02-716

# Cross species analysis of genomics data

**Motif discovery and regulatory interactions**

# Biological Motif Discovery

**Concepts**

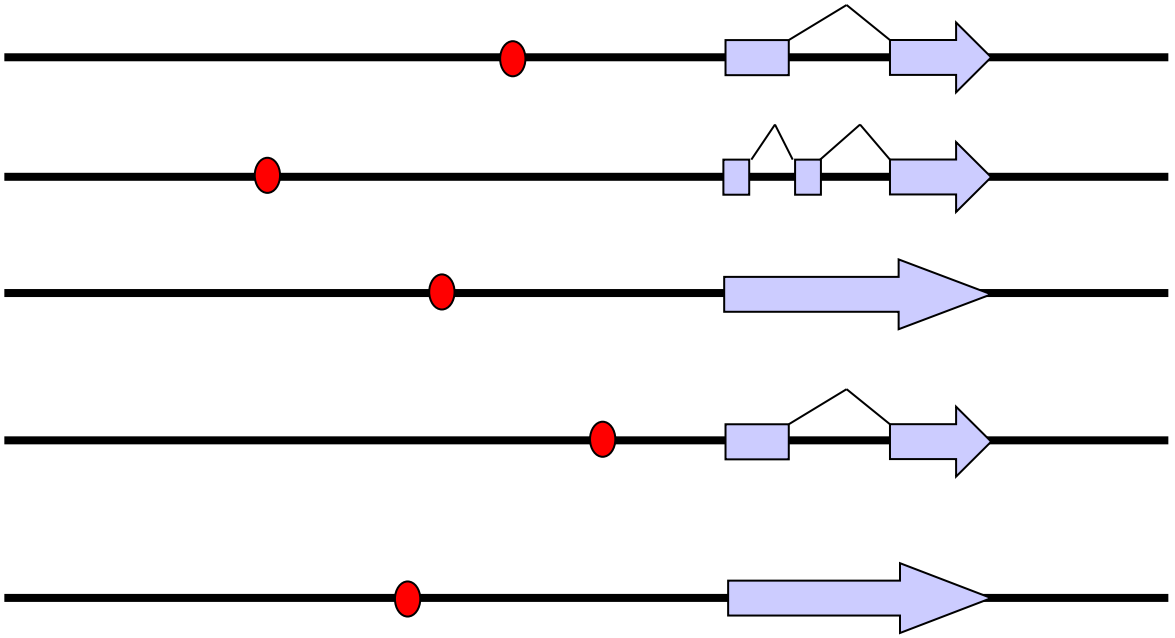
- Motif Modeling and Motif Information
- EM and Gibbs Sampling
- Comparative Motif Prediction

**Applications**

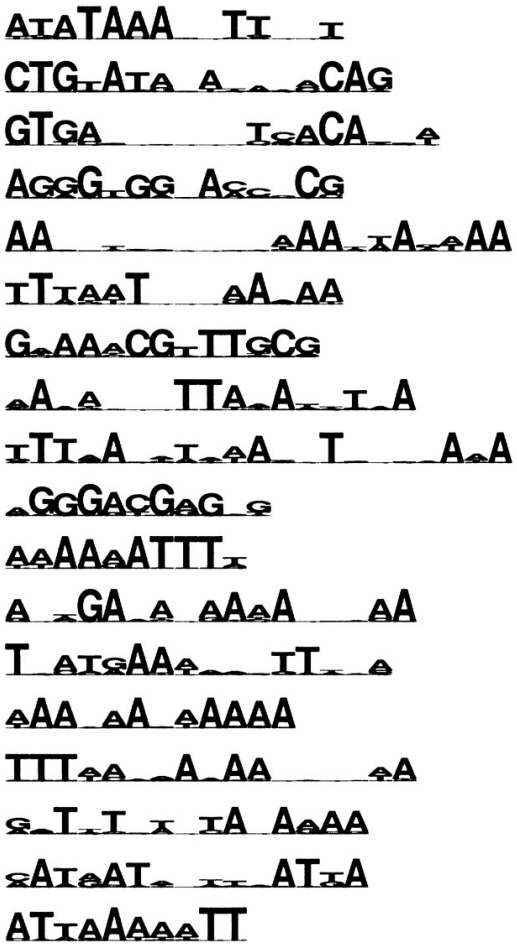
- Transcription Factor Binding Site Prediction

# Regulatory Motifs

**Find promoter motifs associated with co-regulated or functionally related genes**



# Transcription Factor Binding Sites



- Very Short
- Highly Variable
- ~Constant Size
- Often repeated
- Low-complexity-ish

**Slide Credit: S. Batzoglou**

# Other “Motifs”

- Splicing Signals
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)
- Protein Domains
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
- MicroRNA target sites
  - Complementary base pairing

# Essential Tasks

- Modeling Motifs
  - How to computationally represent motifs
- Visualizing Motifs
  - Motif “Information”
- Predicting Motif Instances
  - Using the model to classify new sequences
- Learning Motif Structure
  - Finding new motifs, assessing their quality

# Modeling Motifs

# Consensus Sequences

**Useful for publication**

**Not very amenable to computation**

**IUPAC symbols for degenerate sites**

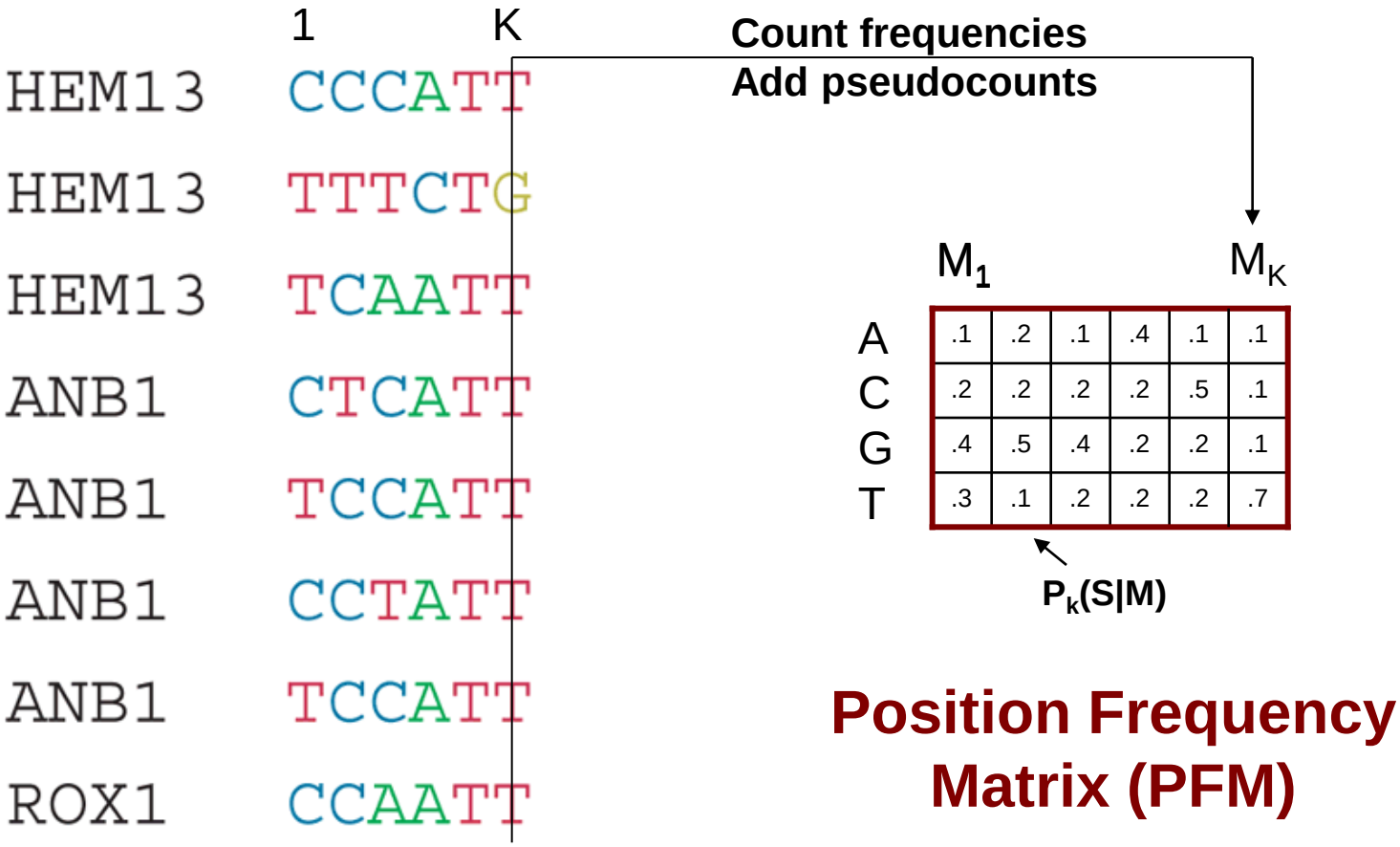
**Symbols:**
**Y = T** or **C** (pyrimidine)
**R** = **G** or **A** (purine)
**H** = **A** or **C** or **T**
**N = A** or **C** or **G** or **T** (any)
etc.

```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG
```

**YCHATTGTTCTC**

*Nature Biotechnology* **24, 423 - 425 (2006)**

# Probabilistic Model




**Position Frequency Matrix (PFM)**

# Scoring A Sequence

**To score a s ompare to a null model**

**Log likelihood ratio**

$$Score = \log \frac{P(S \mid PFM)}{P(S \mid B)}$$

**PFM**

| | | | | | | |
|---|---|---|---|---|---|---|
| A | .1 | .2 | .1 | .4 | .1 | .1 |
| C | .2 | .2 | .2 | .2 | .5 | .1 |
| G | .4 | .5 | .4 | .2 | .2 | .1 |
| T | .3 | .1 | .2 | .2 | .2 | .7 |

**Background DNA (B)**

A: 0.25

T: 0.25

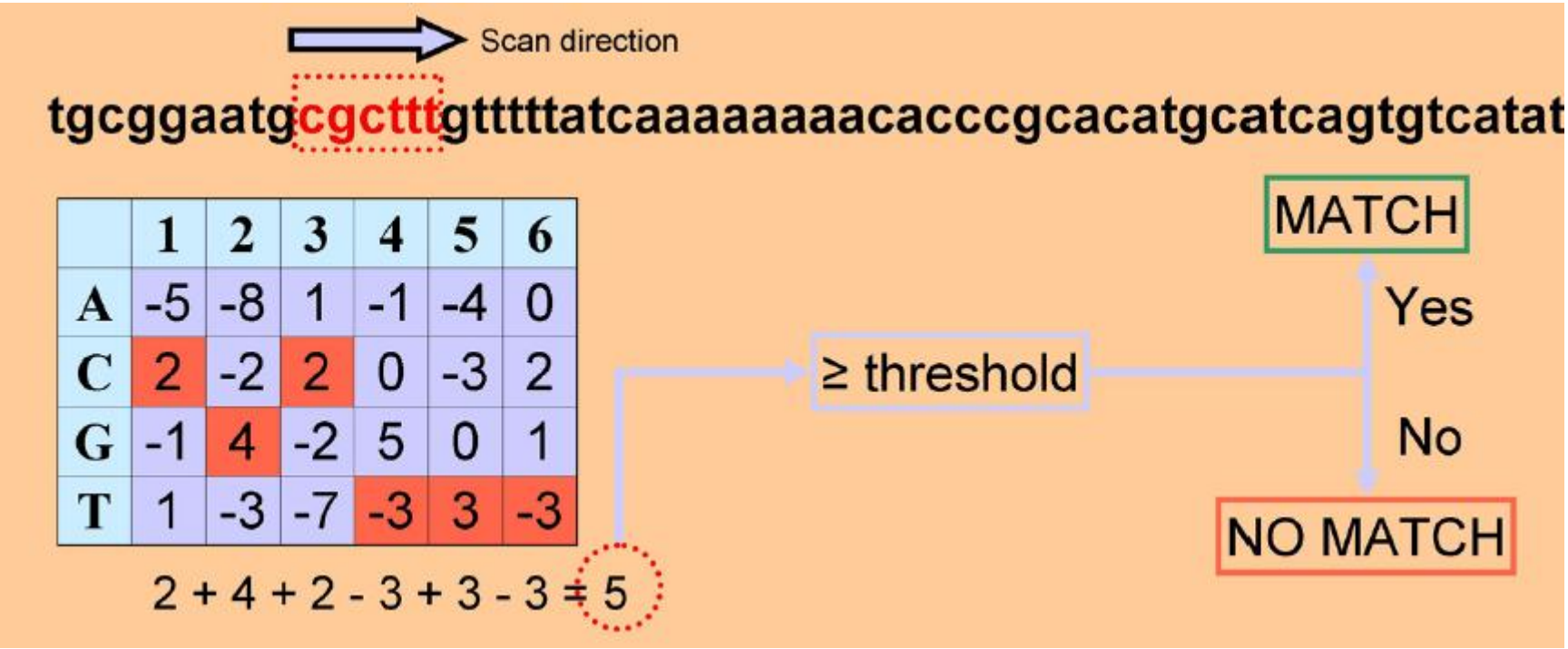
G: 0.25

C: 0.25

**Position Weight Matrix (PWM)**

| | | | | | | |
|---|---|---|---|---|---|---|
| A | -1.3 | -0.3 | -1.3 | 0.6 | -1.3 | -1.3 |
| C | -0.3 | -0.3 | 0.3 | -0.3 | 1 | -1.3 |
| G | 0.6 | 1 | 0.6 | -0.3 | -0.3 | -1.3 |
| T | 0.3 | -1.3 | -0.3 | -0.3 | -0.3 | 1.4 |

# Scoring a Sequence



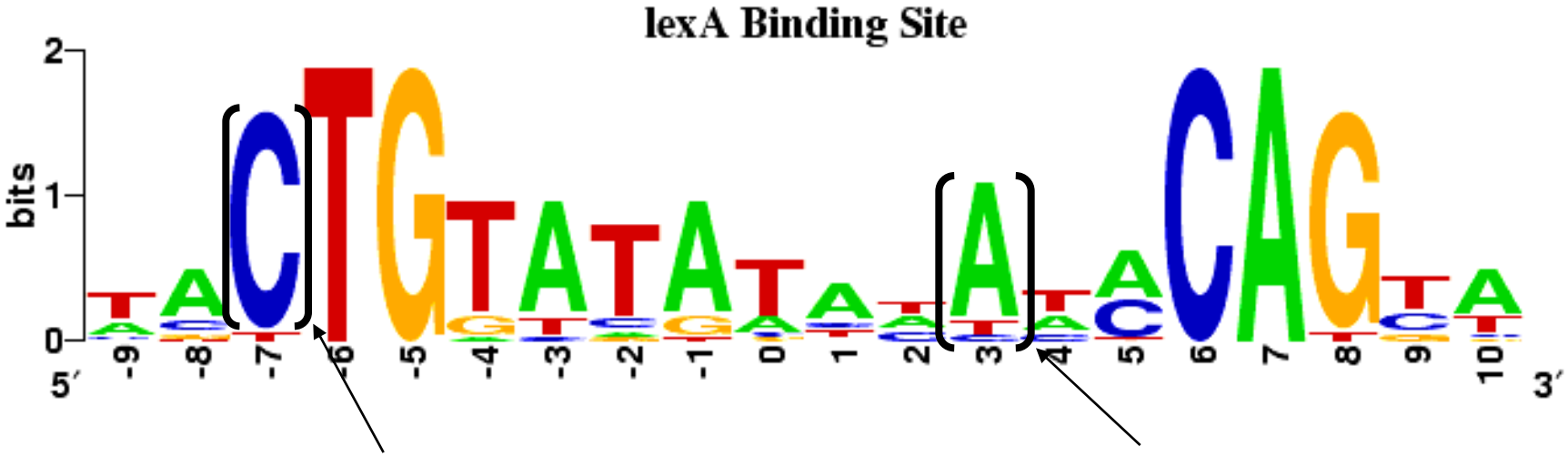

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | -5 | -8 | 1 | -1 | -4 | 0 |
| C | 2 | -2 | 2 | 0 | -3 | 2 |
| G | -1 | 4 | -2 | 5 | 0 | 1 |
| T | 1 | -3 | -7 | -3 | 3 | -3 |

**Common threshold = 60% of maximum score**

**MacIsaac & Fraenkel (2006) PLoS Comp Bio**

# Visualizing Motifs – Motif Logos

**Represent both base frequency and conservation at each position**

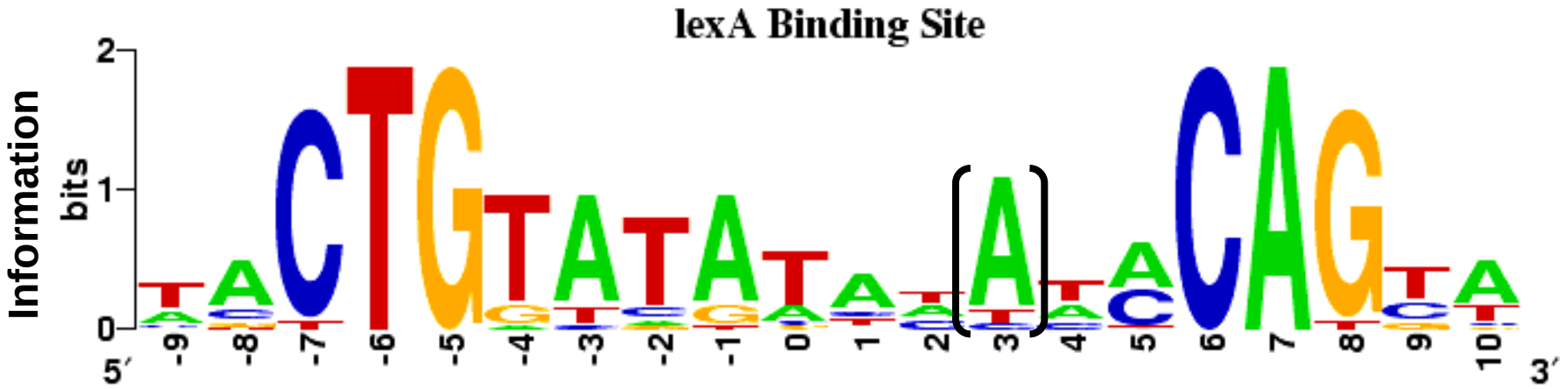



**Height of letter proportional to frequency of base at that position**

**Height of stack proportional to conservation at that position**

# Motif Information

The height of a stack is often called the motif information at that position measured in bits




$$\text{Motif Position Information} = 2 - \sum_{b=\{A,T,G,C\}} -p_b \log p_b$$

***Why is this a measure of information?***

# Online Logo Generation

WEBLOGO

· about · create · examples ·

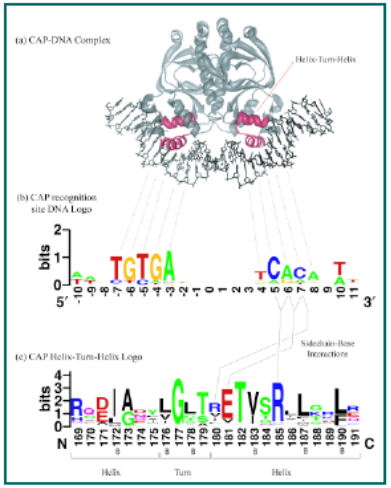
Version 2.8.2 (2005-09-08)

(⇒ WebLogo 3: Public Beta)

## References

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004) [Full Text ]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res. 18*:6097-6100



## Introduction

**WebLogo** is a web based application designed to make the generation of sequence logos as easy and painless as possible. Click **here** to create your own sequence logos.

