Gene regulatory networks

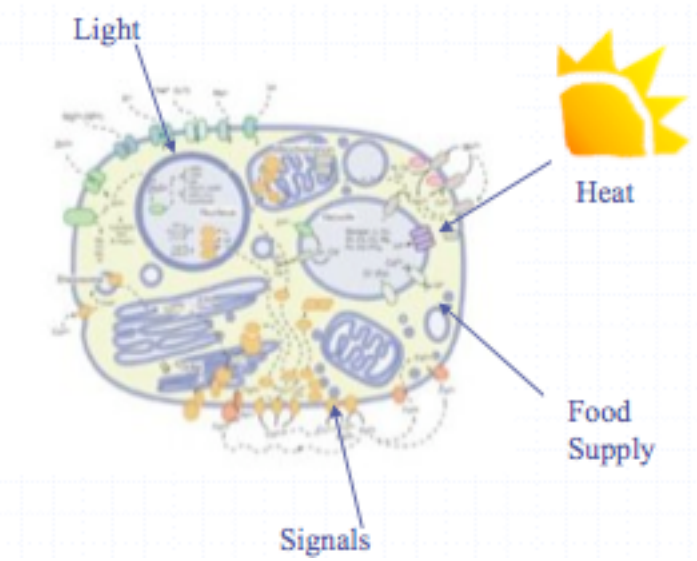
Lecture 1

Cells Must Regulate Internal Processes

- Metabolic processing
- Cell cycle functions
  - Growth
  - DNA replication/repair
  - Mitosis
  - Preparation phases
- These require careful control of gene expression
  - Expression level
  - Timing
  - Coordination

Cells Must Adapt To Their Environment

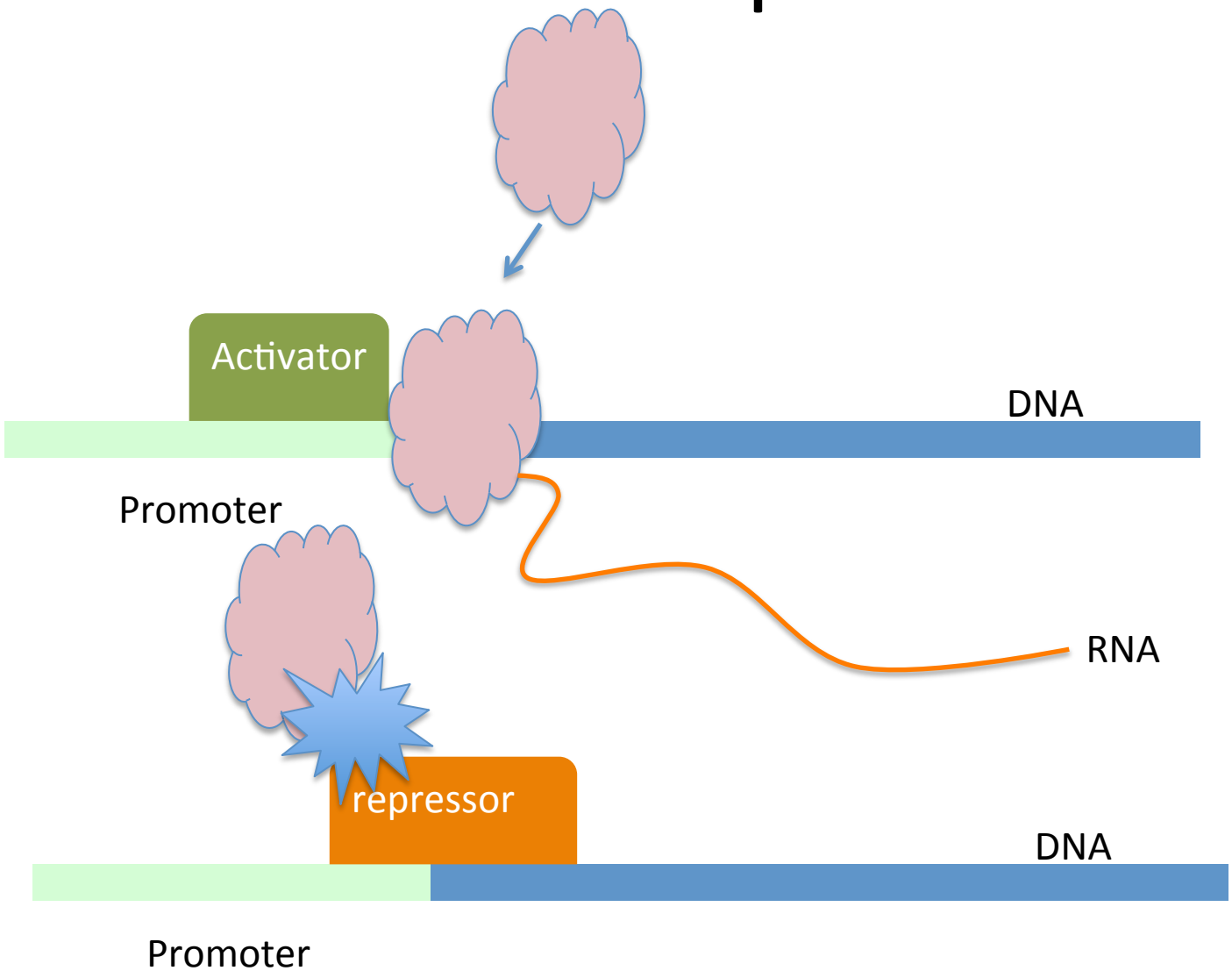
- Getting nutrition supply
- Response for signals
- Response for stress



Regulating Transcription Is A Key Construct

- Transcription factors (TF) regulate transcription
  - Promoters control transcription initiation (cis-regulation)
  - Enhancers control transcription from afar (trans-regulation)
- Most genes are involved in regulation

Activator and repressor



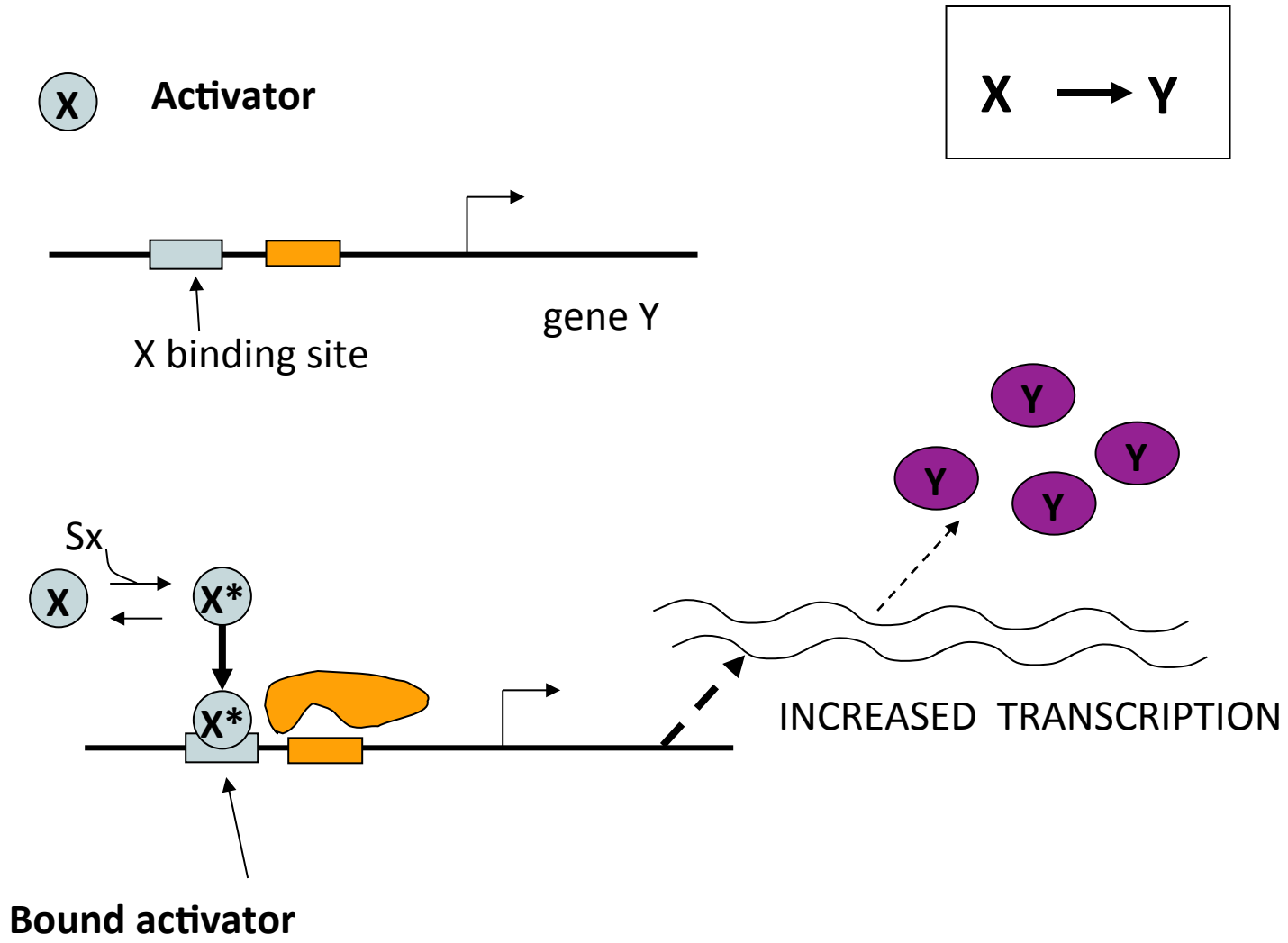


Fig 2.2 (b) An activator X, is a transcription- factor protein that increases the rate of mRNA transcription when it binds the promoter. The activator transits rapidly between active and inactive forms. In its active form, it has a high affinity to a specific site (or sites) on the promoter. The signal Sx increases the probability that X is in its active form X*. Thus, X* binds the promoter of gene Y to increase transcription and production of protein Y. The timescales are typically sub-second for transitions between X and X*, seconds for binding/ unbinding of X to the promoter, minutes for transcription and translation of the protein product, and tens of minutes for the accumulation of the protein,

