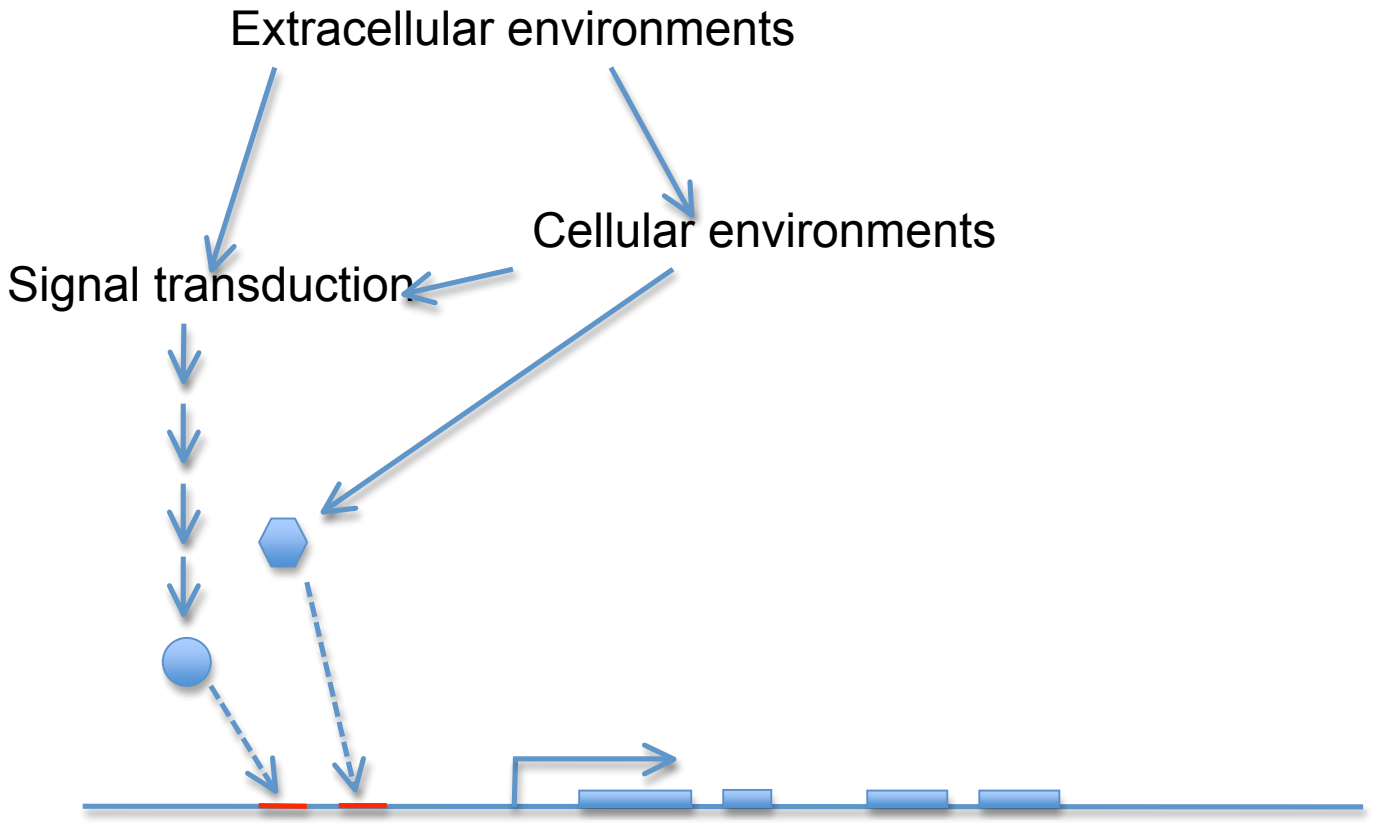
Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*