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

http://weblogo.berkeley.edu/

enoLOGOS

UCSD

matrix or alignment input (select example) C2H2 enoLOGOS form

no input parameters set

enoLOGOS parameters submit (select defaults)

| | | | |
|---|---|---|---|
| weight type | unknown | energy units | none |
| logo title | | logo plot method | relative entropy |
| x-axis label | | scale letters by prob. | ON wts as is |
| y-axis label | bits | log base | 2 |
| y-axis max | 2 | mutual info | OFF reverse-comp OFF |
| x-axis,y-axis | ON ON | column aspect ratio | 3 |

| letters | red | green | blue | %GC (select %GC) |
|---|---|---|---|---|
| A | 0.0 | 0.8 | 0.0 | |
| C | 0.0 | 0.0 | 0.8 | |
| G | 0.8 | 0.8 | 0.1 | |
| T | 0.8 | 0.0 | 0.1 | |

Supported by the National Science Foundation

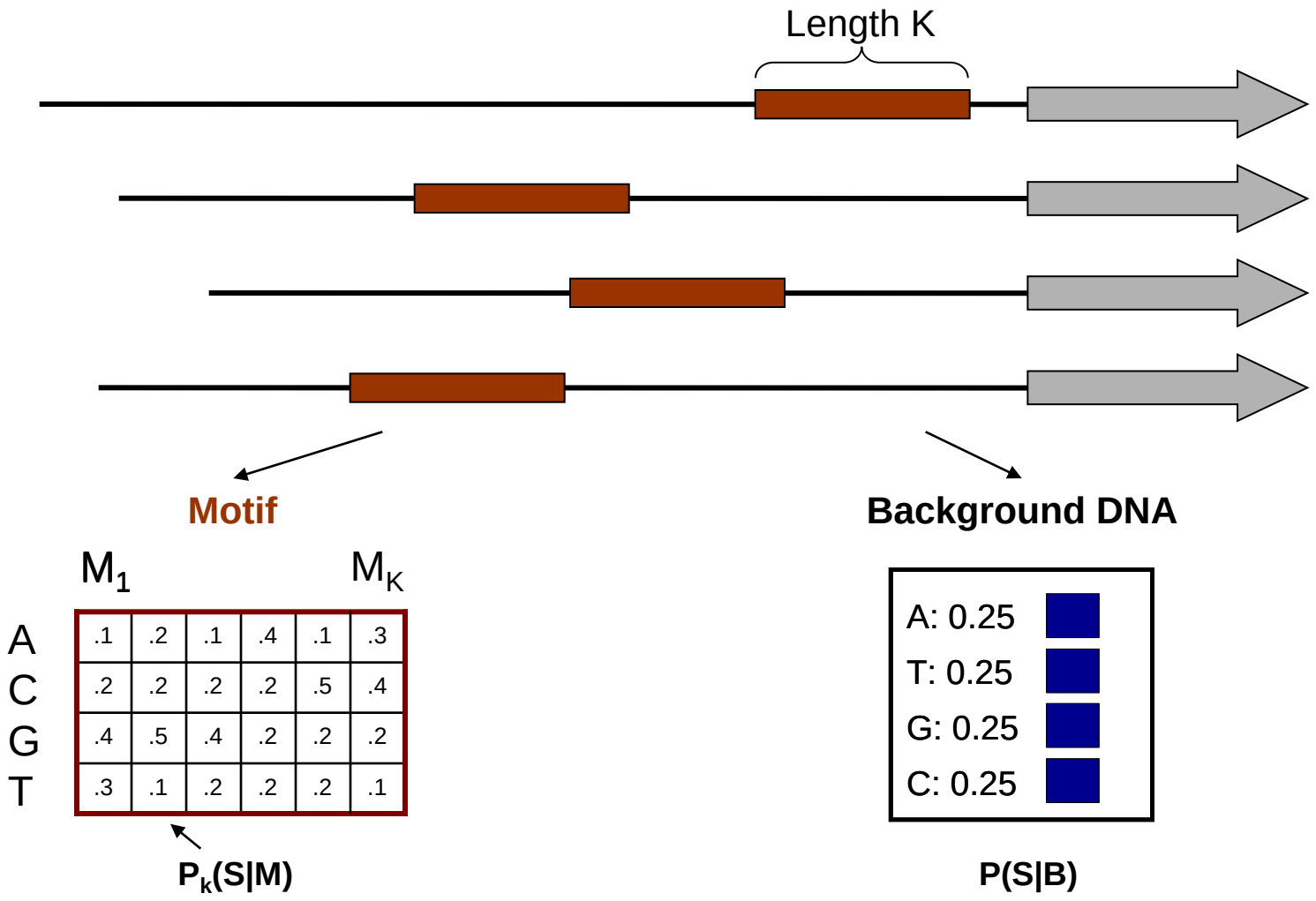
Reference UCSD mirror

http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi

# Finding New Motifs

## Learning Motif Models

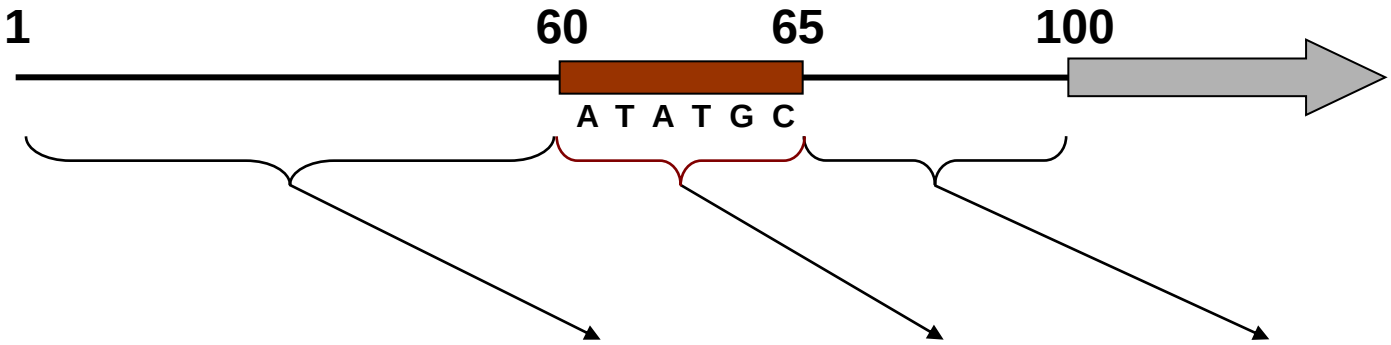
# A Promoter Model




**The same motif model in all promoters**

# Probability of a Sequence

**Given a sequence(s), motif *model* and motif *location***



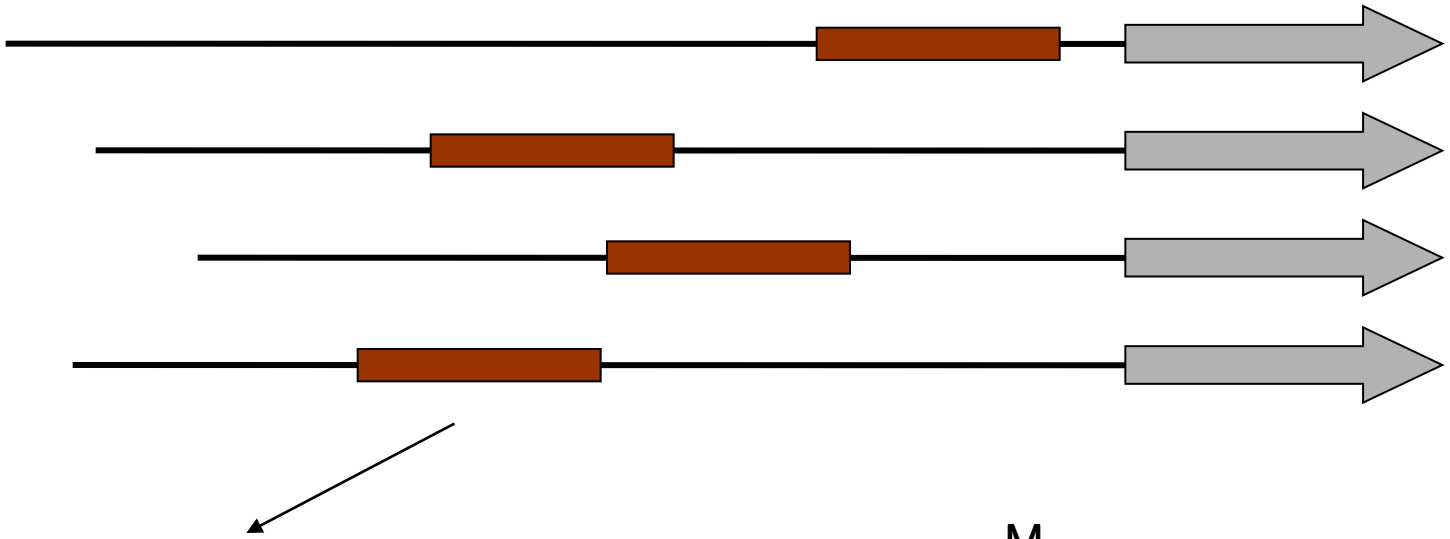

$$P(Seq \mid Mstart = 10, Model) = \prod_{i=1}^{59} P(S_i \mid B) \prod_{k=1}^{6} P_k\left(S_{k+63} \mid M\right) \prod_{i=66}^{100} P(S_i \mid B)$$

$S_i$ = *nucleotide at position i in the sequence*

| | $M_1$ | | | | | $M_K$ |
|---|---|---|---|---|---|---|
| A | .1 | .2 | .1 | .4 | .1 | .3 |
| C | .2 | .2 | .2 | .2 | .5 | .4 |
| G | .4 | .5 | .4 | .2 | .2 | .2 |
| T | .3 | .1 | .2 | .2 | .2 | .1 |

# Parameterizing the Motif Model

**Given multiple sequences and motif locations but no motif *model***



**AATGCG**
**ATATGG**
**ATATCG**
**GATGCA**

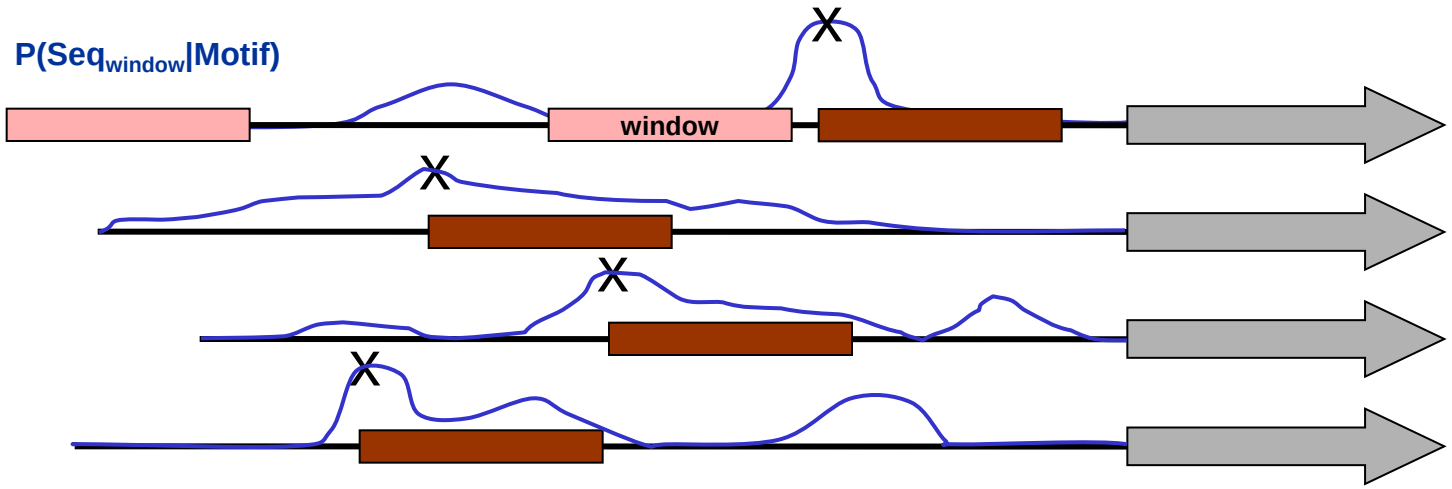
**Count Frequencies**

**Add pseudocounts**

| | $M_1$ | | | | | $M_6$ |
|---|---|---|---|---|---|---|
| A | 3/4 | | | | | |
| C | | ETC… | | | | |
| G | | | | | | 3/4 |
| T | | | | | | |

# Finding Known Motifs

**Given multiple sequences and motif model but no motif *locations***




Calculate P(Seq$_{window}$|Motif) for every starting location

Choose best starting location in each sequence

# Discovering Motifs

Given a set of co-regulated genes, we need to discover with only sequences

*We have neither a motif model nor motif locations*
*Need to discover both*

How can we approach this problem?

# Expectation Maximization (EM)

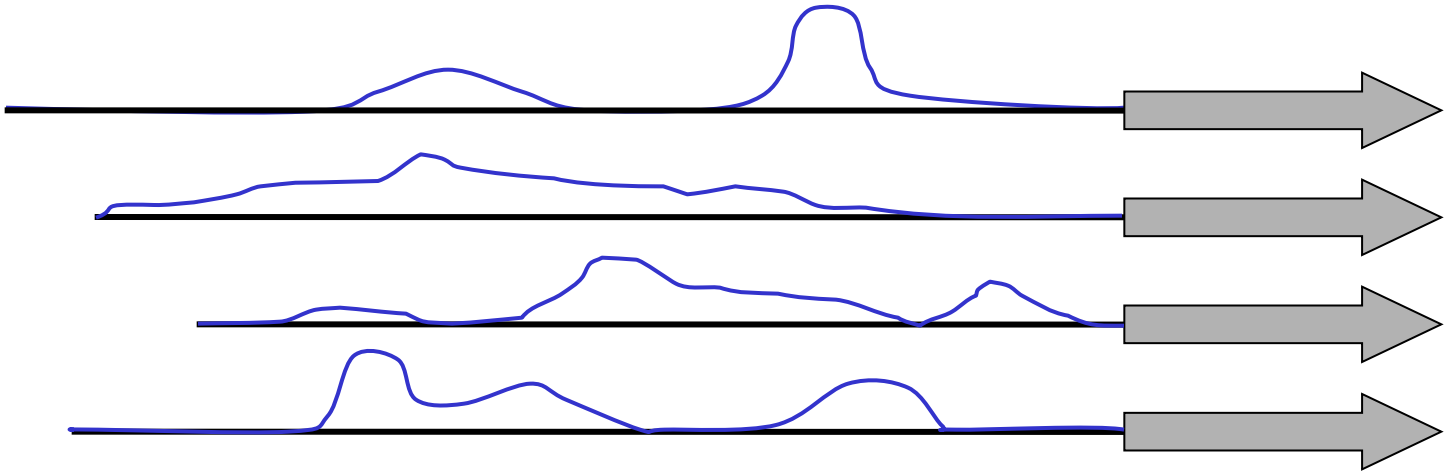
***Remember the basic idea!***

**1.Use model to estimate (distribution of) missing data**
**2.Use estimate to update model**
**3.Repeat until convergence**

Model is the motif model

Missing data are the motif locations

# EM for Motif Discovery



| | | | | | | |
|---|---|---|---|---|---|---|
| A | .1 | .2 | .1 | .4 | .1 | .3 |
| C | .2 | .2 | .2 | .2 | .5 | .4 |
| G | .4 | .5 | .4 | .2 | .2 | .2 |
| T | .3 | .1 | .2 | .2 | .2 | .1 |

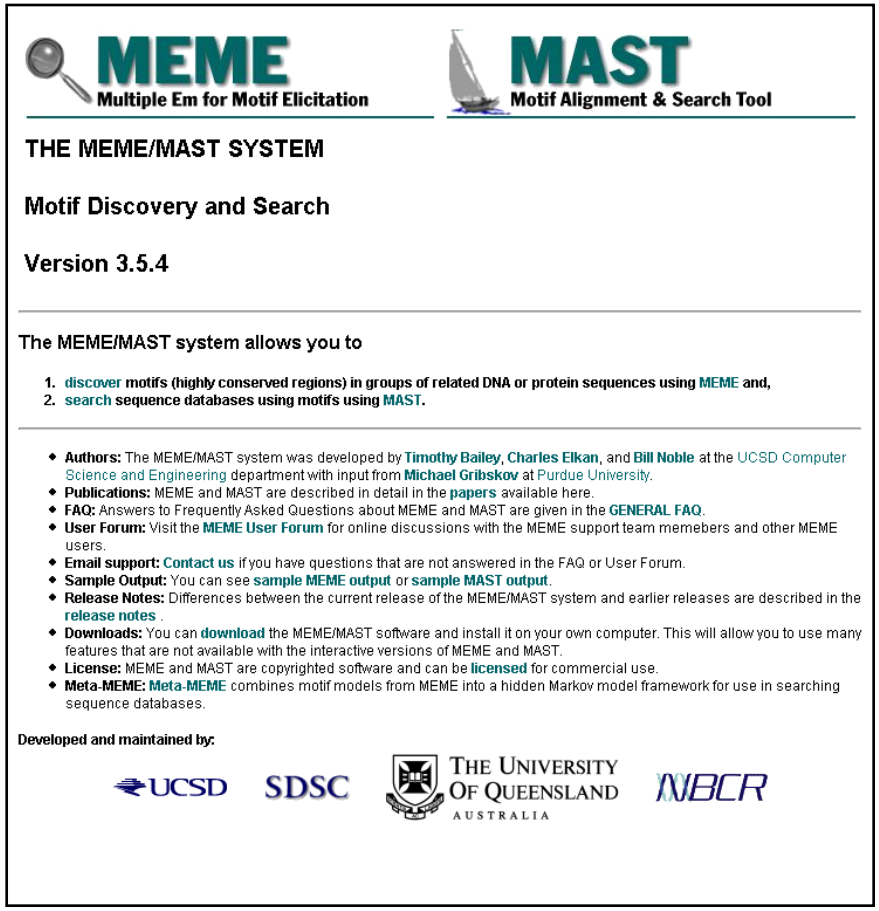
| | | | | | | |
|---|---|---|---|---|---|---|
| A | .1 | .1 | .1 | .1 | .1 | .3 |
| C | .2 | .3 | .2 | .2 | .5 | .1 |

At each iteration, P(Sequences|Model) guaranteed to increase

**ETC…**

# MEME

- MEME - implements EM for motif discovery in DNA and proteins
- MAST – search sequences for motifs given a model