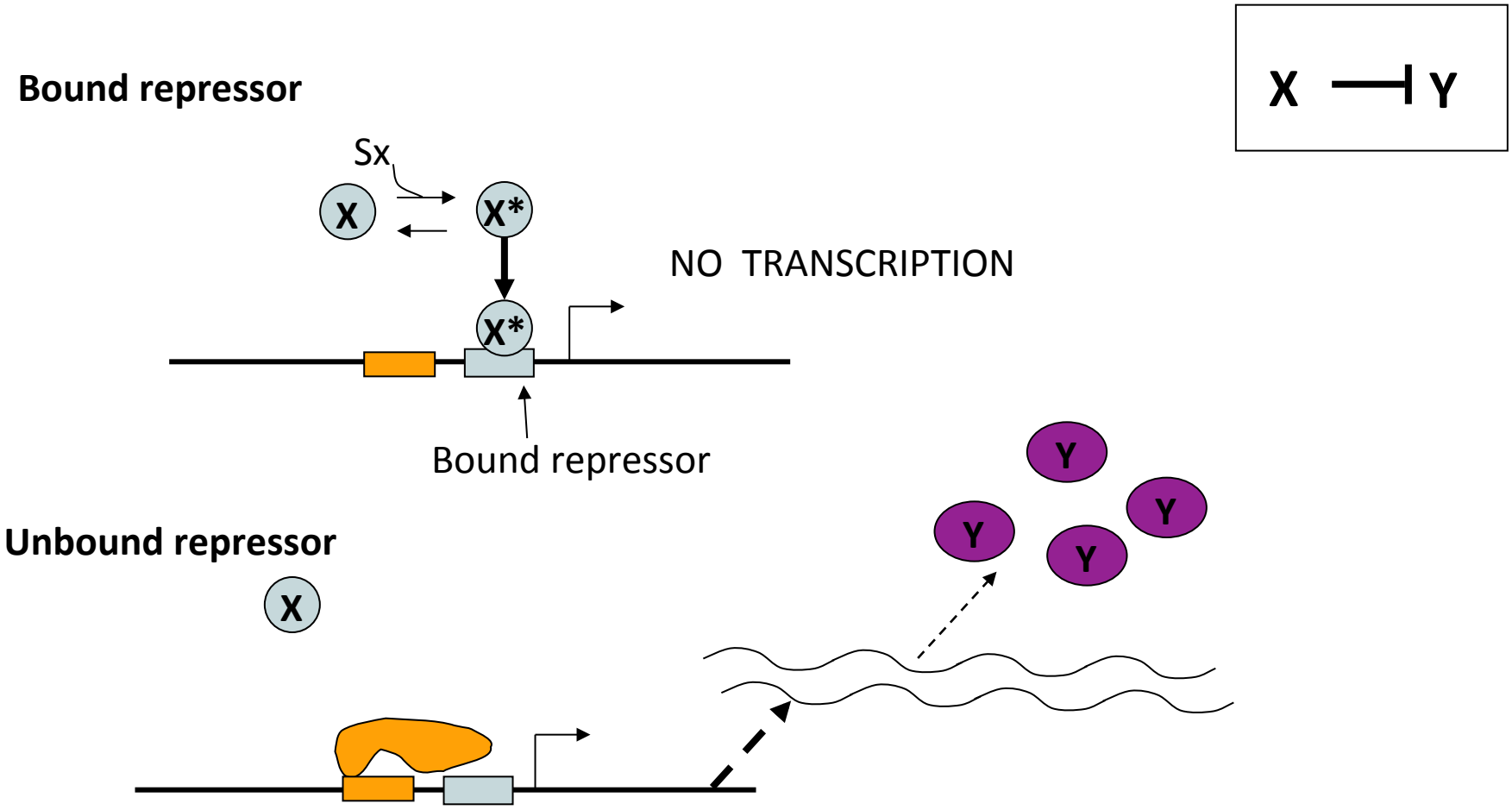
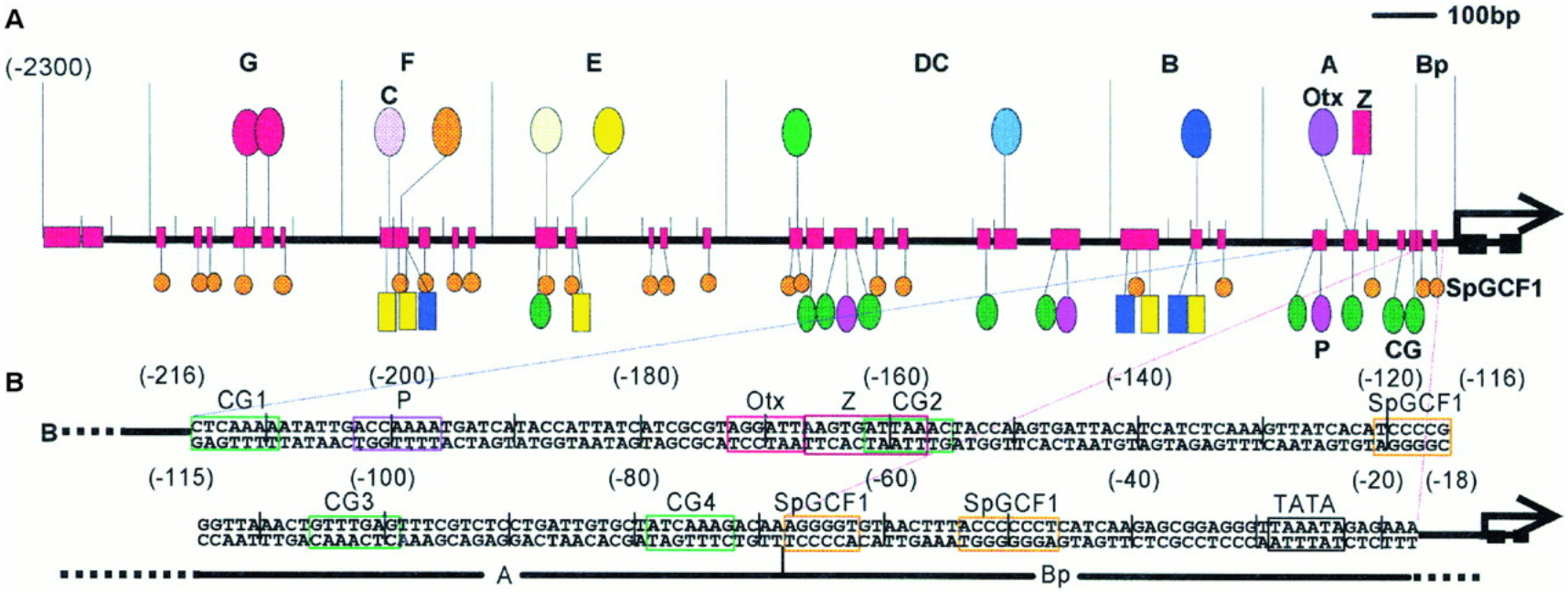


Fig 2.2c A repressor X, is a transcription- factor protein that decreases mRNA transcription when it binds the promoter. The signal Sx increases the probability that X is in its active form X*.
X* binds a specific site in the promoter of gene Y to decrease transcription and production of protein Y.
Many genes show a weak (basal) transcription when repressor is bound.

Architecture Of Cis-Regulation

Yuh et al. Science (1998) 279: 1896-1902



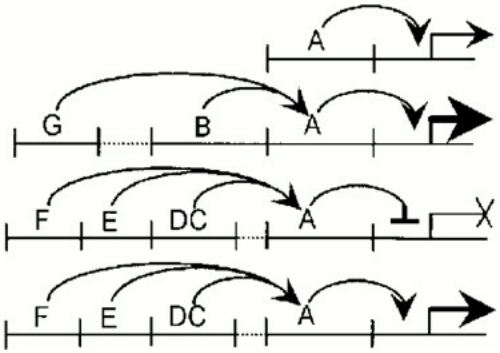
C Module A functions:

Vegetal plate expression in early development:

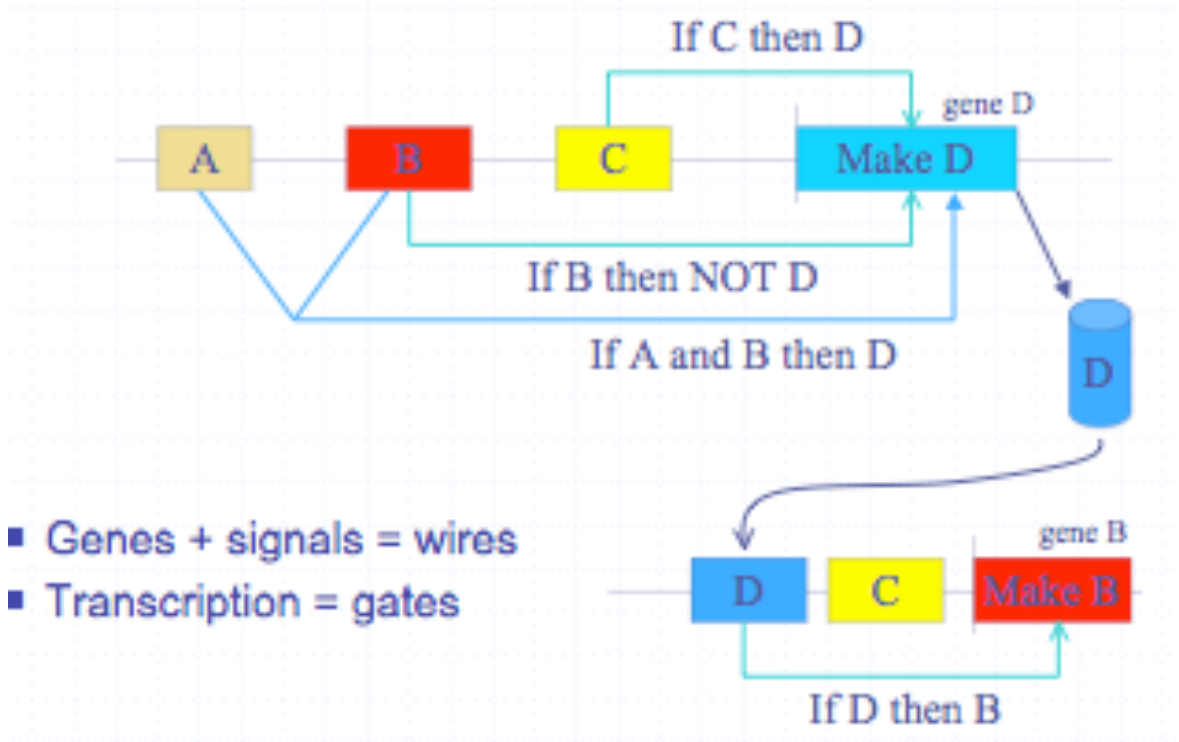
Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:

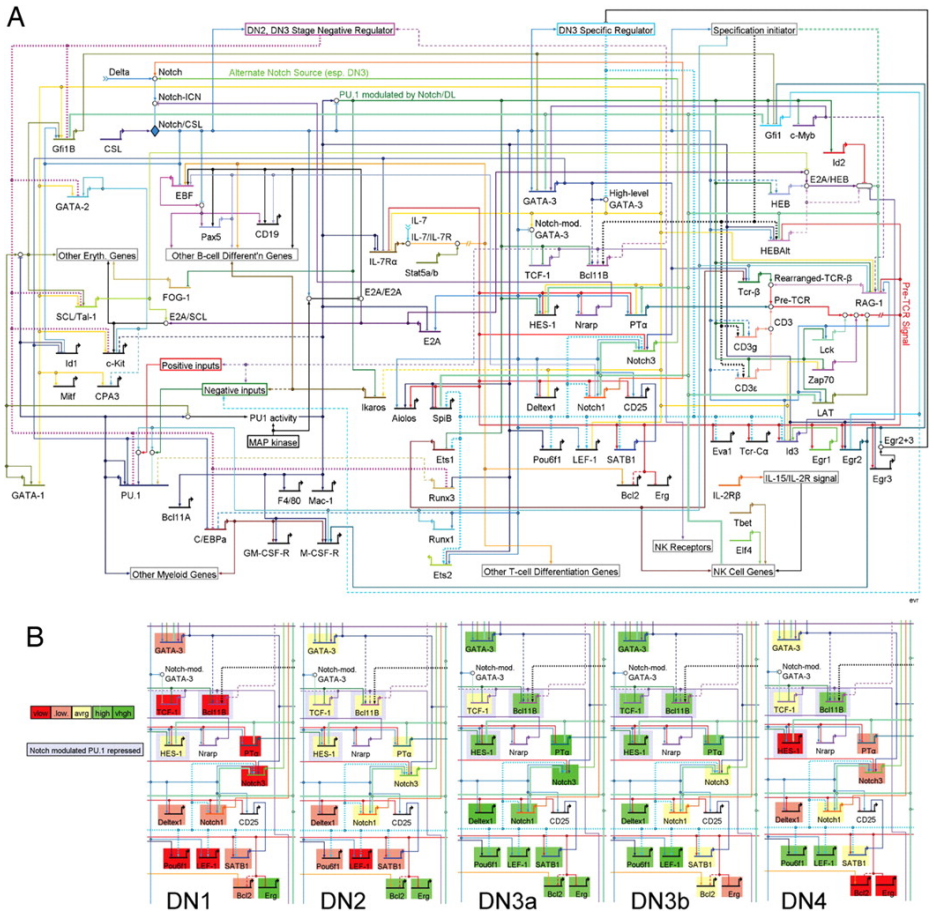


Regulatory Network



Cell can be represented as a regulatory network

The Cell as a regulatory network



Key questions

- How to discover TFs/Binding sites (motifs)? (with DNA sequences)
- How to discover regulatory network pathways? (with expression profiles)

Regulatory Regions

- Every gene contains a regulatory region (RR) typically stretching 100-1000 bp upstream of the transcriptional start site

- Located within the RR are the ***Transcription Factor Binding Sites*** (TFBS), also known as ***motifs***, specific for a given transcription factor

- TFs influence gene expression by binding to a specific location in the respective gene's regulatory region - TFBS

- So finding the same motif in multiple genes' regulatory regions suggests a regulatory relationship among those genes.