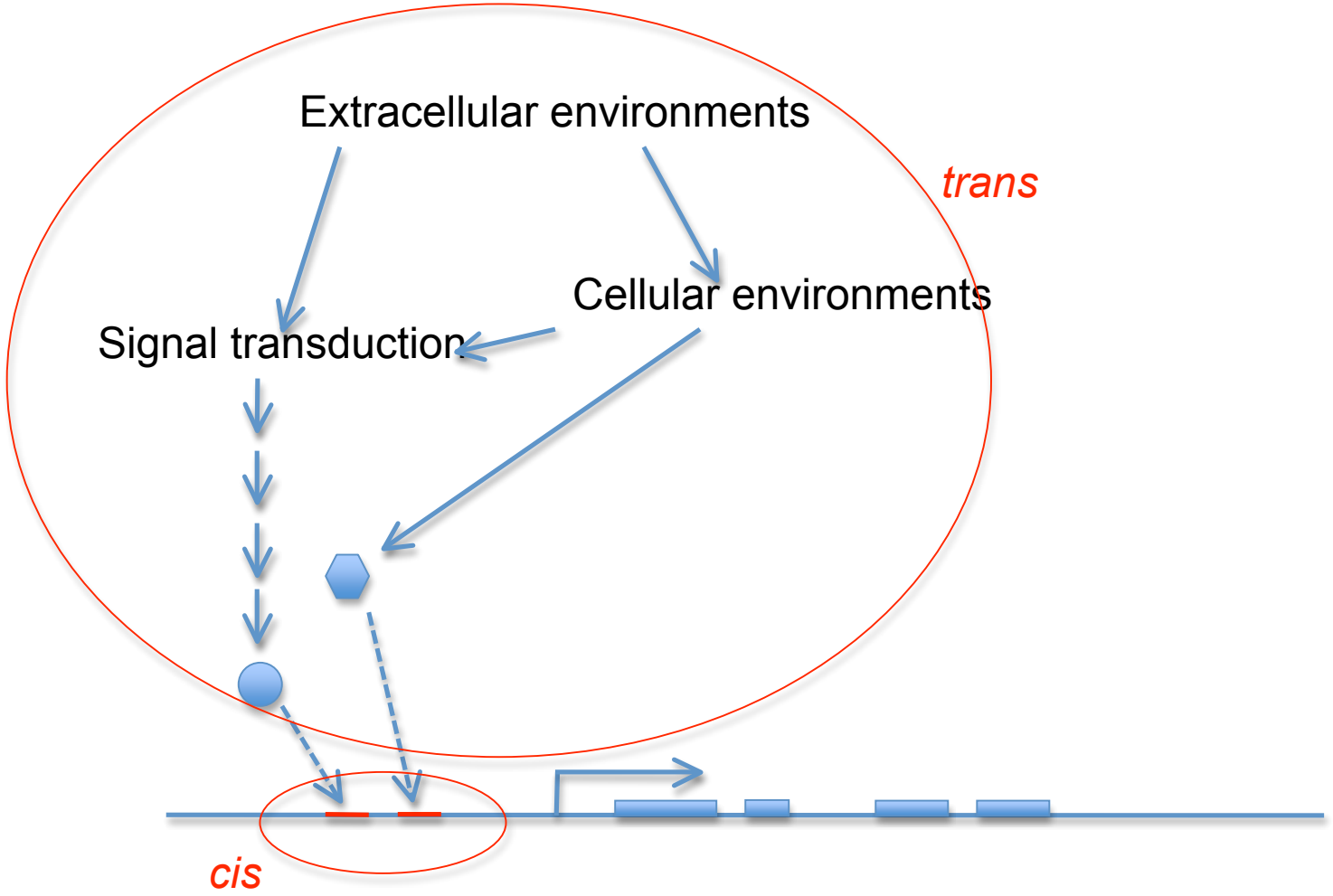
Jason D. Hugher, Preston W. Estep,
Saeed Tavazoie, and George M. Church

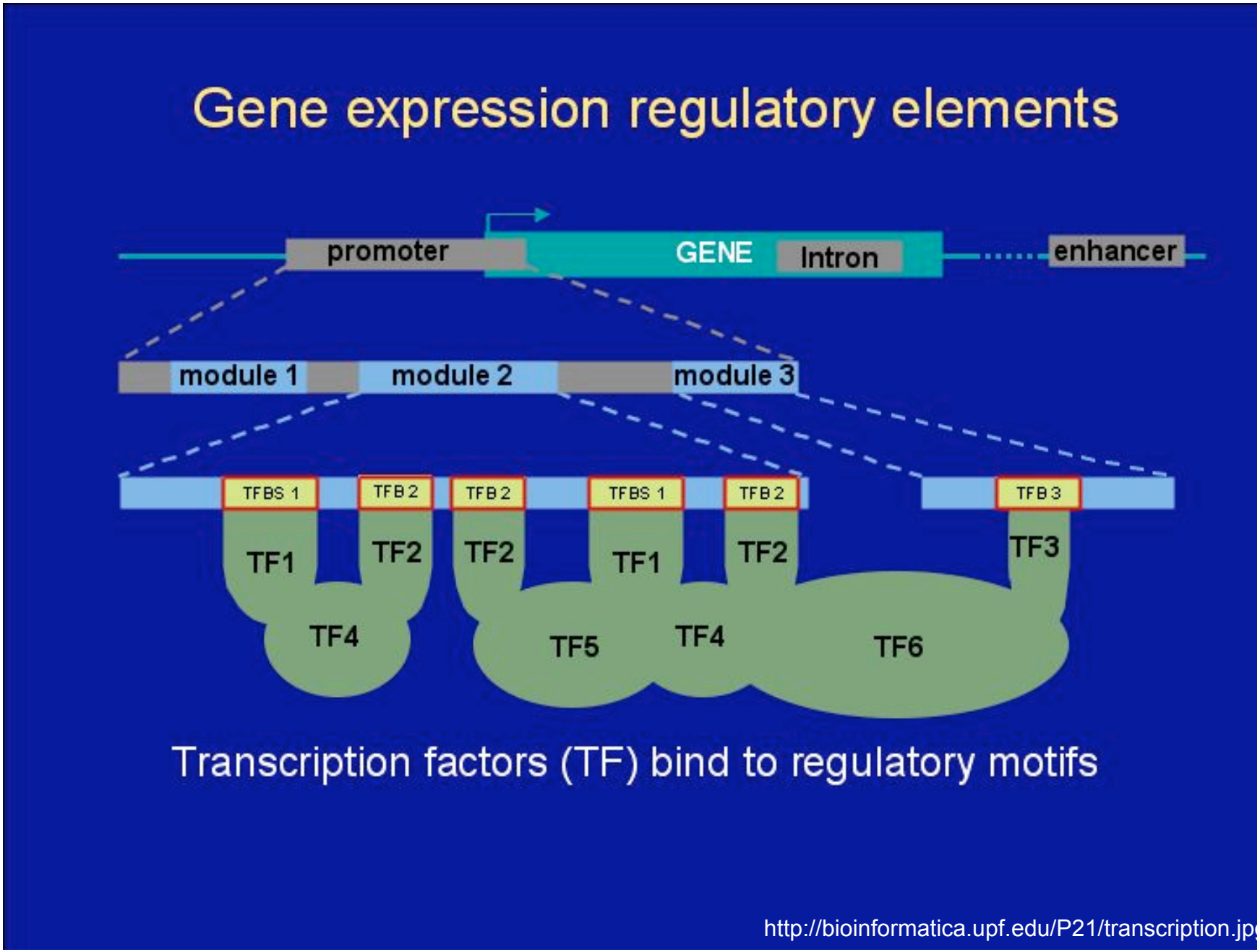
Why study transcription factor binding sites (TFBSs)?

