

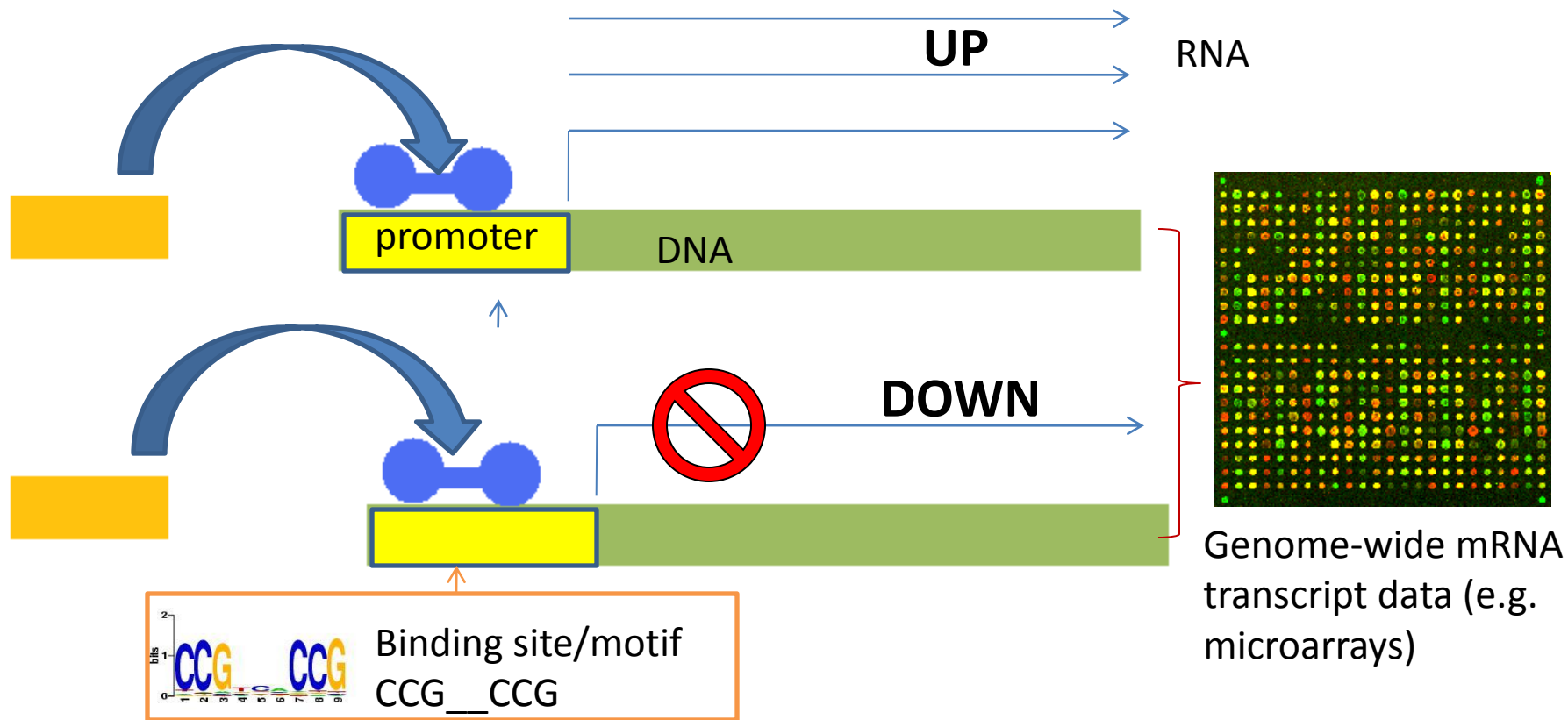
Boosting and ADTs

Tony Boyles and Sweta Sengupta



Transcriptional Regulation

- **Upregulation**- a signal causes the expression of more genes
- **Downregulation**- decreased gene expression