- Note that the same motif is not necessary to be the identical sequences.

Motifs and Transcriptional Start Sites

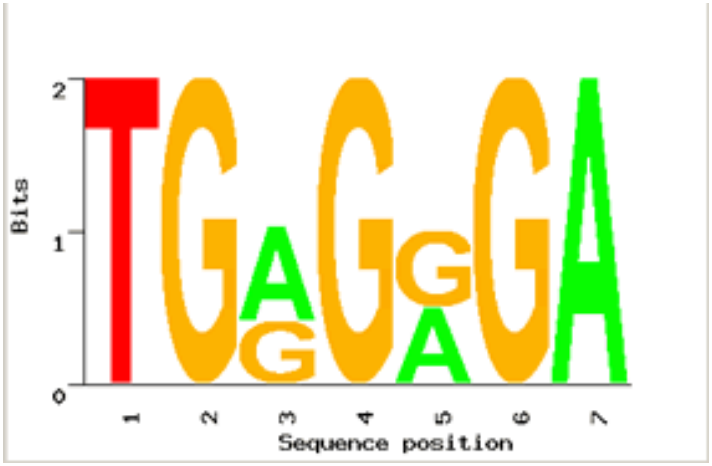
Motif visualization

- Visualizing Motifs
  - Motif “Information”

Motif Logo

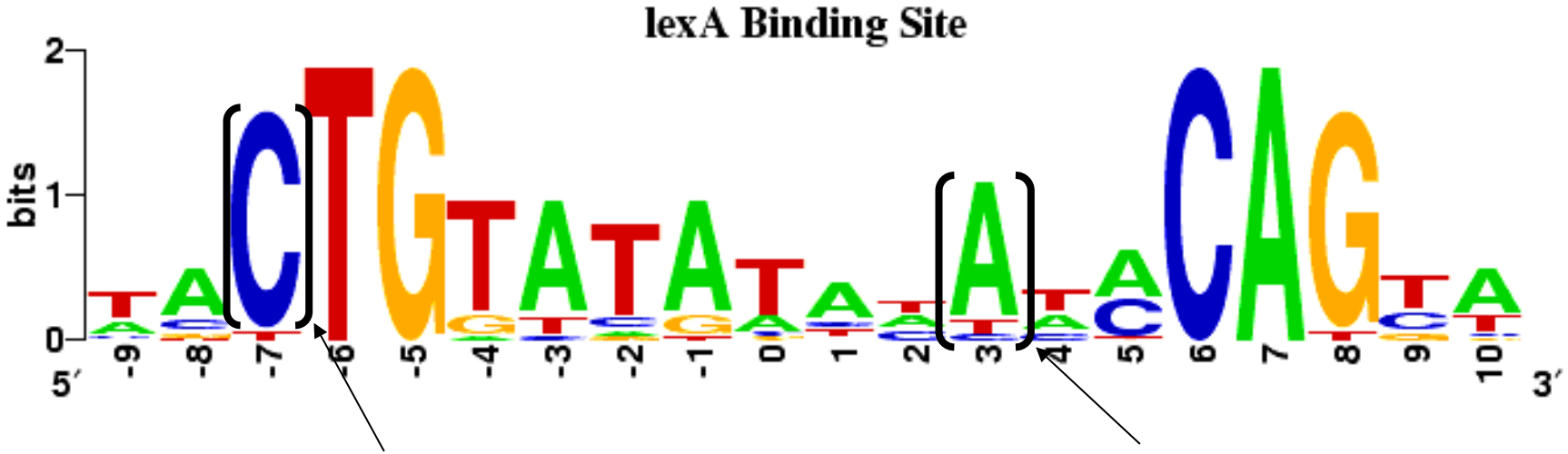
- Motifs can mutate on non important bases
- The five motifs in five different genes have mutations in position 3 and 5
- Representations called *motif logos* illustrate the conserved and variable regions of a motif

TGGGGGA
TGAGAGA
TGGGGGA
TGAGAGA
TGAGGGA



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

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

Online Logo Generation

WEBLOGO

· about · create · examples ·

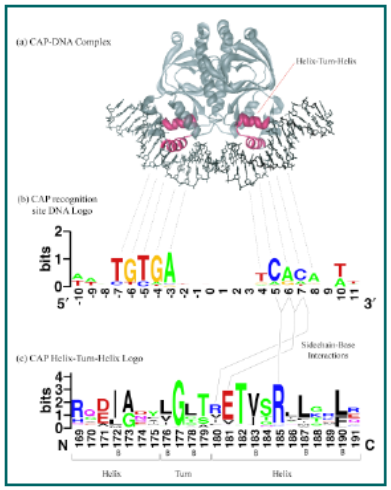
Version 2.8.2 (2005-09-08)

(⇒ 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

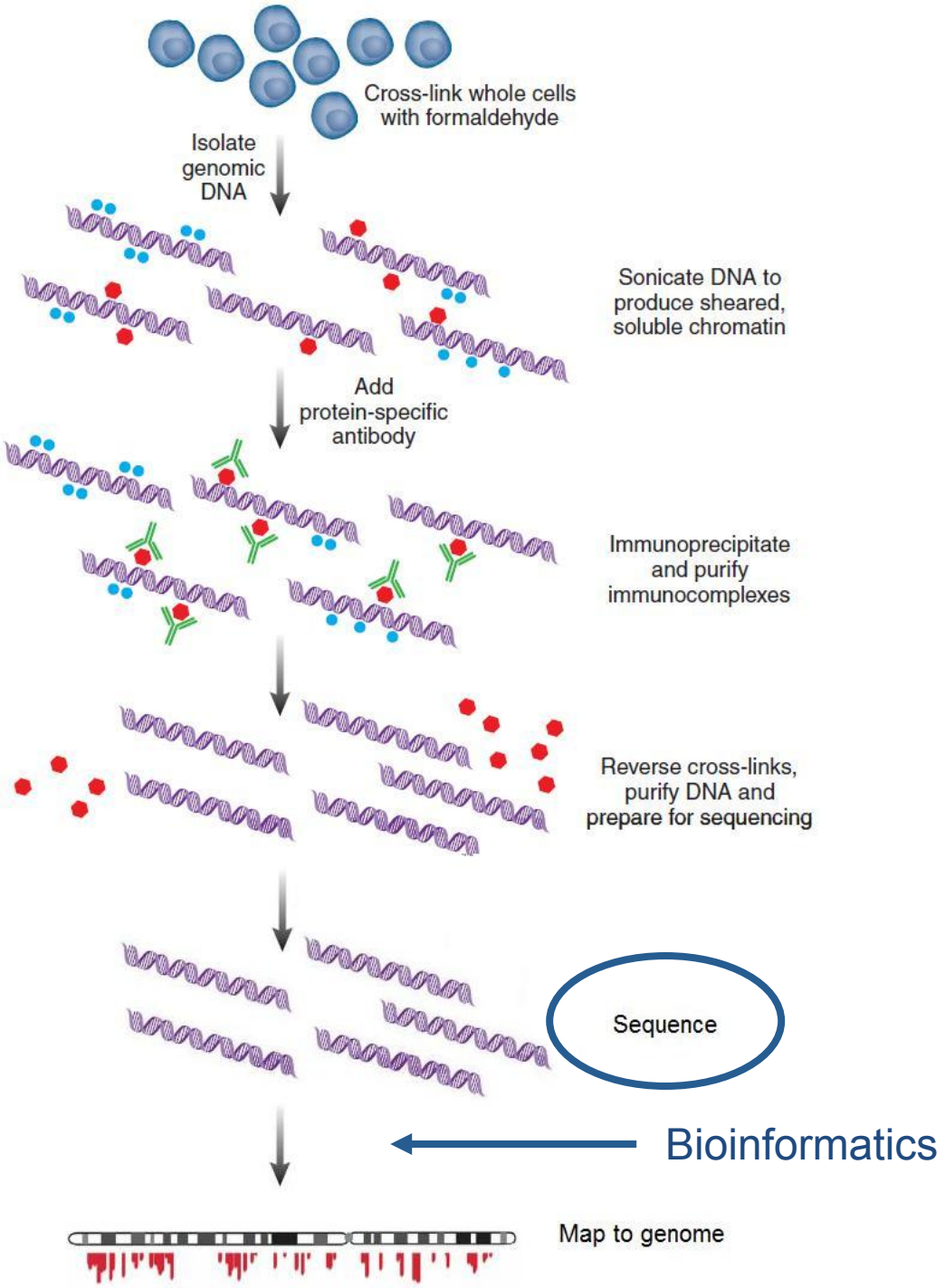
Identifying Motifs

- Regulatory protein (TF) binds to a short DNA sequence called a motif (TFBS)
- So finding the same motif in multiple genes' regulatory regions suggests a regulatory relationship amongst those genes.
- Note that the same motif is not necessary to be the identical sequences.