- All protein coding genes are regulated by transcription factors.



- All protein coding genes are regulated by transcription factors.



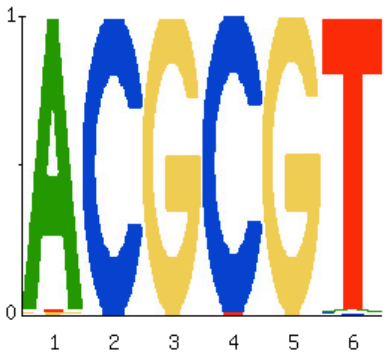


binding sites = binding motifs = regulatory elements = regulatory motifs

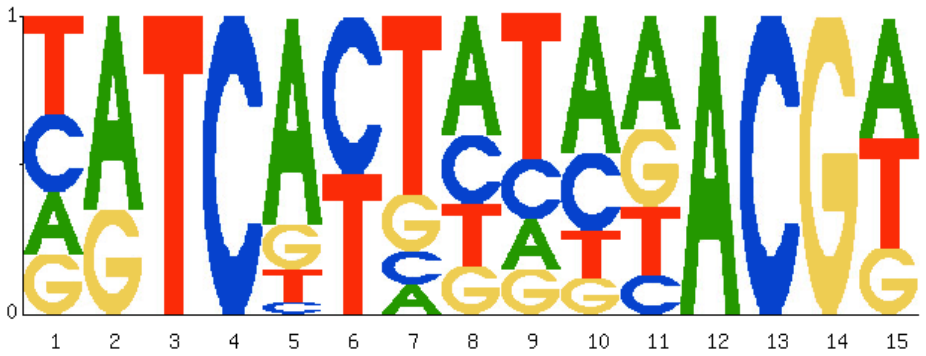
The difficulties

- Transcription factor binding sites (TFBSs) are short and degenerated.
  - Usually 6 ~ 15 bp.

