Prediction of Gene Expression in Yeast using Conserved Sequence Templates

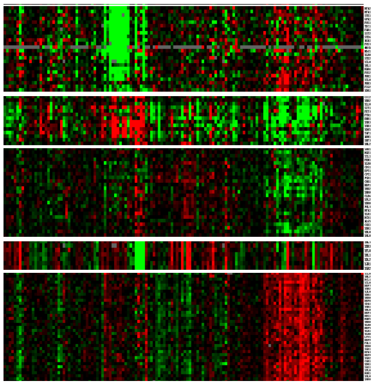
Derek Chiang & Alan Moses
Molecular & Cell Biology, UCB
Eisen Lab

PROBLEM

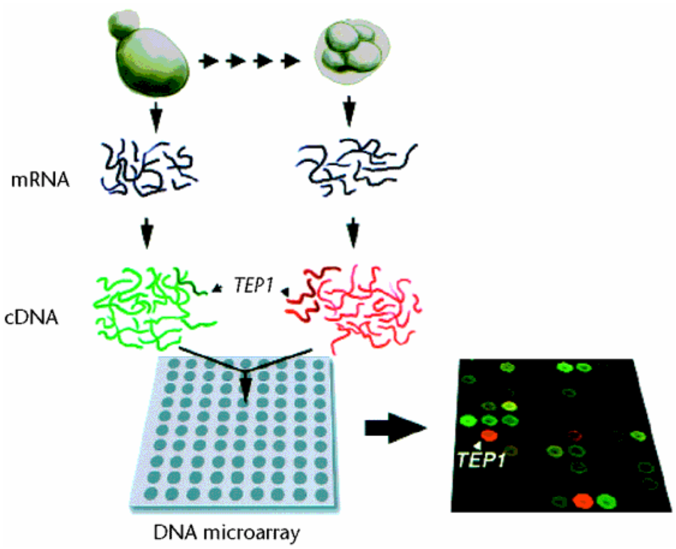
What information controlling gene expression
is encoded in genome sequences?

```
>Saccharomyces cerevisiae chr V
CGTCTCCTCCAAGCCCTGTTGTCTCTTACCC
GGATGTTCAACCAAAAGCTACTTACTACCTT
TATTTTATGTTTACTTTTTATAGATTGTCTT
TTTATCCTACTCTTTCCCACTTGTCTCTCGC
TACTGCCGTGCAACAAACACTAAATCAAAAC
AGTGAAATACTACTACATCAAAACGCATATT
CCCTAGAAAAAAAAATTTCTTACAATATACT
ATACTACACAATACATAATCACTGACTTTCG
TAACAACAATTTCCTTCACTCTCCAACTTCT
CTGCTCGAATCTCTACATAGTAATATTATAT
CAAATCTACCGTCTGGAACATCATCGCTATC
CAGCTCTTTGTGAACCGCTACCATCAGCATG
TACAGTGGTACCTTCGTGTTATCTGCAGCGA
GAACTTCAACGTTTGCCAAATCAAGCCAATG
TGGTAACAACCACACCTCCGAAATCTGCTCC
AAAAGATACTCCAGTTTCTGCCGAAATGTTT
```

