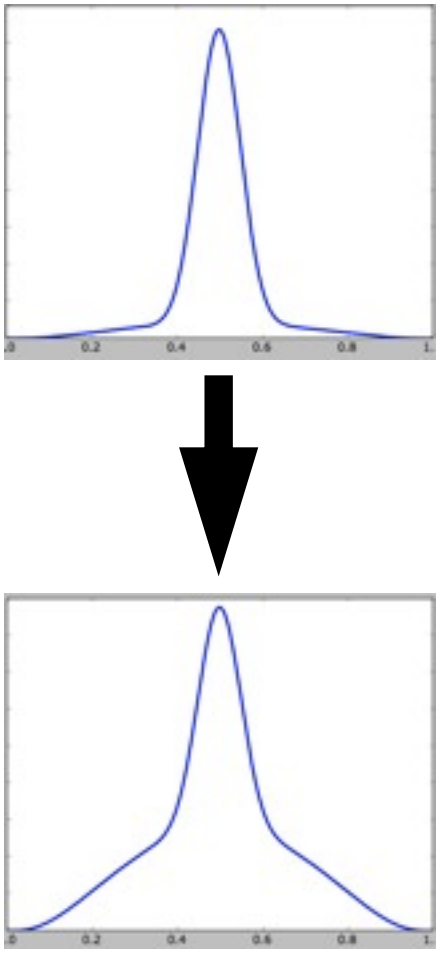
OPTIM?

Comparison

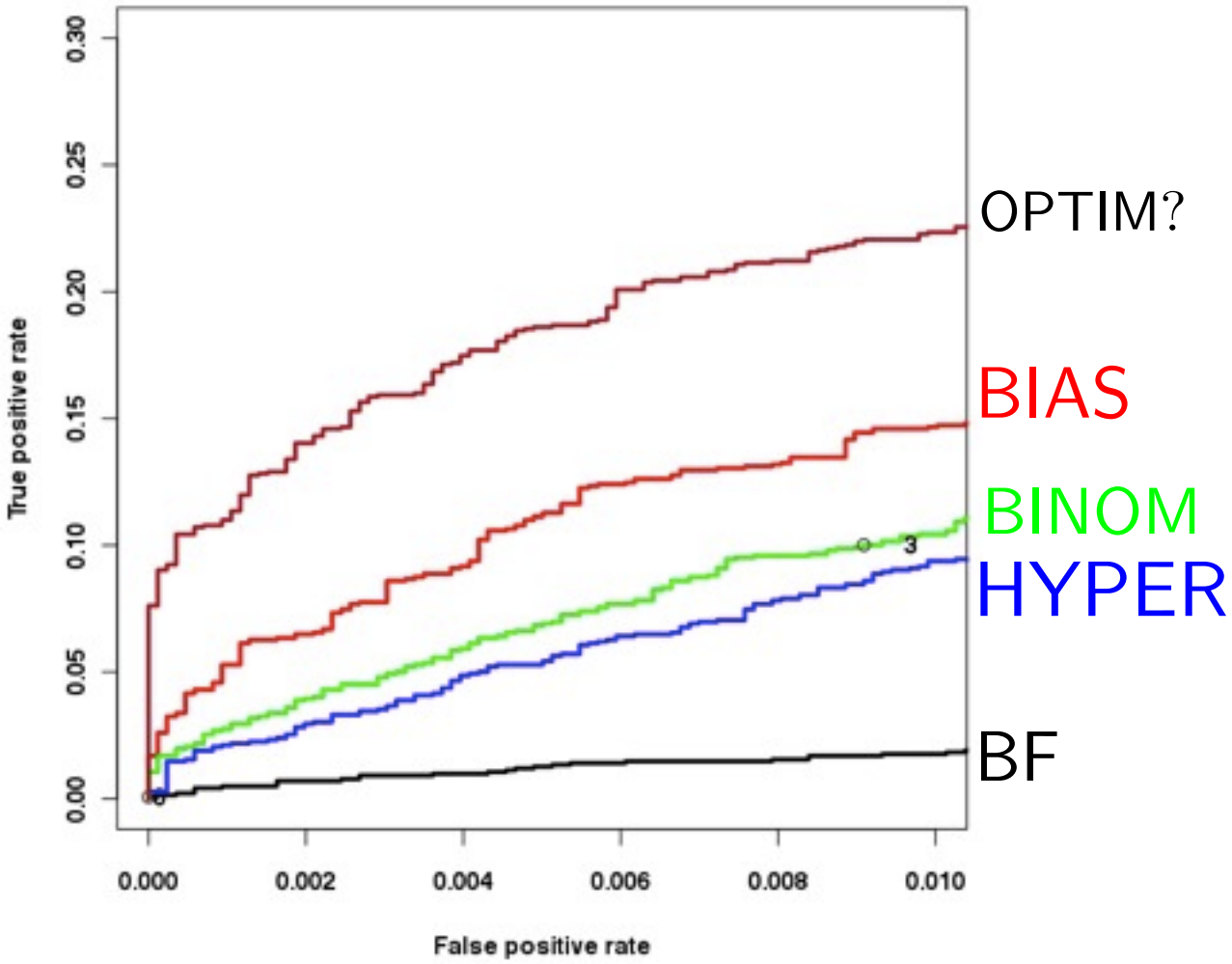

OPTIM?
BF
Random
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
OPTIM?
True positive rate
False positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.000
0.002
0.004
0.006
0.008
0.010
1
2
3
4
5
6

