

# Finding Regulatory Signals in Genomes

## The Computational Problem

**Non-homologous/homologous sequences**

**Known/unknown signal**

**1 common signal/complex signals/additional information**

**Combinations**

## Regulatory signals know from molecular biology

**Different Kinds of Signals**

**Promotors**

**Enhancers**

**Splicing Signals**

**$\alpha$ -globins in humans**

# Weight Matrices & Sequence Logos

## Set of signal sequences:

$f_{b,i}$  b's in position i,  $s(b)$  pseudo count.

$$\text{corrected probability: } p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \text{ nucleo}} s(b')}$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	0	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0
	<b>B</b>	<b>R</b>	<b>M</b>	<b>C</b>	<b>W</b>	<b>A</b>	<b>W</b>	<b>H</b>	<b>R</b>	<b>W</b>	<b>G</b>	<b>G</b>	<b>B</b>	<b>M</b>

## Position Frequency Matrix - PFM

## Consensus sequence:

## Position Weight Matrix - PWM

$$PWM : W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

A	-1.93	.79	.79	-1.93	.45	1.50	.79	.45	1.07	.79	.0	-1.93	-1.93	.79
C	.45	-1.93	.79	1.68	-1.93	-1.93	-1.93	.45	-1.93	-1.93	-1.93	-1.93	.0	.79
G	.0	.45	-1.93	-1.93	-1.93	-1.93	-1.93	.66	-1.93	1.3	1.68	1.07	-1.93	
T	.15	.66	-1.93	-1.93	1.07	.66	.79	.0	.79	-1.93	-1.93	-1.93	.66	-1.93
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

## Score for New Sequence $S = \sum_{i=1}^w W_{b,i}$

<b>T</b>	<b>T</b>	<b>G</b>	<b>C</b>	<b>A</b>	<b>T</b>	<b>A</b>	<b>A</b>	<b>G</b>	<b>T</b>	<b>A</b>	<b>G</b>	<b>T</b>	<b>C</b>
.45	-.66	.79	1.66	.45	-.66	.79	.45	-.66	.79	.0	1.68	-.66	.79

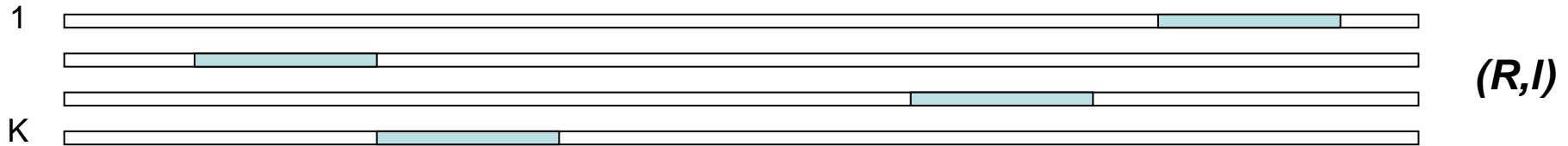
## Sequence Logo & Information content

$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i}$$



# Motifs in Biological Sequences

- 1990 Lawrence & Reilly "An Expectation Maximisation (EM) Algorithm for the identification and Characterization of Common Sites in Unaligned Biopolymer Sequences Proteins 7.41-51.  
 1992 Cardon and Stormo Expectation Maximisation Algorithm for Identifying Protein-binding sites with variable lengths from Unaligned DNA Fragments L.Mol.Biol. 223.159-170  
 1993 Lawrence... Liu "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment" Science 262, 208-214.



$\Theta = (\theta_{1,A}, \dots, \theta_{w,T})$  probability of different bases in the window

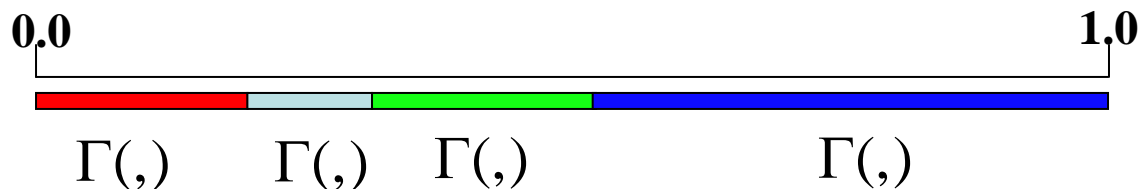
$A = (a_1, \dots, a_K)$  – positions of the windows

$\theta_0 = (\theta_A, \dots, \theta_T)$  – background frequencies of nucleotides.