Transcriptional Regulation

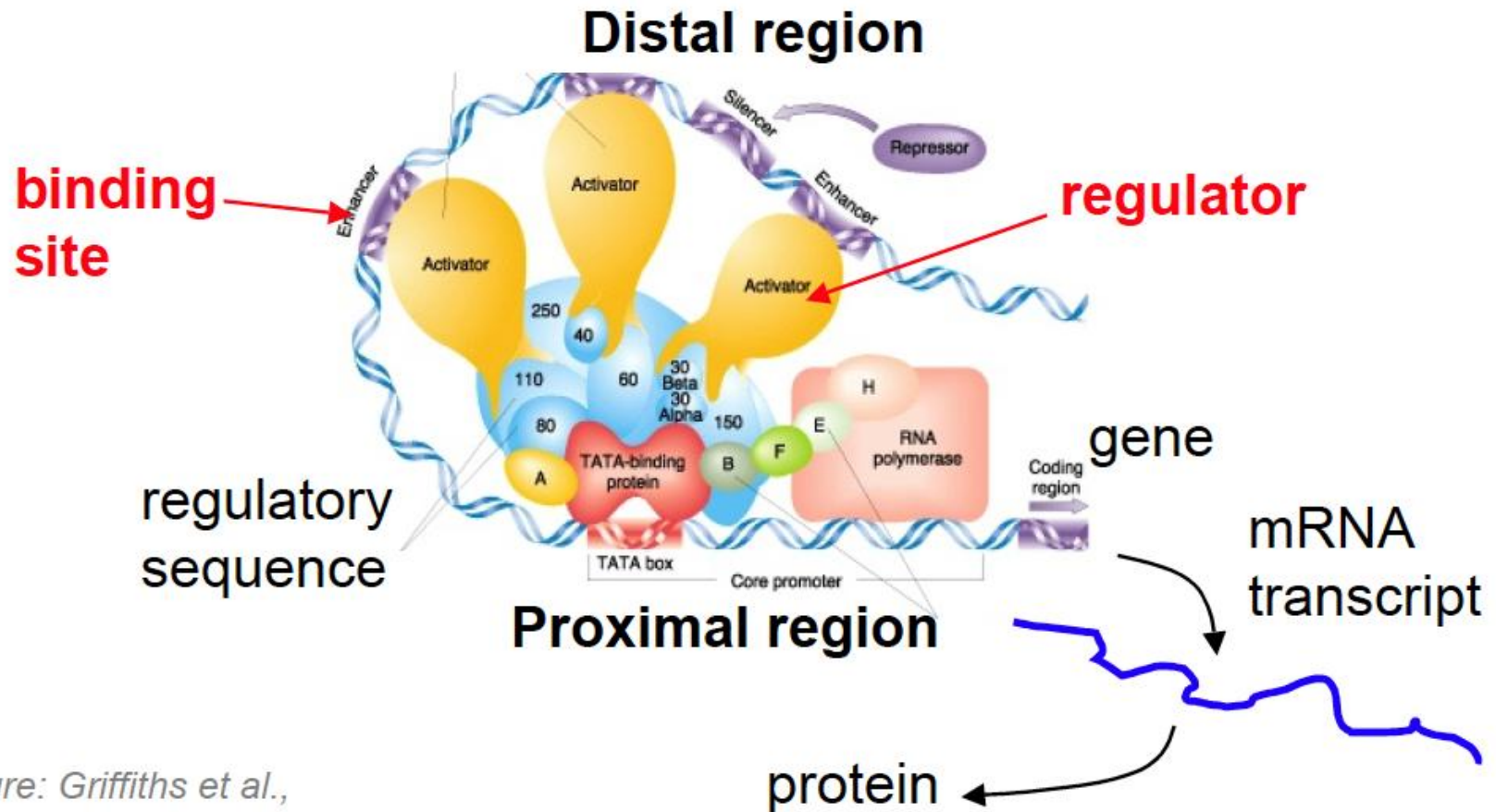
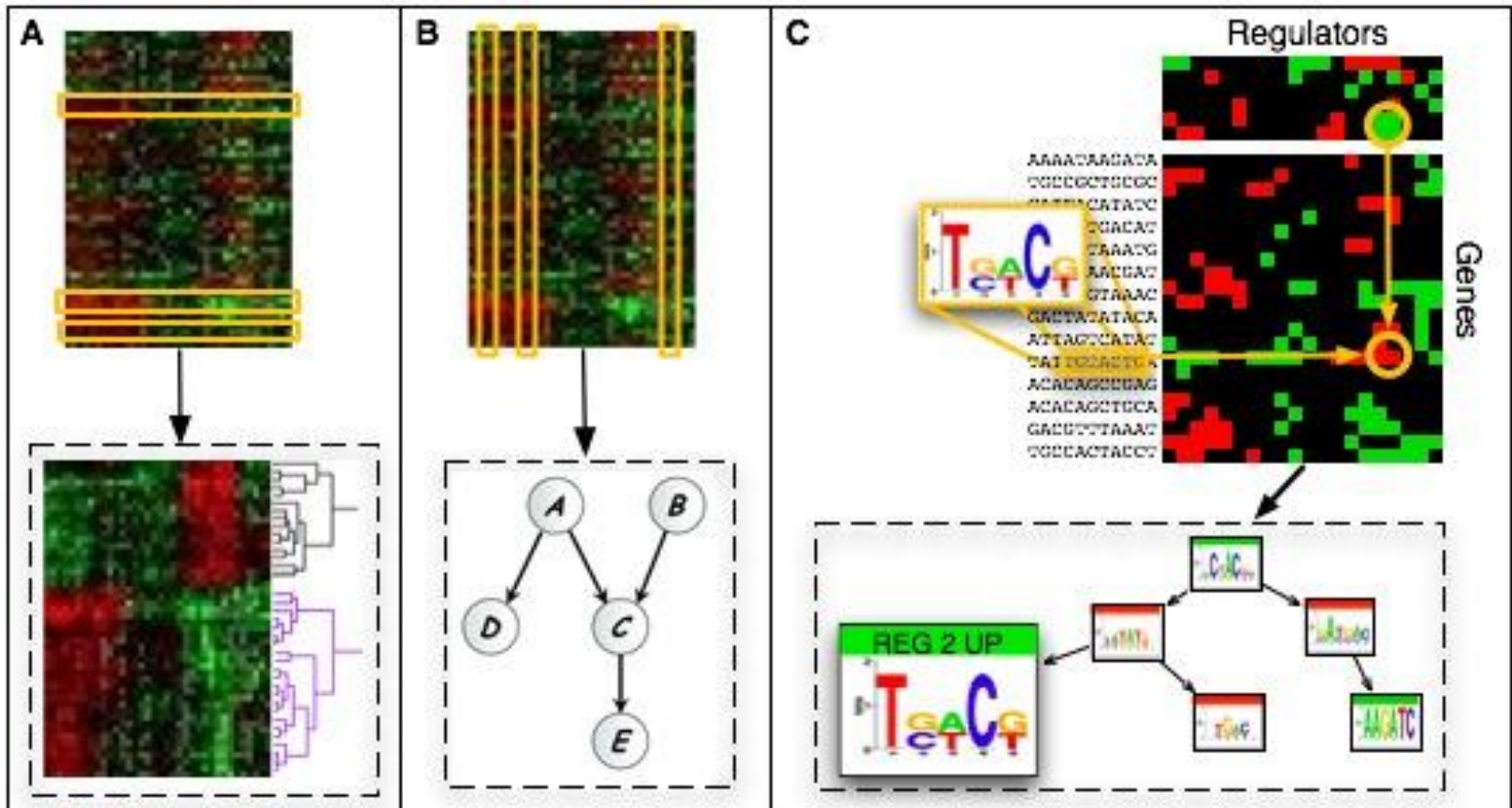


