Graduate Computational Genomics

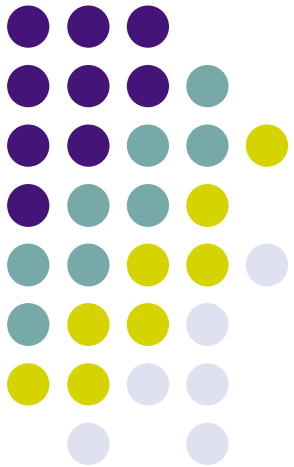
02-710 / 10-810 & MSCBIO2070

Modeling Regulatory Networks (part II)

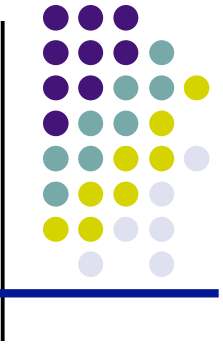
Takis Benos

Lecture #25a, April 24, 2007

Reading: handouts & papers



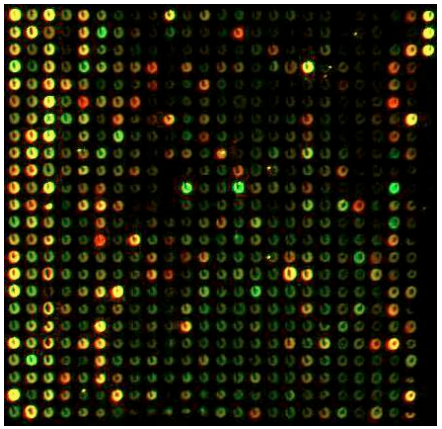
GRNs: the data sources



Gene expression

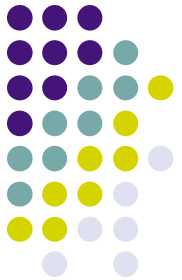
- ✓Microarrays
- SAGE
- Real-time RT-PCR

ChIP-chip



- ✓Promoter arrays
- Tiling arrays

The study that changed everything



Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*

Tong Ihn Lee,[1]* Nicola J. Rinaldi,[1,2]* François Robert,[1]*
Duncan T. Odom,[1] Ziv Bar-Joseph,[3] Georg K. Gerber,[3]
Nancy M. Hannett,[1] Christopher T. Harbison,[1,2]
Craig M. Thompson,[1]† Itamar Simon,[1] Julia Zeitlinger,[1]
Ezra G. Jennings,[1,2] Heather L. Murray,[1] D. Benjamin Gordon,[1]
Bing Ren,[1]‡ John J. Wyrick,[1]§ Jean-Bosco Tagne,[1]
Thomas L. Volkert,[1] Ernest Fraenkel,[1] David K. Gifford,[3]
Richard A. Young[1,2]‖

We have determined how most of the transcriptional regulators encoded in the eukaryote *Saccharomyces cerevisiae* associate with genes across the genome in living cells. Just as maps of metabolic networks describe the potential pathways that may be used by a cell to accomplish metabolic processes, this network of regulator-gene interactions describes potential pathways yeast cells can use to regulate global gene expression programs. We use this information to identify network motifs, the simplest units of network architecture, and demonstrate that an automated process can use motifs to assemble a transcriptional regulatory network structure. Our results reveal that eukaryotic cellular functions are highly connected through networks of transcriptional regulators that regulate other transcriptional regulators.

Lee *et al.*: the method

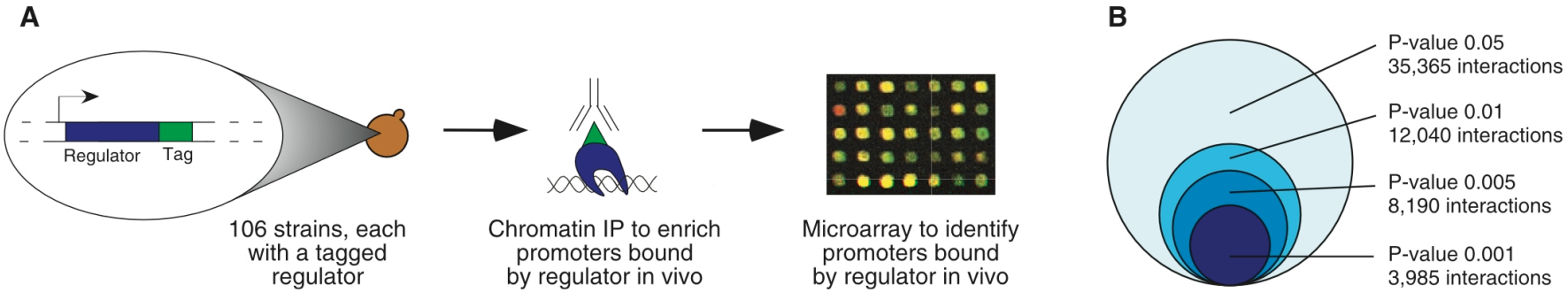
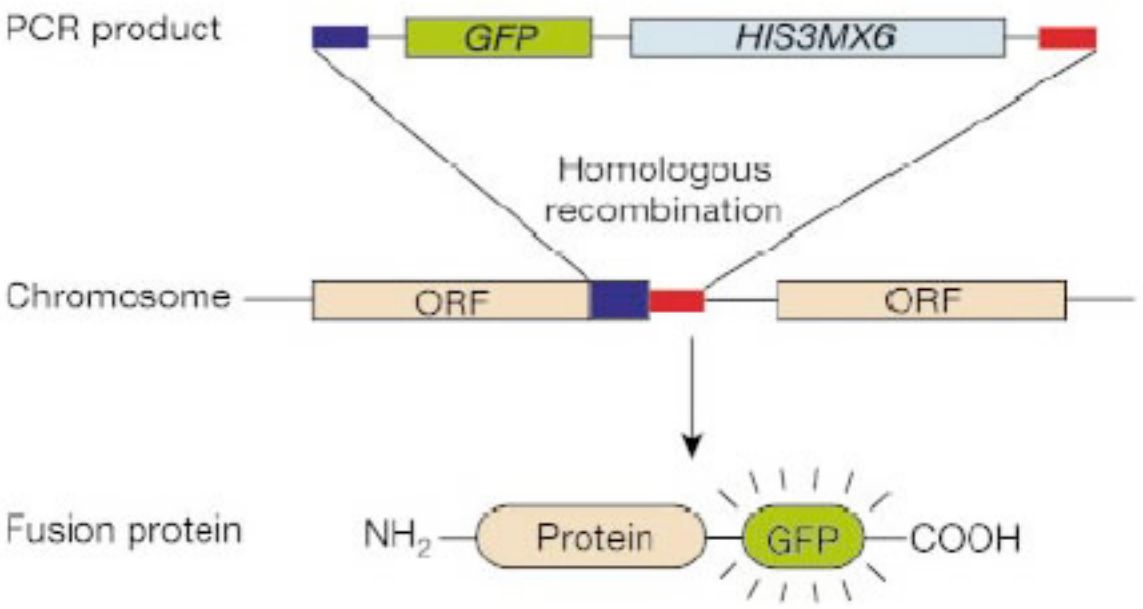
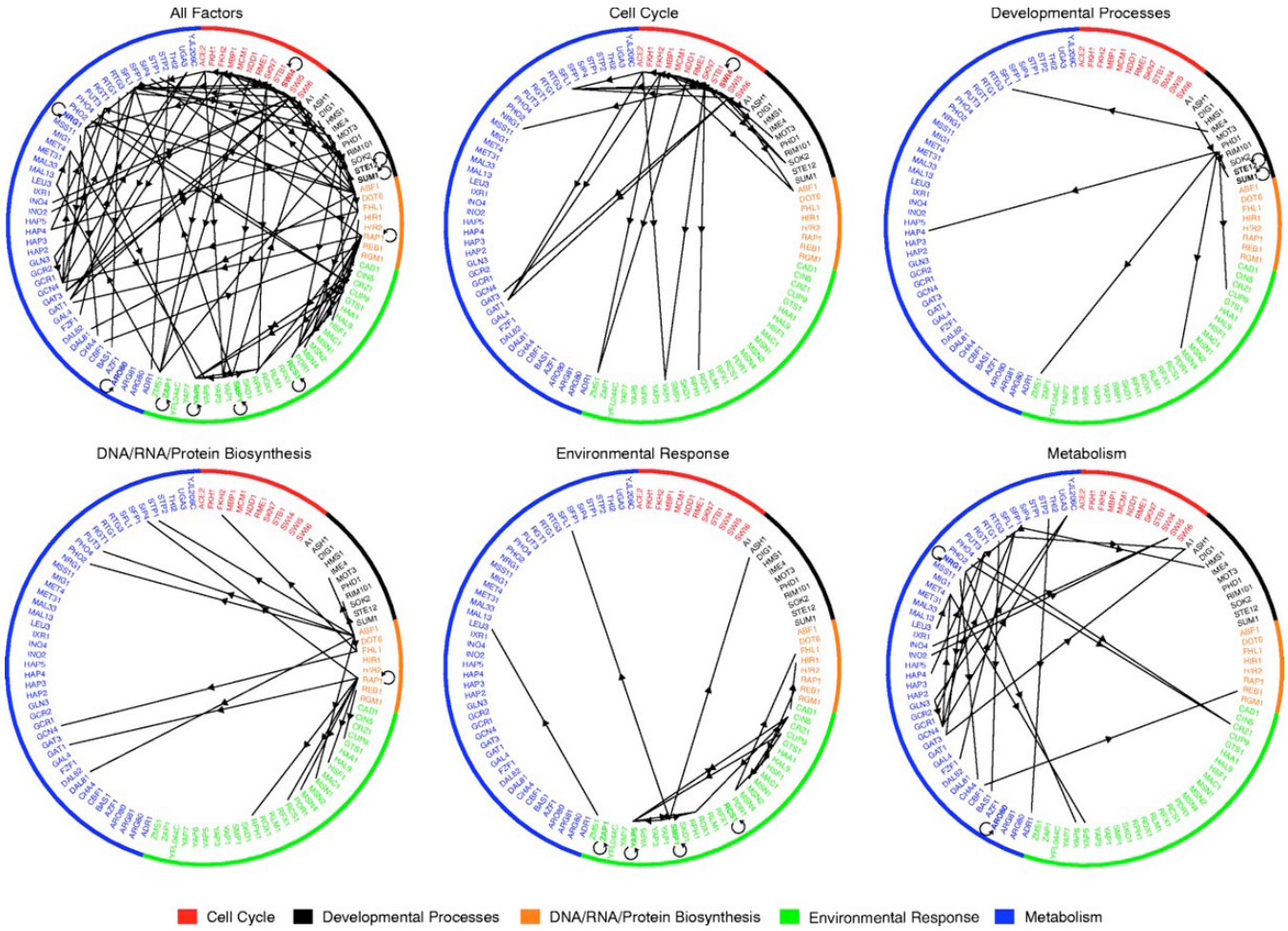


Fig. 1. Systematic genome-wide location analysis for yeast transcription regulators. (**A**) Methodology. Yeast transcriptional regulators were tagged by introducing the coding sequence for a c-*myc* epitope tag into the normal genomic locus for each regulator. Of the yeast strains constructed in this fashion, 106 contained a single epitope-tagged regulator whose expression could be detected in rich growth conditions. Chromatin immunoprecipitation (ChIP) was performed on each of these 106 strains. Promoter regions enriched through the ChIP procedure were identified by hybridization to microarrays containing a genome-wide set of yeast promoter regions. (**B**) Effect of *P* value threshold. The sum of all regulator-promoter region interactions is displayed as a function of varying *P* value thresholds applied to the entire location data set for the 106 regulators. More stringent *P* values reduce the number of interactions reported but decrease the likelihood of false-positive results.