Recent Directions

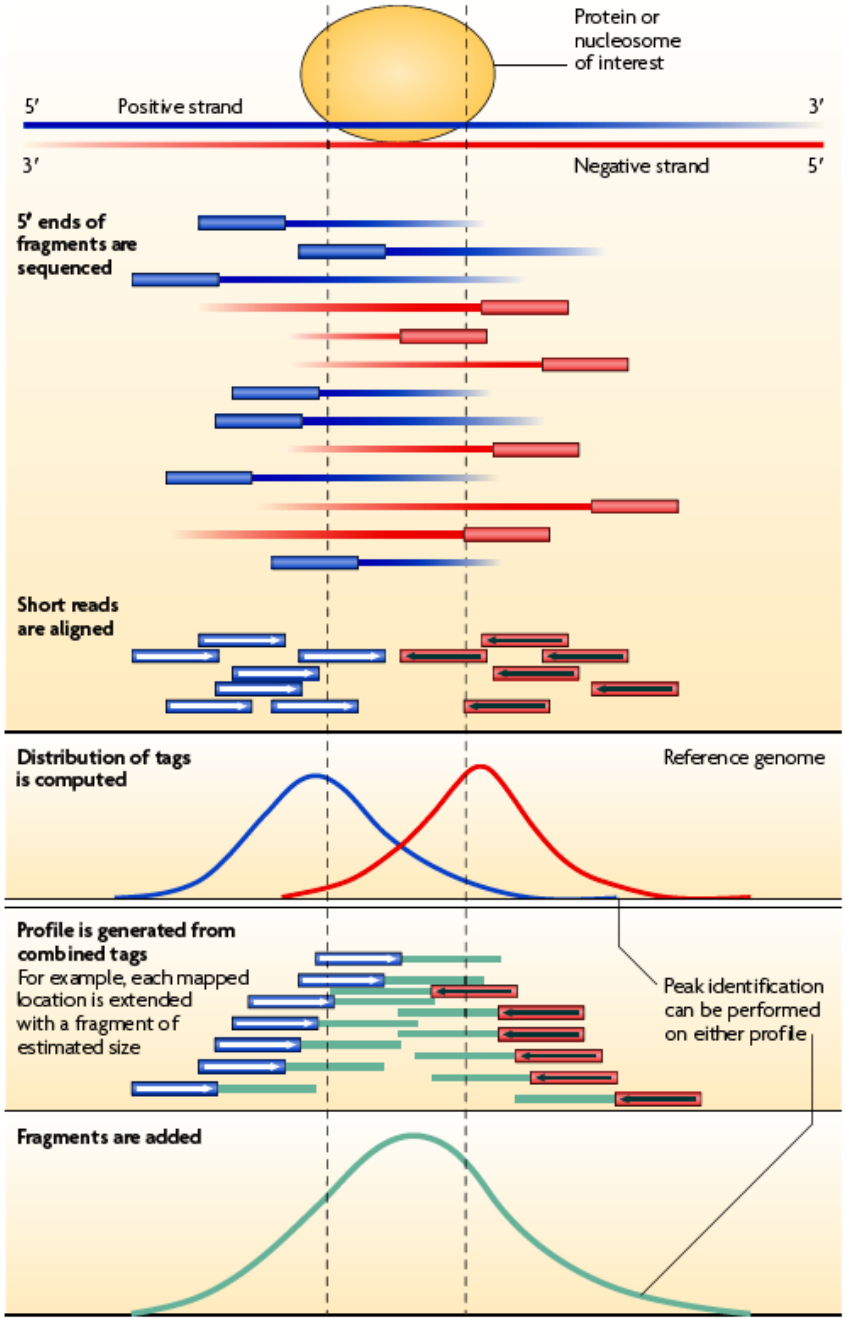
- Experimental Data
  - ChIP-chip
  - ChIP-seq
- Motif identification with bioinformatic approaches

ChIP-seq


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing
Sequence
Bioinformatics
Map to genome

Peaks for TF binding sites --Strand-specific profiles at enriched sites

the fragments are sequenced at the 5' end, and the locations of mapped reads should form two distributions, one on the positive strand and the other on the negative strand, with a consistent distance between the peaks of the distributions.



Recent Directions

- Experimental Data
  - ChIP-chip
  - ChIP-seq
- Motif identification with bioinformatic approaches

Difficulties for bioinformatic approaches

- We do not know the motif sequence.
- We do not know where it is located relative to the genes start.
- Motif sequences can differ slightly from one gene to the next.
- How to discern it from “random” motifs?

The Motif Finding Problem

- Given a random sample of DNA sequences:

```
cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgtttccaaccat
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt
agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgtacgtataca
ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgttaacgtacgtc
```

- Find the pattern that is implanted in each of the individual sequences, namely, the motif

Identifying TFBS

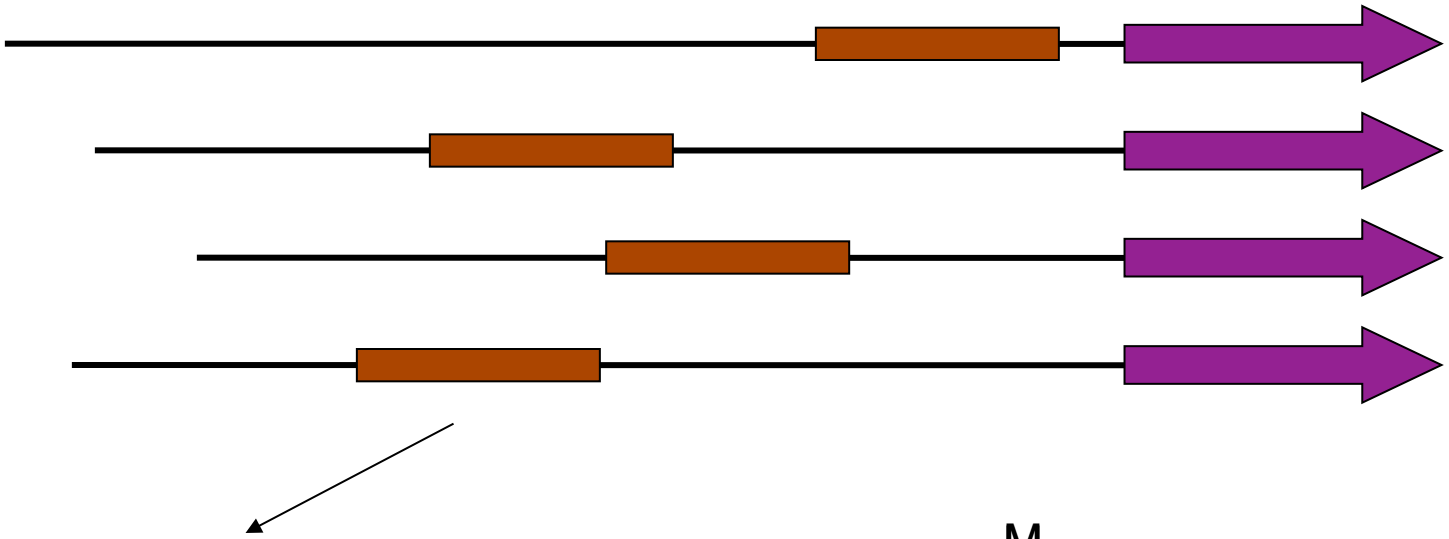
- Knowledge-based methods
  - Probabilistic Model
- *Ab init* methods
  - Gibbs sampling
  - MEME
- Evolution approach

Essential Tasks

- Modeling Motifs (learning procedure)
  - How to computationally represent motifs
- Predicting Motif Instances
  - Using the model to classify new sequences

Parameterizing the Motif Model

Given multiple sequences and motif locations



AATGCG
ATATGG
ATATCG
GATGCA

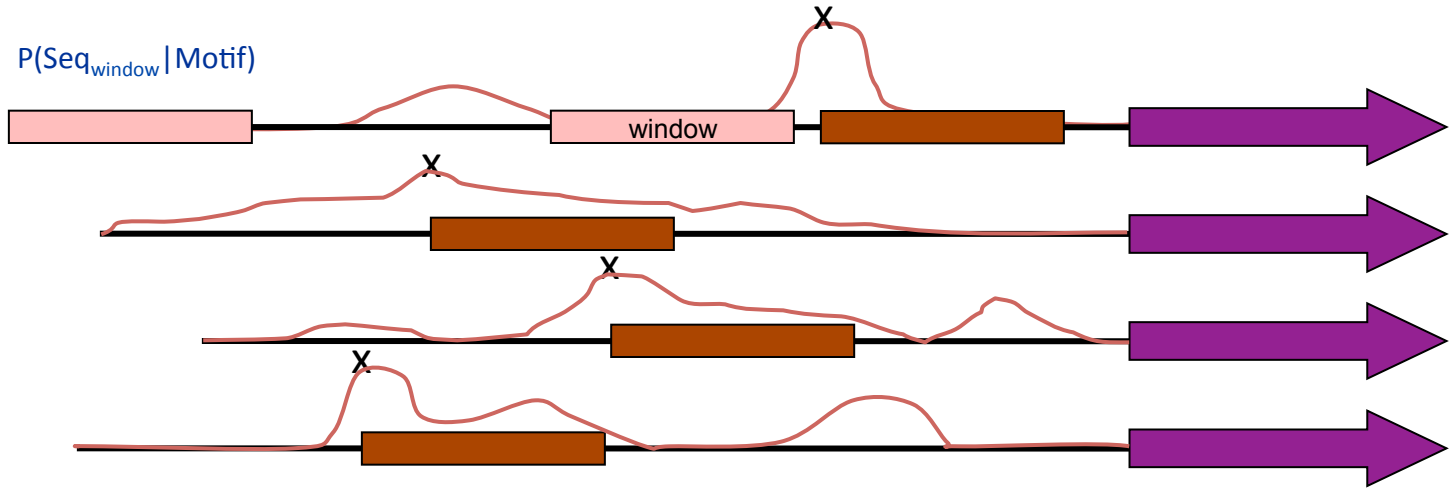
Count Frequencies

Add pseudocounts

	M_1					M_6
A	3/4					
C		ETC…				
G						3/4
T						

Finding Known Motifs

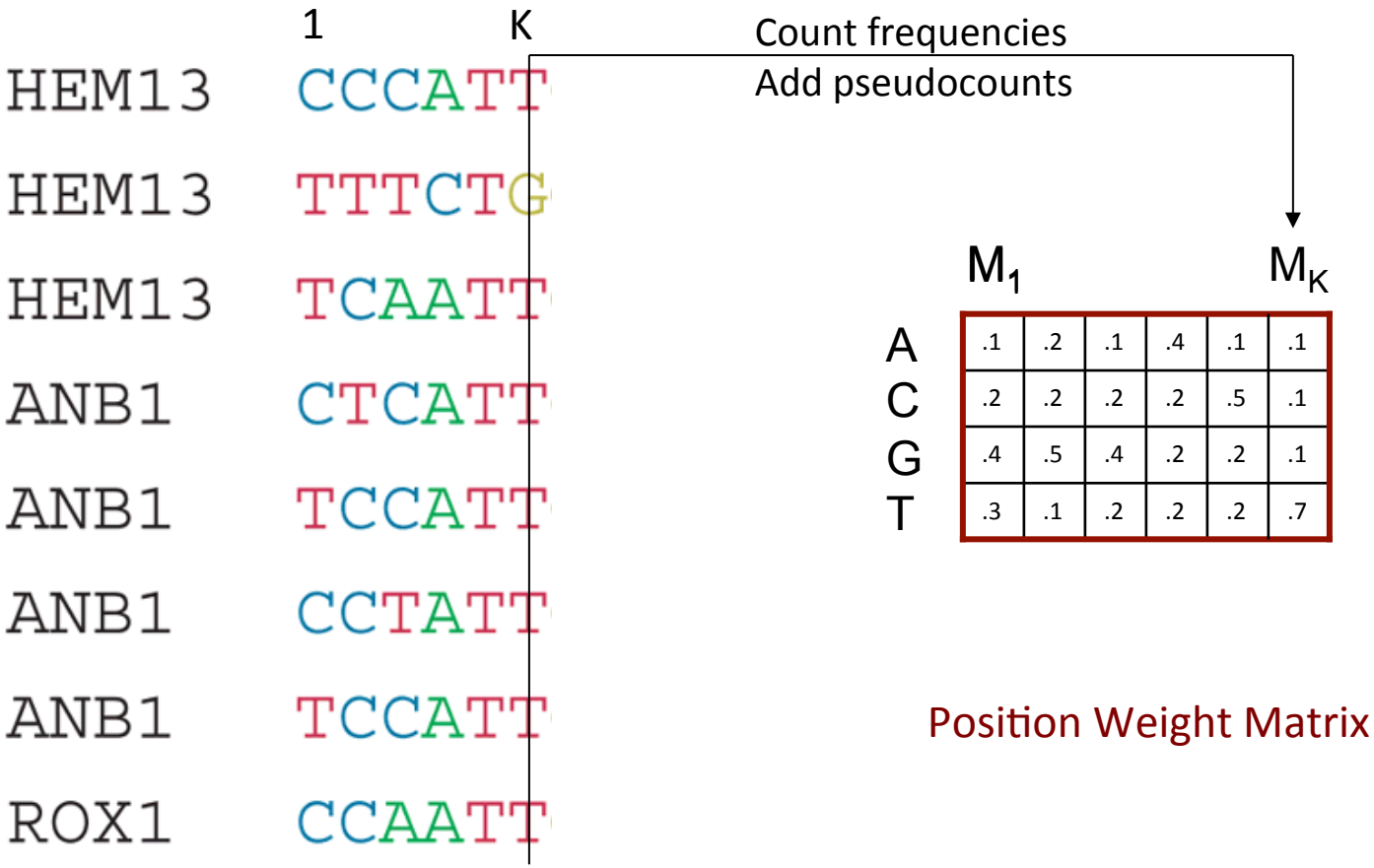
Given multiple sequences and motif model but no motif *locations*