Features? →

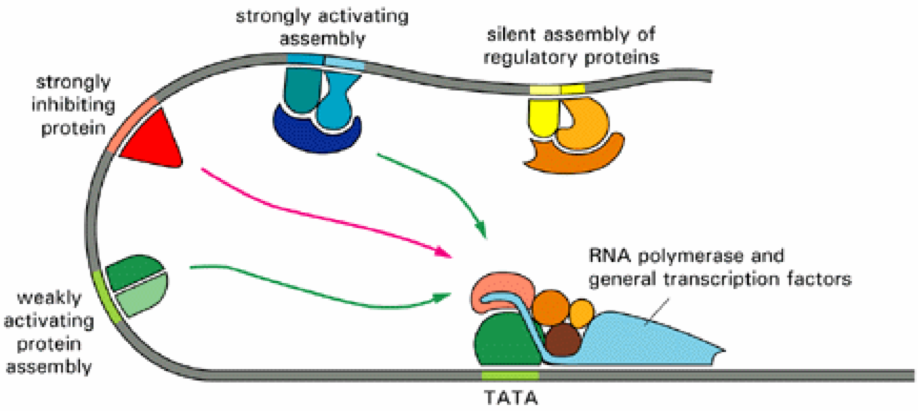


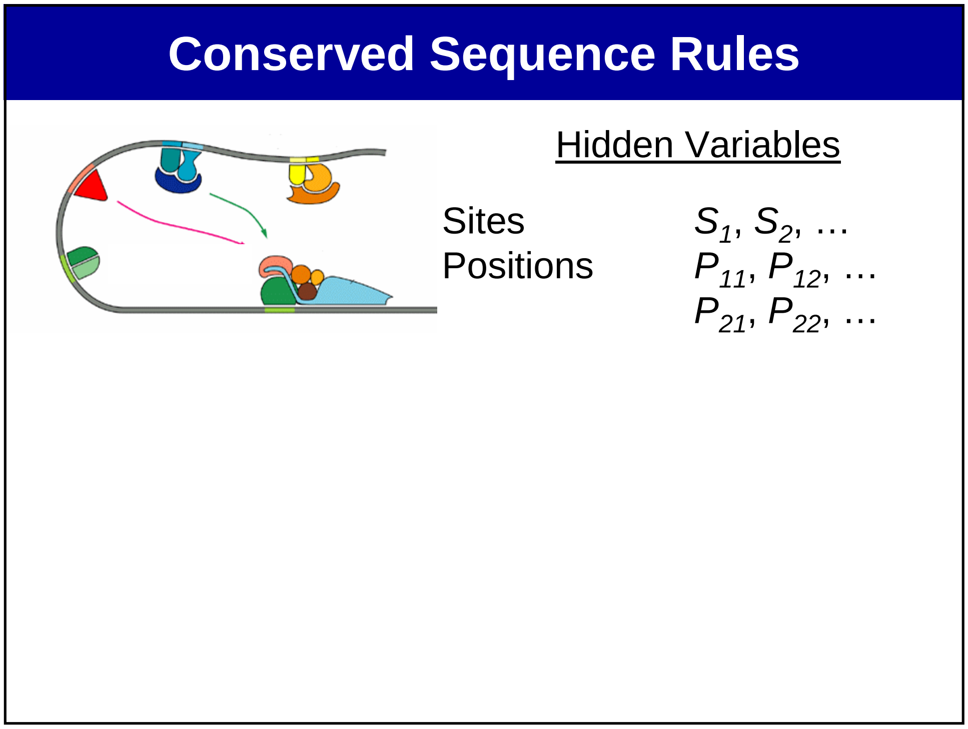
Gene Expression: Experiment



Gene Expression: Mechanism

Promoter Structure


strongly activating assembly
silent assembly of regulatory proteins
strongly inhibiting protein
weakly activating protein assembly
RNA polymerase and general transcription factors
TATA



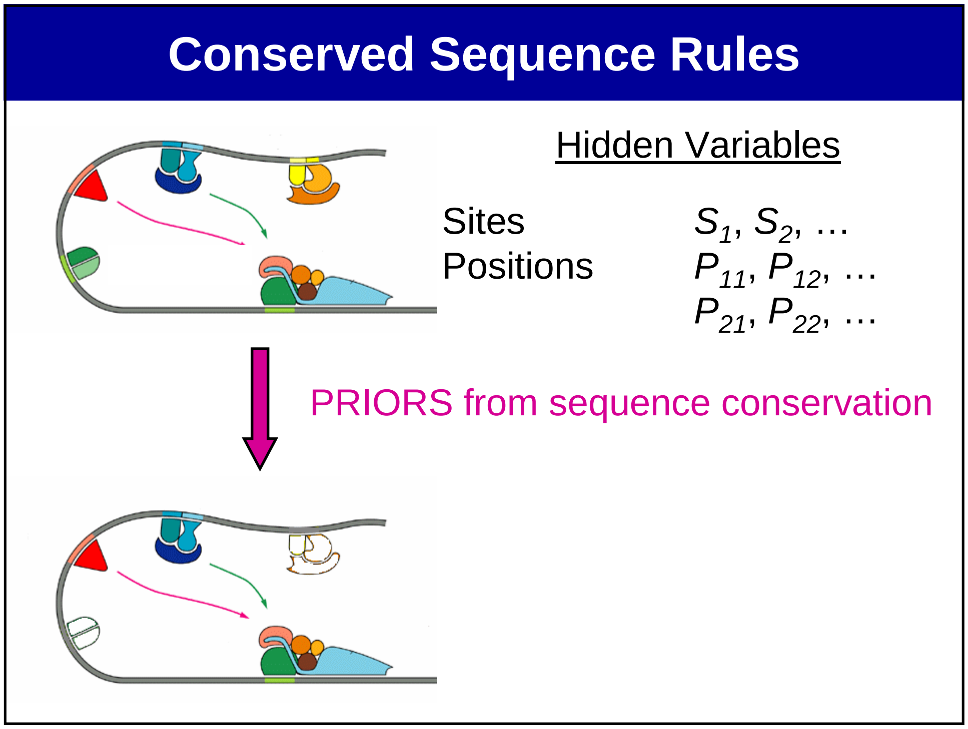
Conserved Sequence Rules

Hidden Variables

Sites $S_1, S_2, \ldots$

Positions $P_{11}, P_{12}, \ldots$

$P_{21}, P_{22}, \ldots$



Conserved Sequence Rules

Hidden Variables

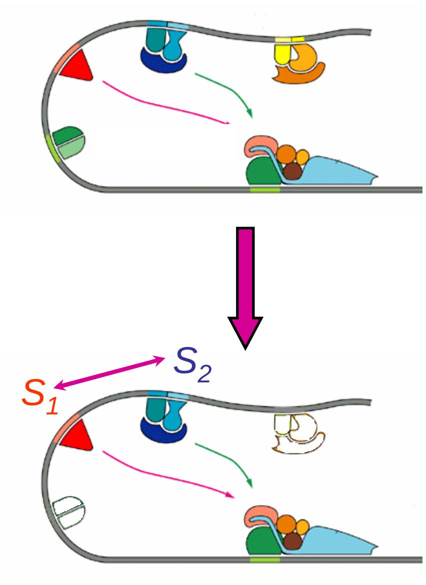
Sites $S_1, S_2, \ldots$

Positions $P_{11}, P_{12}, \ldots$

$P_{21}, P_{22}, \ldots$

PRIORS from sequence conservation

Conserved Sequence Rules



Hidden Variables

Sites	$S_1, S_2, \ldots$
Positions	$P_{11}, P_{12}, \ldots$
	$P_{21}, P_{22}, \ldots$