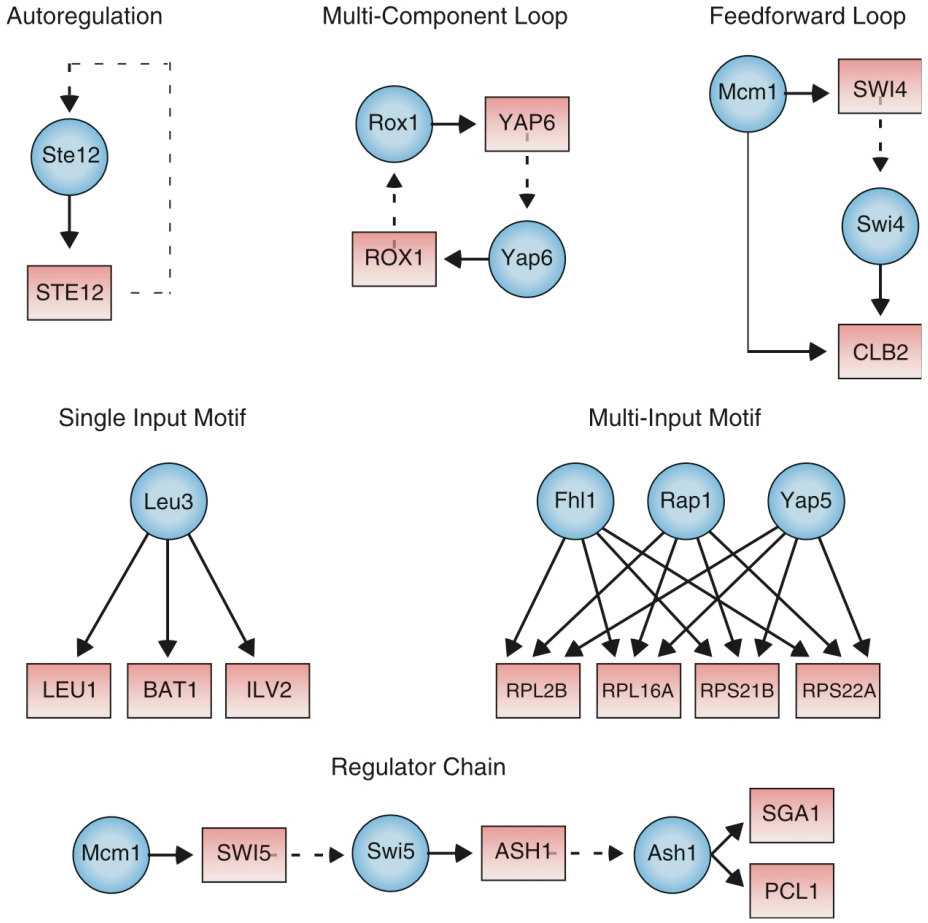
Homologous recombination



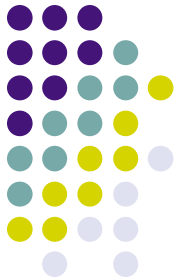
Lee *et al.*: the results



Lee *et al.*: the results *(cntd)*



Gao *et al.*: combining ChIP-chip and microarray data



Research article

Highly accessed Open Access

Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data

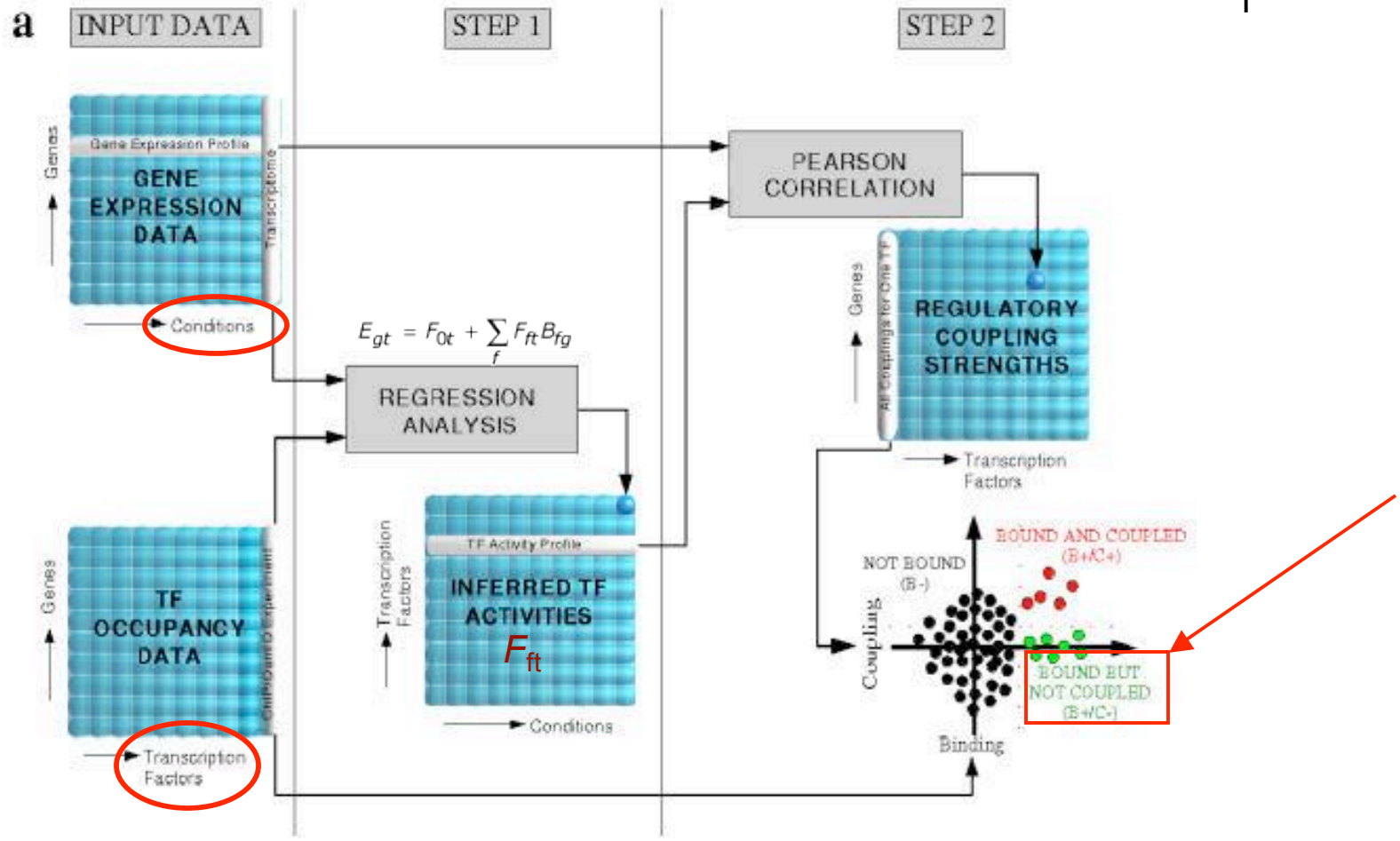
Feng Gao[1], Barrett C Foat[1] and Harmen J Bussemaker[1,2]

[1]Department of Biological Sciences, Columbia University, New York, New York 10027, U.S.A

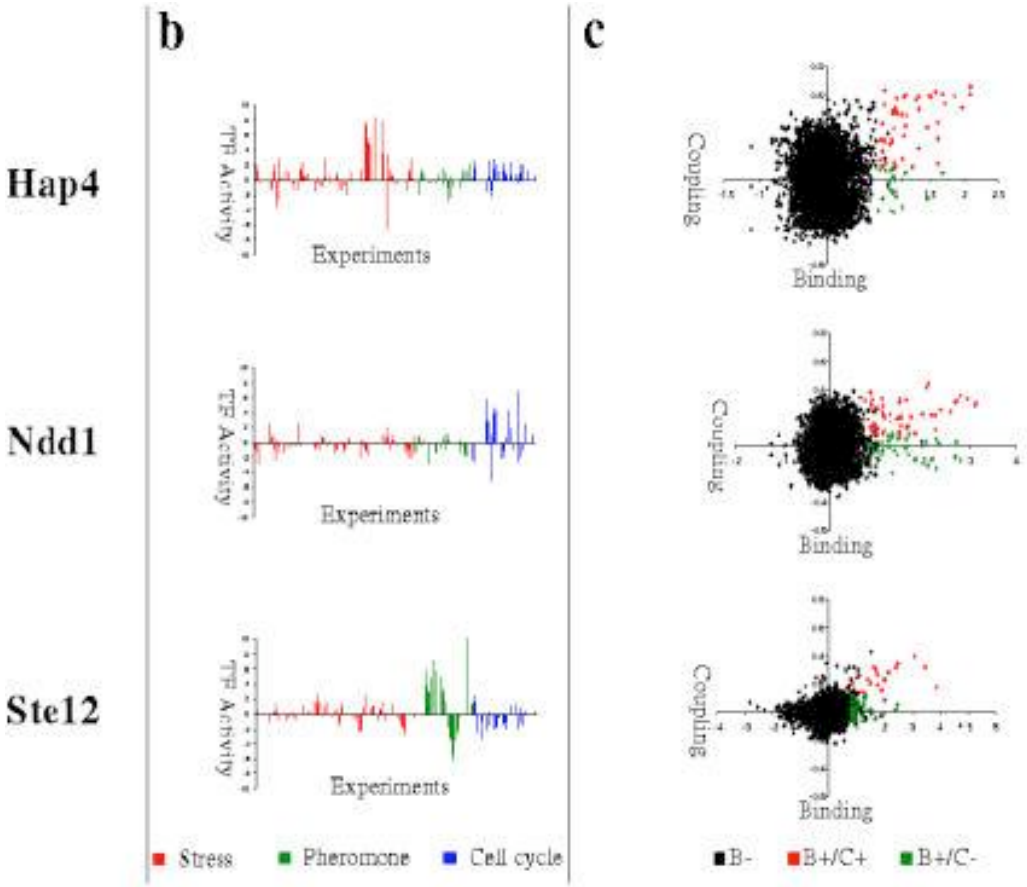
[2]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, U.S.A