- TFBSs can be on either strand of the DNA

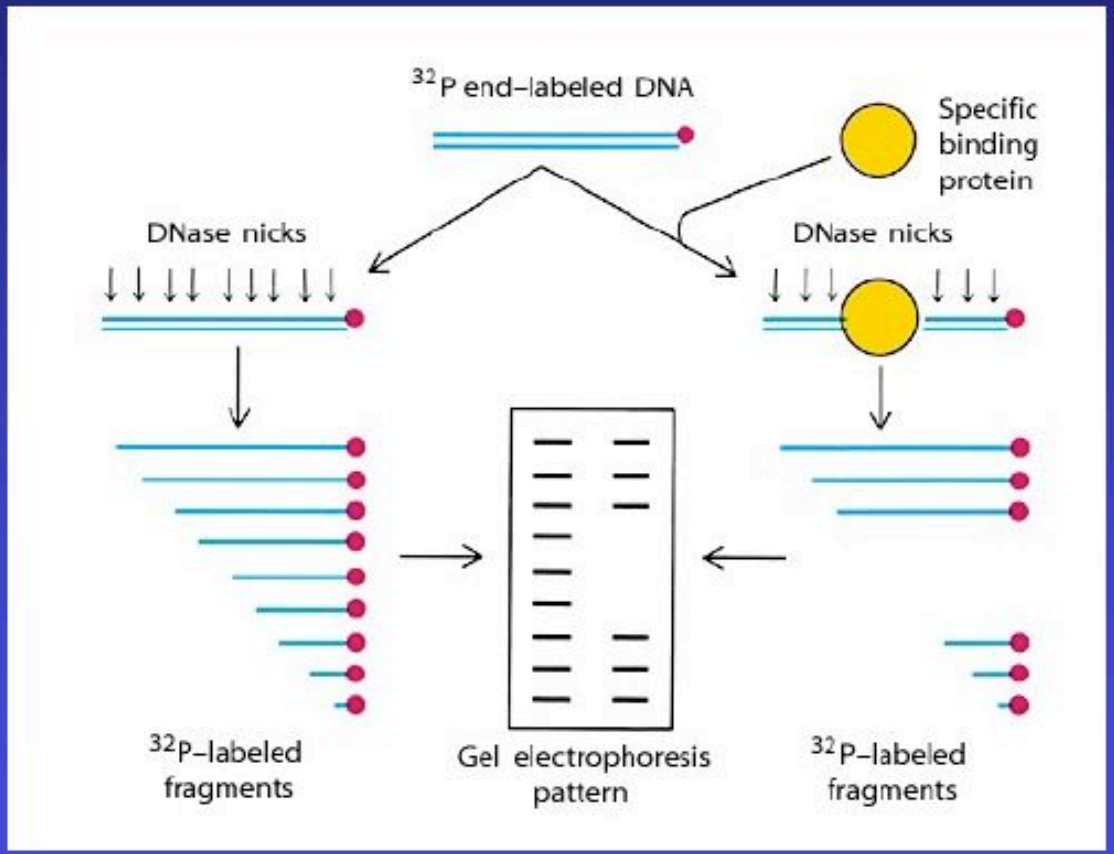


SWI6



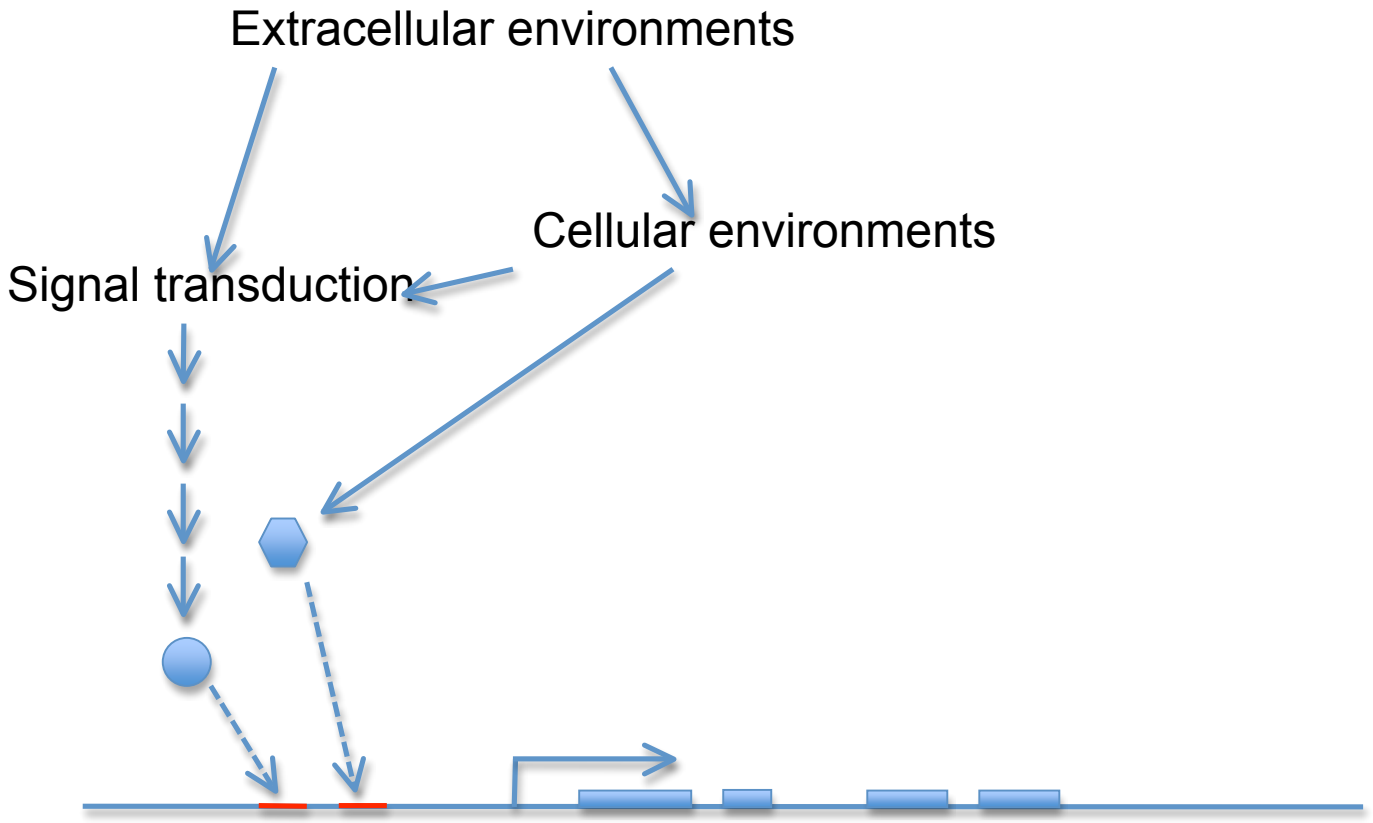
ABF1

Principle of DNA Footprinting



- In yeast, *Saccharomyces cerevisiae:*
  - ~6000 genes
  - >200 TFs

- All protein coding genes are regulated by transcription factors.



Basic idea

- Genes with the same (or similar) functions could be regulated by the same TF, therefore, have the same TFBSs at their promoter regions.