Comparison


OPTIM?
HYPER
BIAS
BINOM
BF
True positive rate
False positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.000
0.002
0.004
0.006
0.008
0.010
2
3
4



Comparison


True positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.000
0.002
0.004
0.006
0.008
0.010
False positive rate
3
OPTIM?
BIAS
BINOM
HYPER
BF

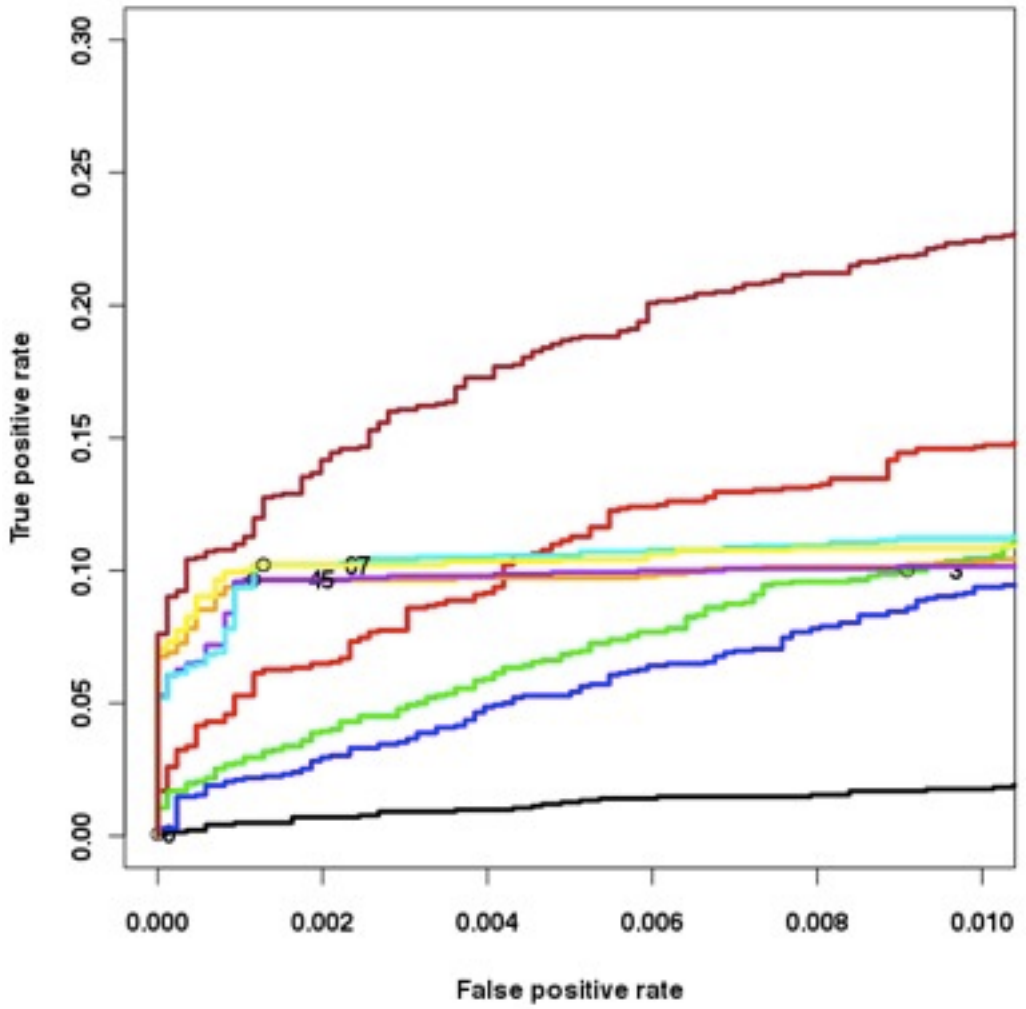
Comparison with combinations

$BIAS^{HYPER}$

$BIAS^{BINOM}$

$HYPER^{BIAS}$

$BINOM^{BIAS}$



OPTIM?