**http://meme.sdsc.edu/meme/**

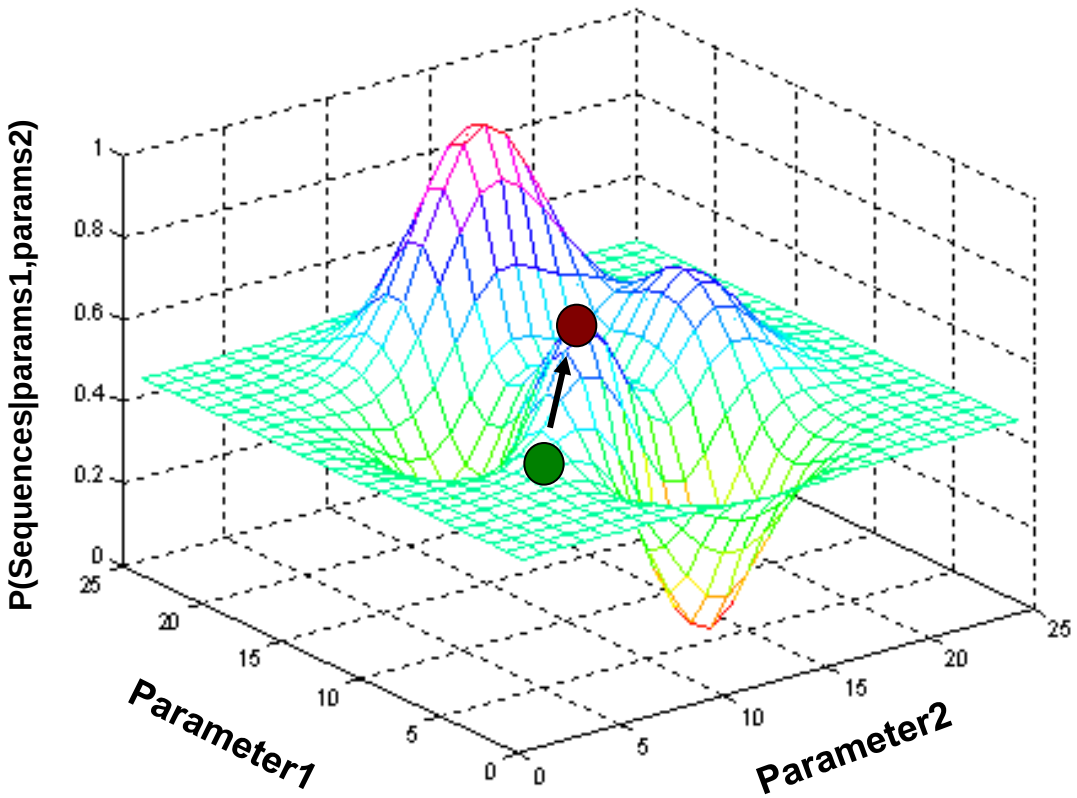
# P(Seq|Model) Landscape

**EM searches for parameters to increase P(seqs|parameters)**

Useful to think of P(seqs|parameters) as a function of parameters

EM starts at an initial set of parameters

And then "climbs uphill" until it reaches a local maximum




***Where EM starts can make a big difference***

# Search from Many Different Starts

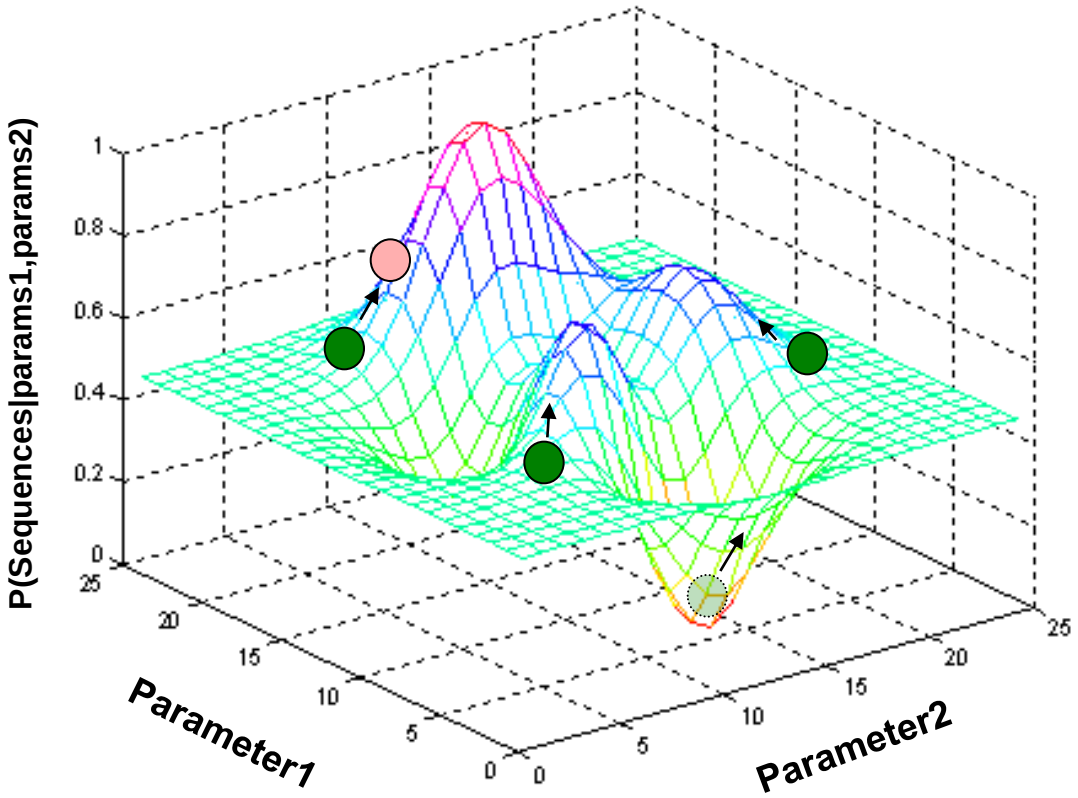
**To minimize the effects of local maxima, you should search multiple times from different starting points**

MEME uses this idea

Start at many points

Run for one iteration

Choose starting point that got the “highest” and continue



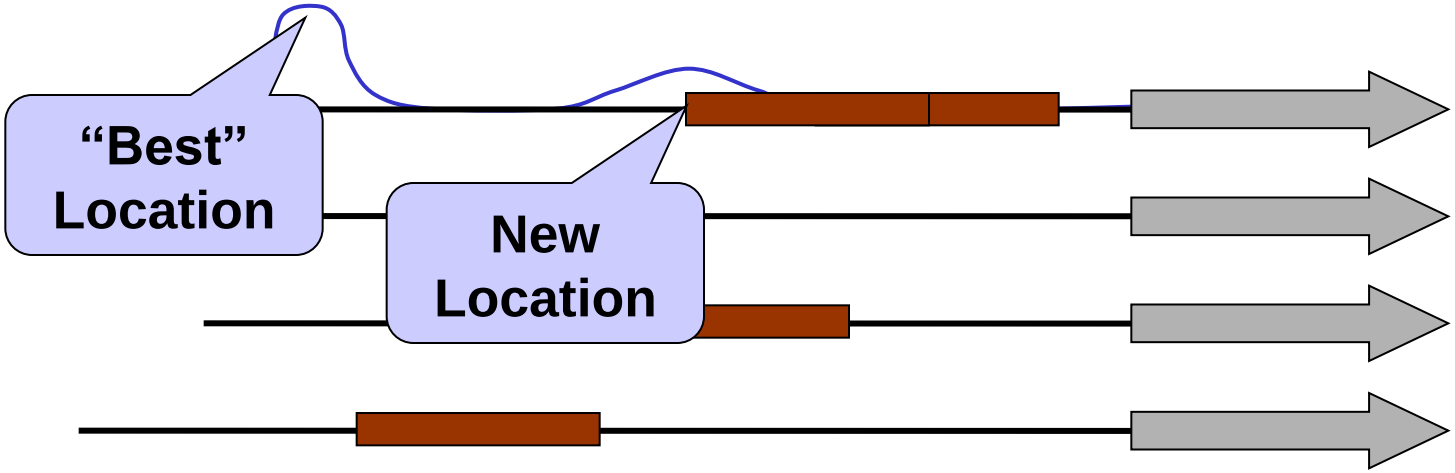
# Gibbs Sampling

**A stochastic version of EM that differs from deterministic EM in two key ways**

**1. At each iteration, we only update the motif position of a single sequence**

**2. We may update a motif position to a "suboptimal" new position**

# Gibbs Sampling



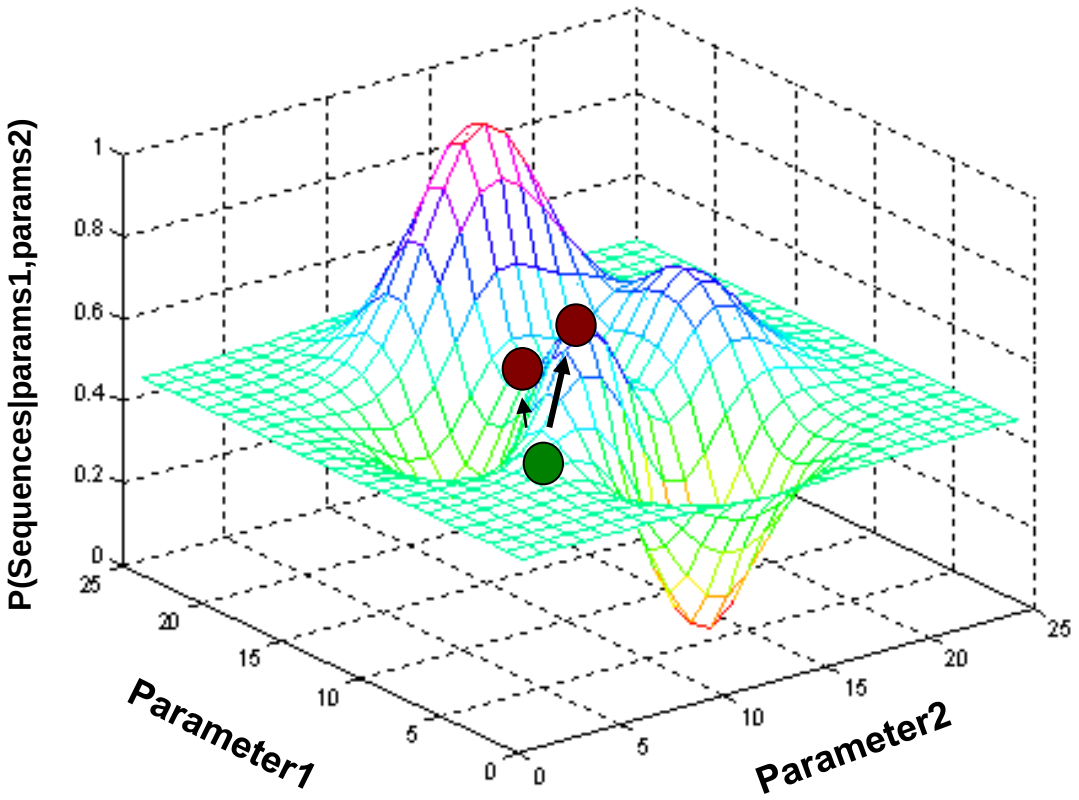

| | | | | | | |
|---|---|---|---|---|---|---|
| A | .1 | .2 | .1 | .4 | .1 | .3 |
| C | .2 | .2 | .2 | .2 | .5 | .4 |
| G | .4 | .5 | .4 | .2 | .2 | .2 |
| T | .3 | .1 | .2 | .2 | .2 | .1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| A | .1 | .1 | .1 | .1 | .1 | .3 |
| C | .2 | .3 | .2 | .2 | .5 | .1 |
| G | .4 | .5 | .4 | .5 | .2 | .1 |
| T | .3 | .1 | .2 | .2 | .2 | .1 |

**ETC…**

# Gibbs Sampling and Climbing

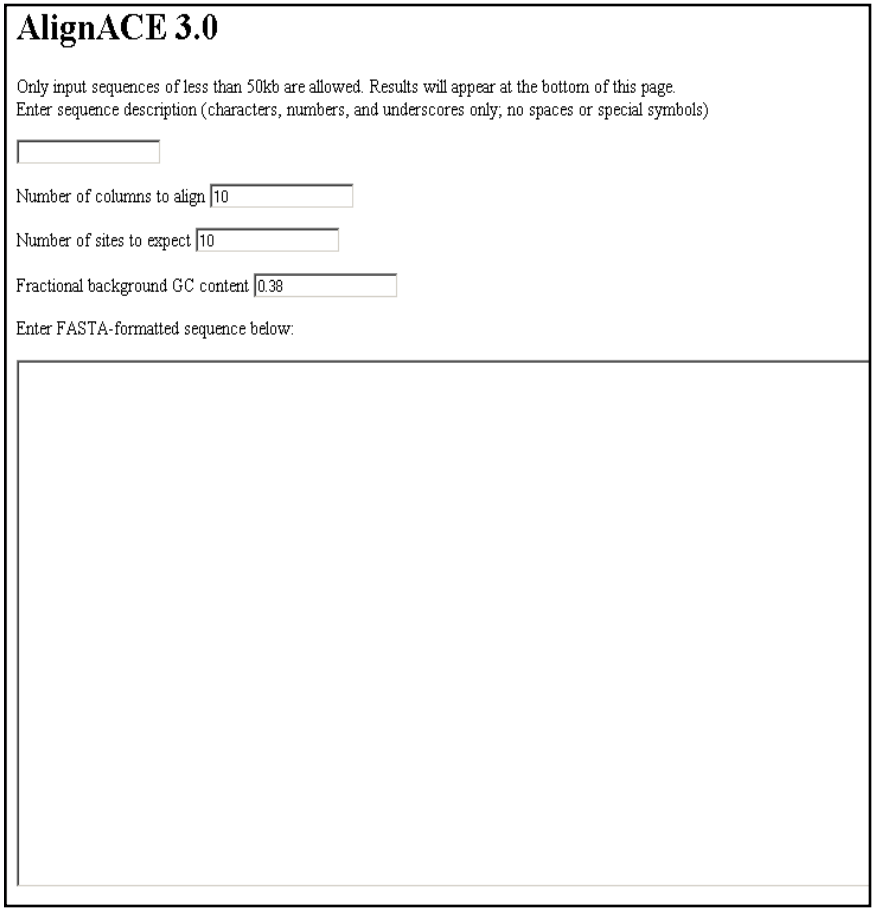
**Because gibbs sampling does not always choose the best new location it can move to another place not directly uphill**




***In theory,* Gibbs Sampling less likely to get stuck a local maxima**

# AlignACE

- **Implements Gibbs sampling for motif discovery**
  - **Several enhancements**
- **ScanAce – look for motifs in a sequence given a model**
- **CompareAce – calculate "similarity" between two motifs (i.e. for clustering motifs)**




http://atlas.med.harvard.edu/cgi-bin/alignace.pl

# Assessing Motif Quality

**Scan the genome for all instances and associate with nearest genes**

- **Category Enrichment** – look for association between motif and sets of genes. Score using Hypergeometric distribution
  - Functional Category
  - Gene Families
  - Protein Complexes
- **Group Specificity** – how restricted are motif instances to the promoter sequences used to find the motif?
- **Positional Bias** – do motif instances cluster at a certain distance from the transcription start site (TSS)?
- **Orientation Bias** – do motifs appear preferentially upstream or downstream of genes?

# Comparative Motif Prediction

# Sequencing and comparison of yeast species to identify genes and regulatory elements

**Manolis Kellis*†, Nick Patterson*, Matthew Endrizzi*, Bruce Birren* & Eric S. Lander*‡**

** Whitehead/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
*† Department of Computer Science and ‡ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

**Identifying the functional elements encoded in a genome is one of the principal challenges in modern biology. Comparative genomics should offer a powerful, general approach. Here, we present a comparative analysis of the yeast *Saccharomyces cerevisiae* based on high-quality draft sequences of three related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*). We first aligned the genomes and characterized their evolution, defining the regions and mechanisms of change. We then developed methods for direct identification of genes and regulatory motifs. The gene analysis yielded a major revision to the yeast gene catalogue, affecting approximately 15% of all genes and reducing the total count by about 500 genes. The motif analysis automatically identified 72 genome-wide elements, including most known regulatory motifs and numerous new motifs. We inferred a putative function for most of these motifs, and provided insights into their combinatorial interactions. The results have implications for genome analysis of diverse organisms, including the human.**

Extracting the complete functional information encoded in a genome—including the genic, regulatory and structural elements—is a central challenge in biological research. Ideally, one would be able to extract this information directly from the DNA sequence itself without recourse to extensive experimentation. At present, however, our ability to directly interpret genomes is rudimentary.

*De novo* identification of the complete set of protein-coding sequences remains imperfect, even in well-studied organisms with compact genomes. The yeast *Saccharomyces cerevisiae*, for example, has enjoyed a complete genome sequence since 1996 (ref. 1), previously been used to identify putative genes or regulatory elements in small genomic regions[11–14]. Light sampling of whole-genome sequence has been studied as a way to improve genome annotation[5,15]. Complete microbial genomes have been compared to identify pathogenic and other genes[16–19]. Genome-wide comparison has been used to estimate the proportion of the mammalian genome under selection[7].