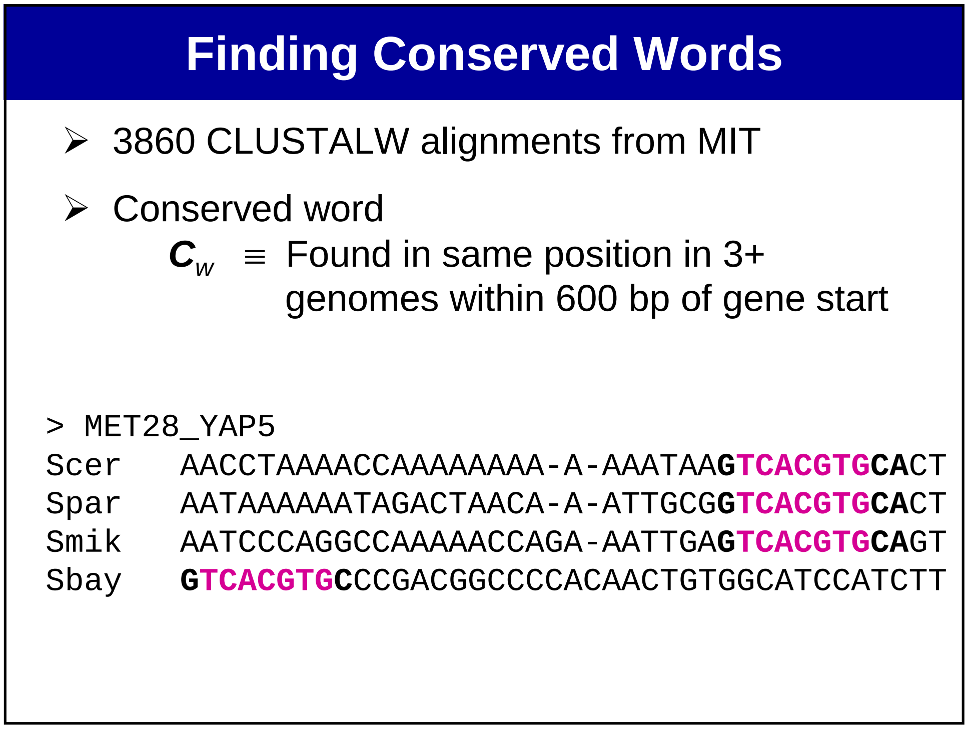
Conserved Sequence Rules

$$\{\ S_1 > 0;\quad S_2 > 0;\quad \min_j(\ |P_{1j} - P_{2j}| < 30\ )\ \}$$

Conserved Word Pairs

- **Finding Conserved Words**
- Evaluate Word Pairs

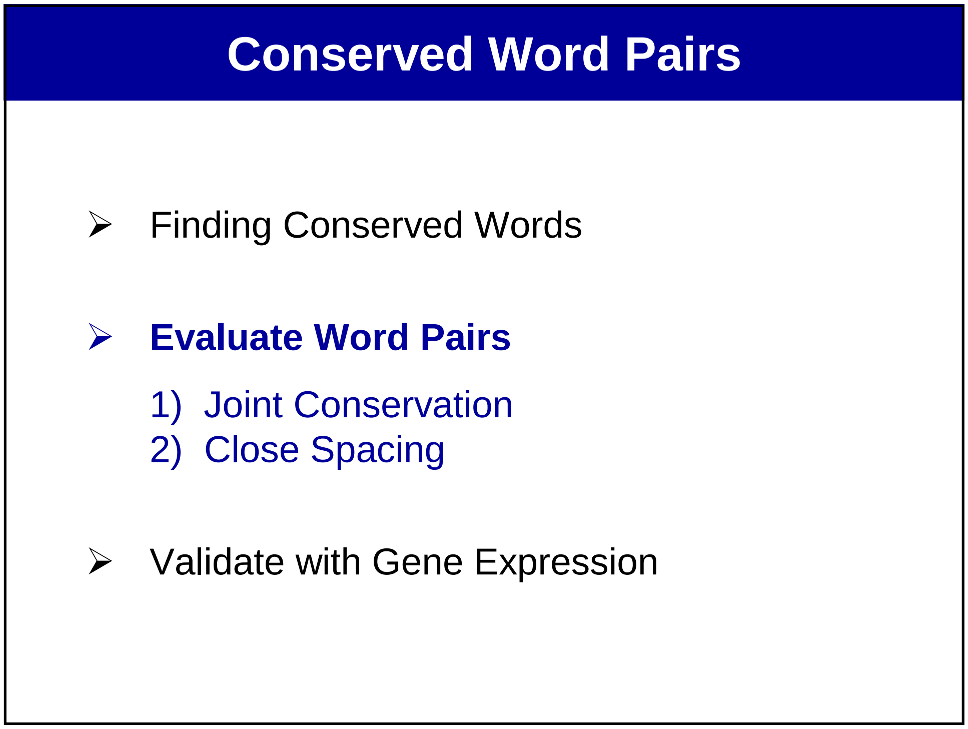
  1) Joint Conservation
  2) Close Spacing
- Validate with Gene Expression