Figure: Griffiths et al.,
"Modern Genetic Analysis"

Clustering, Bayesian analysis, MEDUSA



Horse Gambler



Table 2
Correlations Among Race Horse Characteristics (N = 50)

Characteristics	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Lifetime earnings														
2. Lifetime speed	.36													
3. % Races came in money	.04	-.11												
4. Current jockey ability	.23	.06	.08											
5. Purse size	.75	.28	.05	.49										
6. Position of finish	.27	-.18	.20	.21	.41									
7. Pay off/dollar waged	.05	.13	-.34	.04	.03	-.10								
8. Overall speed	.10	.29	-.16	.01	.14	-.46	-.12							
9. First quarter-mile speed	-.17	-.13	-.07	.14	.00	.09	.15	-.31						
10. Last quarter-mile speed	-.03	-.17	-.05	-.02	.09	.35	-.06	-.16	-.06					
11. Number of moves	.31	.01	-.06	.30	.42	.43	-.10	-.26	.35	.36				
12. Racetrack size	-.01	.20	-.09	-.06	-.01	-.26	.06	.19	.12	-.13	-.10			
13. Track surface condition	-.43	-.18	-.16	-.32	-.52	-.44	-.07	.31	-.20	-.02	-.34	-.01		
14. Jockey ability	.27	.00	.06	.45	.36	.15	.23	.14	.18	-.27	.09	-.19	-.26	
15. Interactive model variable	.68	.36	.18	.57	.89	.38	-.02	.09	.03	-.09	.33	-.05	-.64	.37

Note. Characteristics 1–3 are from the horse’s career; 5–15 are from the prior race.

Ceci, Stephen J.; Liker, Jeffrey K. A day at the races: A study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General*. Vol 115(3), Sep 1986, 255-266.

Adaboost

Given: $(x_1, y_1), \dots, (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{-1, +1\}$

Initialize $D_1(i) = 1/m$.

For $t = 1, \dots, T$:

- Train weak learner using distribution D_t .
- Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error

$$\epsilon_t = \Pr_{i \sim D_t} [h_t(x_i) \neq y_i].$$

- Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$.

Adaboost

- Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$.
- Update:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} \\ &= \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \end{aligned}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$