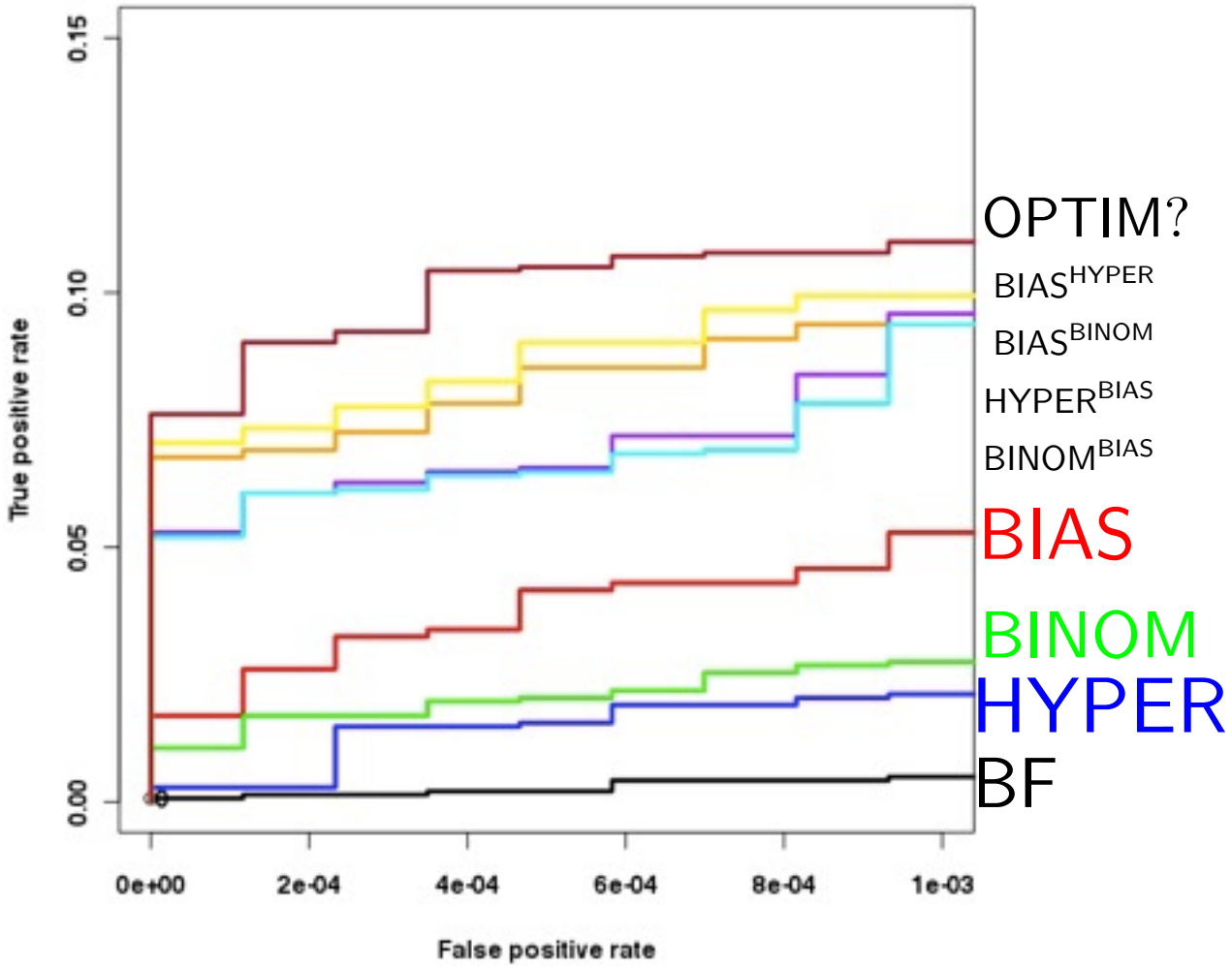
BIAS

BINOM

HYPER

BF

Comparison with combinations



Application in bioinformatics

Cluster 1 - up day 6

Motif logo	Database identifier	Factor family	Conser-vation	Targets in cluster	Enrichment ratio	Enrichment *P*-value
	M01003	IKZF2	0.9	8	10.53	4.51E-07
	M00982	EGR	none	7	10.39	8.40E-06
	M00106	CUX1	1.0	7	8.03	5.26E-06
	M00638	HNF4	0.8	22	4.05	6.09E-08
	M003					5.66E-06
	M010					1.85E-06
	M007					4.55E-06
	M009					8.18E-06
	MA00					6.27E-06
	M00058	HEN1	0.7	44	1.99	6.53E-06
	M01028	REST	0.7	51	1.97	1.34E-06
	M00444	VDR	none	47	1.97	3.41E-06
	M00244	NGFI-C	0.7	51	1.89	2.03E-06
	M00687	αCP1	0.7	46	1.88	9.61E-06
	M01100	LRF	none	57	1.85	1.40E-06
	MA0105	NFκB1	none	53	1.81	7.20E-06

BIAS^{HYPER}

Billon et al. Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. Genome Biol (2010) vol. 11 (8) pp. R80

Summary

- We have presented a general framework for working with hypotheses in bioinformatics
- We have compared some statistics in the context of finding functionally related binary features
- The combined statistics work better on our artificial data
- It is possible to do even better, if more is known about the data generating model

Thanks