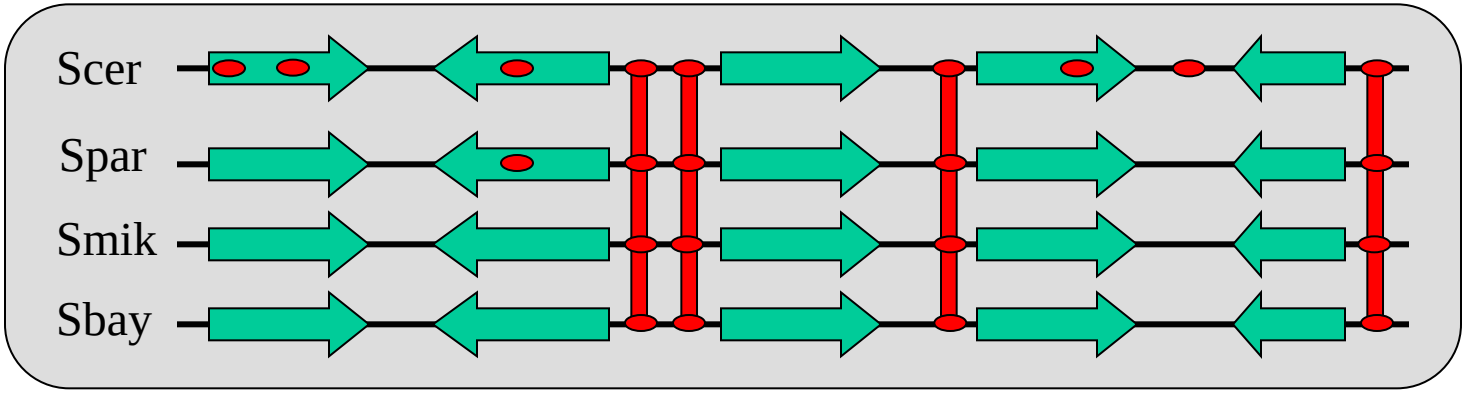
The goal of this paper is to develop and apply general approaches for systematic analysis of protein-coding and regulatory elements within any genome by means of whole-genome comparisons with

**Kellis et al. (2003) Nature**

# Conservation of Motifs


slide credits: M. Kellis

# Genome-wide conservation



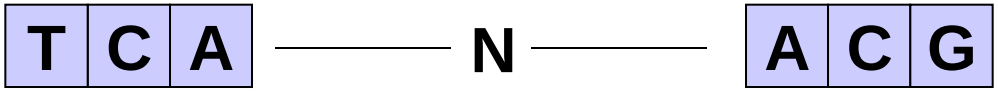

| Evaluate conservation within: | Gal4 | Controls |
|---|---|---|
| (1) All intergenic regions | 22% | 5% |
| (2) Intergenic : coding | 4:1 | 1:3 |
| (3) Upstream : downstream | 12:0 | 1:1 |

A signature for regulatory motifs

# Finding Motifs in Yeast Genomes

M. Kellis PhD Thesis

1. Enumerate all “mini-motifs”

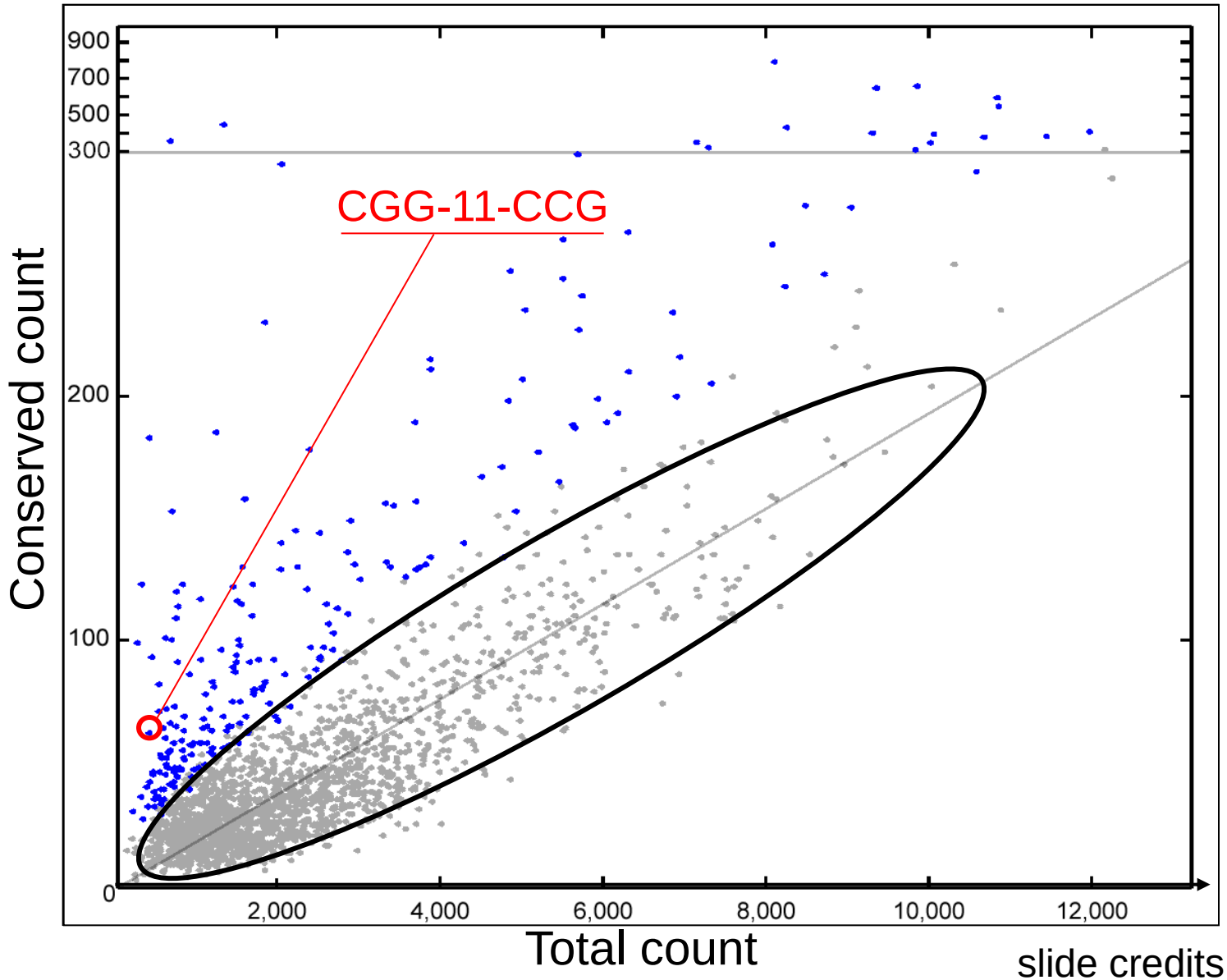



2. Apply three tests

    1. Look for motifs conserved in intergenic regions
    2. Look for motifs more conserved intergenically than in genes
    3. Look for motifs preferentially conserved upstream or downstream of genes

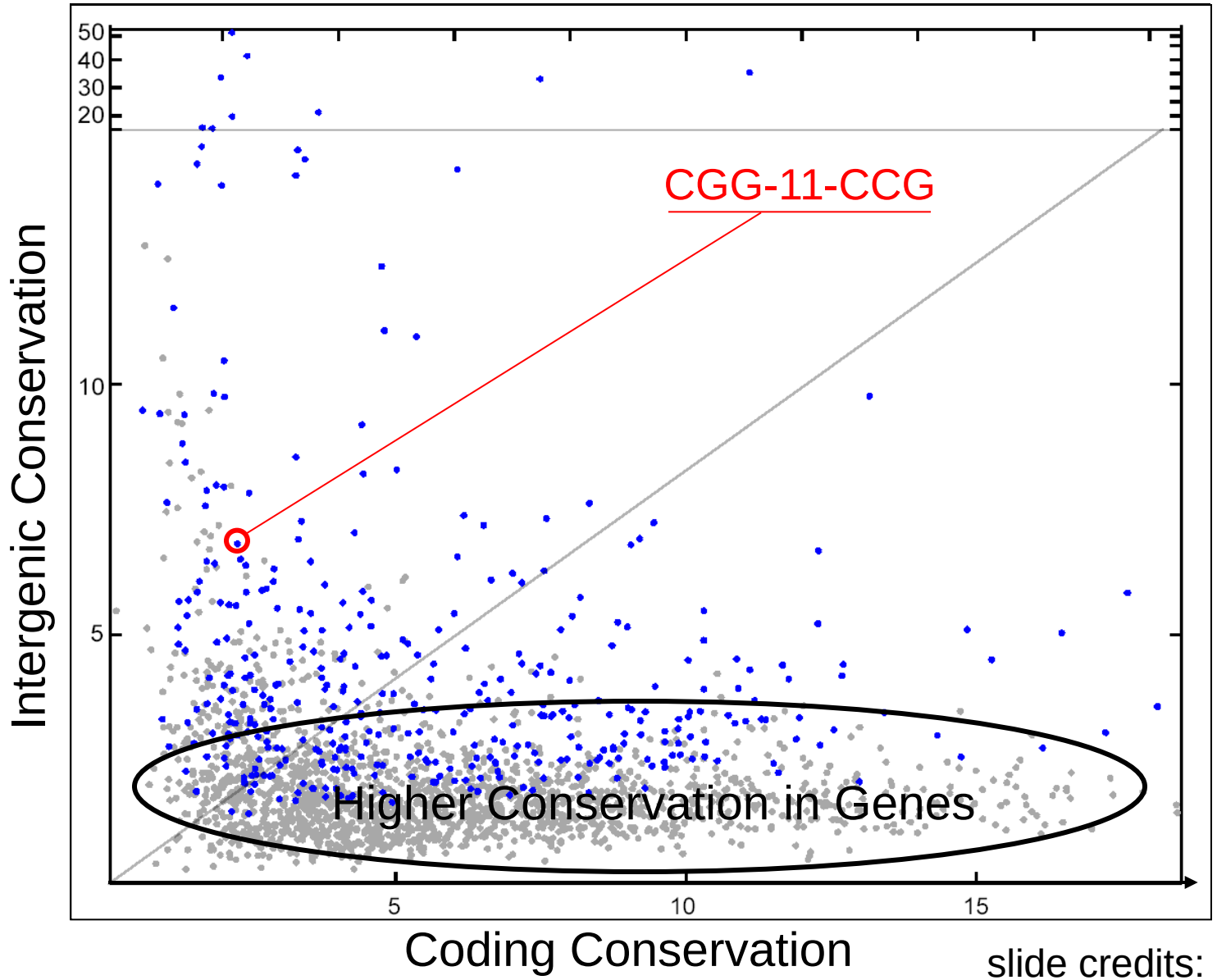
slide credits: M. Kellis

# Test 1: Intergenic conservation



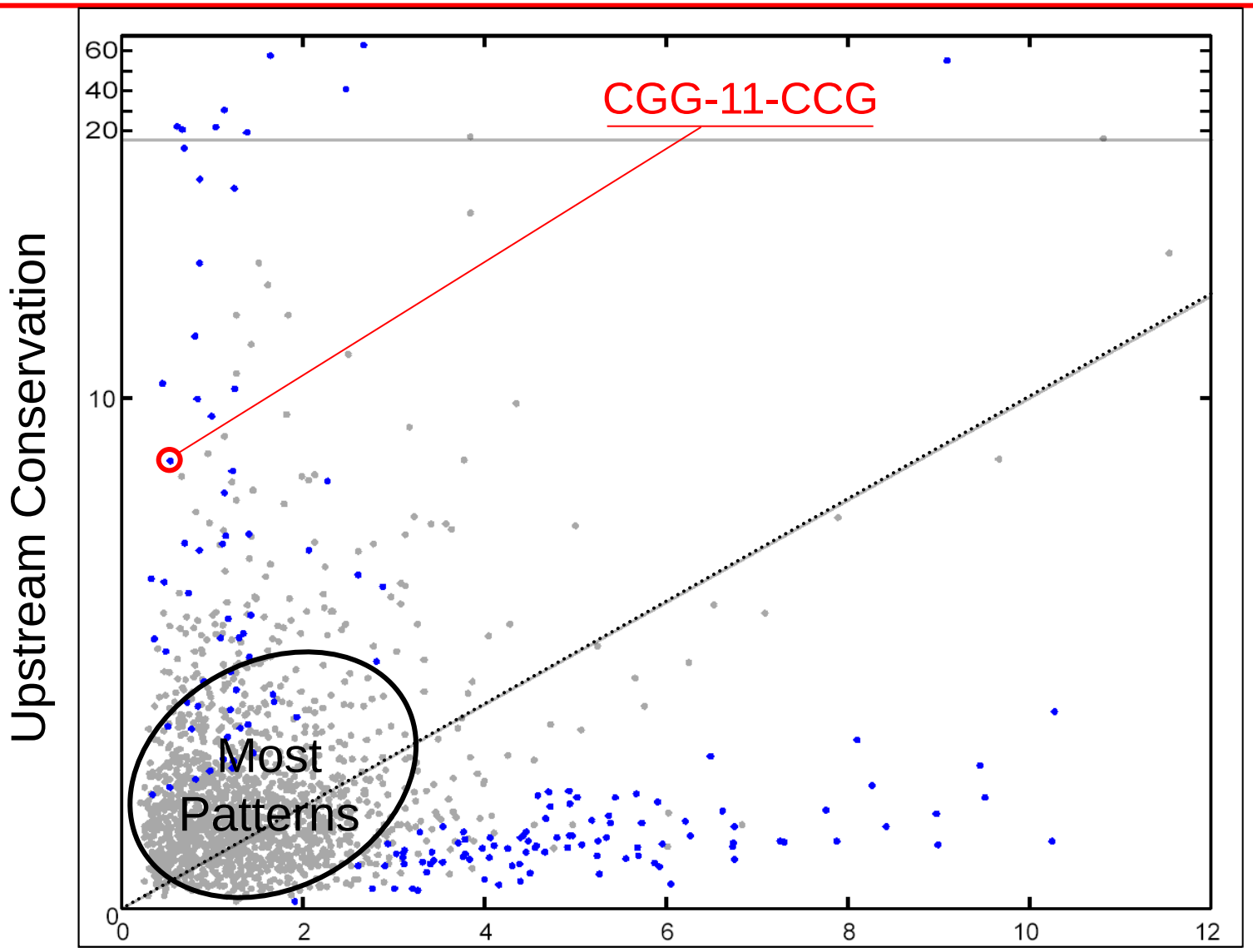

slide credits: M. Kellis

# Test 2: Intergenic vs. Coding



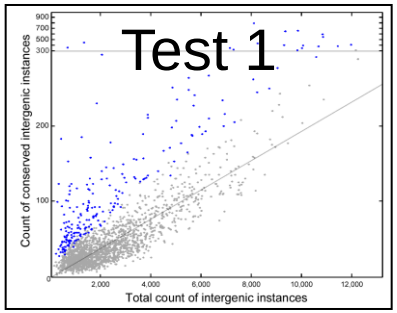

slide credits: M. Kellis

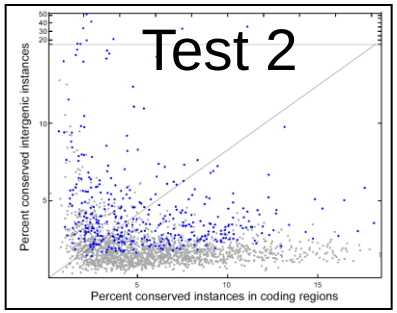
# Test 3: Upstream vs. Downstream

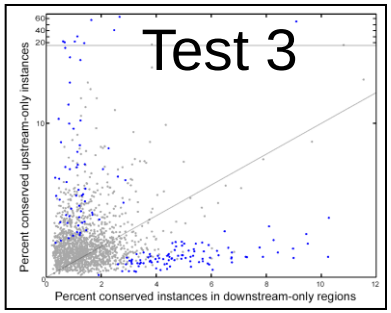



slide credits: M. Kellis

# Constructing full motifs




Extend Extend Extend

Collapse Collapse Collapse

Merge

Full Motifs

2,000
Mini-motifs

72
Full motifs

slide credits: M. Kellis

# Results

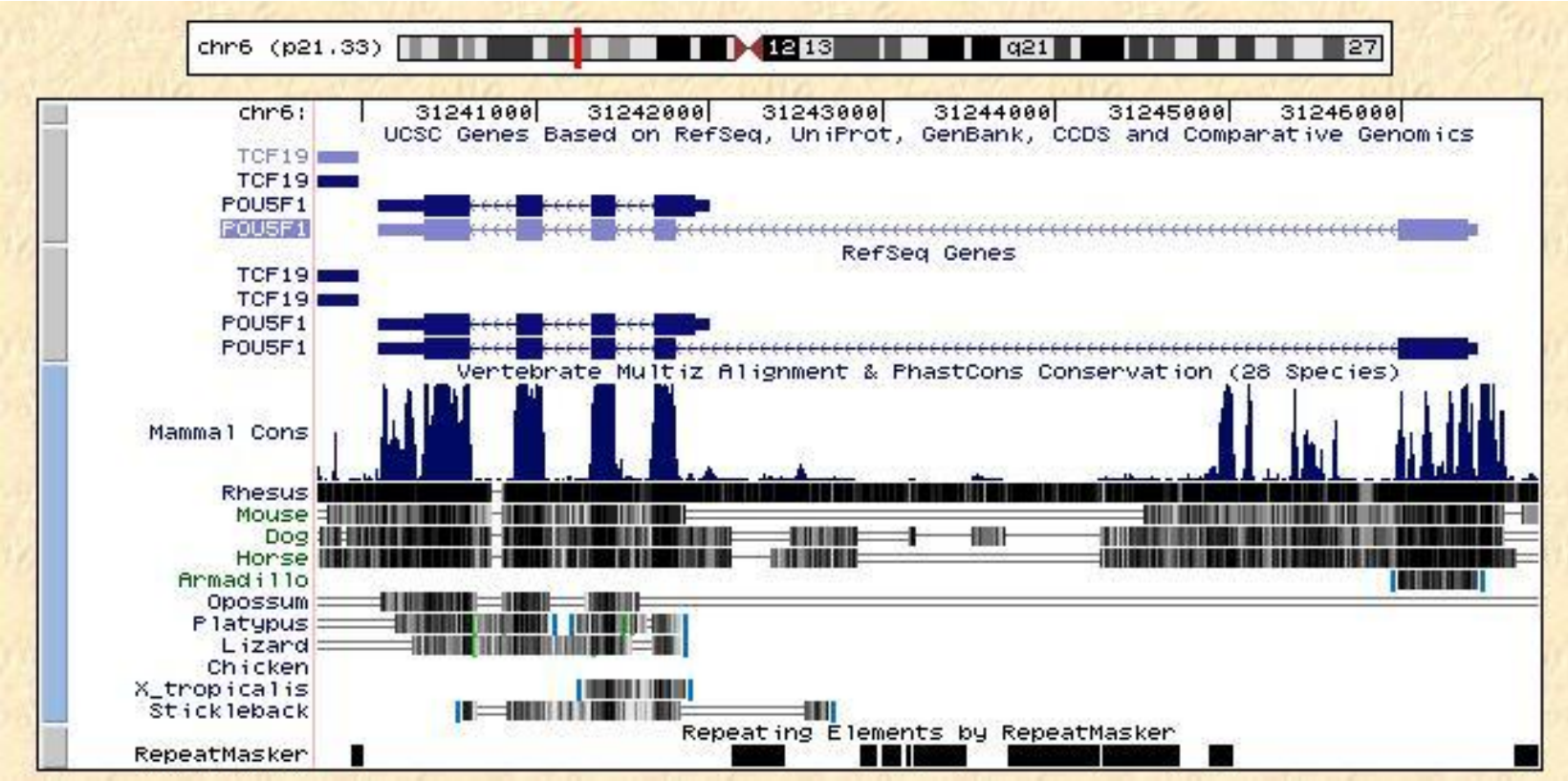
| Rank | Discovered Motif | Known TF motif | Tissue Enrichment | Distance bias |
|---|---|---|---|---|
| 1 | RCGCAnGCGY | NRF-1 | Yes | Yes |
| 2 | CACGTG | MYC | Yes | Yes |
| 3 | SCGGAAGY | ELK-1 | Yes | Yes |
| 4 | ACTAYRnnnCCCR | | Yes | Yes |
| 5 | GATTGGY | NF-Y | Yes | Yes |
| 6 | GGGCGGR | SP1 | Yes | Yes |
| 7 | TGAnTCA | AP-1 | Yes | |
| 8 | TMTCGCGAnR | | Yes | Yes |
| 9 | TGAYRTCA | ATF3 | Yes | Yes |
| 10 | GCCATnTTG | YY1 | Yes | |
| 11 | MGGAAGTG | GABP | Yes | Yes |
| 12 | CAGGTG | E12 | Yes | |
| 13 | CTTTGT | LEF1 | Yes | |
| 14 | TGACGTCA | ATF3 | Yes | Yes |
| 15 | CAGCTG | AP-4 | Yes | |
| 16 | RYTTCCTG | C-ETS-2 | Yes | Yes |
| 17 | AACTTT | IRF1(*) | Yes | |
| 18 | TCAnnTGAY | SREBP-1 | Yes | Yes |
| 19 | GKCGCn(7)TGAYG | | Yes | Yes |