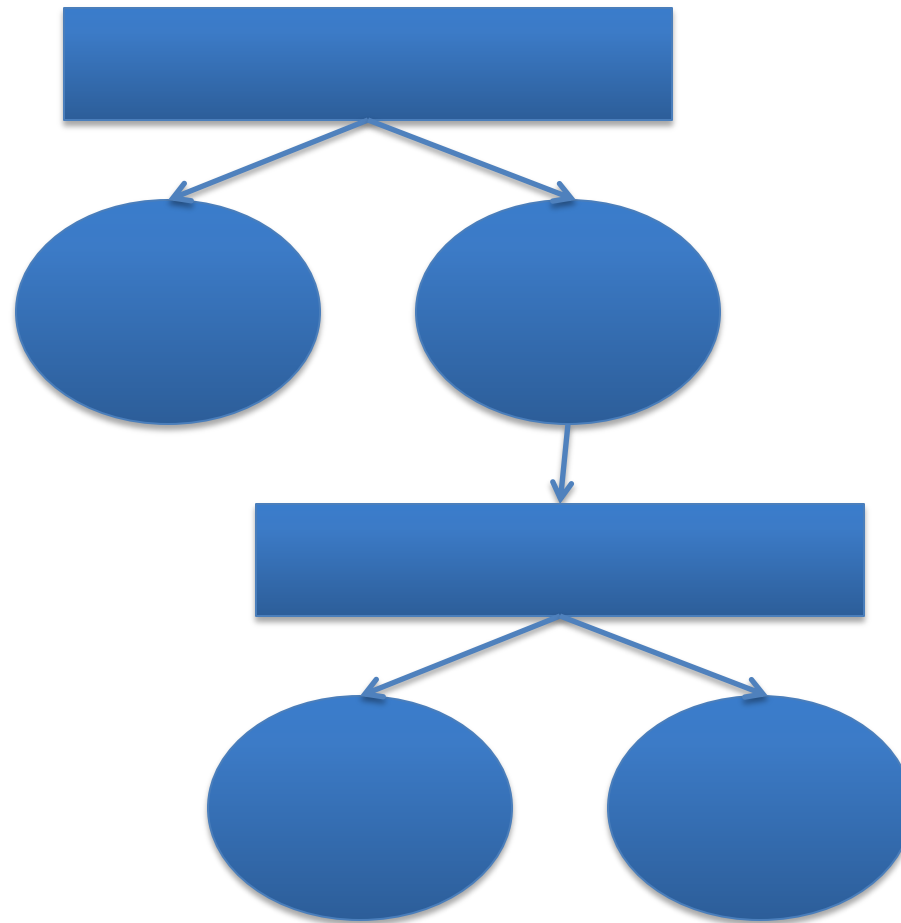
Binary vs. Multiclass Boosting

- y or y' ?



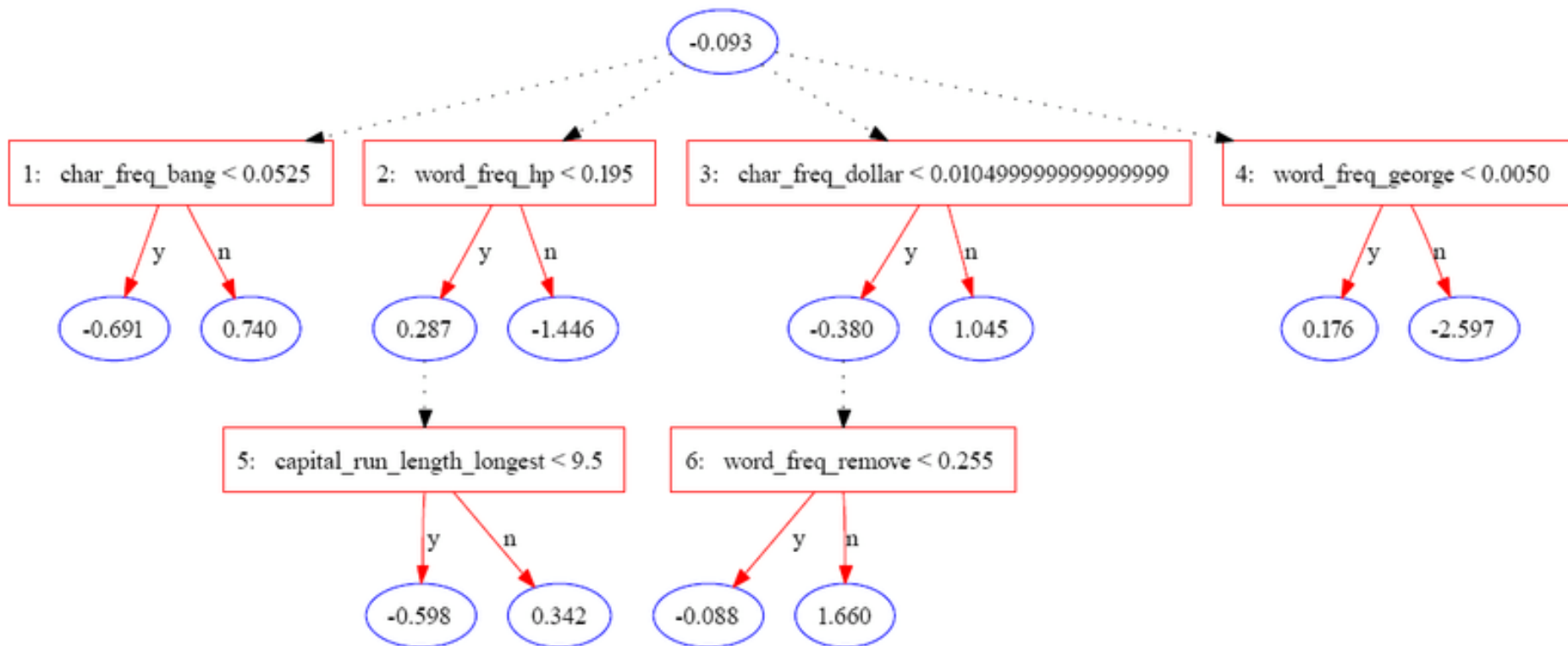
- Noise Problems
- Increased Complexity

Alternating Decision Tree



Freund, Y. and Mason, L. (1999) The alternating decision tree learning algorithm. *Proceedings of the Sixteenth International Conference on Machine Learning, Morgan Kaufmann, pp. 124–133.*

Alternating Decision Tree



MEDUSA

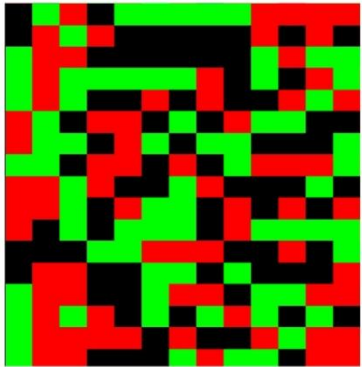
Motif Element Discrimination Using Sequence Agglomeration

- **Medusa is a machine learning algorithm – predict differential expression of target gene**
 - Data Source:
 - mRNA expression
 - promoter sequence
 - ChIP-chip occupancy data
- MEDUSA does not rely on clustering or correlation of expression profiles to infer regulatory
- up/down expression of target genes by identifying condition-specific regulators and discovering DNA motifs, de novo from the promoter sequences, that may mediate their regulation of targets.