To identify where are TFBSs, we need to know:

- 1. What sequences they are
  - Find motifs: AlignACE
- 2. Which TFs will bind to them
  - Assign AlignACE motifs to known *cis*-regulatory elements: CompareACE

- AlignACE

  - Gibbs sampling

- CompareACE

  - Pearson correlation coefficient between the nucleotide base frequencies of two motif alignments

Maximum a priori log likelihood (MAP) score:

$$N \log R$$

N: number of aligned sites
R: degree of over-representation

- Group specificity

$$S = \sum_{i=x}^{\min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$

N: the total number of ORFs

s1, s2: the numbers of ORFs in the group used to find the motif and in the list of target gene, respectively

X: the number of ORFs in the intersection of the two lists

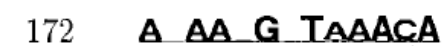
- Position bias

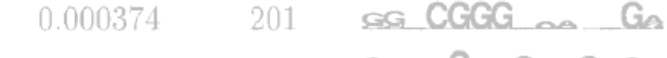
The probability of observing *m* or more sites out of possible *t* in a 50bp window (*w*) of a 600bp region (*s*)

$$P = \sum_{i=m}^{t} \binom{t}{i} \left(\frac{w}{s}\right)^{i} \left(1 - \frac{w}{s}\right)^{t-i}$$

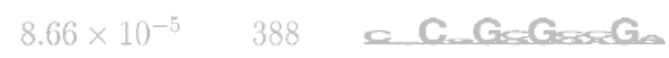
Result 1

- Motifs ranked by specificity score

Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
S1	123	8.6×10^{-48}	4.03×10^{-6}	279		Rap1
S2	128	1.27×10^{-32}	3.02×10^{-10}	125		Rpn4
S3	31.1	5.29×10^{-23}	0.0163	270		Gcn4
S4	55	8.73×10^{-21}	0.0015	169		HSE
S5	26.8	3.3×10^{-18}	0.00305	557		Mig1
S6	30.9	5.36×10^{-18}	0.00159	246		Cbf1
S7	12.4	9.98×10^{-14}	0.000546	241		RPS genes
S8	15.9	2.32×10^{-13}	0.0113	558		Hap2,3,4
S9	25.3	6.13×10^{-13}	1.49×10^{-5}	191		Lys14
S10	39.3	1.43×10^{-12}	4.85×10^{-8}	124		MCB
S11	11.1	1.44×10^{-12}	0.00527	168		RPL genes
S12	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
S13	18.7	2.37×10^{-12}	0.0447	316		STRE
S14	20.6	2.6×10^{-12}	0.00213	171		Met31,32
S15	19.4	3.08×10^{-12}	0.00269	153		Leu3
S16	17.2	3.08×10^{-12}	0.0204	105		Oaf1
S17	14.1	4.17×10^{-12}	0.000374	201		mito. transp.
S18	21	1.14×10^{-11}	0.000413	344		stress
S19	24.6	1.57×10^{-11}	9.29×10^{-6}	308		CSRE
S20	16	2.07×10^{-11}	8.66×10^{-5}	388		TCA cycle
S21	21.2	2.2×10^{-11}	0.000205	172		cytoskel. transp.
S22	20.3	2.27×10^{-11}	0.00241	374		TCA cycle
S23	10.5	2.49×10^{-11}	0.0221	349		Pho4
S24	10.7	4.27×10^{-11}	0.000847	161		Ste12
S25	44.8	8.64×10^{-11}	0.00647	382		Pdr3

Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
S1	123	8.6×10^{-48}	4.03×10^{-6}	279		Rap1
S2	128	1.27×10^{-32}	3.02×10^{-10}	125		Rpn4
S3	31.1	5.29×10^{-23}	0.0163	270		Gcn4
S4	55	8.73×10^{-21}	0.0015	169		HSE
S5	26.8	3.3×10^{-18}	0.00305	557		Mig1
S6	30.9	5.36×10^{-18}	0.00159	246		Cbf1
S7	12.4	9.98×10^{-14}	0.000546	241		RPS genes
S8	15.9	2.32×10^{-13}	0.0113	558		Hap2,3,4
S9	25.3	6.13×10^{-13}	1.49×10^{-5}	191		Lys14
S10	39.3	1.43×10^{-12}	4.85×10^{-8}	124		MCB
S11	11.1	1.44×10^{-12}	0.00527	168		RPL genes
S12	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
S13	18.7	2.37×10^{-12}	0.0447	316		STRE
S14	20.6	2.6×10^{-12}	0.00213	171		Met31,32
S15	19.4	3.08×10^{-12}	0.00269	153		Leu3
S16	17.2	3.08×10^{-12}	0.0204	105		Oaf1
S17	14.1	4.17×10^{-12}	0.000374	201		mito. transp.
S18	21	1.14×10^{-11}	0.000413	344		stress
S19	24.6	1.57×10^{-11}	9.29×10^{-6}	308		CSRE
S20	16	2.07×10^{-11}	8.66×10^{-5}	388		TCA cycle
S21	21.2	2.2×10^{-11}	0.000205	172		cytoskel. transp.
S22	20.3	2.27×10^{-11}	0.00241	374		TCA cycle
S23	10.5	2.49×10^{-11}	0.0221	349		Pho4
S24	10.7	4.27×10^{-11}	0.000847	161		Ste12
S25	44.8	8.64×10^{-11}	0.00647	382		Pdr3