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

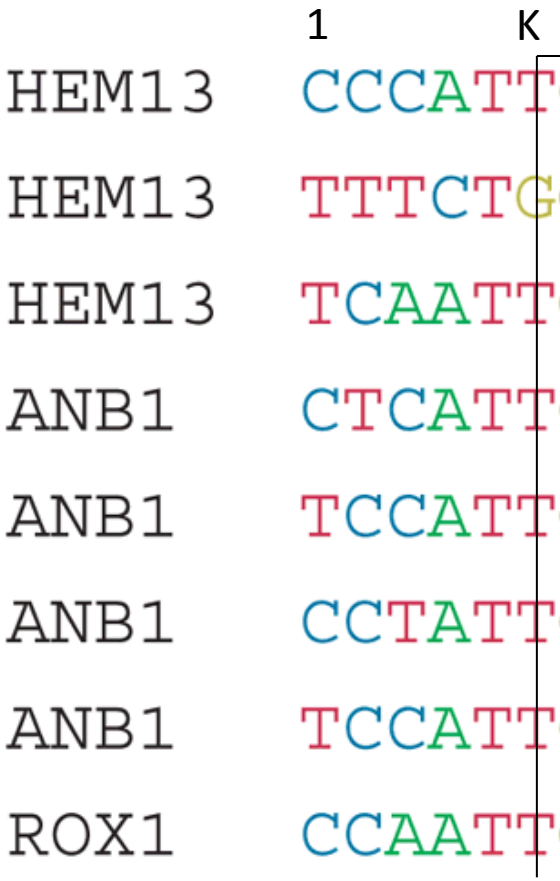
Choose best starting location in each sequence

Probabilistic Model



Position Weight Matrix

Probabilistic Model: step 1



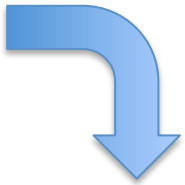
Count frequencies

A	0	0	2	7	0	0
C	4	6	4	1	0	0
G	0	0	0	0	0	1
T	4	2	2	0	8	7

Position Frequency Matrix (PFM)

Probabilistic Model: step 2

A	0	0	2	7	0	0
C	4	6	4	1	0	0
G	0	0	0	0	0	1
T	4	2	2	0	8	7



Position Frequency Matrix (PFM)

$P_k(S|M)$

A	0	0	0.25	0.875	0	0
C	0.5	0.75	0.5	0.125	0	0
G	0	0	0	0	0	0.125
T	0.5	0.25	0.25	0	1	0.875

Scoring A Sequence

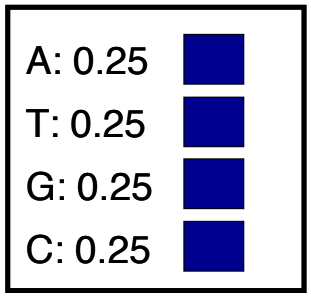
To score a sequence, we compare to a null model

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

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)



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

Probabilistic Model: step 3

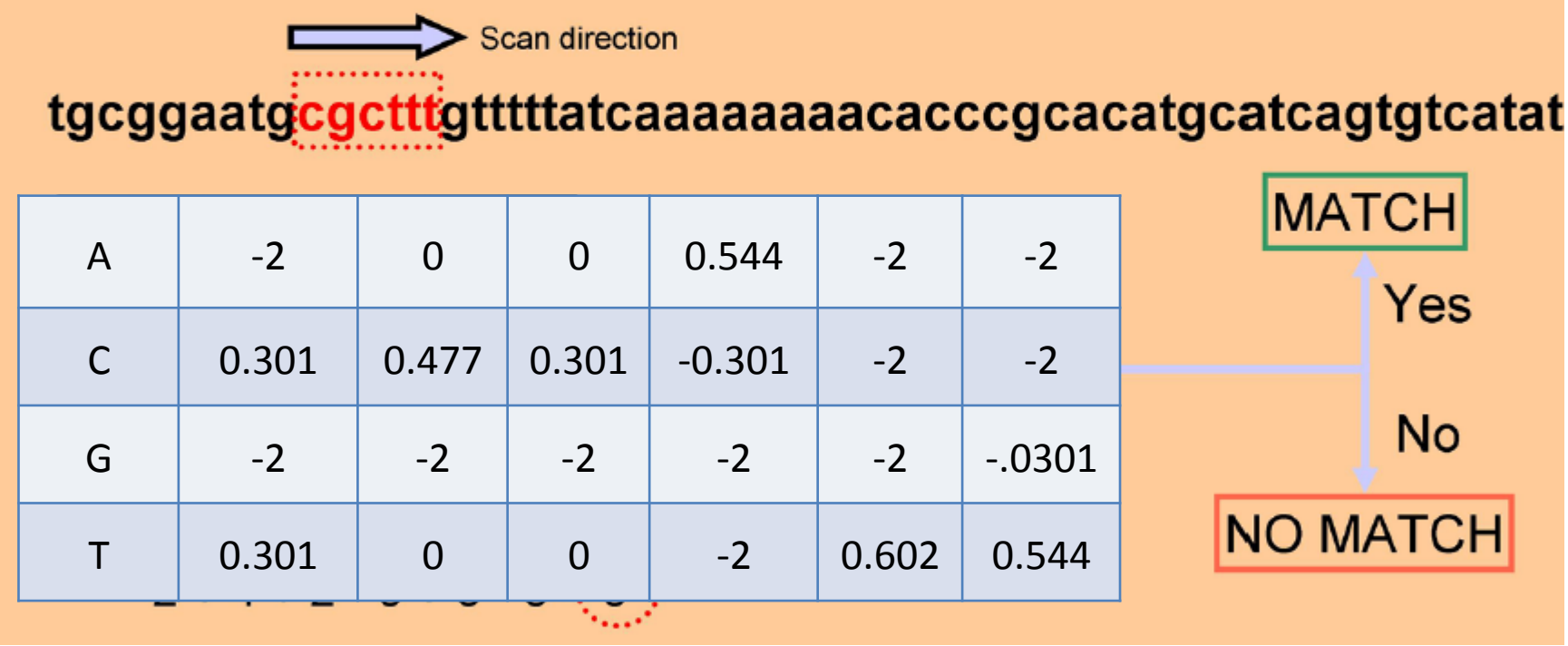
A	0	0	0.25	0.875	0	0
C	0.5	0.75	0.5	0.125	0	0
G	0	0	0	0	0	0.125
T	0.5	0.25	0.25	0	1	0.875

Position Frequency Matrix (PFM)

Position Weight Matrix (PWM)

A	Log(0/0.25+0.01)	0	0	0.544	-2	-2
C	Log(0.5/0.25)	0.477	0.301	-0.301	-2	-2
G	-2	-2	-2	-2	-2	-.0301
T	0.301	0	0	-2	0.602	0.544

Scoring a Sequence



A	-2	0	0	0.544	-2	-2
C	0.301	0.477	0.301	-0.301	-2	-2
G	-2	-2	-2	-2	-2	-.0301
T	0.301	0	0	-2	0.602	0.544

0.301+(-2)+0.301+(-2)+0.602+0.544=-2.252

Common threshold = 60% of maximum score

MacIsaac & Fraenkel (2006) PLoS Comp Bio

Databases

TRANSFAC: http://www.gene-regulation.com/pub/databases.html#transfac

TRANSFAC FACTOR TABLE, Release 7.0 - public - 2005-09-30, (C) Biobase G

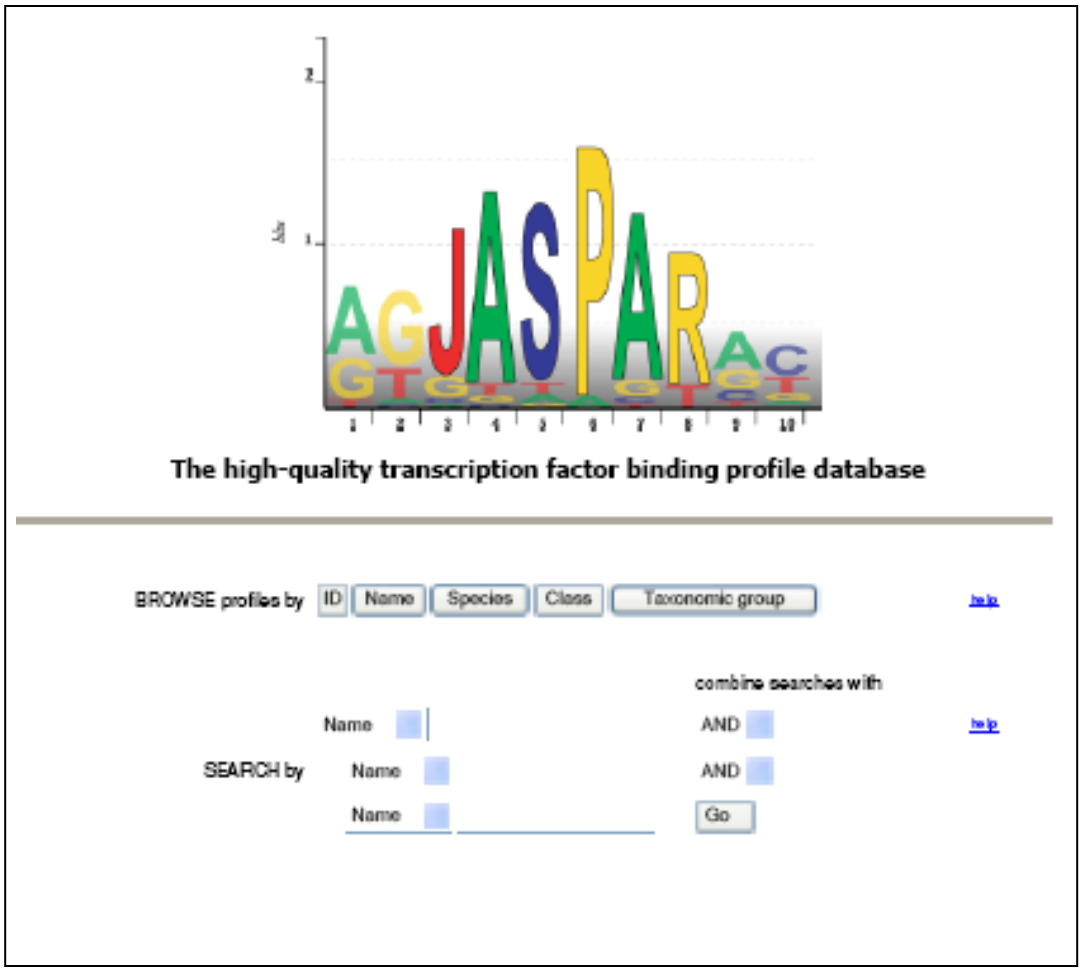
```
AC   T00302
XX
ID   T00302
XX
DT   15.10.1992 (created); ewi.
DT   26.08.2002 (updated); hom.
CO   Copyright (C), Biobase GmbH.
XX
FA   GAL4
XX
SY   GAL4; YPL248C.
XX
OS   yeast, Saccharomyces cerevisiae
OC   Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
OC   Saccharomycetaceae; Saccharomyces.
XX
SQ   MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESR
SQ   LERLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPL
SQ   TLRQHRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDM
SQ   SDGLPFLKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTT
SQ   SRLLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTD
SQ   IDVFYYQNAKSHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLG
SQ   LNRDLPSSFSDSSILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTT
SQ   TGPTIYHGIIETARLLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVWRQAPKFLQ
SQ   MDISTTALTNLLKEHPWLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQS
SQ   YEVKRCSIMLSDAAQRTVMSVSSYMDNHNVTPYFAWNCSYYFFNAVLVPIKTLLSNSKSN
SQ   AENNETAQLLQQINTVLMLLKKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYN

MX   M00049 F$GAL4_01.
MX   M00198 F$GAL4_C.
XX
BS   R00501 AS$GAL4_01; Quality: 1.
BS   R04203 AS$GAL4_02; Quality: 6.
```