Finding Conserved Words

- 3860 CLUSTALW alignments from MIT
- Conserved word
  C_w ≡ Found in same position in 3+ genomes within 600 bp of gene start

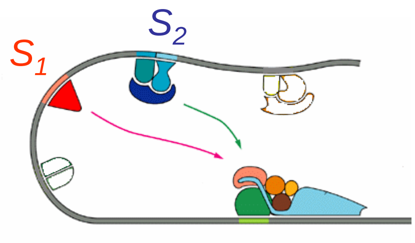
```
> MET28_YAP5
Scer    AACCTAAAACCAAAAAAAA-A-AAATAAGTCACGTGCACT
Spar    AATAAAAAATAGACTAACA-A-ATTGCGGTCACGTGCACT
Smik    AATCCCAGGCCAAAAACCAGA-AATTGAGTCACGTGCAGT
Sbay    GTCACGTGCCCGACGGCCCCACAACTGTGGCATCCATCTT
```



Conserved Word Pairs

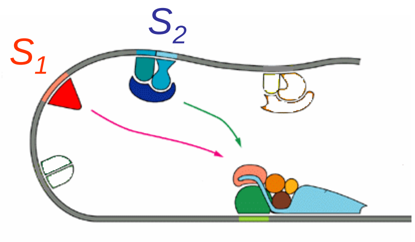
- Finding Conserved Words
- **Evaluate Word Pairs**
  1) Joint Conservation
  2) Close Spacing
- Validate with Gene Expression

TEST: Joint Word Conservation



Word 1 Conserved	Word 2 Conserved: Y	Word 2 Conserved: N
Y		
N		

TEST: Joint Word Conservation



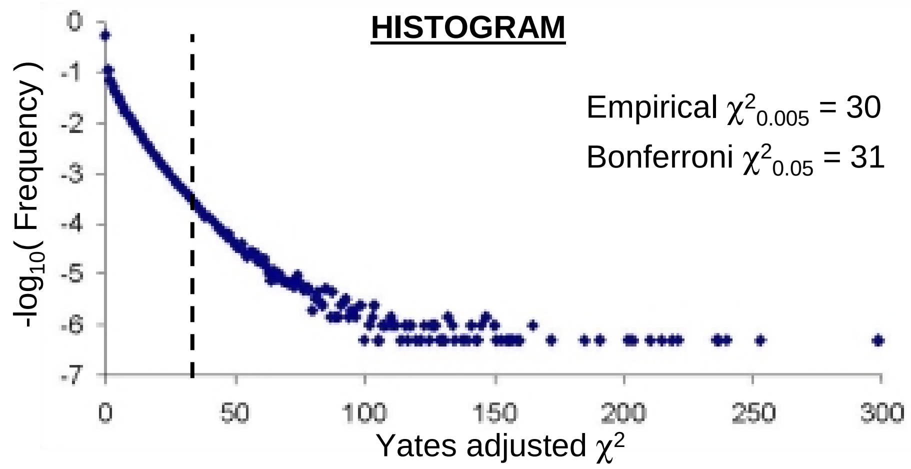
Word 1 Conserved	Word 2 Conserved: Y	Word 2 Conserved: N
Y	32	162
N	134	3226

Chi-square Test for Independence
(Yates adjustment)

$$\chi^2 = \sum_k \frac{(|O_k - E_k| - 1/2)^2}{E_k}$$

TEST: Joint Word Conservation

- 2090 words (Length 6: Word-Rev complement)
- 2.06×10^6 Word PAIRS (Exclude overlap)



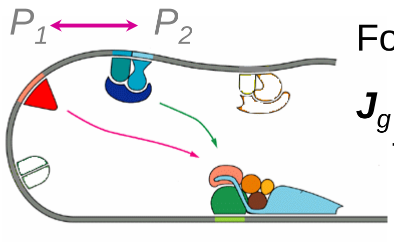
HISTOGRAM

Empirical $\chi^2_{0.005} = 30$

Bonferroni $\chi^2_{0.05} = 31$

Yates adjusted χ^2

TEST: Close Spacing



For conserved genes g : 1 … N

J_g = ($\overrightarrow{P_{1g}}$, $\overrightarrow{P_{2g}}$) sampled jointly from position distributions

Test Statistic

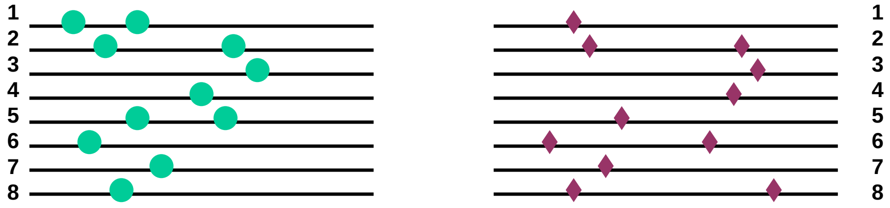
$$X = \frac{1}{N}\sum_{g} \min_{k \in g_i} | P_{1k} - P_{2k} |$$