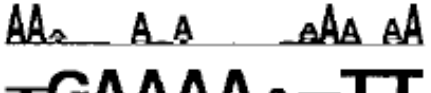
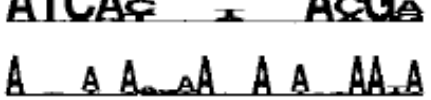
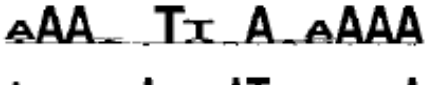
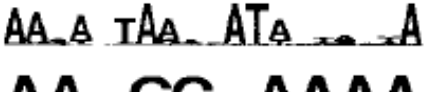
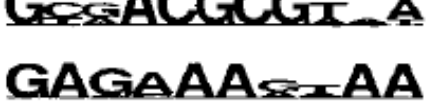
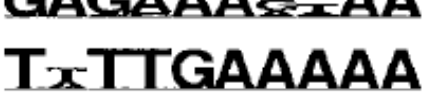
Literature: STRE binding site: AGGGG

Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
S1	123	8.6×10^{-48}	4.03×10^{-6}	279		Rap1
S2	128	1.27×10^{-32}	3.02×10^{-10}	125		Rpn4
S3	31.1	5.29×10^{-23}	0.0163	270		Gcn4
S4	55	8.73×10^{-21}	0.0015	169		HSE
S5	26.8	3.3×10^{-18}	0.00305	557		Mig1
S6	30.9	5.36×10^{-18}	0.00159	246		Cbf1
S7	12.4	9.98×10^{-14}	0.000546	241		RPS genes
S8	15.9	2.32×10^{-13}	0.0113	558		Hap2,3,4
S9	25.3	6.13×10^{-13}	1.49×10^{-5}	191		Lys14
S10	39.3	1.43×10^{-12}	4.85×10^{-8}	124		MCB
S11	11.1	1.44×10^{-12}	0.00527	168		RPL genes
S12	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
S13	18.7	2.37×10^{-12}	0.0447	316		STRE
S14	20.6	2.6×10^{-12}	0.00213	171		Met31,32
S15	19.4	3.08×10^{-12}	0.00269	153		Leu3
S16	17.2	3.08×10^{-12}	0.0204	105		Oaf1
S17	14.1	4.17×10^{-12}	0.000374	201		mito. transp.
S18	21	1.14×10^{-11}	0.000413	344		stress
S19	24.6	1.37×10^{-11}	9.29×10^{-6}	308		CSRE
S20	16	2.07×10^{-11}	8.66×10^{-5}	388		TCA cycle
S21	21.2	2.2×10^{-11}	0.000205	172		cytoskel. transp.
S22	20.3	2.27×10^{-11}	0.00241	374		TCA cycle
S23	10.5	2.49×10^{-11}	0.0221	349		Pho4
S24	10.7	4.27×10^{-11}	0.000847	161		Ste12
S25	44.8	8.64×10^{-11}	0.00647	382		Pdr3

Only half site of known motifs

Result 2

- Positionally biased motifs

Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
P1	637	0.000114	1.55×10^{-166}	140		
P2	15.6	0.0559	2.8×10^{-48}	114		AT repeat
P3	16.4	0.109	8.09×10^{-48}	133		
P4	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
P5	14.3	0.025	2.29×10^{-41}	148		Abf1
P6	23.8	0.0906	1.58×10^{-35}	130		
P7	29.6	0.00888	2.79×10^{-33}	101		PAC
P8	17.6	0.00102	3.53×10^{-31}	148		Reb1
P9	10.2	0.0111	5.66×10^{-21}	35		GT repeat
P10	32.6	1.98×10^{-14}	5.66×10^{-21}	35		Rap1
P11	125	4.08×10^{-28}	1.13×10^{-14}	112		Rpn4
P12	12.5	0.00818	6.6×10^{-11}	123		
P13	13.3	7.5×10^{-6}	7.01×10^{-10}	249		
P14	19.2	0.028	7.45×10^{-10}	114		
P15	10.5	9.72×10^{-5}	5×10^{-9}	141		MCB
P16	11.8	0.000216	5×10^{-9}	37		
P17	20.2	0.00495	9.18×10^{-9}	127		

Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
P1	637	0.000114	1.55×10^{-166}	140		
P2	15.6	0.0559	2.8×10^{-48}	114		AT repeat
P3	16.4	0.109	8.09×10^{-48}	133		
P4	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
P5	14.3	0.025	2.29×10^{-41}	148		Abf1
P6	23.8	0.0906	1.58×10^{-35}	130		
P7	29.6	0.00888	2.79×10^{-33}	101		PAC
P8	17.6	0.00102	3.53×10^{-31}	148		Reb1
P9	10.2	0.0111	5.66×10^{-21}	35		GT repeat
P10	32.6	1.98×10^{-14}	5.66×10^{-21}	35		Rap1
P11	125	4.08×10^{-28}	1.13×10^{-14}	112		Rpn4
P12	12.5	0.00818	6.6×10^{-11}	123		
P13	13.3	7.5×10^{-6}	7.01×10^{-10}	249		
P14	19.2	0.028	7.45×10^{-10}	114		
P15	10.5	9.72×10^{-5}	5×10^{-9}	141		MCB
P16	11.8	0.000216	5×10^{-9}	37		
P17	20.2	0.00495	9.18×10^{-9}	127		