Design

Training data

discretized expression of target genes



discretized expression of regulators



5' UTR promoter sequences

G_1 : ...ACCTAGCTTCA...

G_2 : ...CTAGGCCATAA...

G_3 : ...AATTTAAACGT...

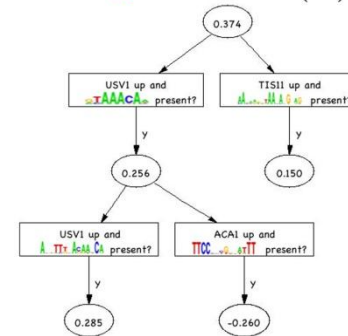
G_n : ...TCCATGGATCA...

Labels +1/-1

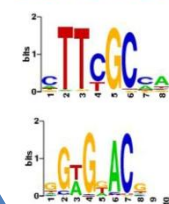
Learning Algorithm (Boosting)

feature vectors \vec{x}

Model of transcriptional regulation $F(\vec{x})$



List of motifs

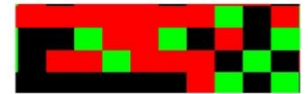


Target gene analysis, important regulators

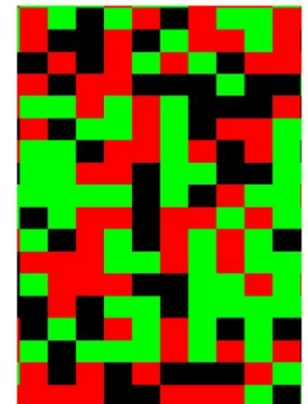
TPK1, USV1, AFR1, XBP1, ...

Test data

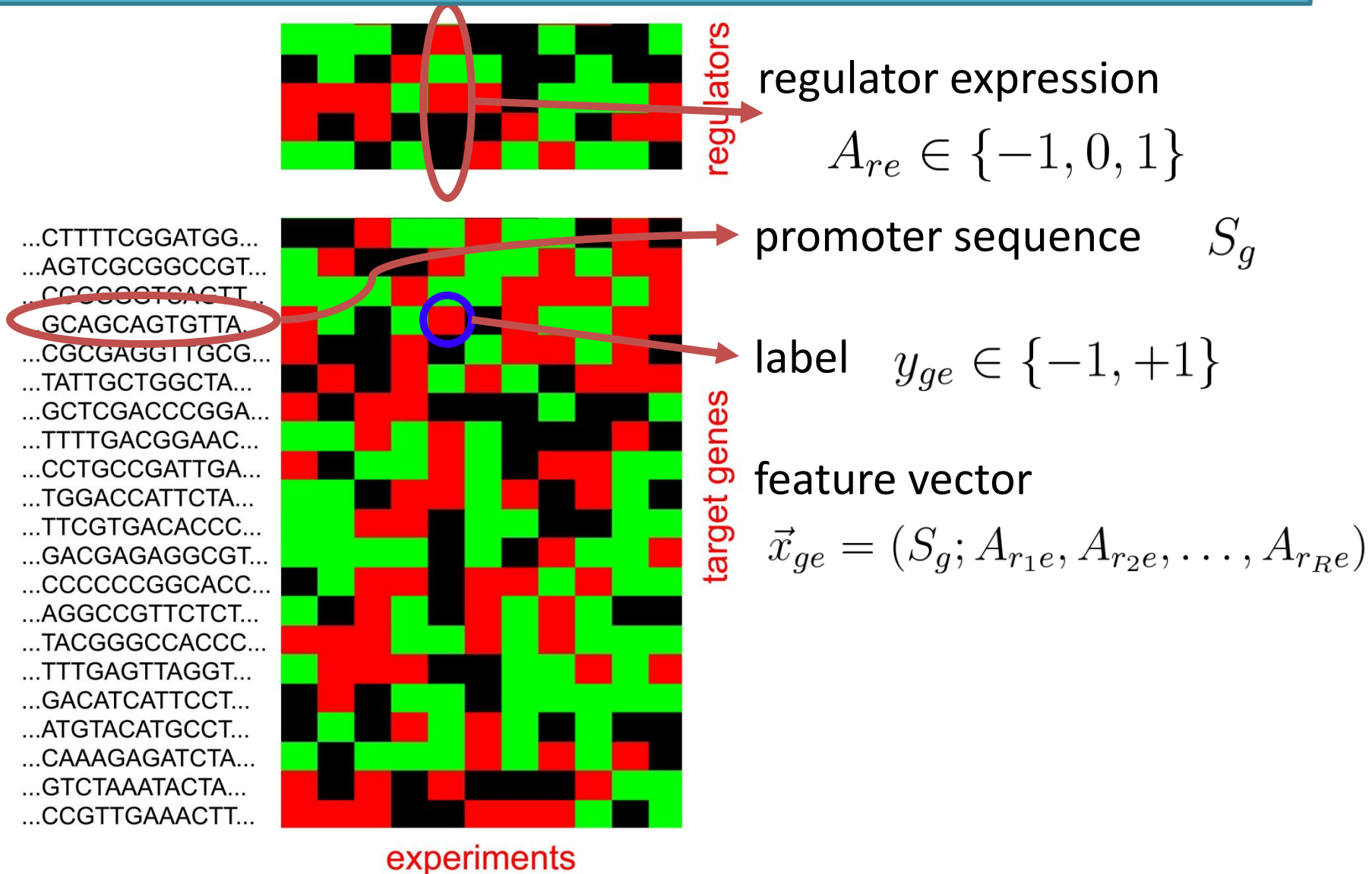
expression of regulators in held-out experiments