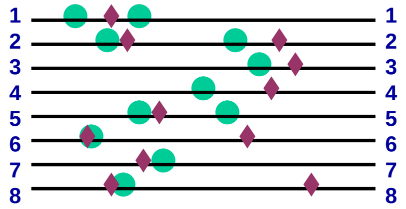
NULL: ***X*** obtained from random sampling

ALT: ***X*** smaller than expected from sampling

TEST: Close Spacing

Nonparametric Bootstrap



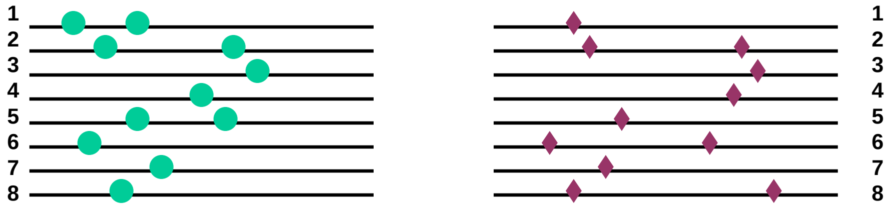


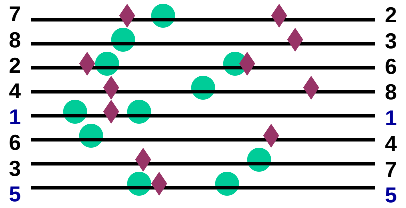
Avg. Min. Distance

$\boldsymbol{X}$ = 32 bp

TEST: Close Spacing

Nonparametric Bootstrap





Bootstrap Distance

$\boldsymbol{X}^{*b}$ = 65 bp

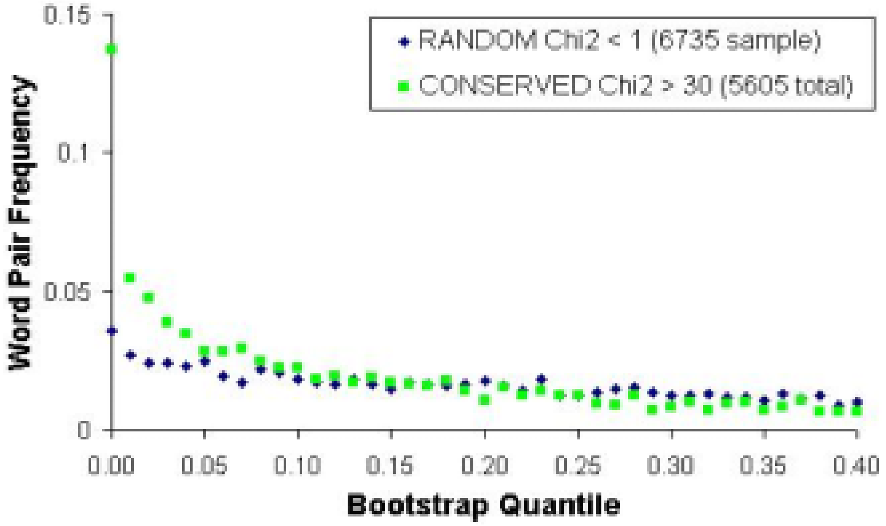
TEST: Close Spacing

Nonparametric Bootstrap

- Position distributions $\Pi_1 = \{ \overrightarrow{\boldsymbol{P}}_{11}, \ldots , \overrightarrow{\boldsymbol{P}}_{1v} \}$
  $\Pi_2 = \{ \overrightarrow{\boldsymbol{P}}_{21}, \ldots , \overrightarrow{\boldsymbol{P}}_{2w} \}$
- Data $\boldsymbol{J}_1 = (\overrightarrow{\boldsymbol{P}}_{11}, \overrightarrow{\boldsymbol{P}}_{21}), \ldots , \boldsymbol{J}_n = (\overrightarrow{\boldsymbol{P}}_{1n}, \overrightarrow{\boldsymbol{P}}_{2n})$
- Bootstrap $\boldsymbol{J}_1^{*b} = (\overrightarrow{\boldsymbol{P}}_{11}{}^{*b}, \overrightarrow{\boldsymbol{P}}_{21}{}^{*b}), \ldots , \boldsymbol{J}_n = (\overrightarrow{\boldsymbol{P}}_{1n}{}^{*b}, \overrightarrow{\boldsymbol{P}}_{2n}{}^{*b})$
  resampled with replacement from Π_1, Π_2
- Record quantile of $\boldsymbol{X}$ in 100000 samples of $\boldsymbol{X}^{*b}$
  (empirical null distribution)

TEST: Close Spacing

HISTOGRAM of Bootstrap Quantiles for Word Pairs

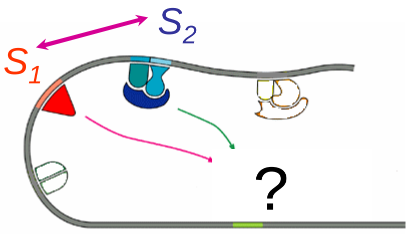


Conserved Word Pairs

- ➢ Finding Conserved Words
- ➢ Evaluate Word Pairs
  1) Joint Conservation
  2) Close Spacing
- ➢ **Validate with Gene Expression**