BMC Bioinformatics 2004, **5**:31 doi:10.1186/1471-2105-5-31

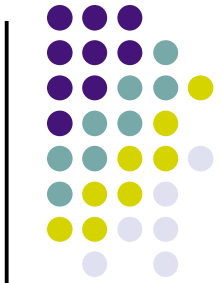
Gao *et al.*: the method



Gao *et al.*: results



Gao *et al.*: results (cntd)



Classification of genes according to ChIP data and inferred regulatory coupling.

TF	B-	B+/C+	B+/C-	Unclassified*
Abf1	5638	138	136	470
Ace2	5843	33	33	473
Arg81	5985	11	10	376
Bas1	5975	28	13	366
Cad1	5854	26	13	489
Dal81	5823	24	16	519
Dig1	5872	20	11	479
Fhl1	5754	146	37	445
Fkh2	5261	61	45	1015
Gal4	5149	20	20	1193
Gat3	5891	40	29	422
Gcn4	5919	58	21	384
Hal9	5948	4	13	417
Hap4	5939	47	21	375
Hir1	5963	19	10	390

Gao *et al.*: results (cntd)

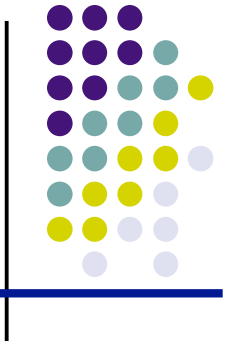


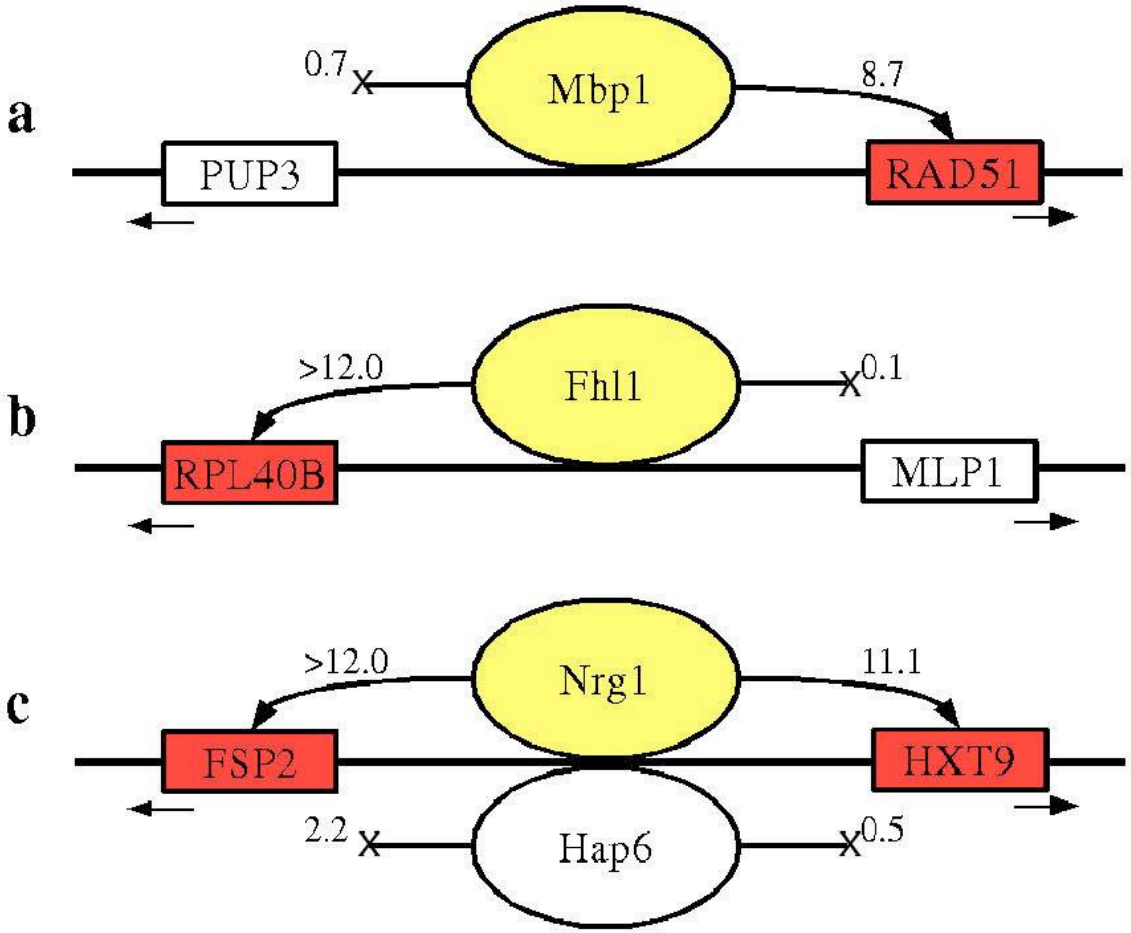
Table 2

Analysis of transcriptional response to transcription factor deletion.

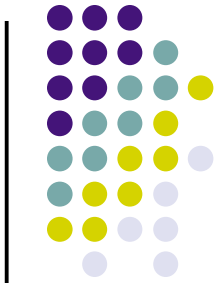
Transcription Factor	Genomewide		B+/C+			B+/C-		
	Mean	SD	Mean	SD	$-\log_{10}(p)$	Mean	SD	$-\log_{10}(p)$
Dig1	0.759	0.198	0.567	0.360	3.76	0.846	0.080	0.04
Gcn4	0.785	0.177	0.294	0.329	28.23	0.793	0.170	0.24
Hir2	0.724	0.218	0.180	0.295	4.00	0.760	0.194	0.14
Mbp1	0.830	0.121	0.541	0.324	27.86	0.770	0.172	2.79
Swi4	0.524	0.338	0.348	0.352	4.89	0.533	0.333	0.24
Swi5	0.744	0.204	0.583	0.373	5.98	0.759	0.186	0.17
Yap1	0.756	0.201	0.471	0.376	7.66	0.587	0.337	2.67

The mean and the standard deviation of gene expression log-ratio between mutant and wild type as obtained in Hughes *et al.* were calculated for all genes in the genome as well as for the B+/C+ and B+/C- groups [21]. A sample t-test was performed to determine the significance of the change in expression. The B+/C+ genes show a significant change in mRNA expression for the 7 transcription factors for which deletion and ChIP data is available. By contrast, the response of the B+/C- genes is insignificant for most transcription factors.

Gao *et al.*: results (cntd)



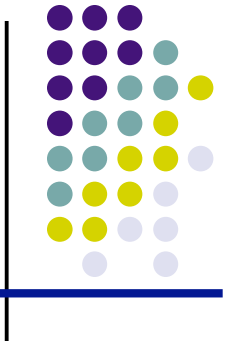
Gao *et al.*: Conclusions



- A new method that combines both ChIP and gene expression data to assess transcription regulation.

- Half of the TF targets resulting from the ChIP assay ($p<10^{-3}$) are non-functional. *(why?)*
  - ChIP false positives
  - Competition with other factors
  - Chromatin structure

For some reading



- Lee *et al.*, "*Transcriptional regulatory networks in Saccharomyces cerevisiae*" (2002) *Science* **298**: 799-804.

- Gao *et L.*, "*Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data*" (2004) *BMC Bioinformatics* **5**: 31.