Binding Sites

More Databases

http://jaspar.genereg.net/



Species-specific:

SCPD (yeast) http://rulai.cshl.edu/SCPD/

DPInteract (e. coli) http://arep.med.harvard.edu/dpinteract/

Drosophila DNase I Footprint Database (v2.0) http://www.flyreg.org/

Identifying TFBS

- Knowledge-based methods
  - Consensus sequences
  - Probabilistic Model
- *Ab init* methods (find new motifs)
  - MEME
  - Gibbs sampling
- Evolution approach

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? (Hint: start with a random motif model)

Expectation-Maximization (EM) Algorithm

- MEME
  - Missing data problem: Expectation-Maximization (EM) Algorithm to obtain maximum likelihood estimates
- EM Algorithm

  Initialization: Set frequency matrix p and p_0

  Iterations:
  - E-step: Calculate probability of motif start-positions

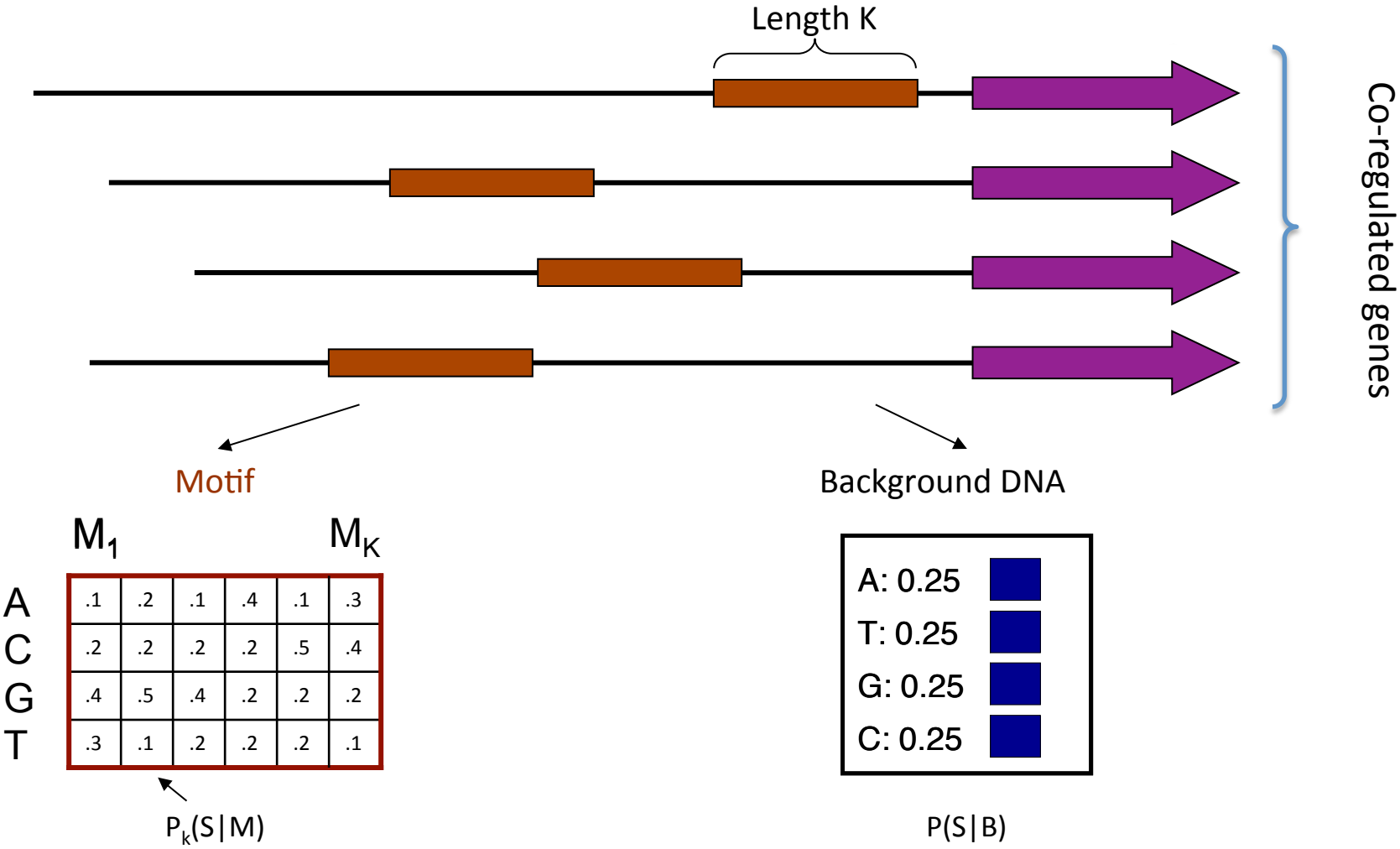
$$For\ each\ sequence\ k\ and\ position\ j$$

$$W_{kj} = Pr(motif\ start\text{-}position = j \mid p)$$

  - M-step: Update frequency matrix estimate

$$\hat{p}_i(b) = \frac{\sum_k \sum_j W_{kj}\, 1(sequence\ k, position\ j + i - 1 = b)}{N}, \quad b = \texttt{A,C,G,T}$$

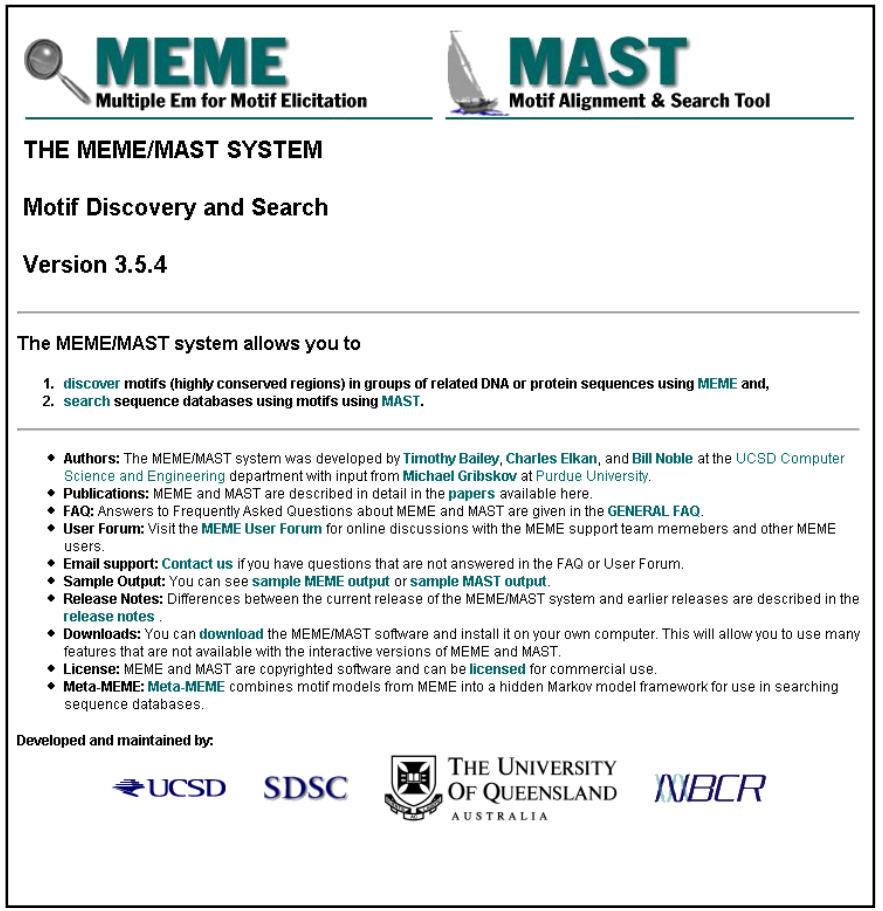
A Promoter Model



The same motif model in all promoters

MEME

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



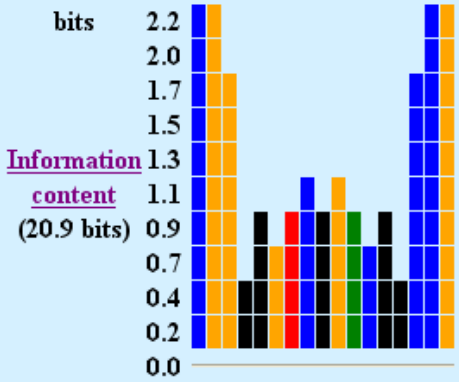
http://meme.sdsc.edu/meme/

MEME Output

PN **MOTIF 1** **width = 17** **sites = 14** **llr = 203** **E-value = 2.7e-025**

Simplified pos.-specific probability matrix

```
A : : : 5: 1615: : 1: 1: : :
C a: 1152: 7: 246549a:
G : a945642: 7: 2511: a
T : : : 1: 1: : 5161: 5: : :
```



Multilevel consensus sequence

```
CGGAGGACTGTCCTCCG
 GCCG  A  CGGC
```

NAME	START	P-VALUE	SITES		
YBR020W_74	93	1.11e-09	AGCCGCCGAG	CGGGCGACAGCCCTCCG	ACGGAAGACT
YBR020W_153	153	1.11e-09	AGCCGCCGAG	CGGGCGACAGCCCTCCG	ACGGAAGACT
YBR020W_154	136	1.11e-09	AGCCGCCGAG	CGGGCGACAGCCCTCCG	ACGGAAGACT
YBR019C_94	176	1.44e-09	AGTCTTCCGT	CGGAGGGCTGTCGCCCG	CTCGGCGGCT