predicting target gene expression in held-out experiments



Training Data Input



Weak learner

...AGCTATGCCATCGACTGCTCCAGTCGCACACACAAAGATTTGAG
GCTATAGCTACTTTATAAAGGGGCTACGGCAAATT...

k-mers ($k \leq 7$)

AGCTATG
GCTATGC
CTATGCC
•
•
•

dimers (gapped elements)

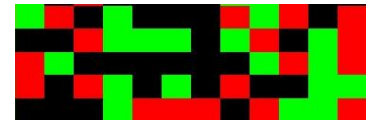
TTT_AAA
GCTA_GCTA
•
•
•



Is AGCTATG present and USV1 up?
Is AGCTATG present and USV1 down?
Is GCTATGC present and USV1 up?
Is GCTATGC present and TPK1 up? ...

*try all motif-regulator
pairs as weak rules ...*

Regulator expression



Weak learner

...AGCTATGCCATCGACTGCTCCAGTCGCACACACAAAGATTTGAG
GCTATAGCTACTTTATAAAGGGGCTACGGCAAATT...

k-mers ($k \leq 7$)

AGCTATG

GCTATGC

CTATGCC



minimizes boosting loss

Is GCTATGC present and USV1 up?

dimers (gapped elements)

TTT_AAA

GCTA_GCTA



Is AGCTATG present and USV1 up?
Is AGCTATG present and USV1 down?
Is GCTATGC present and USV1 up?
Is GCTATGC present and TPK1 up? ...

*try all motif-regulator
pairs as weak rules ...*

Regulator expression



Agglomeration

boosting loss



- Is GCTATGC present and USV1 up?
- Is GCAATGC present and USV1 up?
- Is TCTATGC present and USV1 up?
- Is GCTTTGC present and USV1 up?
- ...

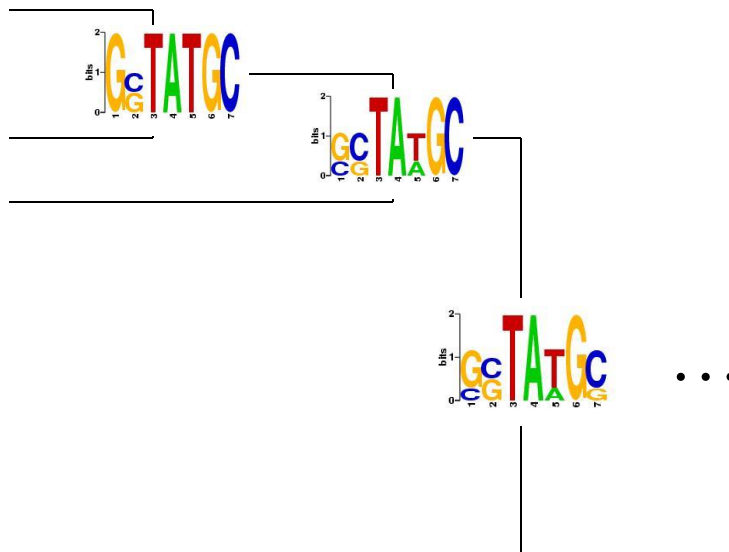
Agglomerate

GCTATGC
GCAATGC
GGTATGC
CCTAAGC
GCTATTT

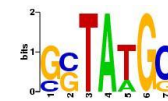
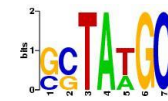
...

...

GGTATGG



PSSMs



...

Agglomeration

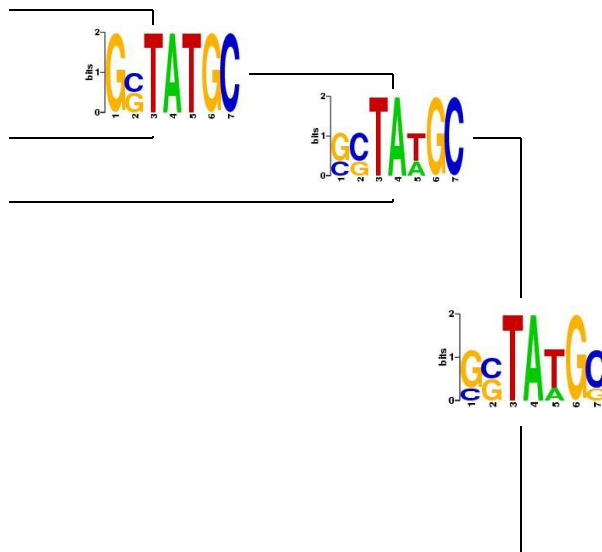
boosting loss ↓

- Is GCTATGC present and USV1 up?
- Is GCAATGC present and USV1 up?
- Is TCTATGC present and USV1 up?
- Is GCTTTGC present and USV1 up?
- ...