- 174 promoter motifs
  - ✓ 70 match known TF motifs
  - ✓ 60 show positional bias
  - ➔ 75% have evidence
- Control sequences
  - < 2% match known TF motifs
  - < 3% show positional bias
  - ➔ < 7% false positives

slide credits: M. Kellis

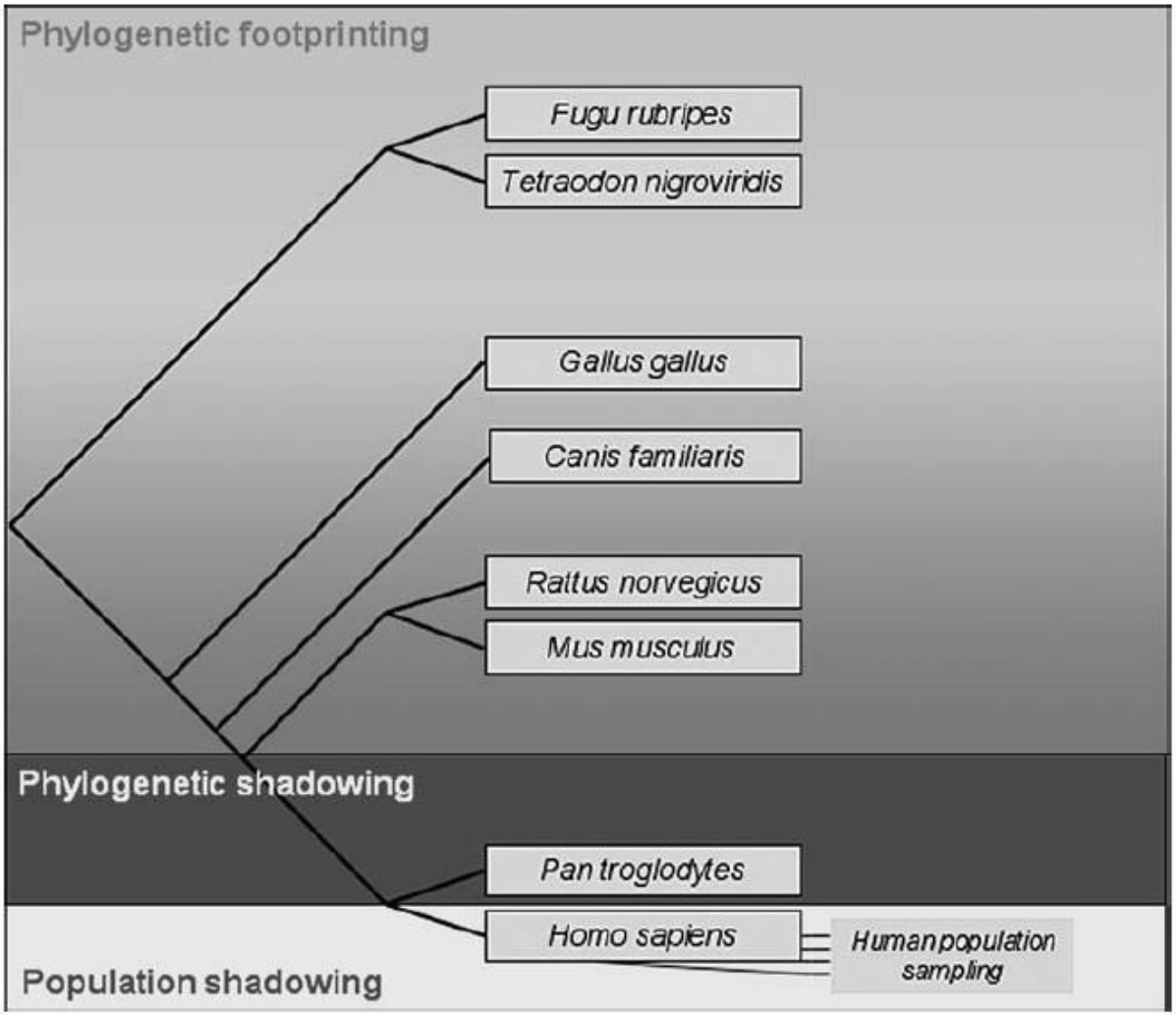
# Phylogenetic Footprinting

# Functional elements tend to be conserved across species




**For example, exons are conserved due to the selection pressure. Introns and intergenic regions are less likely to be conserved.**

# Phylogenetic footprinting




Miller et al. Annu. Rev. Genomics Hum. Genet. 2004

# Incorporating cross-species conservation into motif discovery

- A threshold method (Wasserman et al. Nature Genetics, 2000)

STEP1: construct cross-species alignment

STEP2: compute conservation measure from the alignment

STEP3: Non-conserved regions are filtered out

STEP4: Gibbs motif sampler is applied to conserved regions of the target genome

# Phylogenetic footprinting & motif discovery

- CompareProspector (Liu Y. et al. Genome Res. 2004)

STEP1: construct cross-species alignment

STEP2: compute conservation measure (window percent identity, WPID) from the alignment

STEP3: multiply the likelihood ratio at a position by the corresponding WPID, thus likelihood landscape is changed to favor conserved sites

STEP4: apply a Gibbs motif sampler based algorithm

# Phylogenetic footprinting & motif discovery

- Evolutionary model based approach

  EMnEM (Moses et al. 2004)

  PhyME (Sinha et al. 2004)

  PhyloGibbs (Siddharthan et al. 2005)

  Tree Sampler (Li and Wong, 2005)

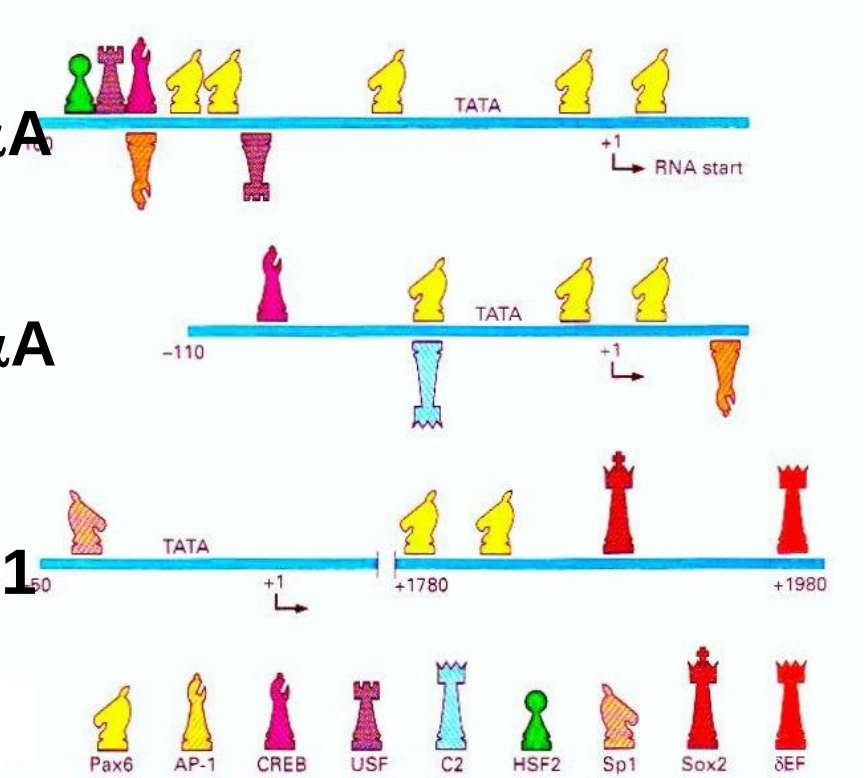
# Evolution of Regulation

# Outline of Questions

- Conserved and evolved properties
- Mechanisms of conservation and evolution
- Bridging Genotype and Phenotype

# *Cis*-Regulatory Factors

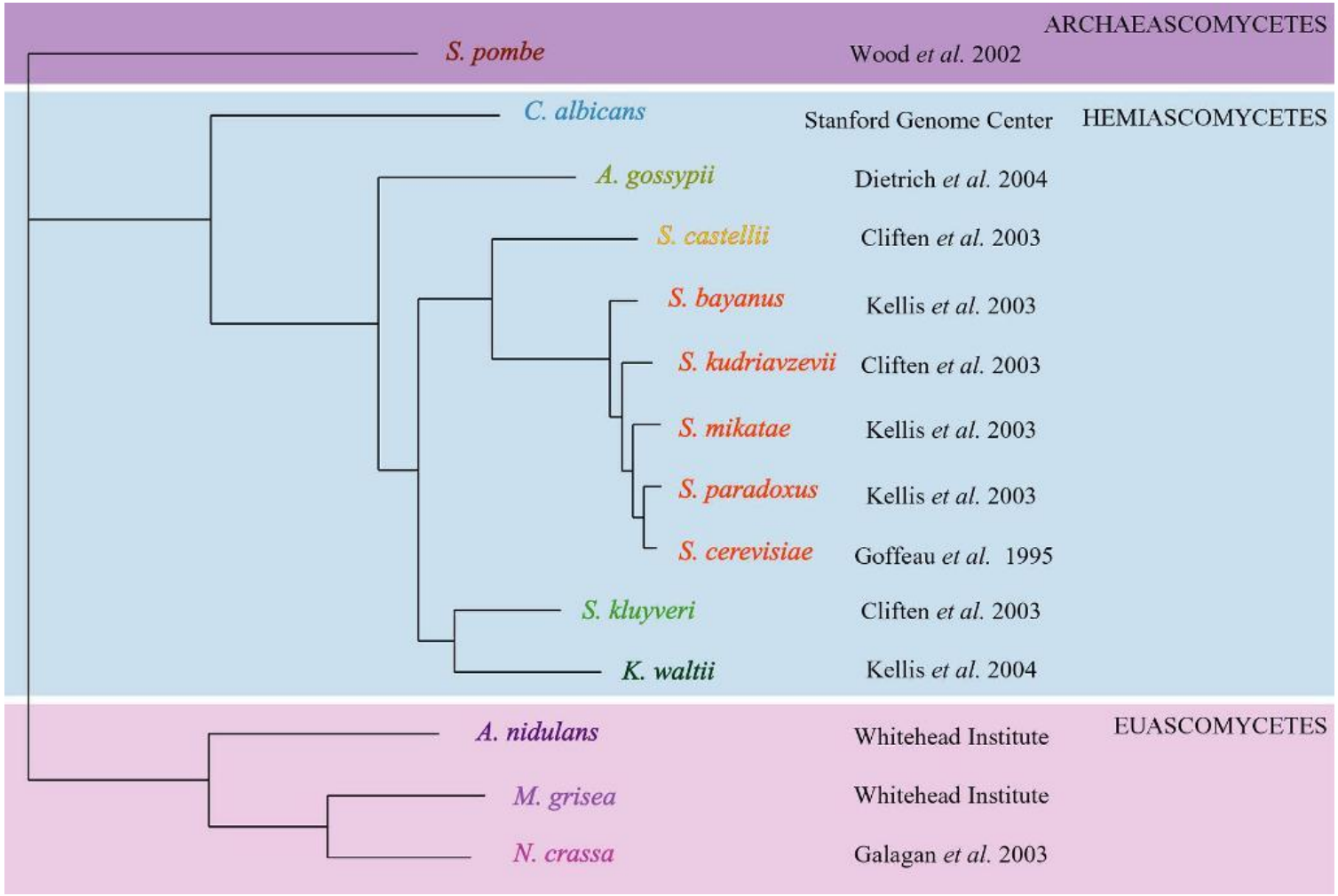
- Compositior
- Location
- Modules…




**Gene control regions for eye lens chrystallins**

***Molecular Biology of the Cell*, Alberts et al., 4th ed.**

# The Fungul Family




**A. P. Gasch *et al.*, *PLoS Biol*., 2004**

*Open access, freely available online* PLOS BIOLOGY

# Conservation and Evolution of *Cis*-Regulatory Systems in Ascomycete Fungi

**Audrey P. Gasch[1¤*], Alan M. Moses[2], Derek Y. Chiang[3], Hunter B. Fraser[3], Mark Berardini[4], Michael B. Eisen[1,3*]**

**1** Genome Sciences Department, Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **2** Graduate Group in Biophysics, University of California, Berkeley, California, United States of America, **3** Department of Molecular and Cell Biology, University of California, Berkeley, California, United States of America, **4** Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

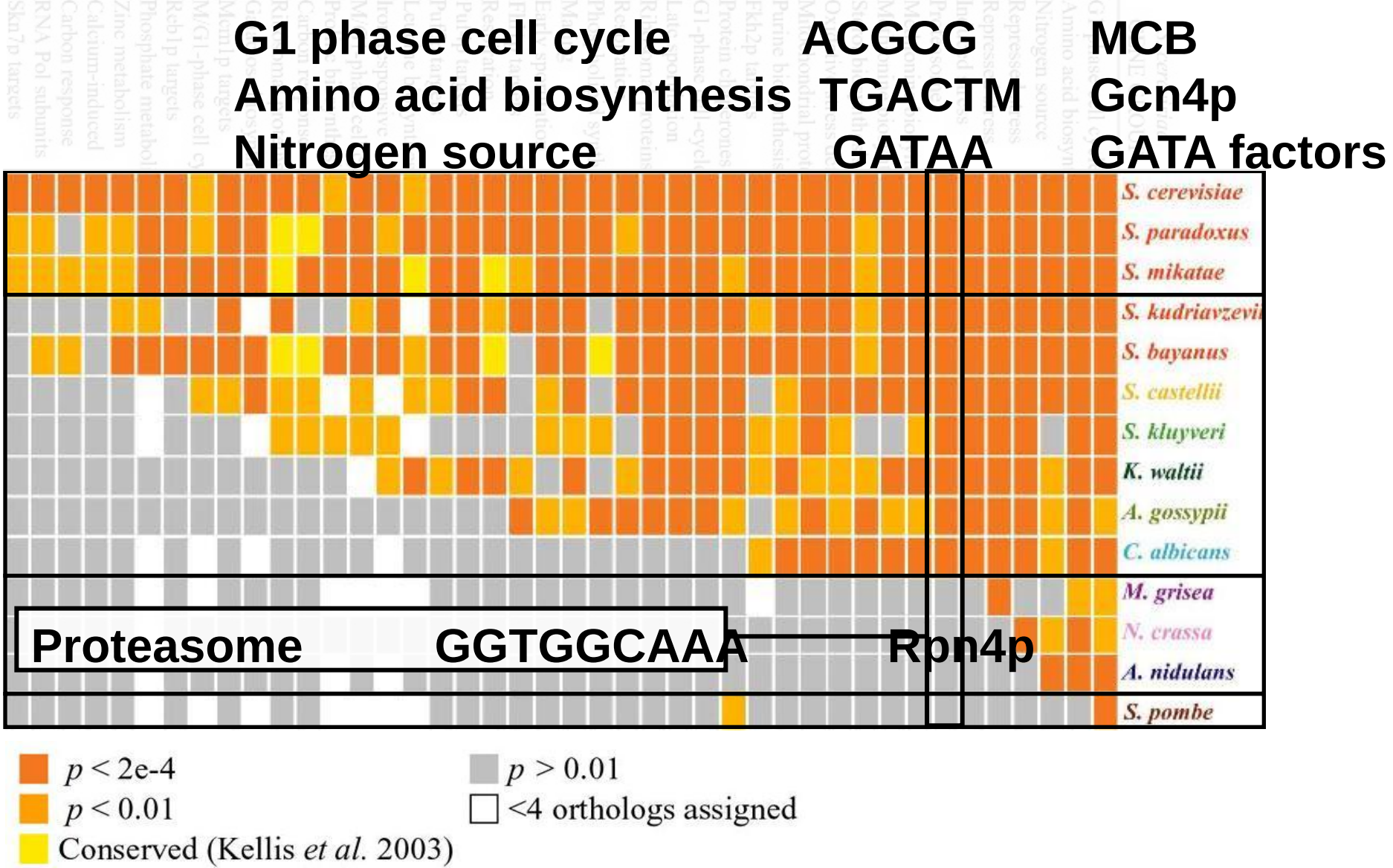
## Goal

- Explore evolution of gene expression regulation
  - Mainly through examination of *cis-elements*

# Large Scale Analysis

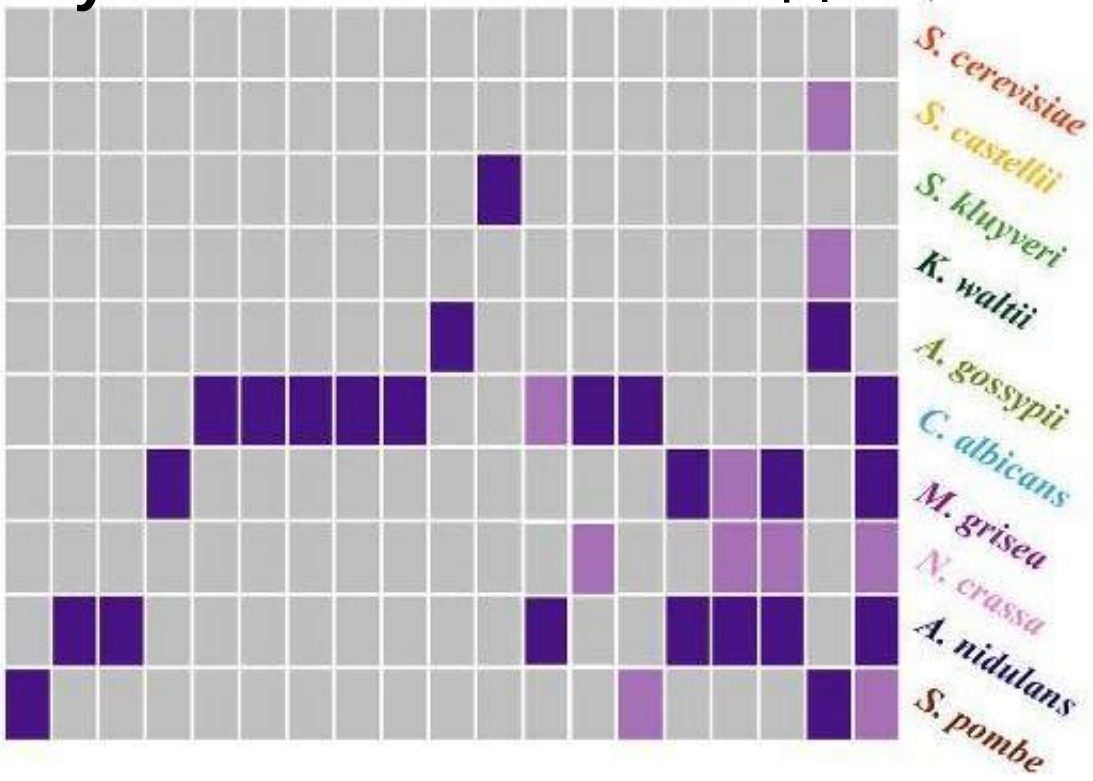
- Identify 264 co-regulated gene groups in *S. serevisiae*
- Putative *cis*-regulatory elements
  - 80 known consensus binding sites
  - 597 elements by motif search with MEME
- Score enrichment of genes containing each putative element
  - 42 cis-elements in 35 unique groups
- Orthologous modules in other species
- Enrichment of orthologous modules