Negative controls

	Random groupings motifs		Functional category motifs	
Spec. score cutoff	MAP > 0	MAP > 10	MAP > 0	MAP > 10
1	3692(1063)	1792(205)	3311(1324)	1234(208)
10^{-1}	2766(1038)	1047(202)	2713(1284)	815(194)
10^{-2}	2026(978)	553(181)	2198(1201)	530(179)
10^{-3}	1416(831)	285(149)	1622(1016)	337(153)
10^{-4}	935(641)	151(104)	1109(753)	226(121)
10^{-5}	554(425)	72(56)	750(543)	160(90)
10^{-6}	329(290)	31(29)	446(329)	122(67)
10^{-7}	151(143)	15(15)	270(199)	91(47)
10^{-8}	60(59)	9(9)	164(118)	73(35)
10^{-9}	37(36)	6(6)	97(62)	60(28)
10^{-10}	14(14)	5(5)	69(38)	54(25)

Positive controls

- Search 29 literature reported motifs
  - 21 of 29 were found

Conclusion

The method (AlignACE + group specificity + position bias) can be applicable to groups of genes other than functional categories.

After that …

- Microarray
- Chromatin Immunoprecipitation chip (ChIP-chip)

ChIP-chip

Transcriptional regulatory code of a eukaryotic genome

Christopher T. Harbison[1,2*], D. Benjamin Gordon[1*], Tong Ihn Lee[1], Nicola J. Rinaldi[1,2], Kenzie D. Macisaac[3], Timothy W. Danford[3], Nancy M. Hannett[1], Jean-Bosco Tagne[1], David B. Reynolds[1], Jane Yoo[1], Ezra G. Jennings[1], Julia Zeitlinger[1], Dmitry K. Pokholok[1], Manolis Kellis[1,3,4], P. Alex Rolfe[3], Ken T. Takusagawa[3], Eric S. Lander[1,2,4], David K. Gifford[3,4], Ernest Fraenkel[1,3] & Richard A. Young[1,2,4]