Validating Expression Subsets

Conserved Sequence (SUBSETTING) Rules



Genome (6000 genes)

$\{\ S_1 > 0;\quad S_2 > 0;\quad \min_j(\,|\,P_{1j} - P_{2j}\,| < d\,)\ \}$

SUBSET (*N* genes)

Assess gene expression

Validating Expression Subsets

Some Nonparametric Tests

1) Subset mean (Mann-Whitney)

2) Subset distribution (Kolmogorov-Smirnov)

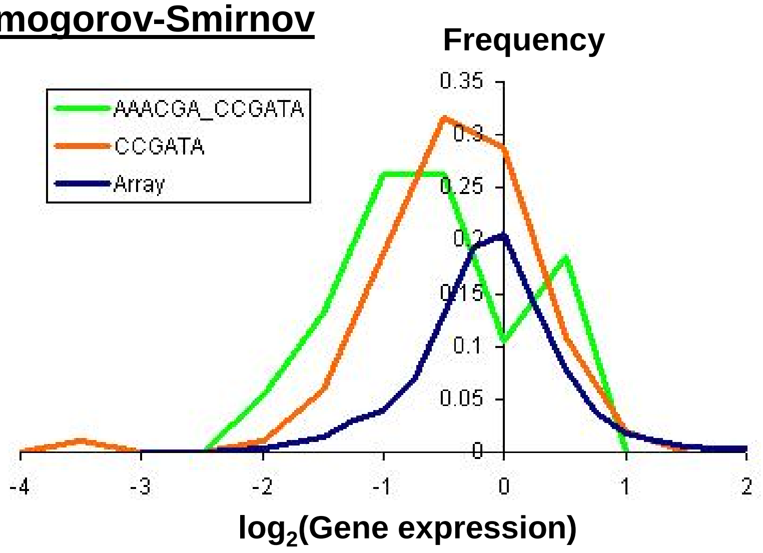
3) Subset weighted correlation (?)

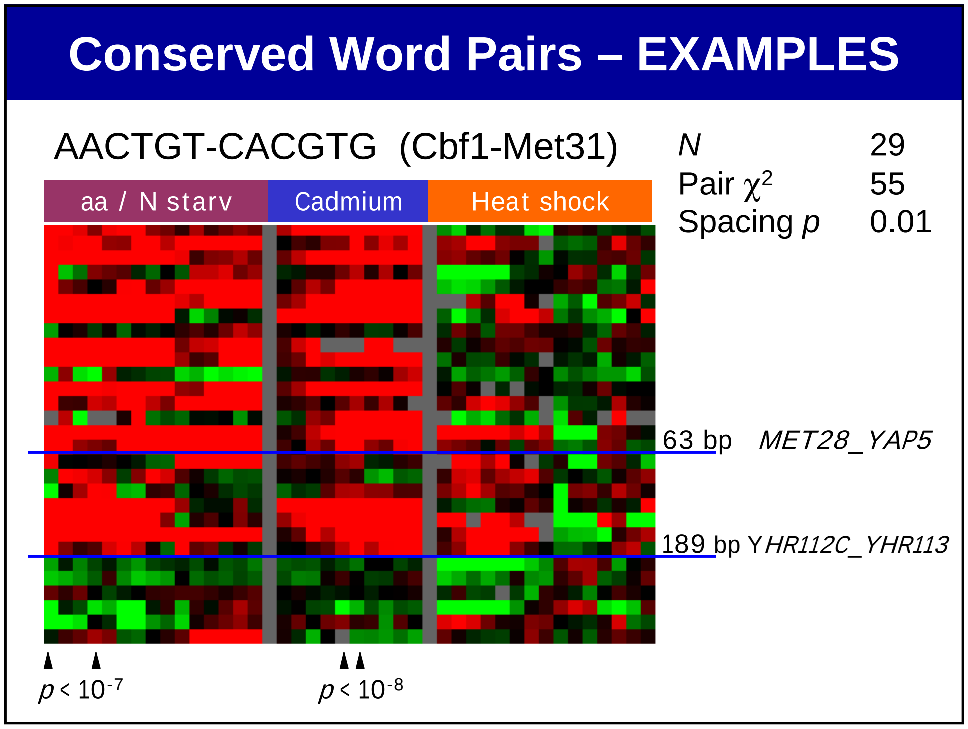
4) Kernel density classification (Mixture of normals)