Graduate Computational Genomics

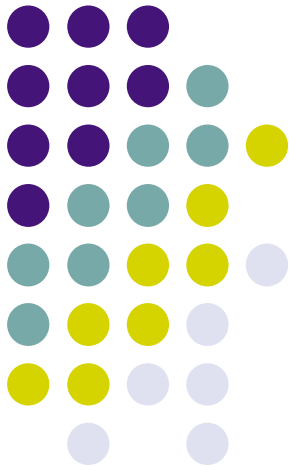
02-710 / 10-810 & MSCBIO2070

Evolution of regulatory regions and binding motifs

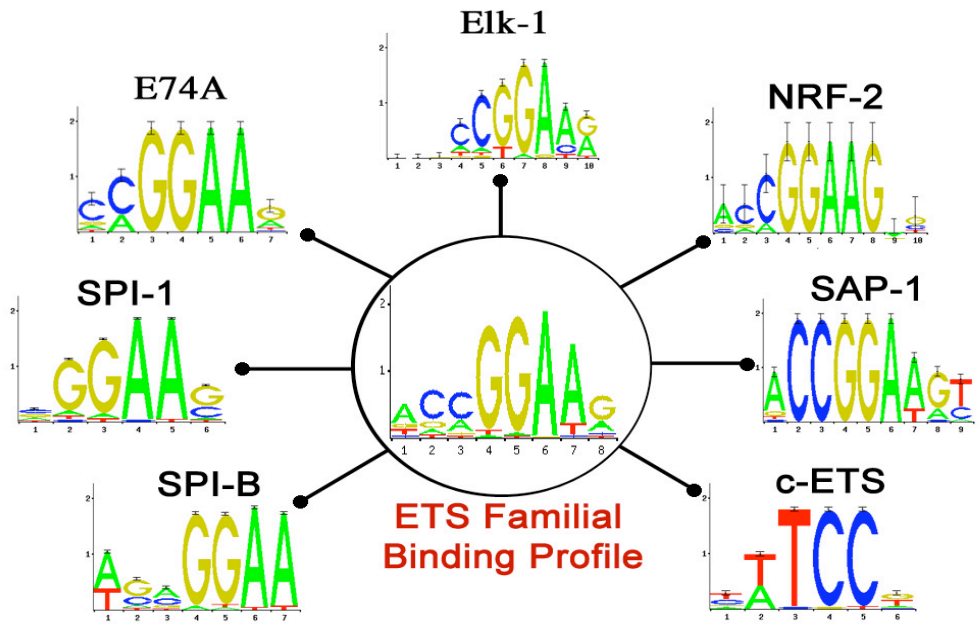
Takis Benos

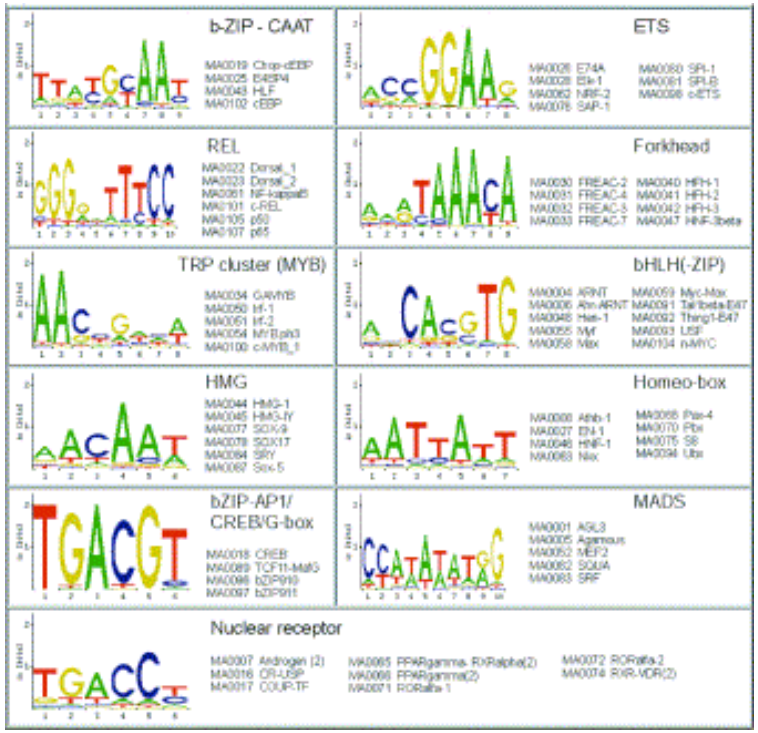
Lecture #25b, April 24, 2007

Reading: handouts & papers

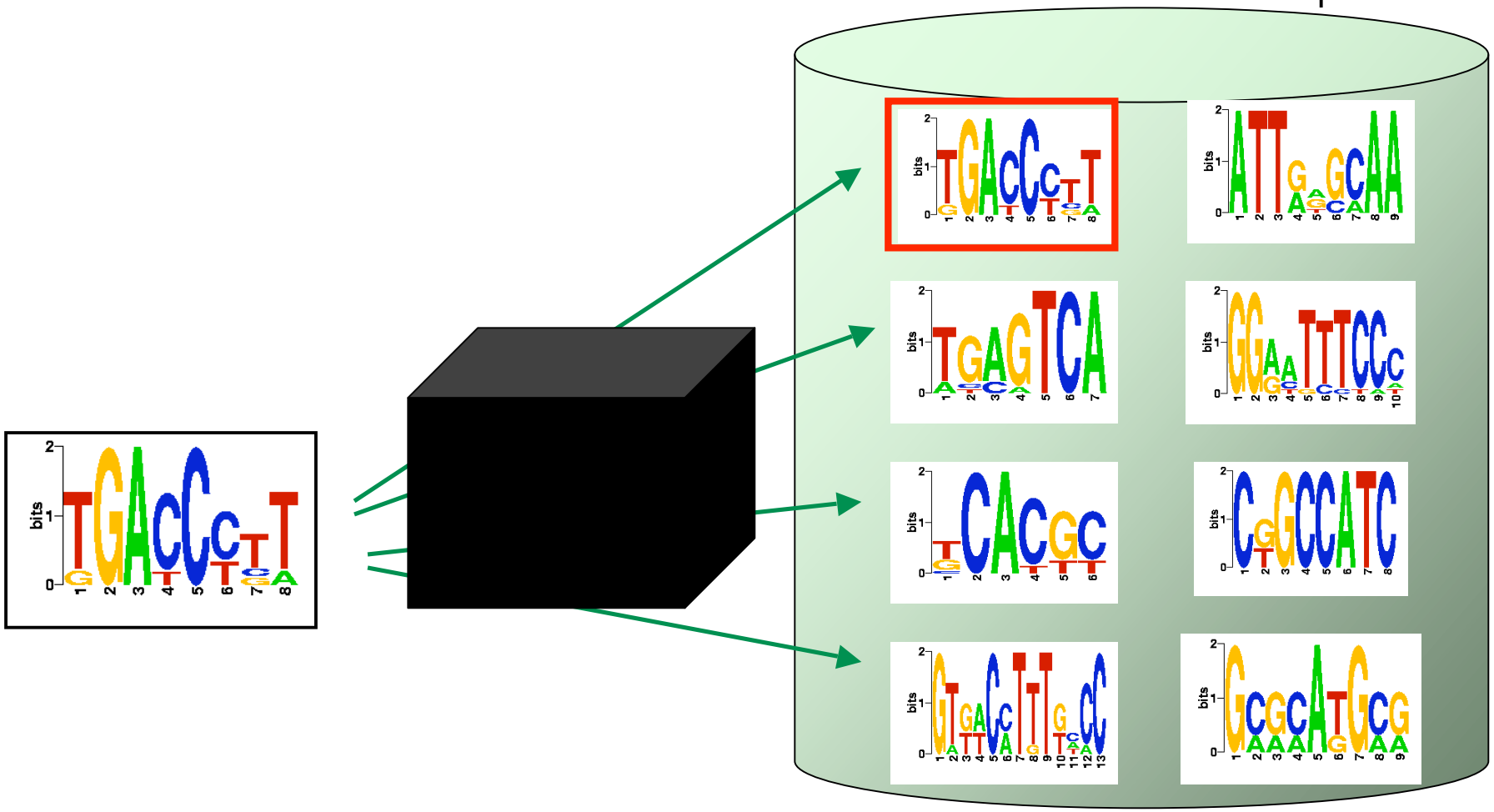


Familial binding profiles (FBS)

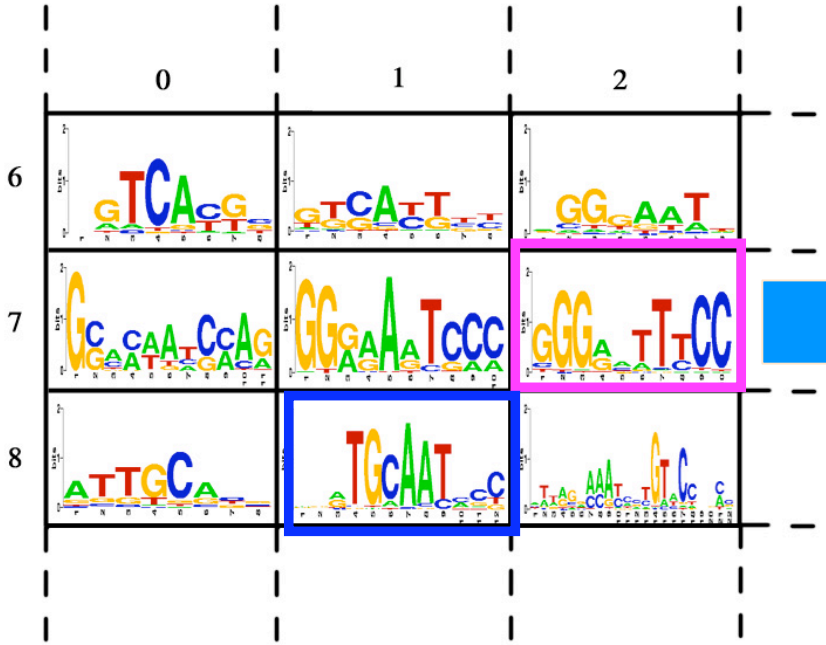


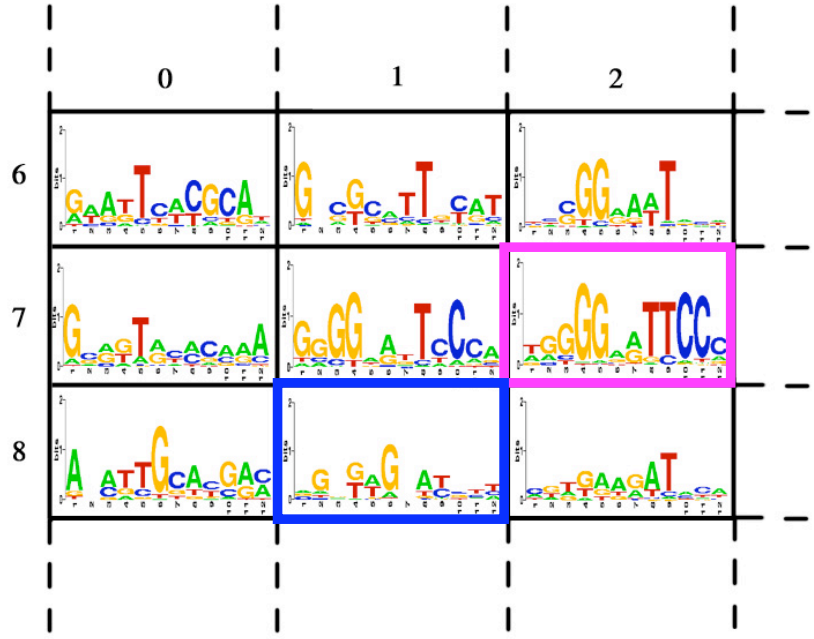


Comparing motifs



SOMBRERO & FBP Priors





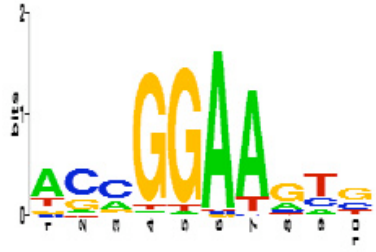
Start: BP-SOM trained on 257 mammalian PSSMs (incl. REL)

Finish: SOMBRERO used to find motifs of NF-κB (REL-class).