$$p(R | \theta_0, \Theta, A) = \theta_0^{h(R_{\{A\}^c})} \prod_{j=1}^w \theta_j^{h(R_{A+j-1})} = \theta_0^{h(R)} \prod_{j=1}^w \left( \frac{\theta_j}{\theta_0} \right)^{h(R_{A+j-1})}$$

**Priors** A has uniform prior

$\Theta_j$  has Dirichlet( $N_0 \alpha$ ) prior –  $\alpha$  base frequency in genome.  $N_0$  is pseudocounts



# The Gibbs Sampler

$\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)})$  for iteration  $t$ . At iteration  $t + 1$

For  $i=1, \dots, d$ : Draw  $x_i^{(t+1)}$  from conditional distribution  $\pi(\cdot | \mathbf{x}_{[-i]}^{(t)})$  and leave remaining components unchanged, i.e.  $\mathbf{x}_{[-i]}^{(t+1)} = \mathbf{x}_{[-i]}^{(t)}$

Both random & systematic scan algorithms leaves the true distribution invariant.

$$\pi(x_i^{t+1} | \mathbf{x}_{[-i]}^t) \times \pi(\mathbf{x}_{[-i]}^t) = \pi(\mathbf{x}_{[-i]}^t, x_i^{t+1})$$

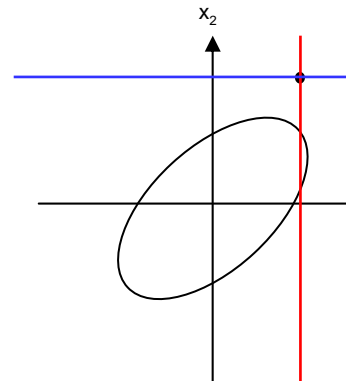
## An example:

Target Distribution is  $x = (x_1, x_2)$  is  $N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right\}$  distributed.

The conditional distributions are then:  $x_2^{t+1} | x_1^{t+1} \sim N\{\rho x_1^{t+1}, (1 - \rho)^2\}$ ,  
 $x_1^{t+1} | x_2^{t+1} \sim N\{\rho x_2^{t+1}, (1 - \rho)^2\}$ ,

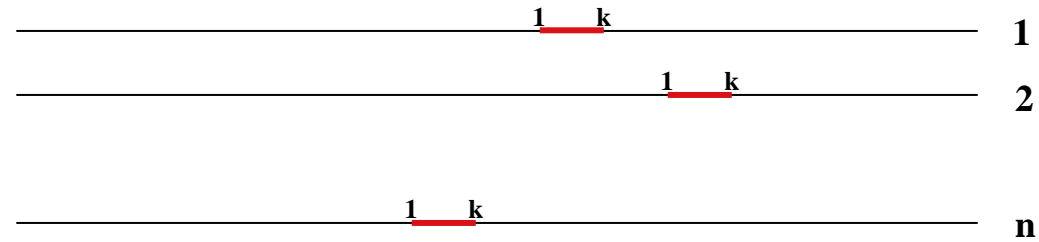
The approximating distribution after  $t$  steps of a systematic GS will be:

$$\begin{pmatrix} x_1^t \\ x_2^t \end{pmatrix} \sim N\left\{\begin{pmatrix} \rho^{2t-1} x_2^0 \\ \rho^{2t} x_2^0 \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4t-2} & \rho - \rho^{4t-1} \\ \rho - \rho^{4t-1} & 1 - \rho^{4t} \end{pmatrix}\right\}$$



# The Gibbs sampler

**Objective**: Find conserved segment of length  $k$  in  $n$  unrelated sequences

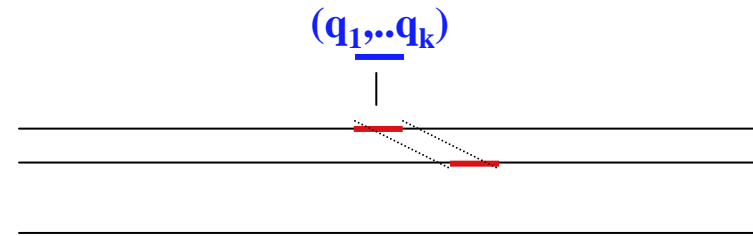


## **Gibbs iteration**:

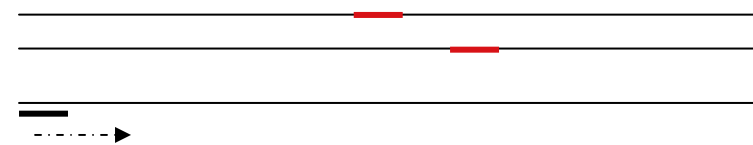
Remove one at random -  $s_j$



Form profile of remaining  $n-1$



Let  $p_i$  be the probability with which  $s_j[i..i+k-1]$  fits profile. Including pseudo-counts. Choose to start replacement at  $i$  with probability proportional to  $p_i$