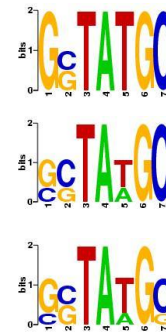
Agglomerate

Optimize over offsets when merging *k*-mers/PSSMs:
 -- GCTATGC
 GCTATTT --

GCTATGC
 GCAATGC
 GGTATGC
 CCTAAGC
 GCTATTT



PSSMs



...
 ...
 GGTATGG

Agglomeration

boosting loss



- Is GCTATGC present and USV1 up?
- Is GCAATGC present and USV1 up?
- Is TCTATGC present and USV1 up?
- Is GCTTTGC present and USV1 up?
- ...



2 PSSMs p and q

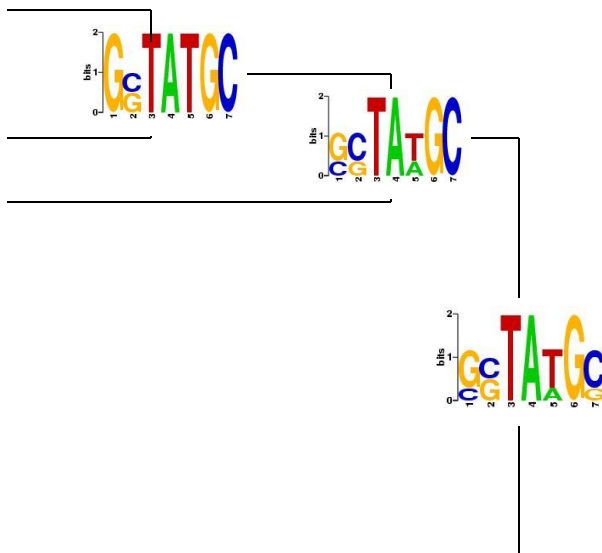
$$d(p, q) \equiv \min_{\text{offsets}} [w_1 D_{KL}(p || w_1 p + w_2 q) + w_2 D_{KL}(q || w_1 p + w_2 q)]$$

GCTATGC
 GCAATGC
 GGTATGC
 CCTAAGC
 GCTATTT

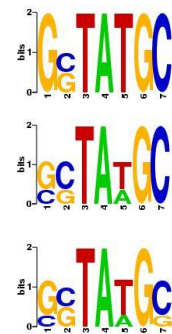
...

...

GGTATGG



PSSMs



...




Agglomeration

boosting loss



- Is GCTATGC present and USV1 up?
- Is GCAATGC present and USV1 up?
- Is TCTATGC present and USV1 up?
- Is GCTTTGC present and USV1 up?
- ...



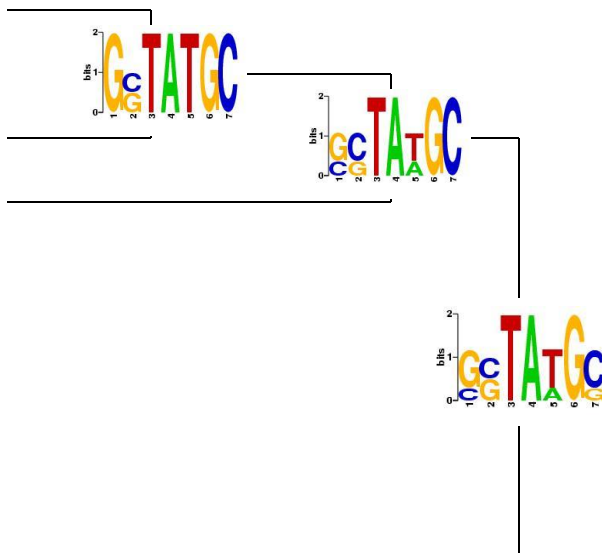
- Is  present and USV1 up?
- Is  present and USV1 up?
- Is  present and USV1 up? ...

GCTATGC
 GCAATGC
 GGTATGC
 CCTAAGC
 GCTATTT

...

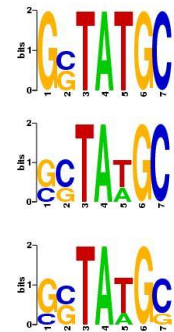
...

GGTATGG



...

PSSMs






...