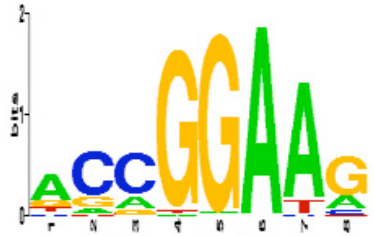
Binding Profile SOM (BP-SOM)

bHLH Tal1beta-E47S bHLH Thing1-E47	NUCLEAR Androgen HMG HMG-1	REL Dorsal_1	TRP-CLUSTER MYB.ph3 TRP-CLUSTER c-MYB_1 TRP-CLUSTER GAMYB
FORKHEAD HFH-2 FORKHEAD HNF-3b FORKHEAD HFH-1 FORKHEAD HFH-3 FORKHEAD FREAC-2 FORKHEAD FREAC-4	MADS AGL3 MADS MEF2 FORKHEAD FREAC-7	NUCLEAR RORalfa-1 NUCLEAR RORalfa-2	ETS SAP-1 ETS SPI-B ETS NRF-2 ETS SPI-B ETS E74A ETS c-ETS ETS Elk-1
MADS Agamous	MADS SRF	HMG SOX-9 HMG Sox-5 HMG SOX17 HMG SRY	HOMEO Nkx HOMEO S8
bZIP HLF bZIP E4BP4 bZIP cEBP bZIP CREB	bHLH-ZIP Max bHLH-ZIP USF bHLH-ZIP n-MYC bHLH-ZIP Myc-Max bHLH ARNT bZIP bZIP911 bZIP TCF11 bZIP bZIP910	TRP_CLUSTER Irf-1 TRP_CLUSTER Irf-2	bHLH Hen-1 bHLH Ahr-ARNT FORKHEAD FREAC-3
MADS SQUA HOMEO HNF-1 HMG HMG-IY	bHLH Myf	NUCLEAR CFI-USP NUCLEAR COUP-TF NUCLEAR RXR-VDR NUCLEAR PPARgamma NUCLEAR PPARgamma-rxr	REL p65 REL NF-kB REL p50 REL Dorsal_2 REL c-REL HOMEO EN-1 bZIP Chop-cEBP

ETS

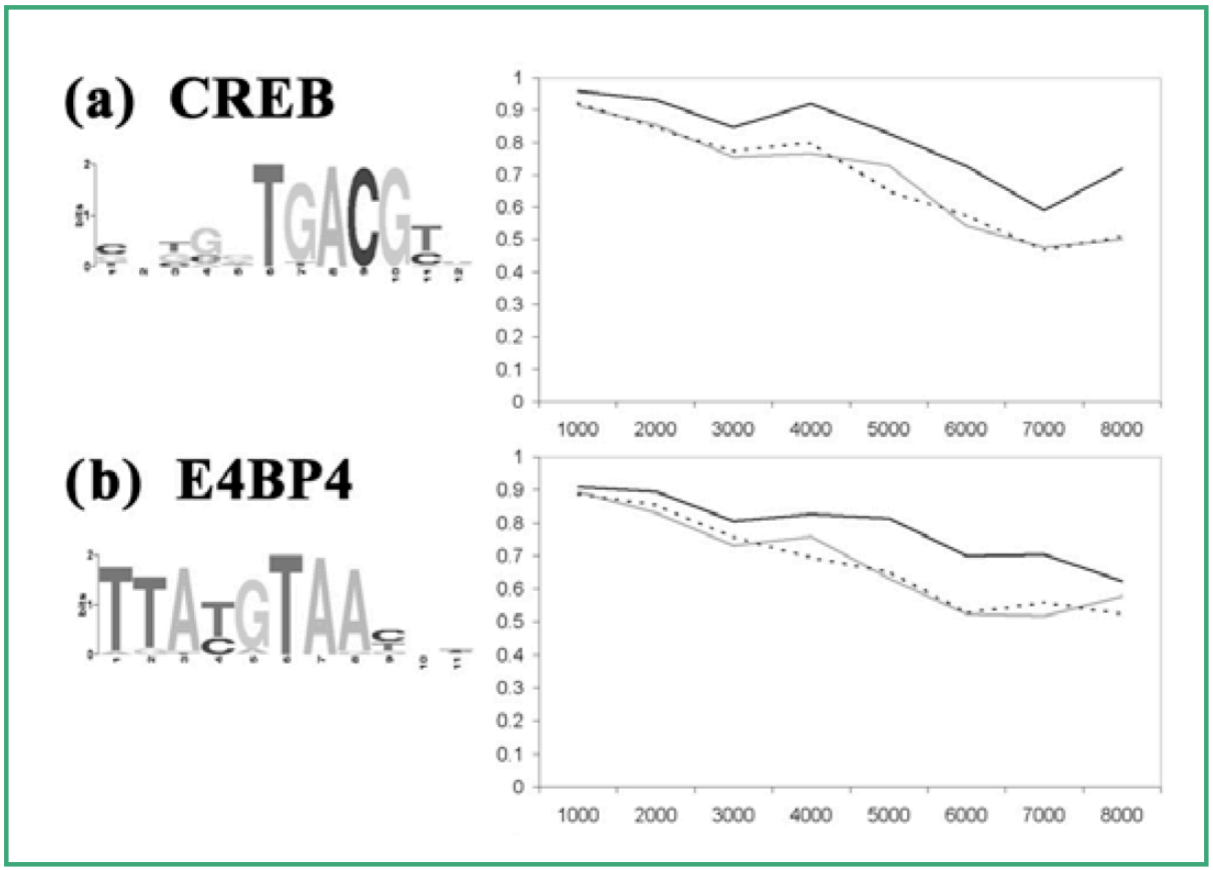


BP-SOM

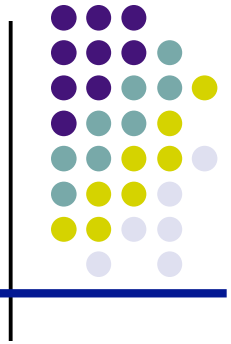


Sandelin & Wasserman

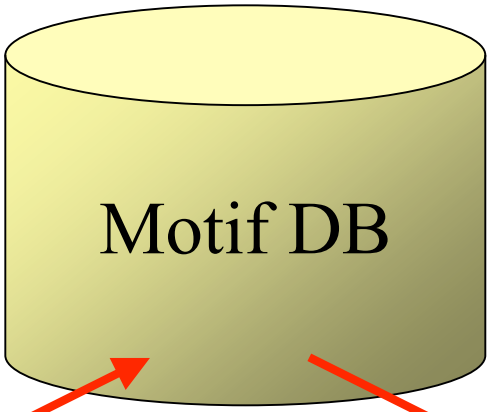
SOMBRERO: results



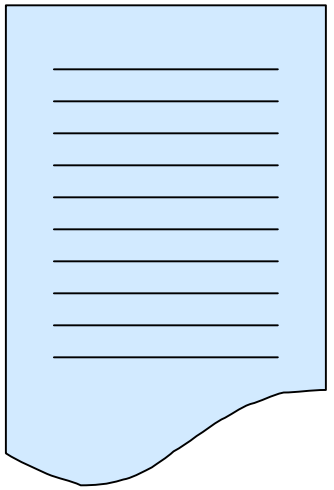
BLAST-like motif searches



Input motif



List of motif "hits"

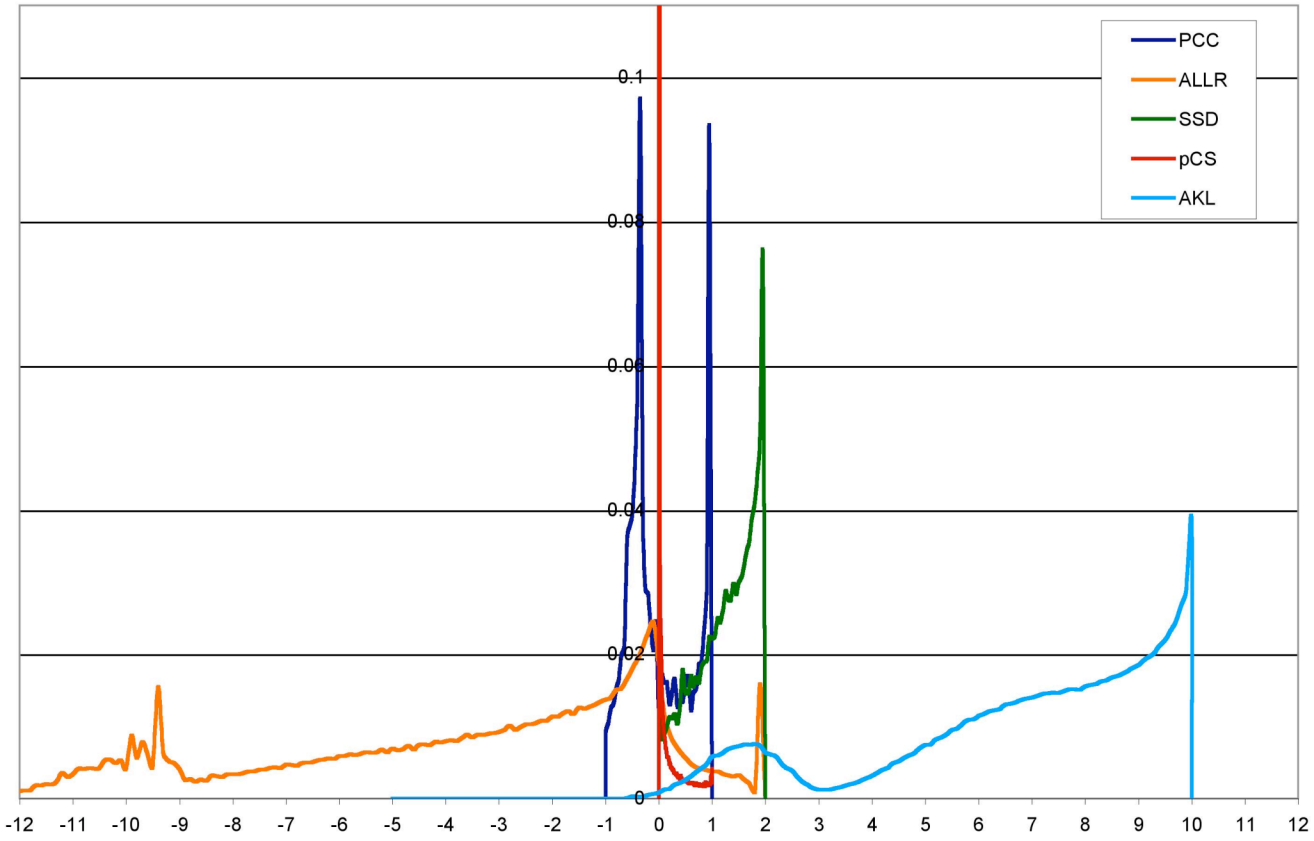


		Accuracy	
Name	**No. of TFs**	**Best-hit**	**Bayesian learning**
bZIP	93	**0.94**	0.92
C2H2	74	0.76	0.77
C4	52	**0.98**	0.91
Homeo	50	0.82	0.85
Forkhead	49	**0.90**	0.83
bHLH	37	0.81	0.88
Average		***0.87***	*0.86*