Find optimal word pair distance ***d**** for each test …

Validating Expression Subsets

Kolmogorov-Smirnov

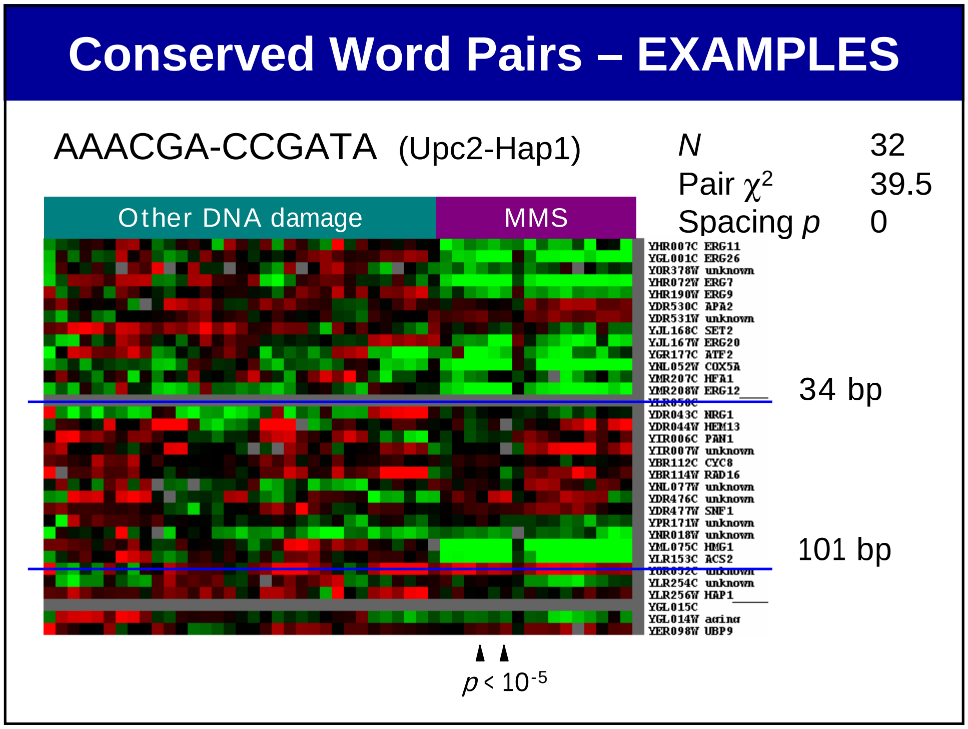




Conserved Word Pairs – EXAMPLES

AACTGT-CACGTG (Cbf1-Met31)

N	29
Pair χ^2	55
Spacing *p*	0.01

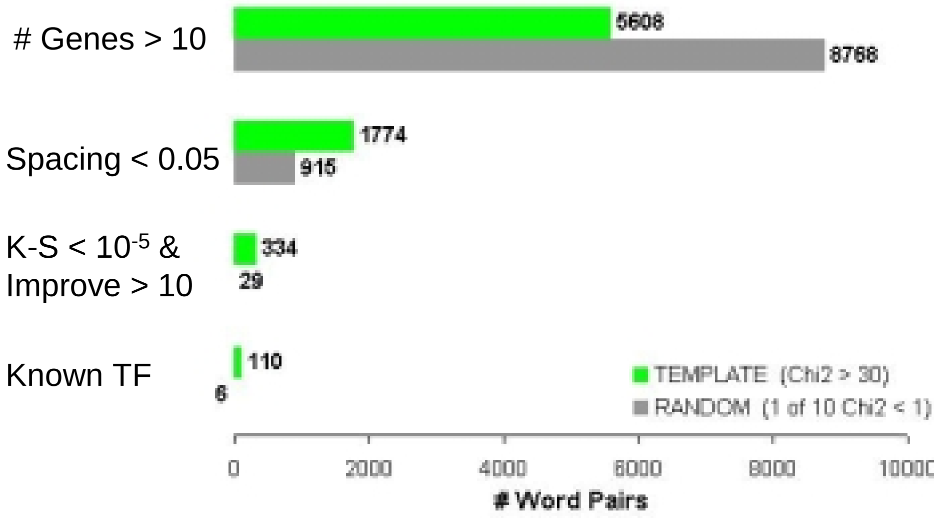


Conserved Word Pairs – EXAMPLES

AAACGA-CCGATA (Upc2-Hap1)

N	32
Pair χ^2	39.5
Spacing *p*	0

Pipeline Summary



Future Directions

- Better gene expression subset tests (Timecourse)
- More flexible sequence models (IUPAC, Self-dimer)
- Automate distance cutoff (Distance *d*)
- Parameter optimization: 8 threshold values!
  (Conserved: # aligned genomes & # upstream bp,
  Joint conservation χ^2, Bootstrap quantile, K-S probs,
  Min gene #, Distance *d*)