Gibbs Motif Sampling

- Gibbs Motif Sampler
  - Bayesian model, prior distribution
- Algorithm (*Markov Chain Monte Carlo* (*MCMC*) and *Gibbs sampling*)

  Initialization: Randomly select motif start-positions in each sequence

  Iterations:

  Remove randomly selected sequence k'

  - Update frequency matrix
  - Randomly select a motif start-position j for k' proportional to:

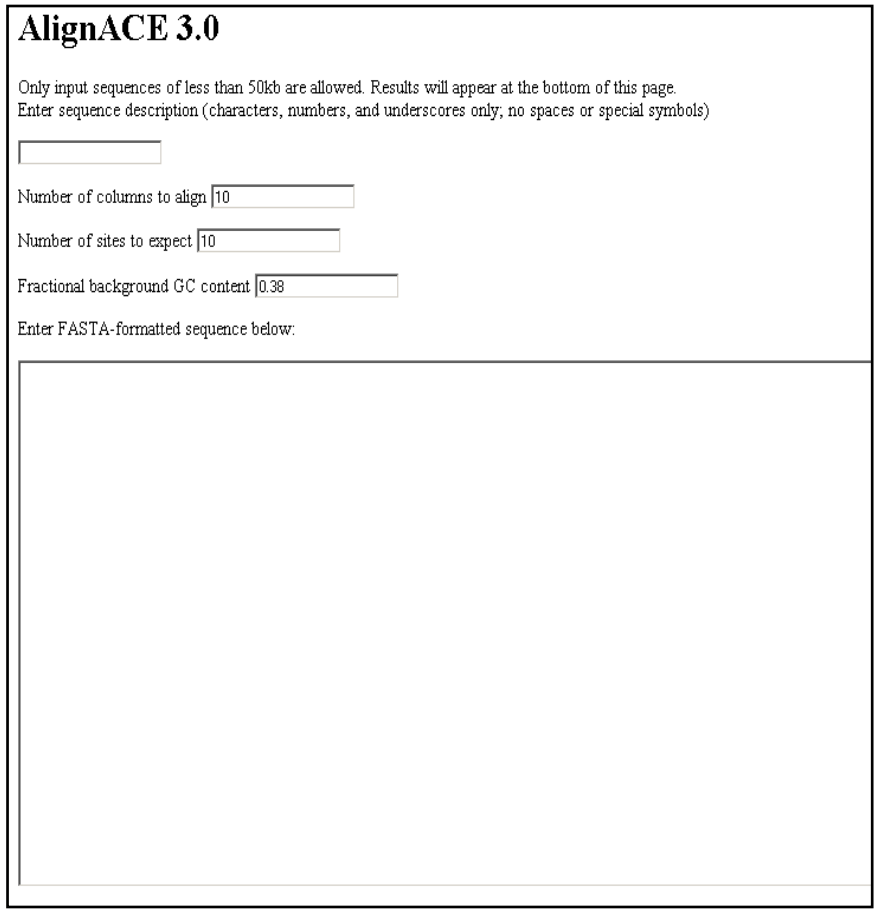
$$\frac{\text{Probability under motif model}}{\text{Probability under background model}} = \prod_i \frac{p_i(b^k_{j+i-1})}{p_0(b^k_{j+i-1})}$$

Gibbs Motif Sampler

http://bayesweb.wadsworth.org/gibbs/gibbs.html

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

Identifying TFBS

- Knowledge-based methods
  - Consensus sequences
  - Probabilistic Model
- Ab init methods
  - MEME
  - Gibbs sampling
- Evolution approach

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.

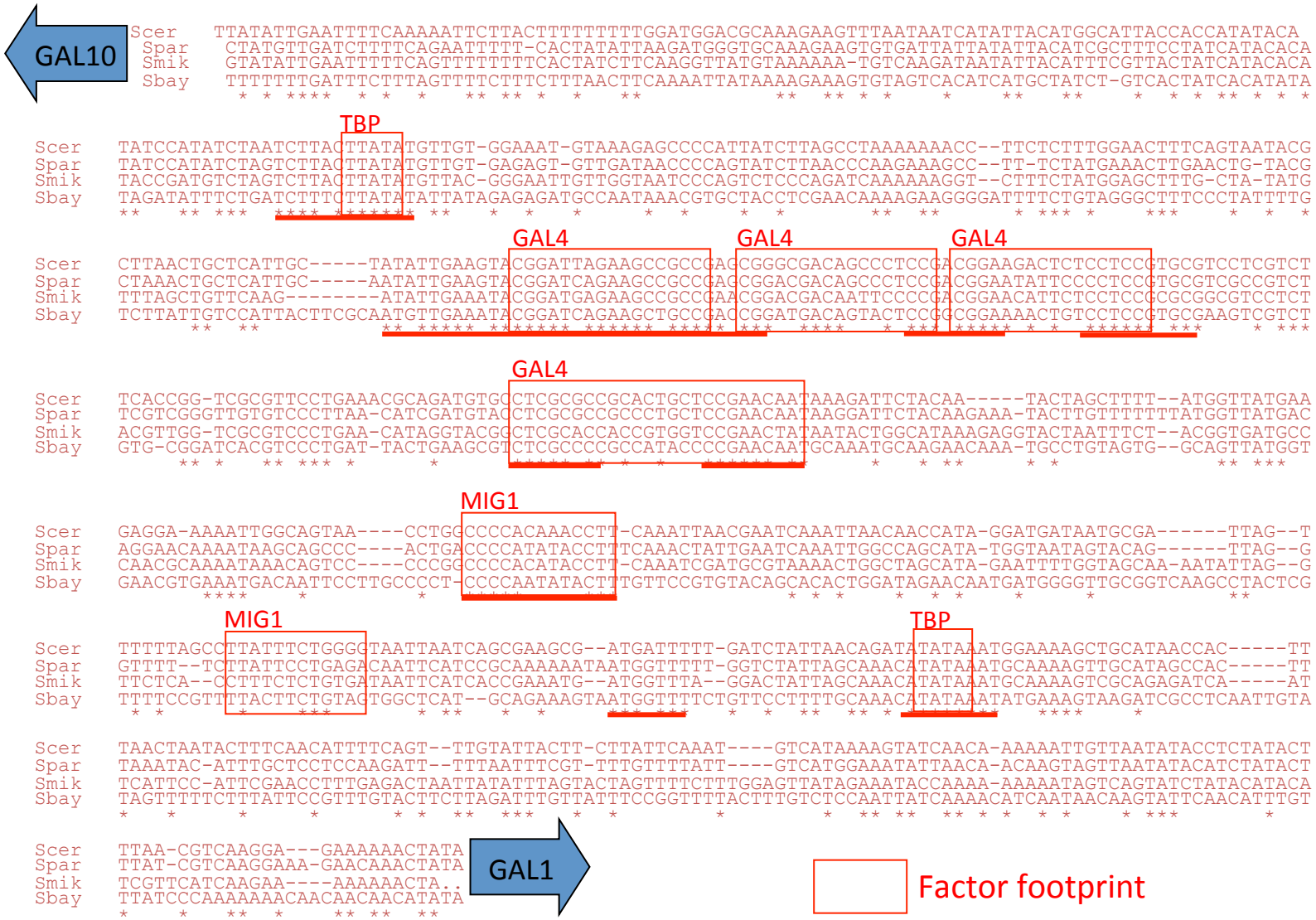
Extracting the complete functional information encoded in a genome—including the genic, regulatory and structural elements—is a central challenge in biological research. Ideally, one would be able to extract this information directly from the DNA sequence itself without recourse to extensive experimentation.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

Key questions

- How to discover TFs/Binding sites (motifs)?
- How to discover regulatory network pathways? (with expression profiles)

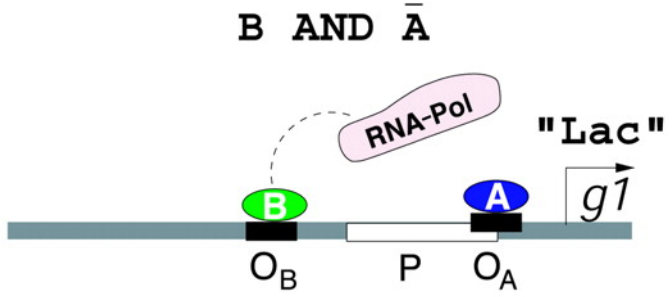
Boolean model

(a) Some possible gene responses (ON or OFF) according to the specific activation patterns of two TFs, A and B, as denoted by their cellular concentrations (high or low)

a

A	B	"Lac" *g1*	AND *g2*	OR *g3*	NAND *g4*	XOR *g5*	EQ *g6*
low	low	OFF	OFF	OFF	ON	OFF	ON
high	low	OFF	OFF	ON	ON	ON	OFF
low	high	ON	OFF	ON	ON	ON	OFF
high	high	OFF	ON	ON	OFF	OFF	ON