Agglomeration

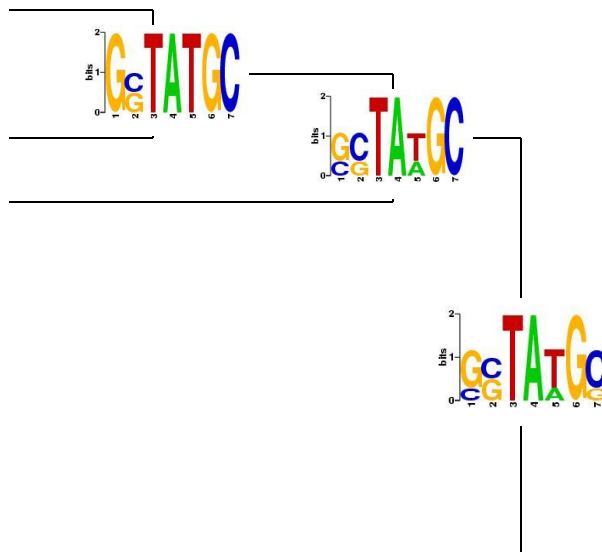
boosting loss

- Is GCTATGC present and USV1 up?
- Is GCAATGC present and USV1 up?
- Is TCTATGC present and USV1 up?
- Is GCTTTGC present and USV1 up?
- ...

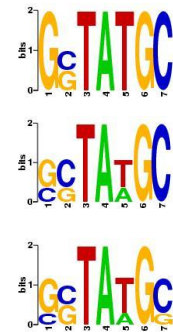
minimize boosting loss
 \Rightarrow *final weak rule*

- Is  present and USV1 up?
- Is  present and USV1 up?
- Is  present and USV1 up? ...

GCTATGC
 GCAATGC
 GGTATGC
 CCTAAGC
 GCTATTT



PSSMs



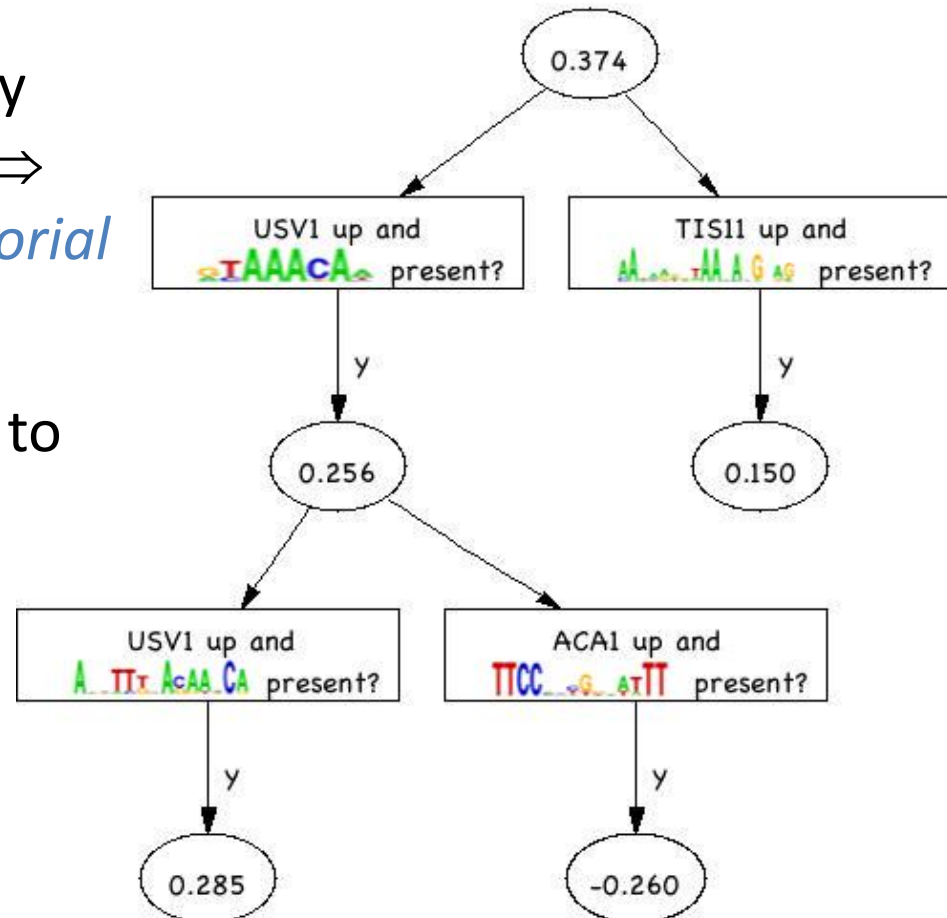
...
 ...

GGTATGG

...

Result

- *Combine weak rules* into an ADT
- Lower nodes are conditionally dependent on higher nodes \Rightarrow can possibly reveal *combinatorial interactions*
- Able to reveal motifs specific to *subsets of target genes*



Margin-based score

- *margin-based score:*
- Measure the importance of the regulator R

$$\sum_{g \in T, e \in E} y_{ge} \left(F(\vec{x}_{ge}) - F_{f-}(\vec{x}_{ge}) \right)$$

Problems with MEDUSA

- $(N_{k\text{-mers}} + N_{\text{dimers}} + N_{\text{PSSMs}}) * N_{\text{reg}} * 2$
= possible weak rules at every node
- Binary (sort of) evaluation:
Up-, non-, or down-regulation

Problems with MEDUSA

- $(N_{k\text{-mers}} + N_{\text{dimers}} + N_{\text{PSSMs}}) * N_{\text{reg}} * 2$
= possible weak rules at every node
- Binary (sort of) evaluation:
Up-, non-, or down-regulation
- Fixing either exacerbates the other