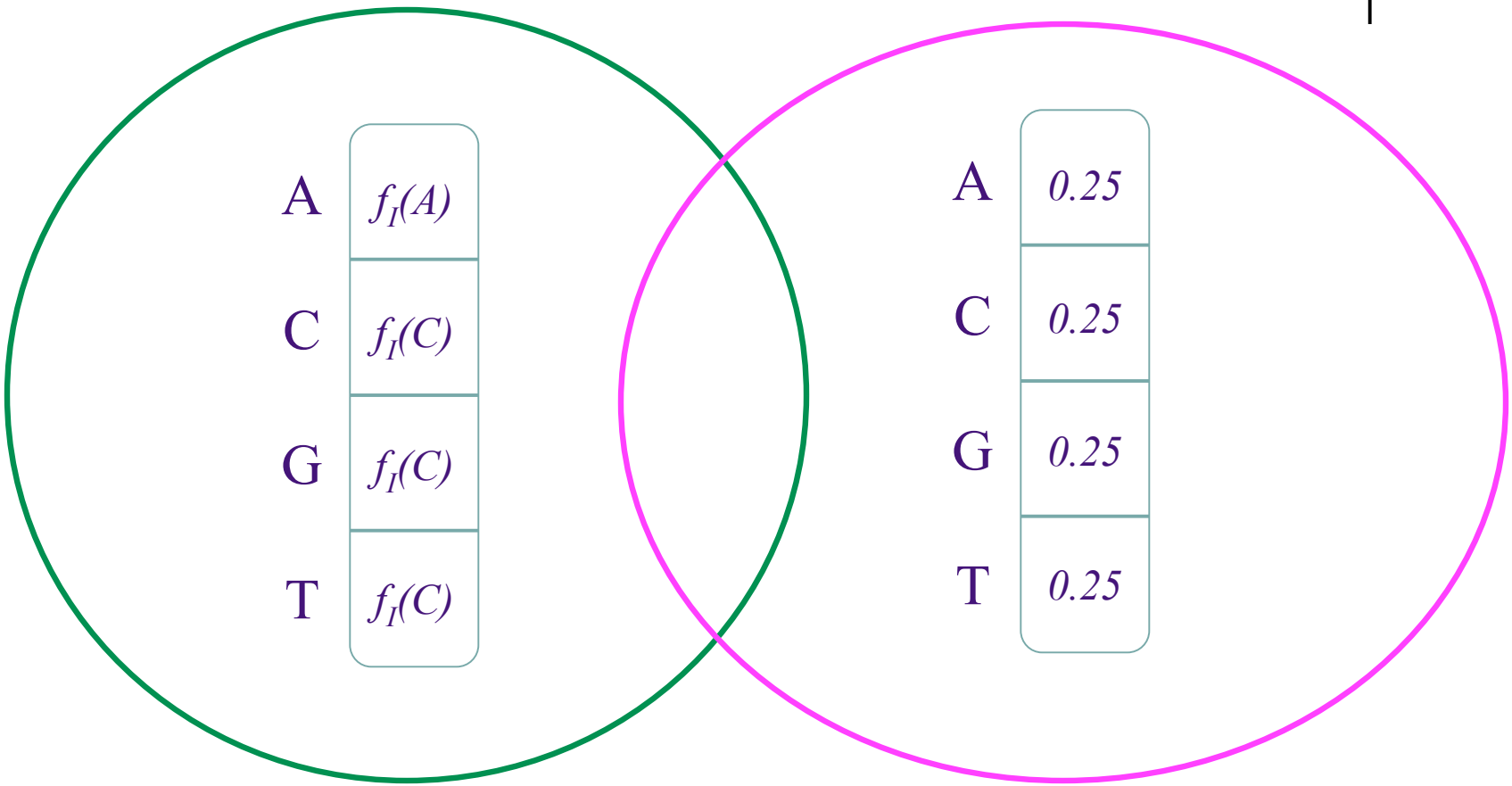
The tools: distance metrics

Similarity metric	Formula
Pearson Correlation Coefficient (**PCC**)	$PCC(X,Y) = \frac{\sum_{b=A}^{T}(f_X(b)-\overline{f_X})\cdot(f_Y(b)-\overline{f_Y})}{\sqrt{\sum_{b=A}^{T}(f_X(b)-\overline{f_X})^2\cdot\sum_{b=A}^{T}(f_Y(b)-\overline{f_Y})^2}}$
Chi-square (**pCS**) (1 - *p-value of*)	$\chi_3^2(X,Y) = \sum_{K=\{X,Y\}}\sum_{b=A}^{T}\frac{(n_K(b)-n_K^e(b))^2}{n_K^e(b)}$
Average Kullback-Leibler (**AKL**)	$AKL(X,Y) = 10 - \frac{\sum_{b=A}^{T} f_X(b)\cdot\log\frac{f_X(b)}{f_Y(b)} + \sum_{b=A}^{T} f_Y(b)\cdot\log\frac{f_Y(b)}{f_X(b)})}{2}$
Sum of squared distances (**SSD**)	$SSD(X,Y) = 2 - \sum_{b=A}^{T}(f_X(b)-f_Y(b))^2$
Average Log-likelihood Ratio (**ALLR**)	$ALLR(X,Y) = \frac{\sum_{b=A}^{T} n_X(b)\cdot\log\frac{f_Y(b)}{p_{ref}(b)} + \sum_{b=A}^{T} n_Y(b)\cdot\log\frac{f_X(b)}{p_{ref}(b)}}{\sum_{b=A}^{T}(n_X(b)+n_Y(b))}$
ALLR with Lower Limit (**ALLR_LL**)	Same as above, but a lower limit of -2 is imposed on the score (see text)

Distance metrics (cntd)

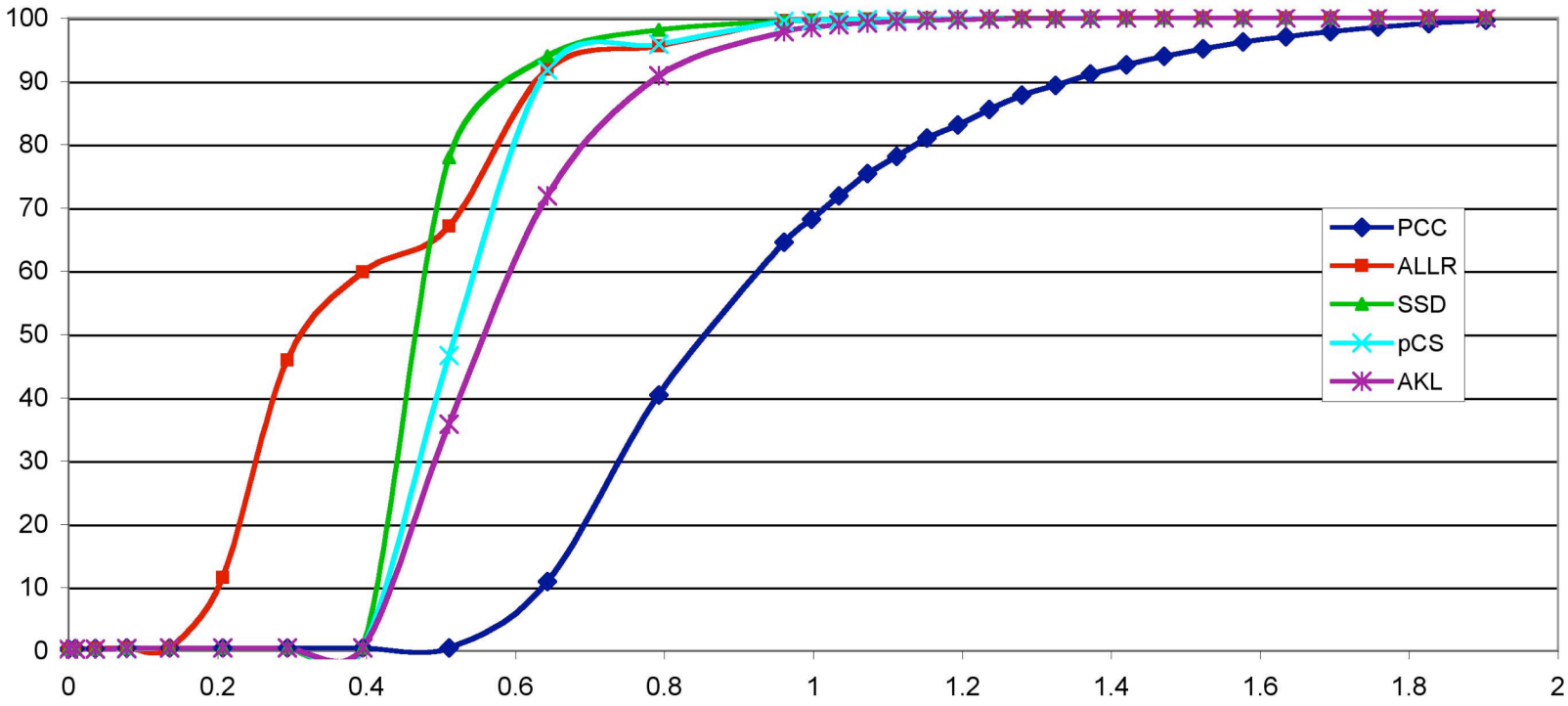


Distance metrics (cntd)