# Conservation of *S. cerevisiae* motifs



# Novel Sequences

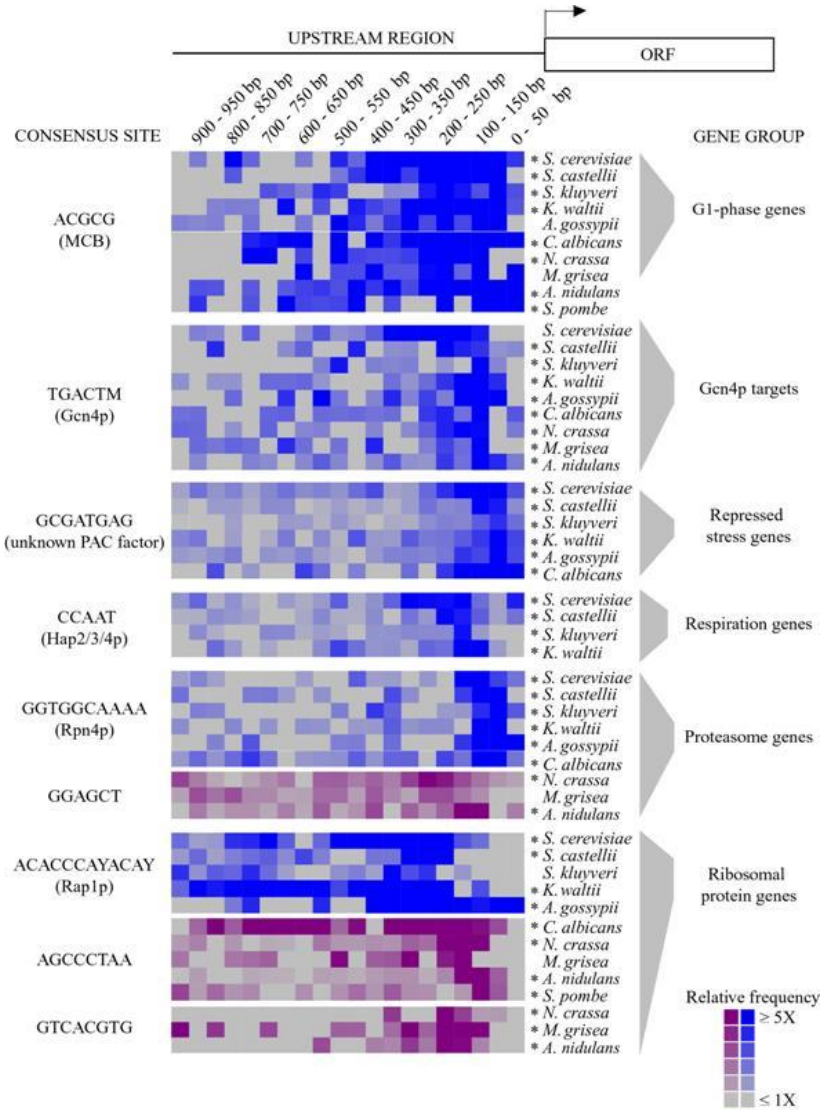
**Ribosomal proteins** AGCCCTAA
**Ribosomal proteins** GTGACTGT
**tRNA synthetases** TGACTCAN



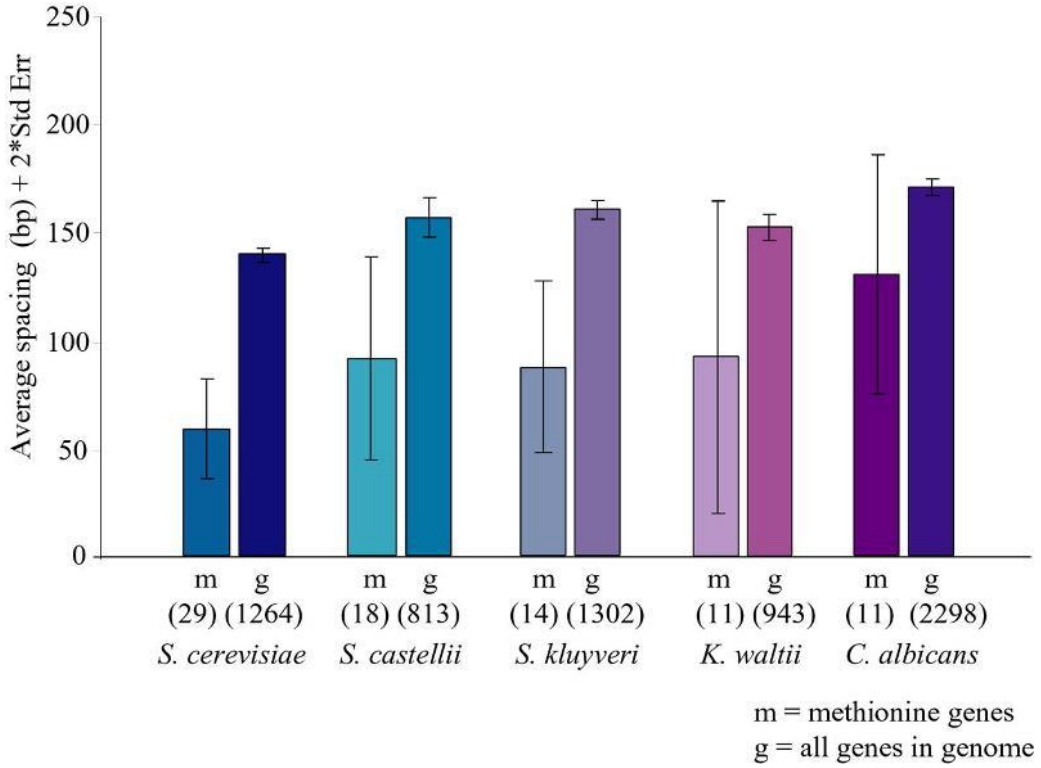

Enrichment $p$ < 1e-5
Enrichment $p$ < 0.01
Enrichment $p$ > 0.01

# Positions of binding sites

- Non random distribution
- Similar across species



# Spacing between binding sites in Methionine Biosynthesis genes




- Small distance between Cbf1p and Met31/32p
- Conserved across species
- Independent of exact positions

# Zooming In – The Proteasome

- The proteasome is regulated by Rpn4p
- Consensus sequence enriched across all hemiascomycete
- Slight differences between *C. Albicans* and *S. Serevisiae*

➔Exploring evolution of the sequence

PNAS

# Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast

Amos Tanay*†, Aviv Regev†‡§¶, and Ron Shamir*¶

*School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel; ‡Bauer Center for Genomics Research, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; and §The Broad Institute of Harvard and Massachusetts Institute of Technology, 320 Charles Street, Cambridge, MA 02141

Communicated by Eric S. Lander, Whitehead Institute for Biomedical Research, Cambridge, MA, March 28, 2005 (received for review January 9, 2005)

**Transcriptional modules of coregulated genes play a key role in regulatory networks. Comparative studies show that modules of coexpressed genes are conserved across taxa. However, little is known about the mechanisms underlying the evolution of module**

boring both cis-elements in a tightly coupled configuration. The full spectrum of evolutionary events we discovered, encompassing both conservation and divergence, provides a general framework for the study of the evolution of transcription regulation and highlights the

## Goals

- Learn mechanisms of evolution of regulation systems
- Integrating comparative expression and sequence analysis

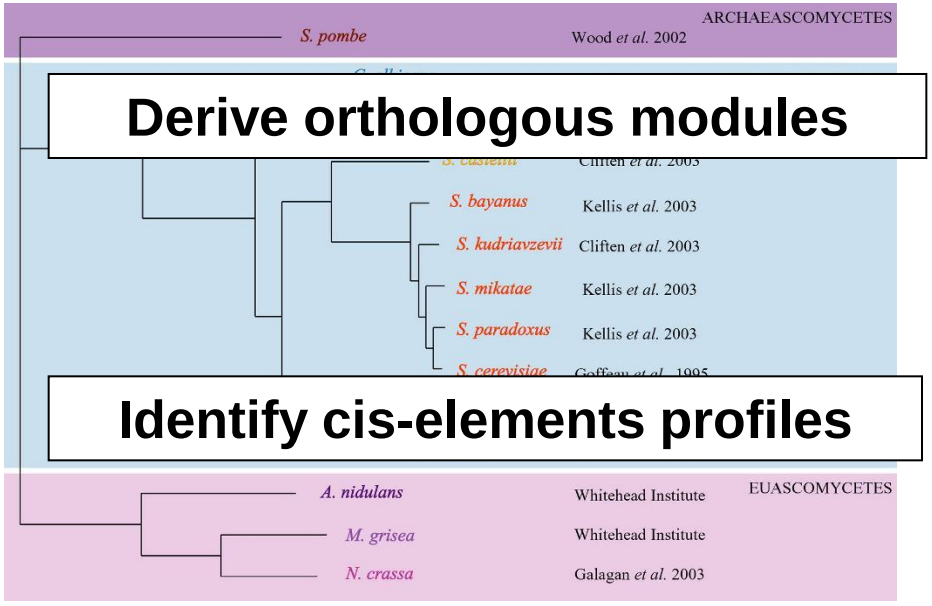
# Computational framework

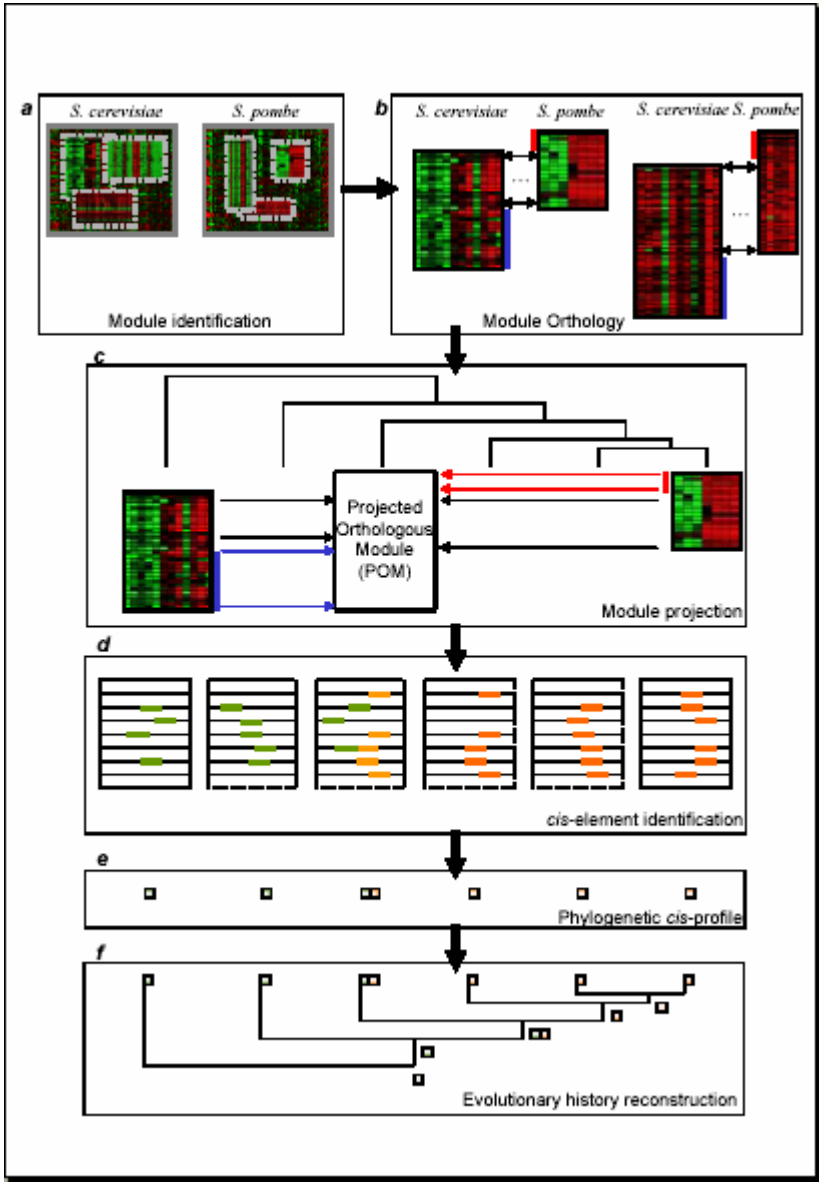
**Identify conserved modules**

**Derive orthologous modules**

**Identify cis-elements profiles**

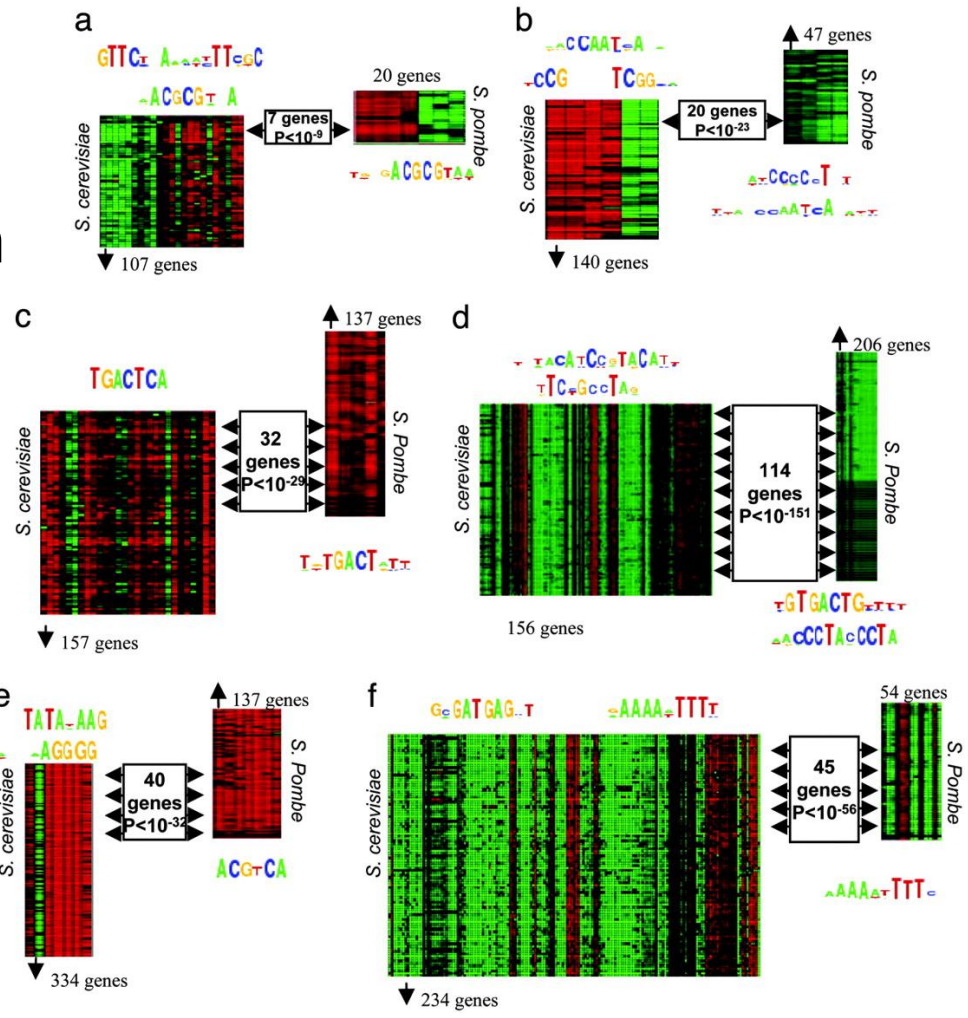
**Reconstruction of evolution**



# Orthologous Modules in Yeast

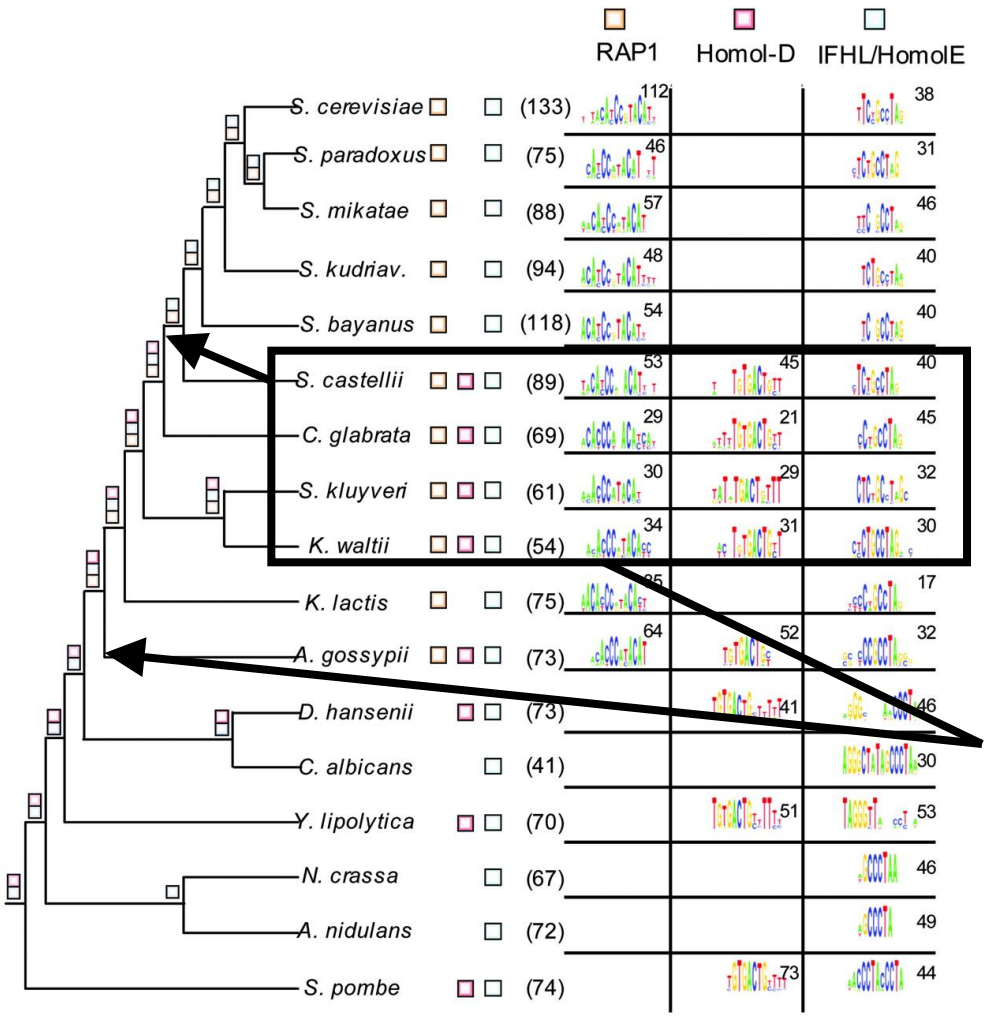
a. **S-phase module**
b. **Respiration**
c. **Amino acid metabolism**
d. **Ribosomal proteins synthesis**
e. **Stress**
f. **Ribosome biogenesis**