Acknowledgements

Michael Eisen

Eisen Lab

Justin Fay

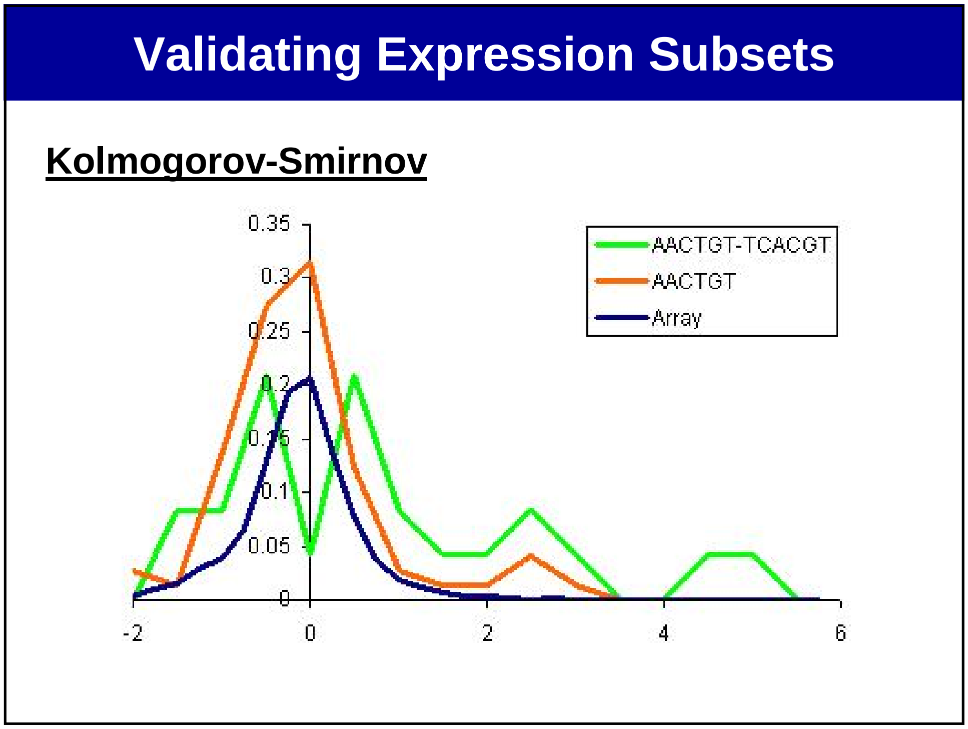
Audrey Gasch

Hunter Fraser

Venky Nandagopal

Dan Pollard

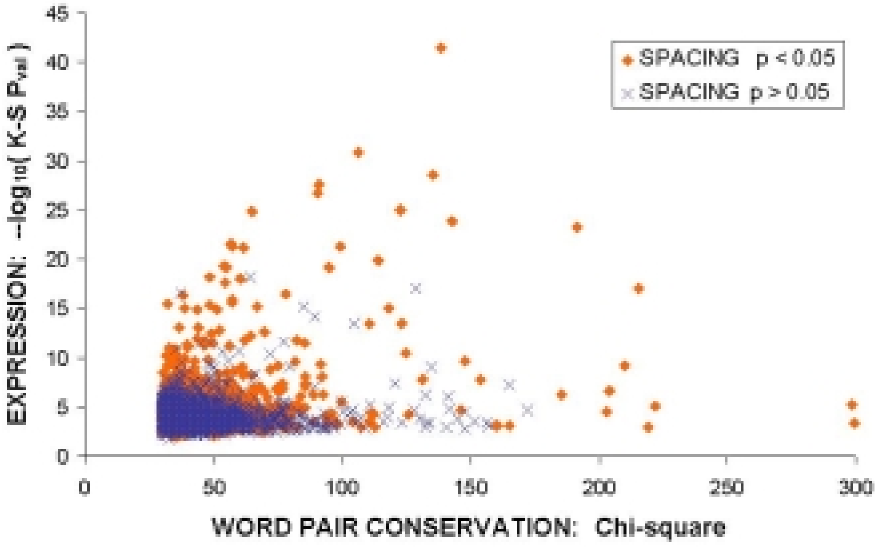
Ben Lewis



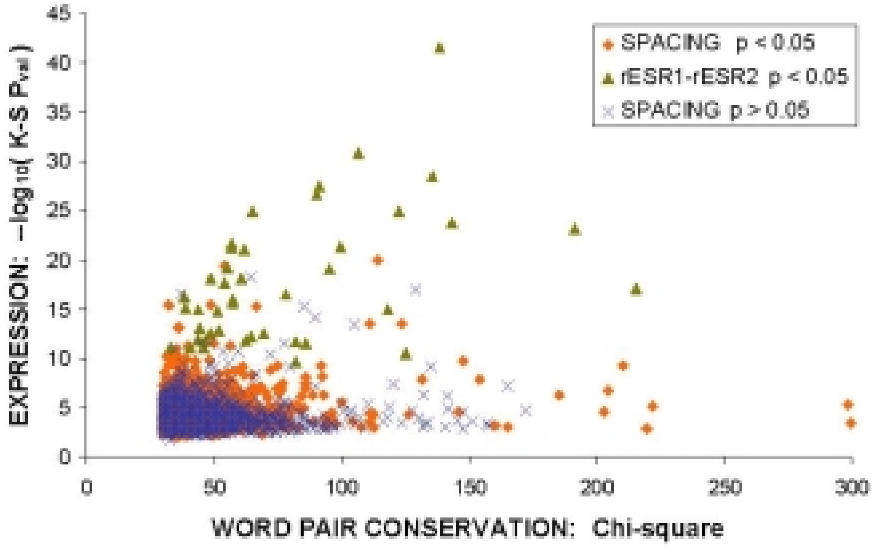
Validating Expression Subsets

Kolmogorov-Smirnov

Comparing Expression Samples



Comparing Expression Samples


SPACING p < 0.05
rESR1-rESR2 p < 0.05
SPACING p > 0.05
EXPRESSION: −log10(K-S Pval)
WORD PAIR CONSERVATION: Chi-square