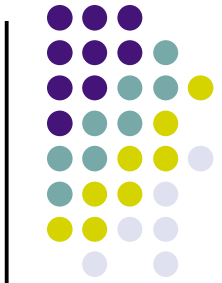


$$H = \sum_{b=A}^{T} f_I(b) \cdot \log_2 f_I(b)$$

Distance metrics (cntd)



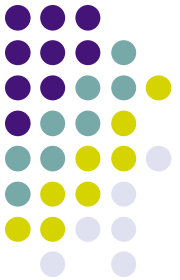
The tools: alignments and trees



- Alignment methods:
  - Needleman-Wunsch
  - Smith-Waterman

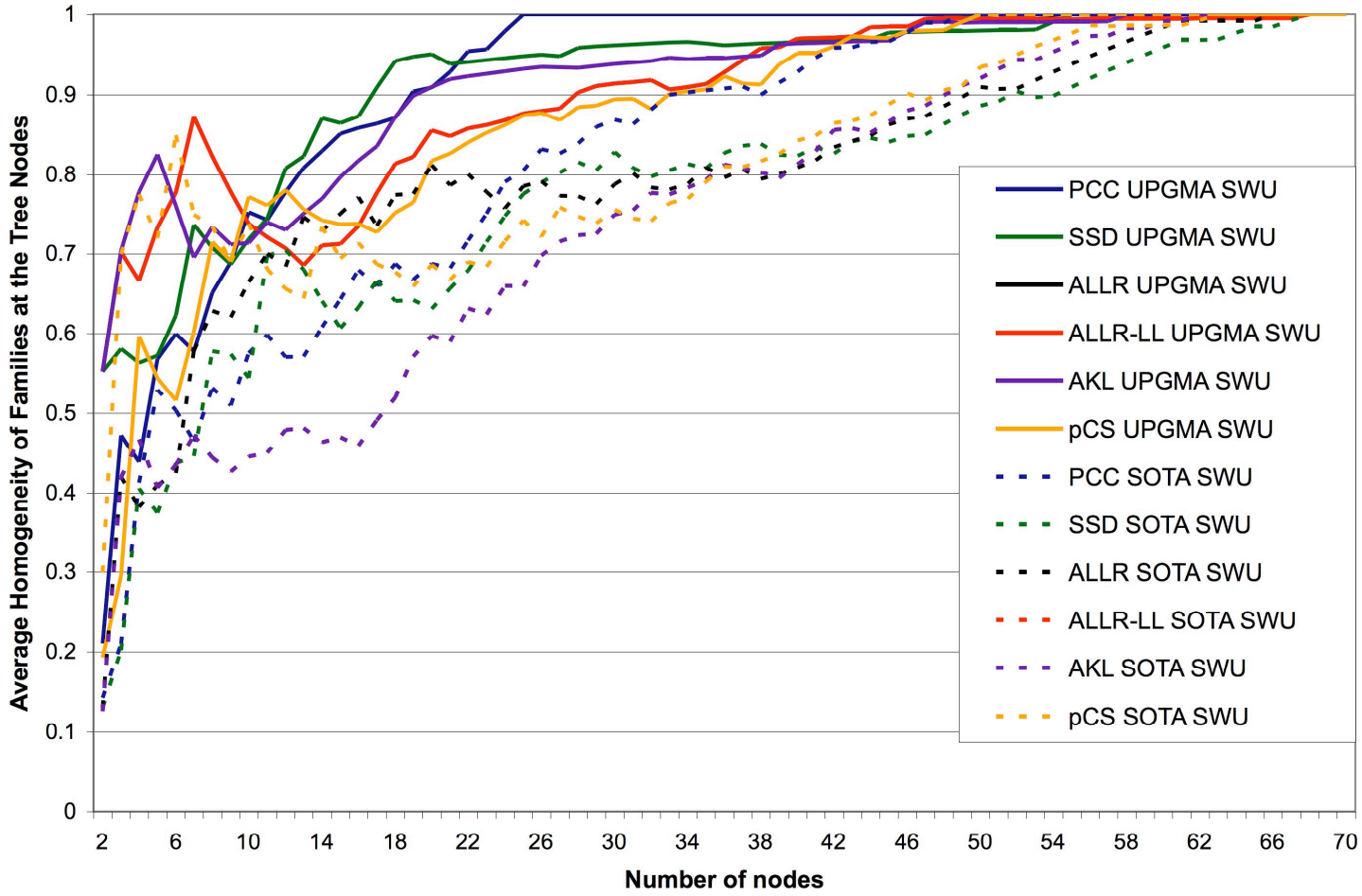
- Tree-building methods:
  - UPGMA
  - SOTA

Motif alignments

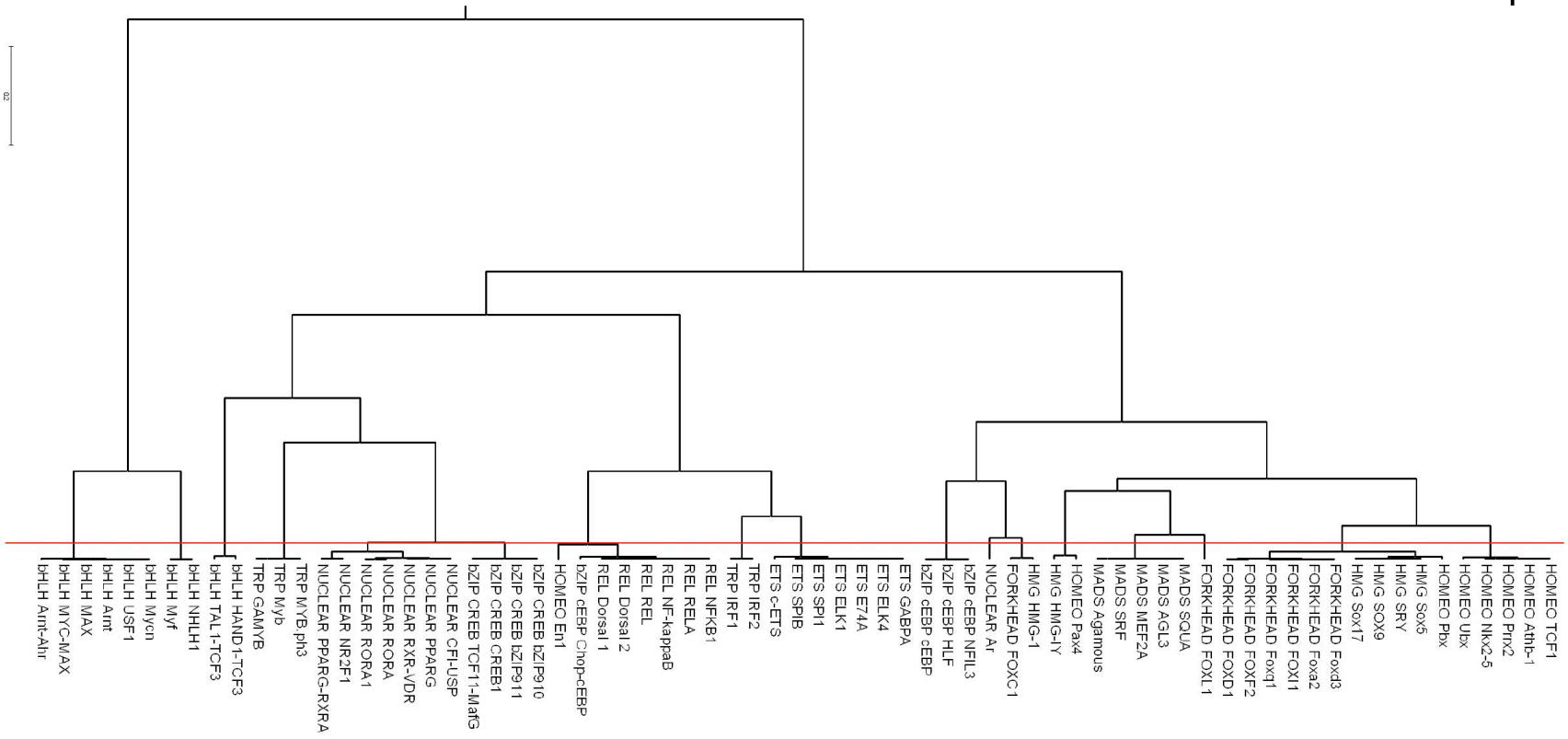


Alignment algorithm	Similarity metric	Gap open	JASPAR non-ZNF	JASPAR ZNF	TRANSFAC non-ZNF	TRANSFAC ZNF	*Avg*
SW	SSD	1.00	0.845	0.480	0.833	0.788	**0.811**
SW (ovrlp)	SSD	1.00	0.845	0.480	0.833	0.788	**0.811**
SW	SSD	0.75	0.845	0.440	0.831	0.794	**0.809**
SW (ovrlp)	SSD	0.75	0.845	0.440	0.831	0.794	**0.809**
SW	SSD	0.50	0.845	0.520	0.824	0.794	**0.808**
SW (ovrlp)	SSD	0.50	0.845	0.520	0.824	0.794	**0.808**
SW	PCC	1.50	0.845	0.520	0.817	0.800	**0.805**
SW	SSD	0.25	0.859	0.560	0.808	0.806	**0.804**
SW (ovrlp)	SSD	*0.25*	0.859	0.560	0.808	0.806	**0.804**
SW	PCC	1.00	0.831	0.600	0.815	0.788	**0.802**
SW (ovrlp)	PCC	1.00	0.831	0.600	0.812	0.788	**0.801**
SW (ungap)	PCC	N/A	0.887	0.600	0.812	0.763	**0.801**
SW	PCC	1.50	0.845	0.520	0.810	0.800	**0.801**
NW	SSD	1000	0.859	0.480	0.817	0.781	**0.801**
SW	SSD	1000	0.859	0.480	0.817	0.781	**0.801**
…	…	…	…	…	…	…	**…**
NW	pCS	0.75	0.662	0.600	0.657	0.675	**0.660**
NW	pCS	1.00	0.662	0.440	0.653	0.688	**0.654**
NW	pCS	0.50	0.690	0.560	0.653	0.650	**0.652**
NW	pCS	0.25	0.634	0.640	0.650	0.656	**0.650**
NW	ALLR	5.00	0.634	0.480	0.648	0.619	**0.633**
NW	AKL	4.00	0.704	0.400	0.594	0.600	**0.600**
NW	pCS	2.00	0.563	0.400	0.596	0.656	**0.600**
NW	SSD	1.00	0.634	0.480	0.585	0.581	**0.585**
NW	PCC	1.50	0.606	0.520	0.580	0.581	**0.581**
NW	PCC	2.00	0.521	0.480	0.423	0.500	**0.453**
NW	ALLR	10.00	0.394	0.440	0.343	0.400	**0.365**
NW	SSD	2.00	0.437	0.400	0.324	0.375	**0.350**
NW	ALLR	15.00	0.268	0.400	0.261	0.381	**0.295**
NW	PCC	4.00	0.282	0.480	0.282	0.269	**0.286**
NW	ALLR	20.00	0.268	0.440	0.232	0.306	**0.261**