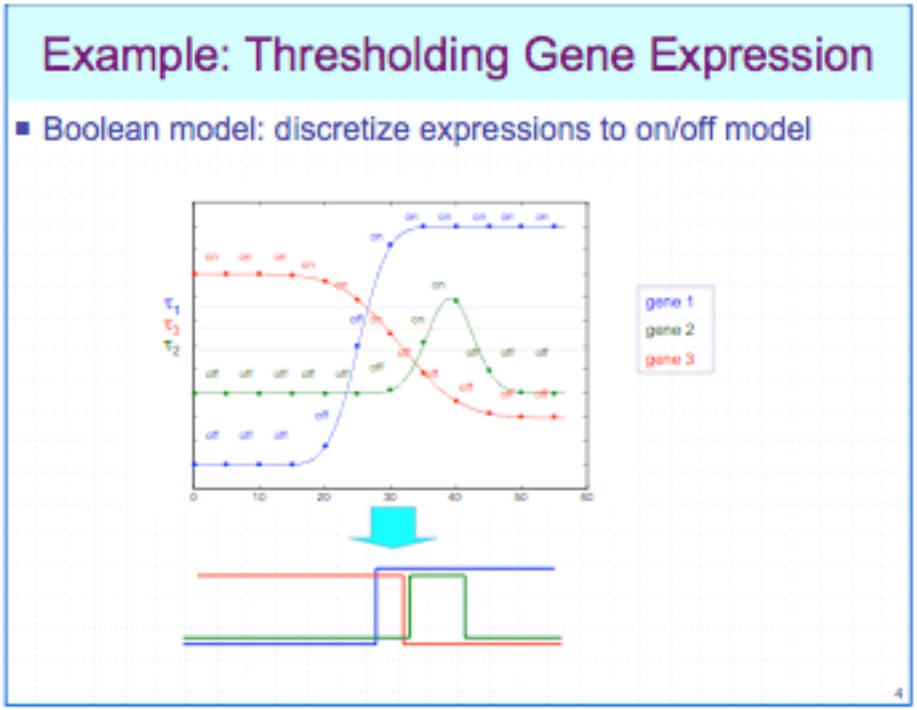
b



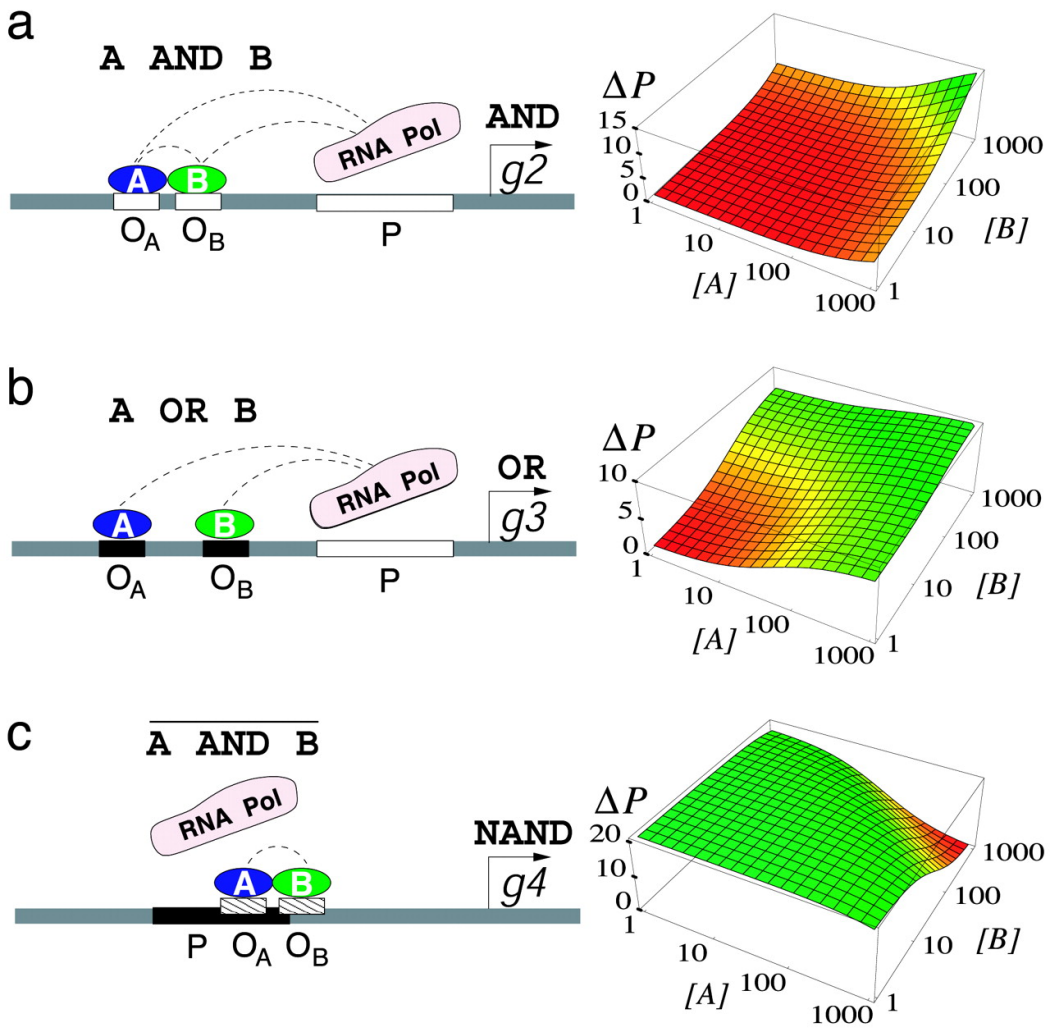
Buchler N E et al. PNAS 2003;100:5136-5141

Discretization

• Given expression profiles, values could be 1 or 0 under threshold.

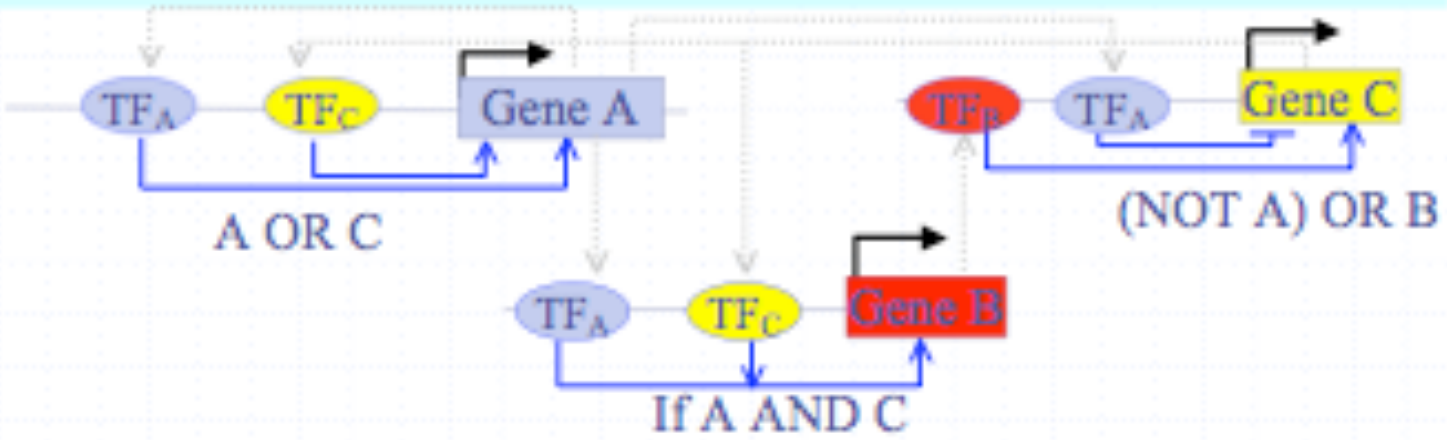


Cis-regulatory constructs and response characteristics of the AND (a), OR (b), and NAND (c) gates

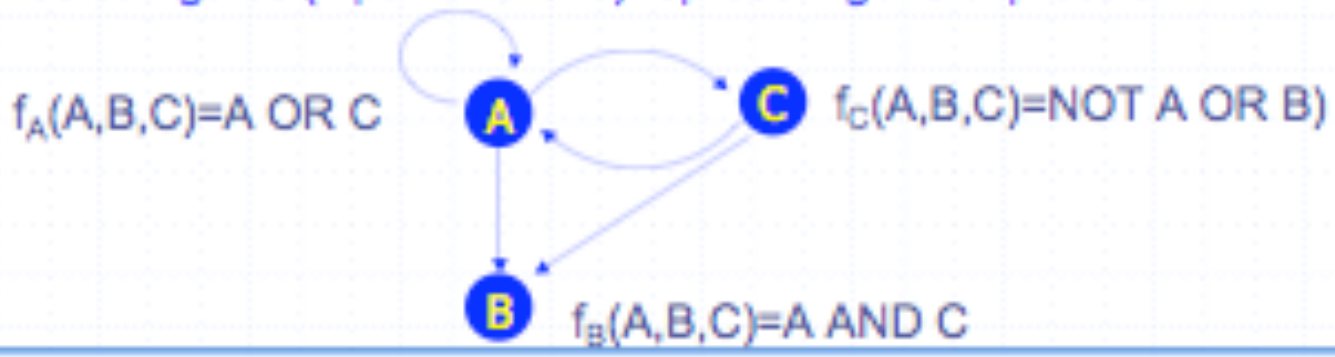


Buchler N E et al. PNAS 2003;100:5136-5141

Boolean Network Model



- A Boolean Network Model:
  - Nodes represent transcription factors
  - Edges represent regulatory input
  - Boolean gates (input functions) represent gene expression

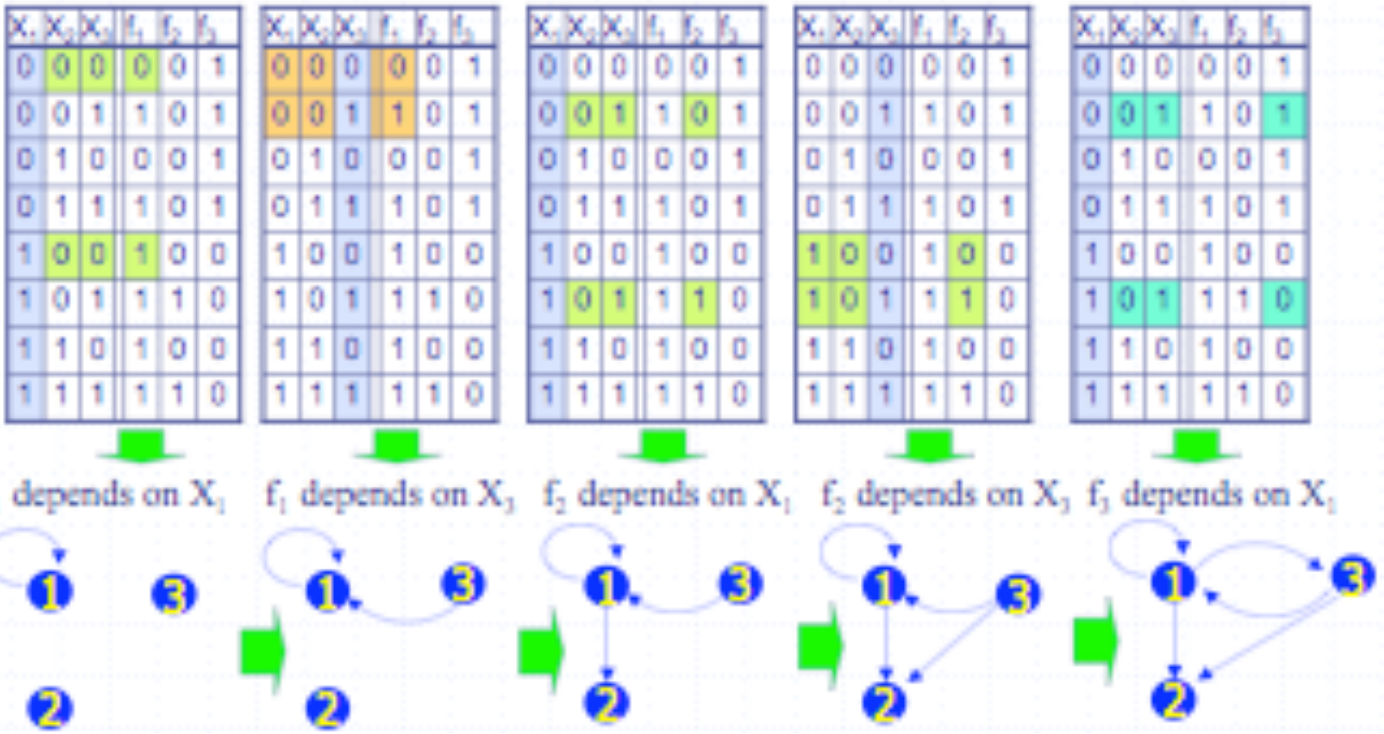


Algorithms

- Intuitive algorithm
- REVEAL algorithm
- Buchler-Hwa algorithm

An Intuitive Algorithm

- Repeat for all X_i and f_k :
  - Scan M to find a dependency of f_k on X_i; if found then add an X_i=>f_k edge to G
  - Else (no dependency found) then f_k is independent of X_i



How Good Are Boolean Models?

- Advantages
  - Provide good qualitative interpretation of regulation
  - Particularly important for switching behaviors
    - Such systems are "robust" than using exact expression values
  - Useful connection with evolutionary behaviors

- Disadvantages
  - Boolean abstraction is poor fit to real expression data
  - Cannot model important features:
    - Amplification of a signal; subtraction and addition of signals
    - Handling smoothly varying environmental parameter (e.g. temperature, nutrients)
    - Temporal performance behavior (e.g. cell cycle period)
    - Negative feedback control (Boolean model oscillates vs. stabilize)

Software

- dChip: gene expression and genome variation
- EXPANDER: sequence and expression
- Galaxy: sequence and expression
- CisGenome: sequence and expression
- WGCNA: co-expression network
- Cytoscape: TN, signaling network