# Conserved and Diverged Regulatory Mechanisms

- Conserved cis-elements – Respiration module
  - Mlu cell cycle element (`ACGCGT`)
  - Bound by MBF complexes in both species
- Diverged - Ribosomal protein synthesis
  - RAP1 and IFHL in *S.* cerevisae (`TACATCCGTACAT & TCCGCCTAG` resp.)
  - Homol-D box and Homol-E site in *S. pombe* (`TGTGACTG & ACCCTACCCTA`)
- Conserved and diverged – Ribosome biogenesis module

# The evolution of the Ribosomal Regulatory Program



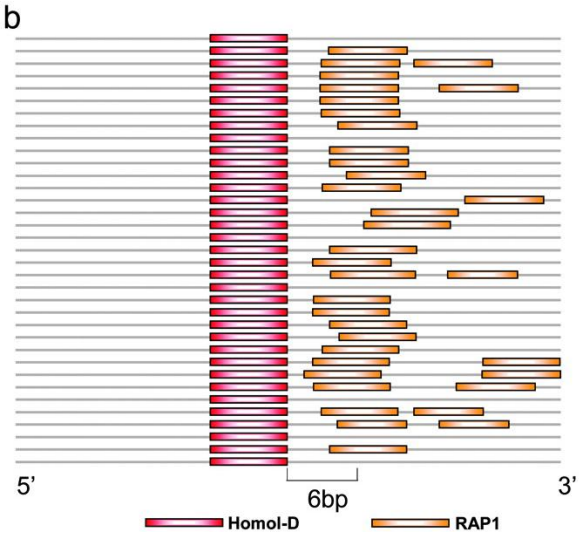

**Gradual Evolution in the IFHL Box**

**Apparent redundancy of binding sites**

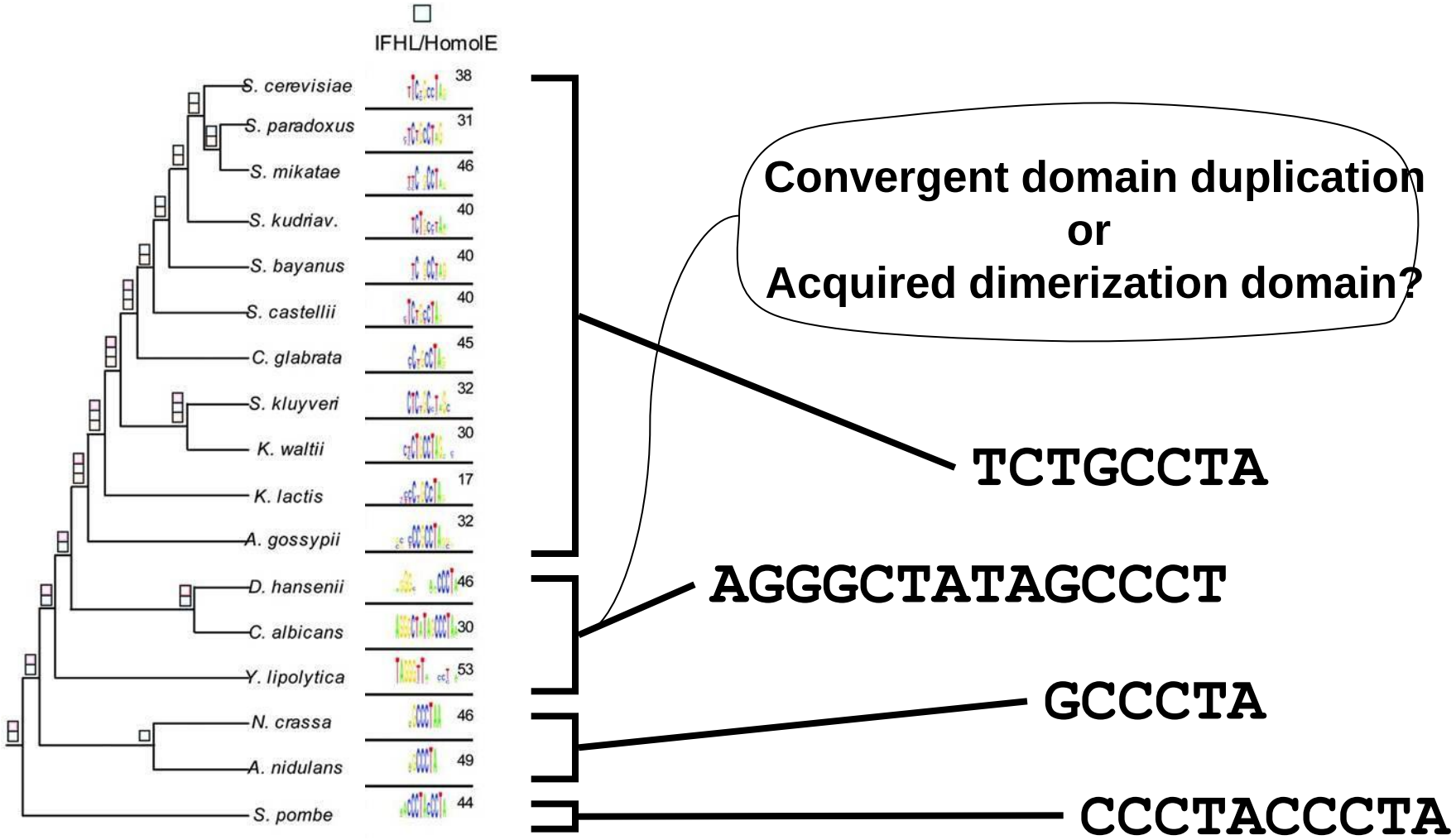
**Switching from Homol-D to RAP1**

# Evidence for Regulator Switching

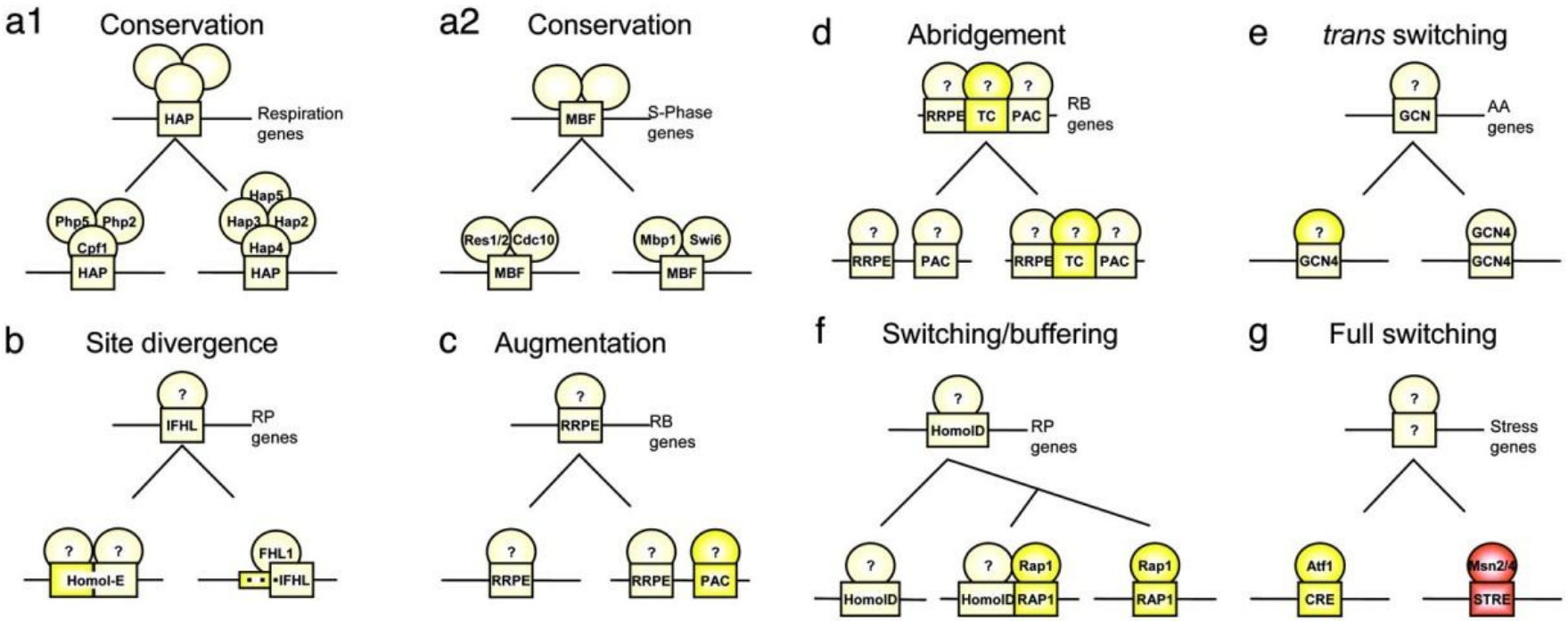
- Evolution of Transcription factor
  - RAP1 is a submotif of telomeric repeat
  - Rap1p regulates telomer length
  - Association of events:
    Rap1p Gained of Trans activation domain, RAP1 joined RP module
- Interacation of RAP1 & Homol-D?
  - usually 2-6 base pairs apart
  - Conserved order



# Gradual Evolution of IFHL Box



# Intermediate Summary - Modes of Evolution




- Conserved modules with diverged regulatory mechanisms
- Some changes via redundant intermediate programs

# Rewiring of the Yeast Transcriptional Network Through the Evolution of Motif Usage

Jan Ihmels,[1] Sven Bergmann,[1] Maryam Gerami-Nejad,[2] Itai Yanai,[1] Mark McClellan,[2] Judith Berman,[2] Naama Barkai[1]*

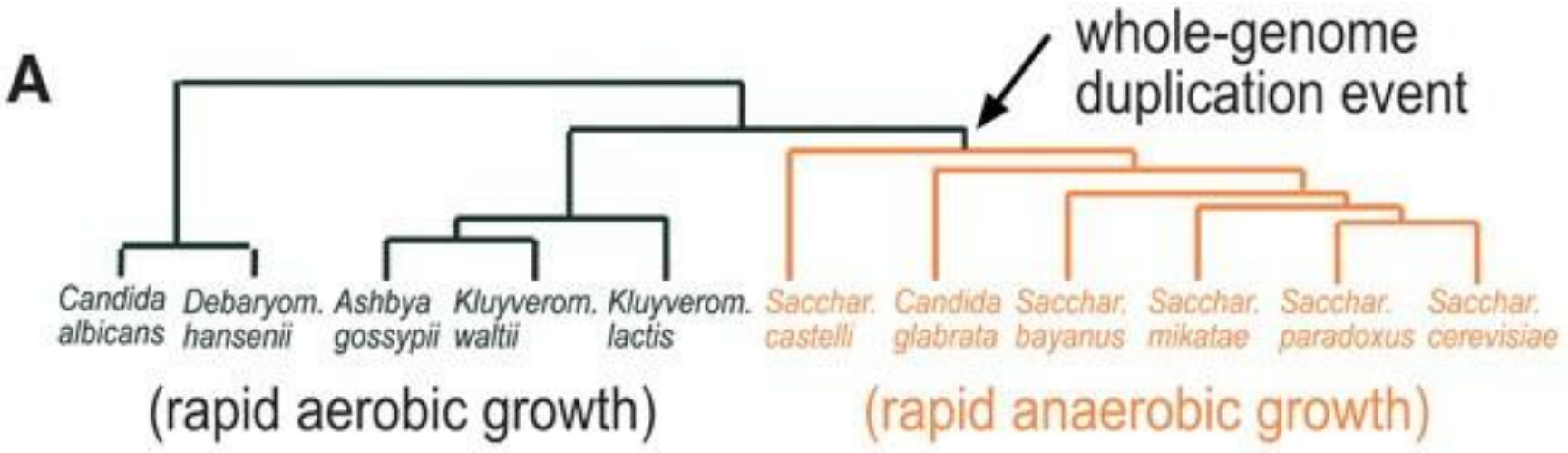
5 AUGUST 2005 VOL 309 SCIENCE www.sciencemag.org

## Goals

- Genotype and Phenotype

# Background – Yeast Growth

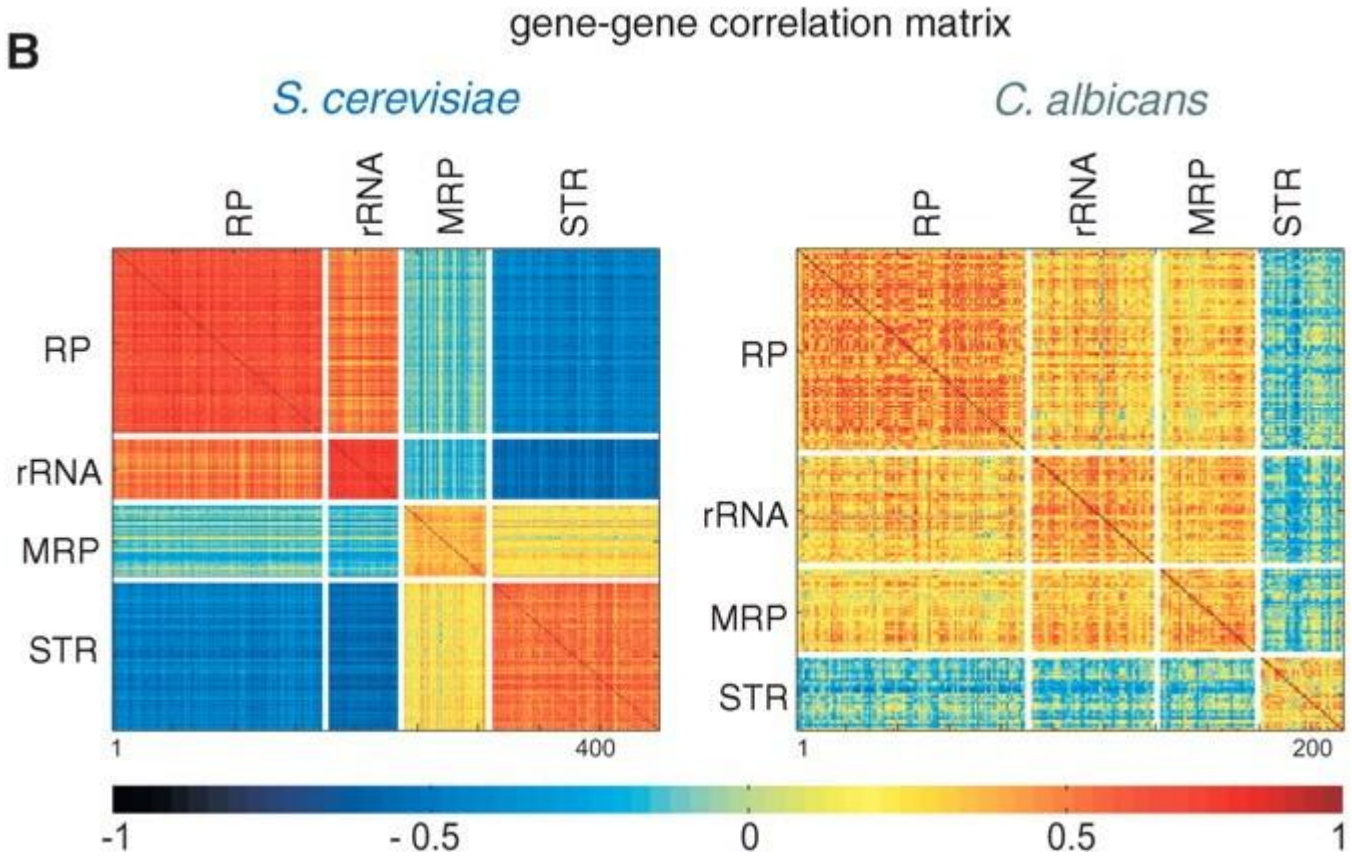
- Some Yeast species are fermentative
- Others can employ only respiration
- Connection to whole genome duplication



# Analysis of Genetic Basis of Phenotypic Diversity

- Examine gene expression program
  - *S. Cerevisiae* with 1000 expression profiles
  - *C. Albicans* with 198 profiles
- Motif search in related modules
- Validation of motif role
- Comparison of motif and phenotype evolution

# Transcriptional Wiring Differences



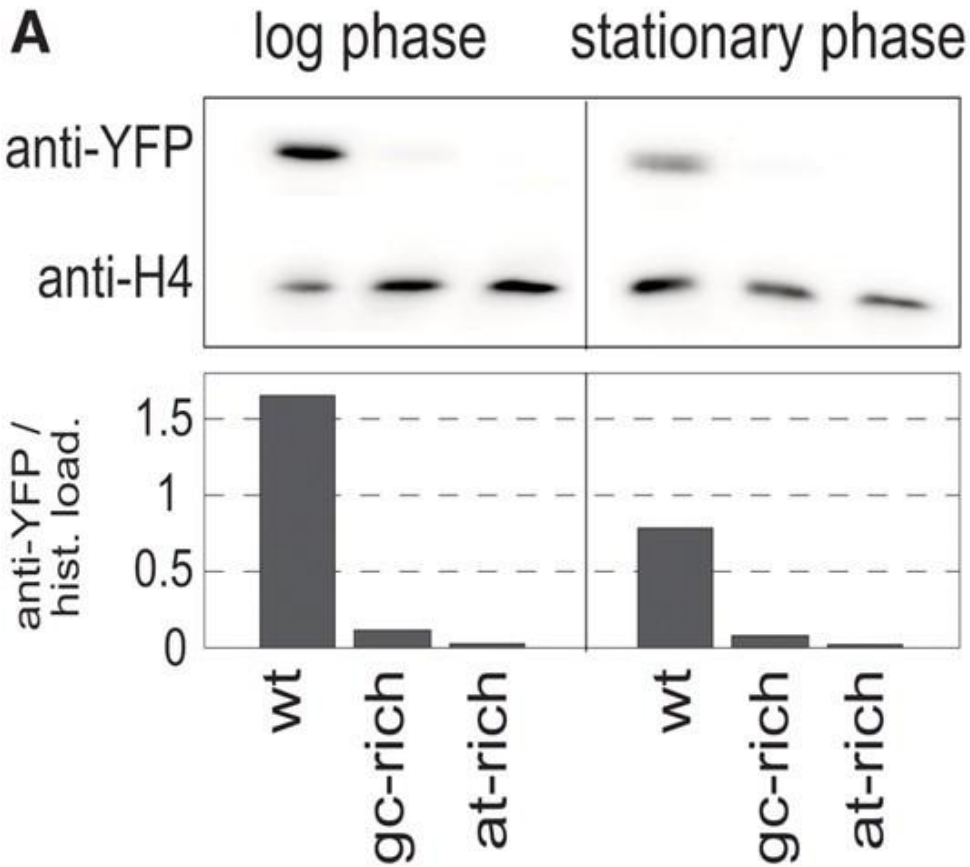

**RP** **Cytoplasmic ribosomal proteins**
**rRNA** **rRNA processing genes**
**MRP** **Mitochondrial RP**
**STR** **Environmental Stress response**

# Cis or Trans?

***Cis* element search:**

- PAC in rRNA of both species
- None for MRP in *S. Serevisiae*
- `AATTTT` in MRP, putative rRNA regulator