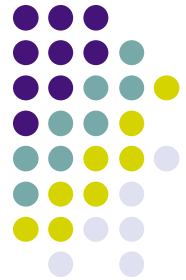
Tree-building



Tree-building (cntd)

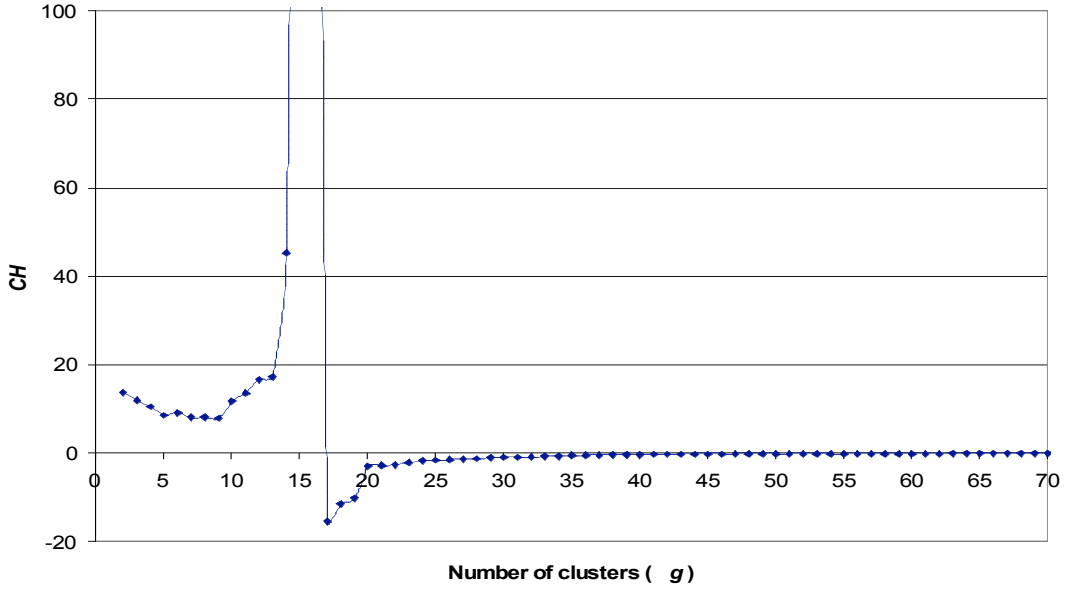


Automatic clustering statistics



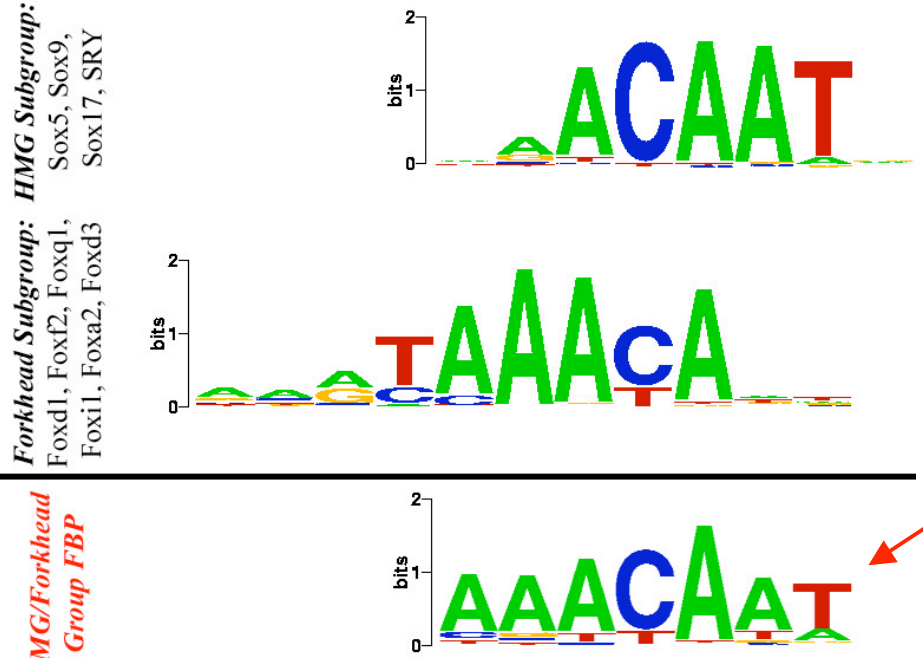
Statistic	Formula
Hastie, Tibshirani, Walter (***HTW***)	$Gap_n(g) = E_n^*(\log(W_g)) - \log(W_g)$
Calinski & Harabasz (***CH***)	$CH = \dfrac{\mathbf{B}/(g-1)}{\mathbf{W}/(n-g)}$
Log-modified Calinski & Harabasz (***CH$_{log}$***)	$CH_{\log} = \dfrac{\log(\mathbf{B})/(g-1)}{\log(\mathbf{W})/(n-g)}$

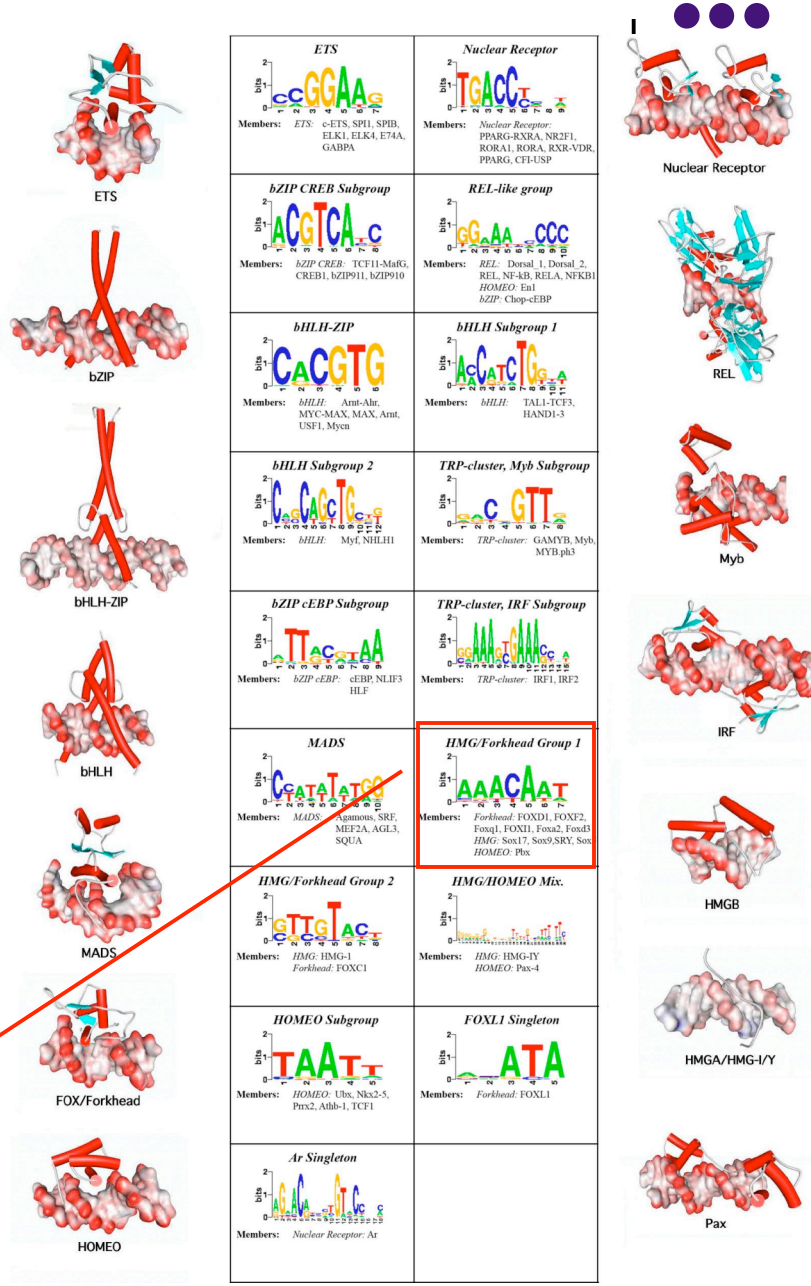
Automatic motif clustering



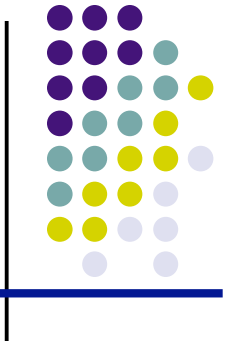
Dataset	Jacknife cross-validation	
	STAMP	Sandelin & Wasserman
Full set (79 motifs)	91% (72/79)	76% (60/79)
Non-zinc fingers (71 motifs)	94% (67/71)	87% (62/71)

New FBPs!


Forkhead Subgroup: Foxd1, Foxf2, Foxq1, Foxi1, Foxa2, Foxd3
HMG Subgroup: Sox5, Sox9, Sox17, SRY
HMG/Forkhead Group FBP
bits


ETS
bZIP
bHLH-ZIP
bHLH
MADS
FOX/Forkhead
HOMEO
ETS
Nuclear Receptor
bZIP CREB Subgroup
REL-like group
bHLH-ZIP
bHLH Subgroup 1
bHLH Subgroup 2
TRP-cluster, Myb Subgroup
bZIP cEBP Subgroup
TRP-cluster, IRF Subgroup
MADS
HMG/Forkhead Group 1
HMG/Forkhead Group 2
HMG/HOMEO Mix.
HOMEO Subgroup
FOXL1 Singleton
Ar Singleton
Nuclear Receptor
REL
Myb
IRF
HMGB
HMGA/HMG-I/Y
Pax

For some more reading



- Mahony, Auron, Benos, "*DNA Familial Binding Profiles made easy: comparison of various motif alignment and clustering strategies*" (2007) *PLoS Comput Biol* **3**: e61.