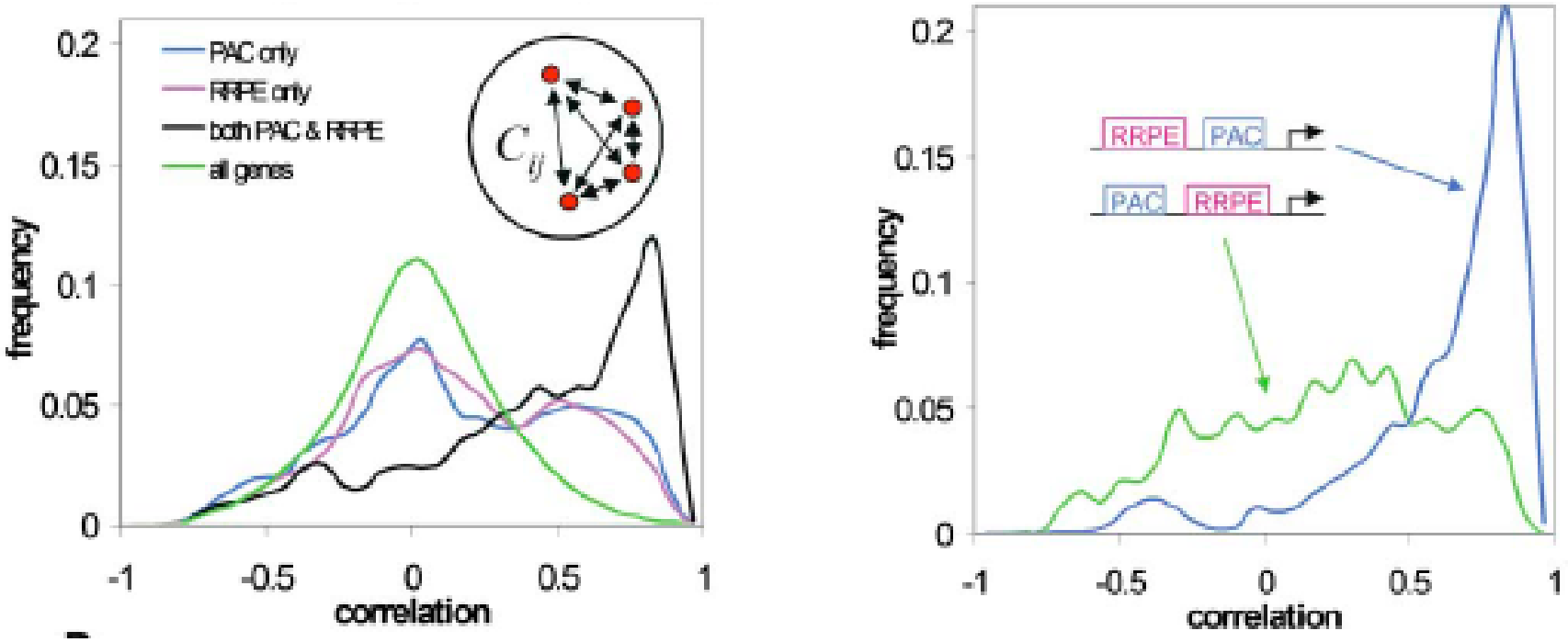
Position bias

Cell, Vol. 117, 185–198, April 16, 2004, Copyright ©2004 by Cell Press

Predicting Gene Expression from Sequence

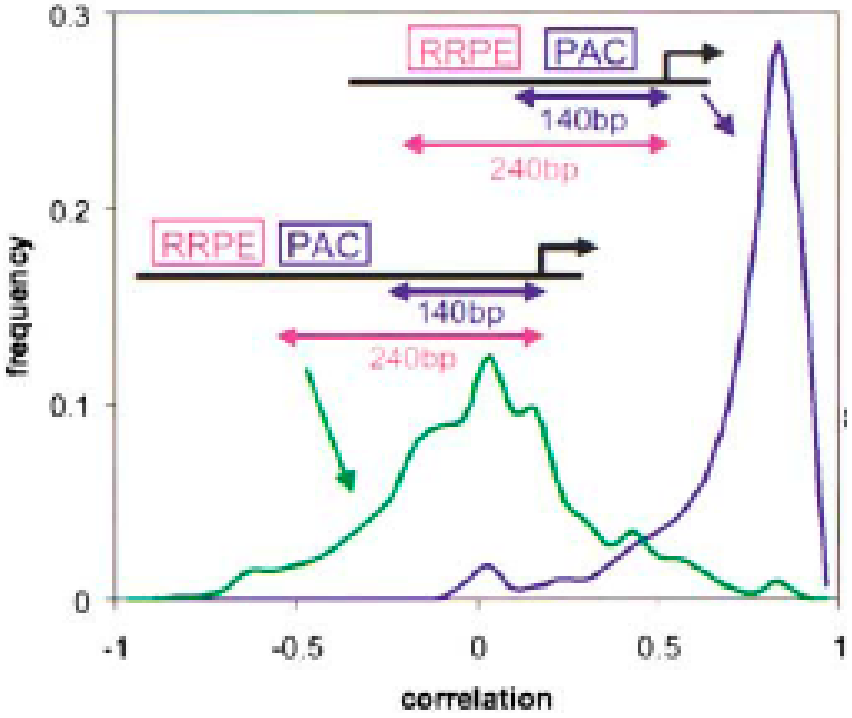
Michael A. Beer and Saeed Tavazoie*
Lewis-Sigler Institute for Integrative Genomics
and Department of Molecular Biology
Princeton University
Princeton, New Jersey 08544

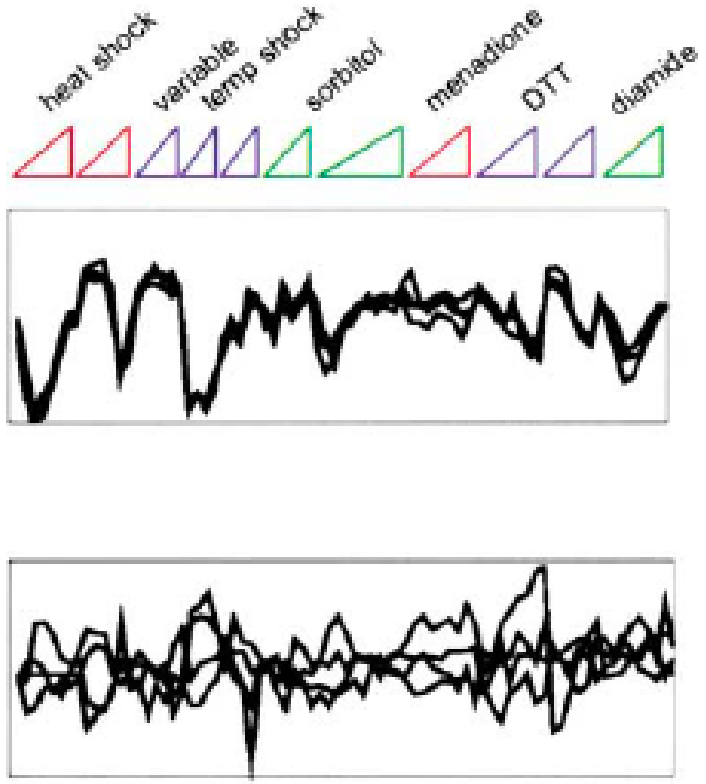
Cis-rule



Cell (2004) 117:185-198

Cis-rule





Cell (2004) 117:185-198

Dealing with gapped motifs

Discovering gapped binding sites of yeast transcription factors

Chien-Yu Chen*, Huai-Kuang Tsai†, Chen-Ming Hsu‡, Mei-Ju May Chen§, Hao-Geng Hung¶, Grace Tzu-Wei Huang‖, and Wen-Hsiung Li**††

- Questions