Validation of cis-regulatory Role of `AATTTT` in MRP



# Spatial Configuration of AATTTT

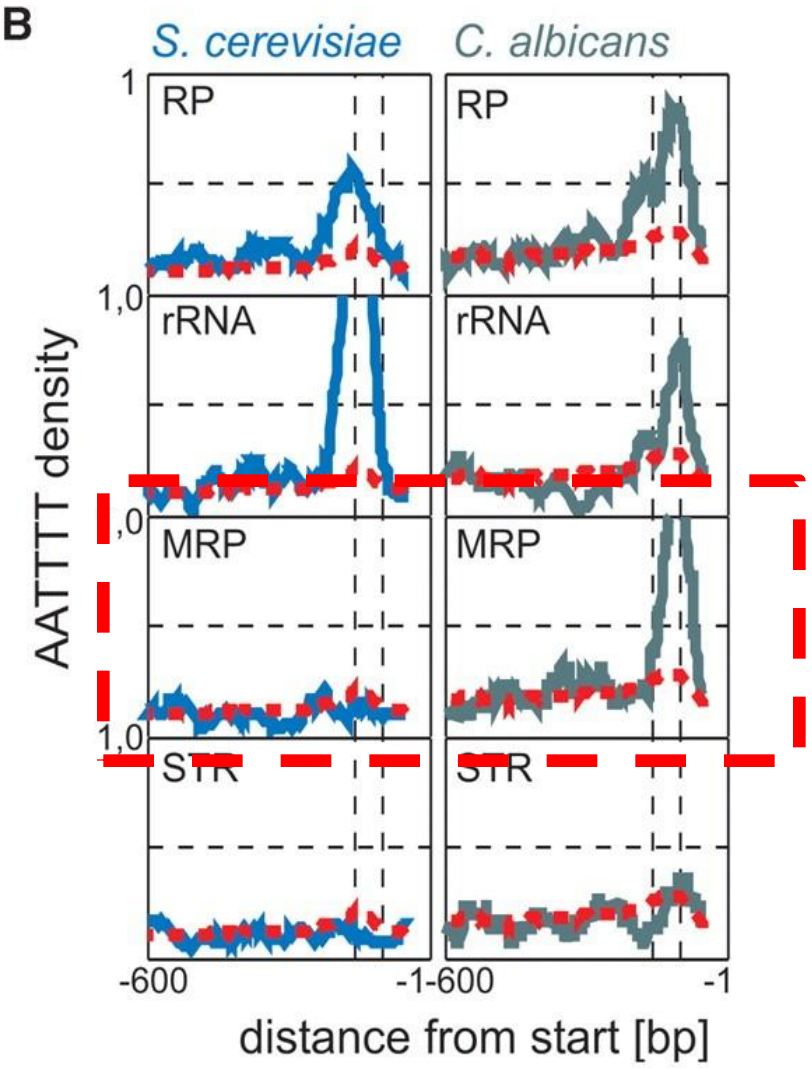
**Both species:**

- Position is confined in RP and rRNA
- Not represented in STR

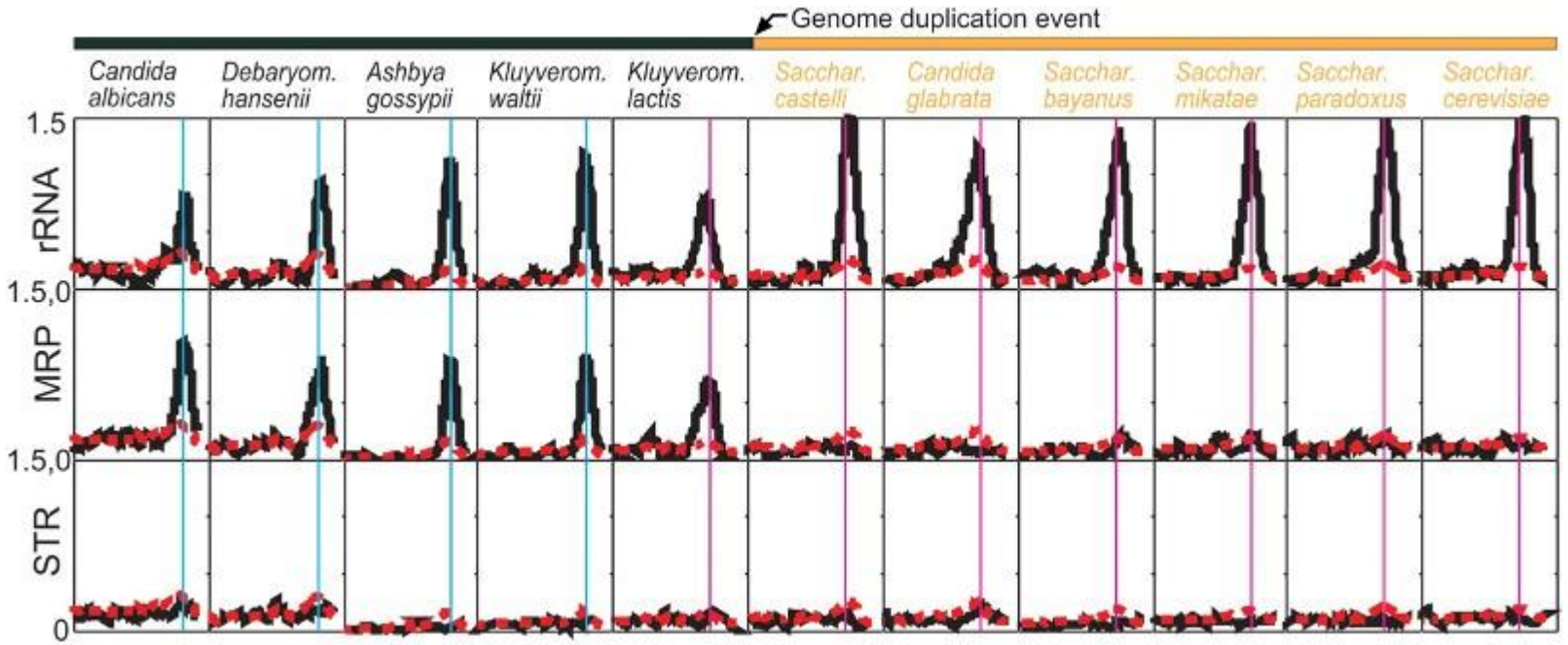
***C. Albicans*:**

- **AATTTT** Regulates also MRP

➡ **Rapid growth element**



# Loss or Gain of Binding Site?




- Loss of binding site in MRP associated with whole genome duplication

# Evolution of Genome and Phenotype

- “Gene duplication can facilitate the evolution of new function”
  - by specialization of new coding sequences
  - Also by facilitating the evolution of gene expression

# Summary of methodologies

- Integration of sequence and gene-expression
  - Finding orthologous modules
  - Finding orthologous binding motifs
- Exploiting phylogenetic trees
  - Find genotype change rules
  - Associate phenotype and genotype changes
  - Exploit gene expression data of extremes

# Summary of principles

- Conservation of regulatory programs
  - Binding site conservation
  - Position and spacing
- Conservation of module with evolution of control
  - Loss and gain
  - Drift
  - Switching, could be explained via redundancy
- Evolution of regulatory programs -> Evolution of phenotype
  - Addition of gene targets
  - Cooption or loss of TF
  - Also facilitated by whole genome duplication

# Uncertainty and probability

**Uncertainty is related to our surprise at an event**

**"The sun will rise tomorrow"** **Not surprising (p~1)**

**"The sun will *not* rise tomorrow"** ***Very*** **surprising (p<<1)**

**Uncertainty is inversely related to probability of event**

$$\text{Uncertainty} \propto \log \frac{1}{p_{event}}$$

# Average Uncertainty

**Two possible outcomes for sun rising**

**A "The sun will rise tomorrow"** $P(A)=p_1$

**B "The sun will *not* rise tomorrow"** $P(B)=p_2$

**What is our *average uncertainty* about the sun rising**

# Entropy

Entropy measures average uncertainty

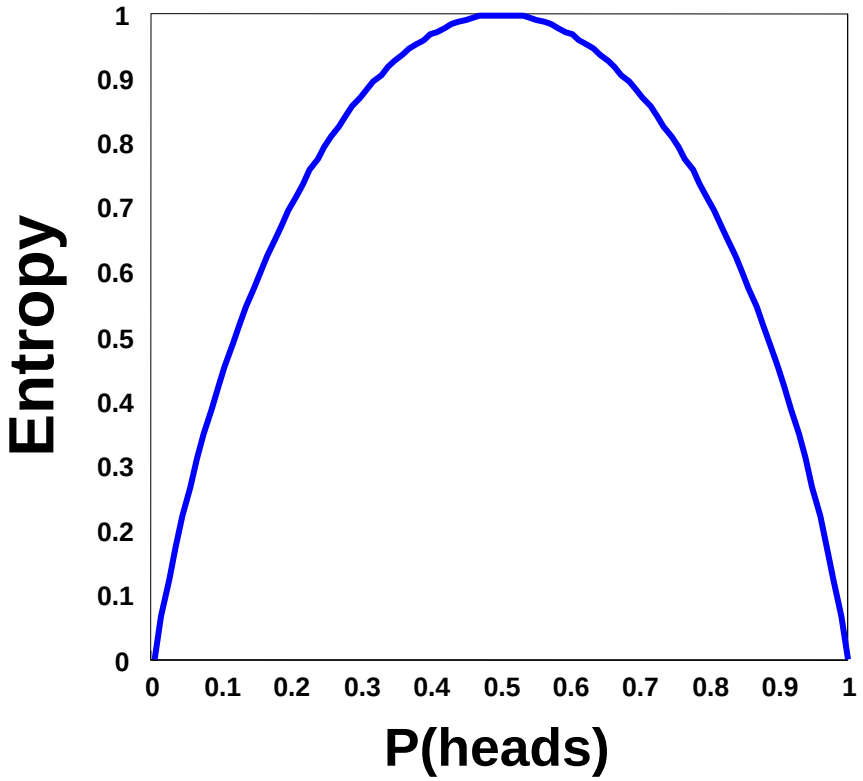
Entropy measures randomness

$$H(X) = -\sum_i p_i \log_2 p_i$$

If log is base 2, then the units are called bits

# Entropy versus randomness

**Entropy is maximum at maximum randomness**

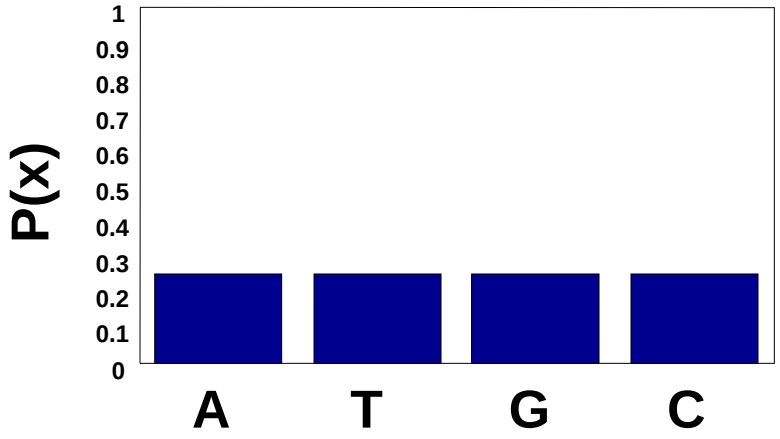



**Example: Coin Toss**

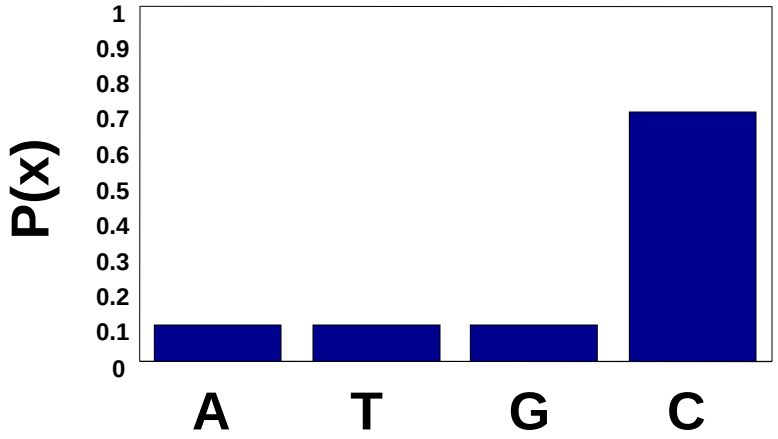
**P(heads)=0.1 Not very random**
**H(X)=0.47 bits**

**P(heads)=0.5 Completely random**
**H(X)=1 bits**

# Entropy Examples




$$H(X) = -[0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25)]$$
$$= 2 \text{ bits}$$




$$H(X) = -[0.1\log(0.1) + 0.1\log(0.1) + 0.1\log(0.1) + 0.75\log(0.75)]$$
$$= 0.63 \text{ bits}$$

# Information Content

Information is a *decrease in uncertainty*

**Once I tell you the sun will rise, your uncertainty about the event decreases**

**Information = $H_{before}(X)$ - $H_{after}(X)$**

***Information is difference in entropy after receiving information***

# Motif Information

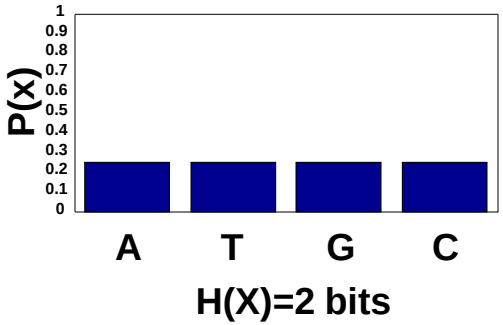
**Motif Position Information =** $2 - \sum_{b=\{A,T,G,C\}} -p_b \log p_b$

$H_{background}(X)$

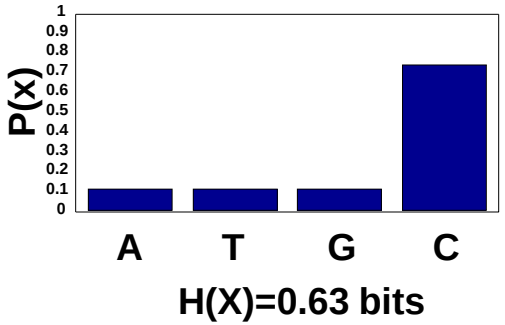
$H_{motif\_i}(X)$

**Prior uncertainty about nucleotide**

**Uncertainty after learning it is position i in a motif**



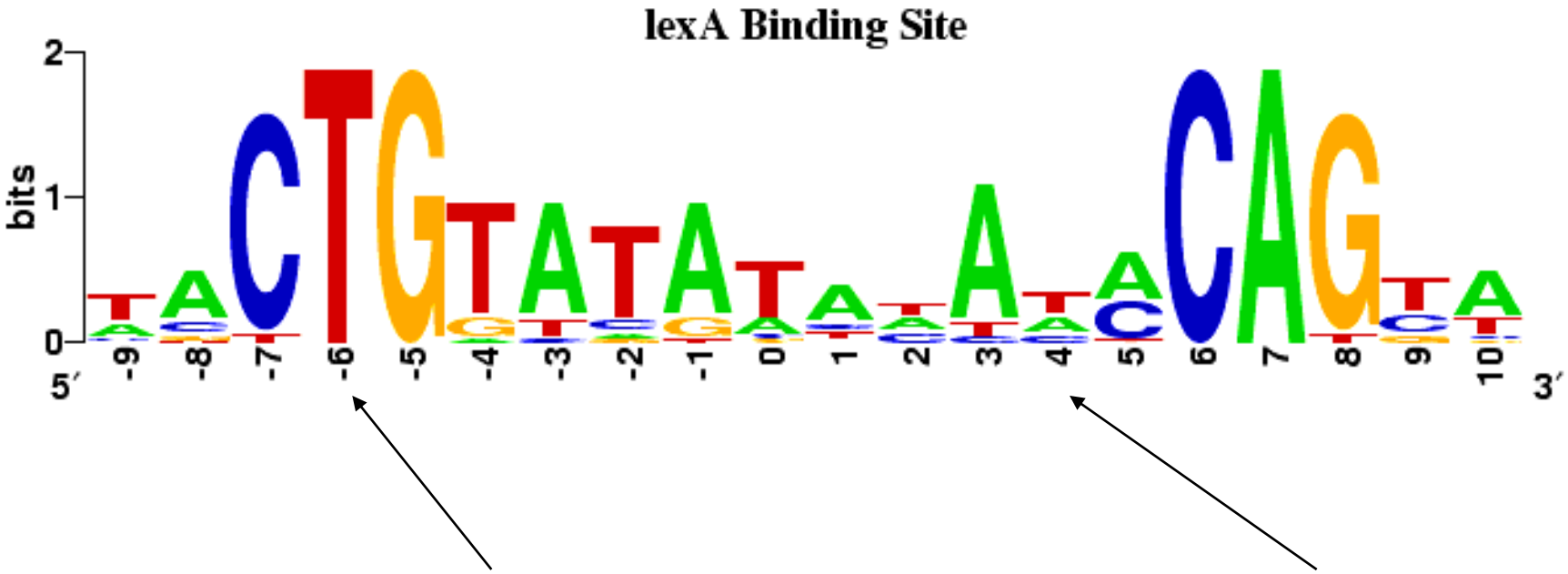

**H(X)=2 bits**




**H(X)=0.63 bits**

**Uncertainty at this position has been reduced by 0.37 bits**

# Motif Logo




**Conserved Residue**
**Reduction of uncertainty of 2 bits**

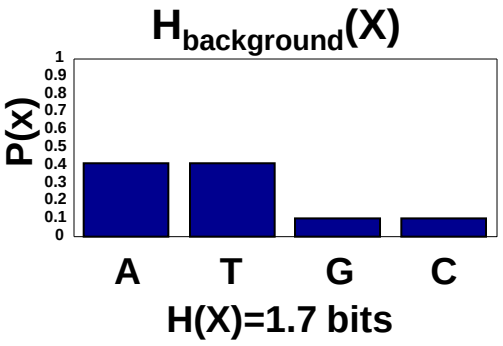
**Little Conservation**
**Minimal reduction of uncertainty**

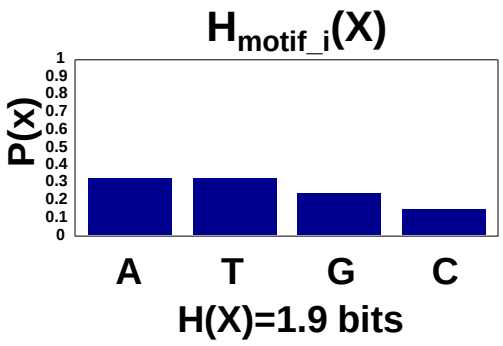
# Background DNA Frequency

The definition of information assumes a uniform background DNA nucleotide frequency

What if the background frequency is not uniform?

**(e.g. Plasmodium)**




**Motif Position Information =** **1.7** **-** $\sum_{b=\{A,T,G,C\}} -p_b \log p_b$ **= -0.2 bits**

Some motifs could have negative information!