# The Gibbs sampler: example

## Observed pattern in aligned sequences

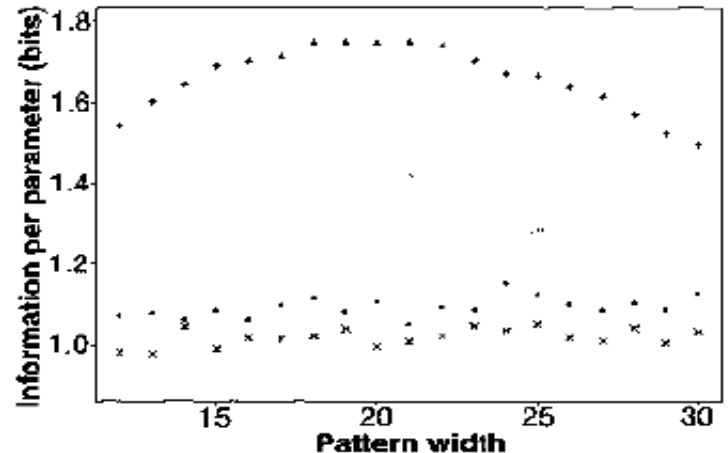
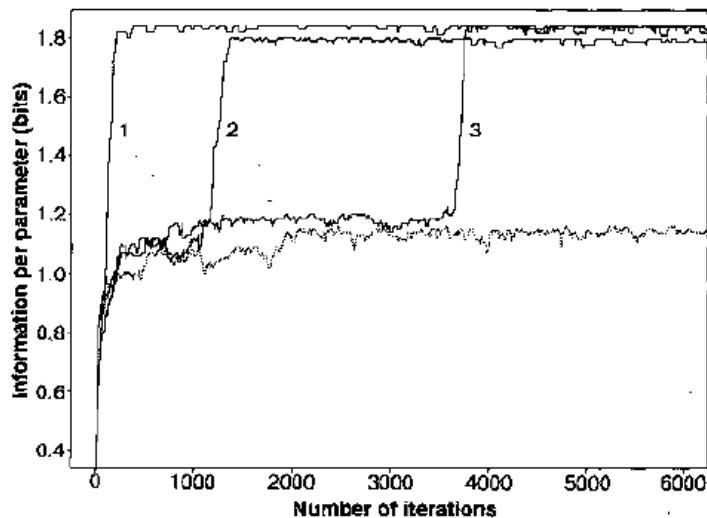
Sigma-37	223	IIDLTYIQNK	SQKETSGLIGISQMHVSR	LQRKAVEKLR	240	A25944
SpoIIIC	94	REGDLEKKEK	TQREIAKELGISRSYVSR	IEKRALMKNF	111	A28627
Nahr	22	VVFNGLVDR	RVSTIAENLGTQPAVSN	ALKRLPTSLQ	39	A32837
Antennapedia	126	EHFNRYLTR	RRIEIAHALCLTERQIKI	WFQNRMKWK	343	A21450
NrcC (Brady.)	449	LTAALAAATR	NQIRAAQLLGNRRITLAK	KIRDLLEIQVY	466	B24699
Dica	22	IVYRKNLKH	TORSIAKALKTISHVSVS	WERGDSEPTG	39	B24328 (BYEEDA)
MerD	5	MMAY	TVSRLLADAGVSVHVRD	YLRGLLRLEV	22	C29010
Fis	73	LMMVQYTRG	NQIRAAALPMGIRNGTLRK	KLKRYGMN	90	A32142 (DNECF5)
MAT a1	99	FRRKGLNSK	EREVAKKCGITPLQVRV	WFINRMRSK	116	A90983 (JEHV1)
Lambda cII	25	SALLNRIAML	GTERIAEAVGVDRSQISR	WKNWIPKFS	42	A03579 (QCBP2L)
Crp (CAP)	169	THFDQMQIKI	TRGEIGIVGCSRETVSR	LHMLEDQNL	186	A03553 (QBECC)
Lambda Cro	15	ITLKVYANRF	GQRTIARDLGVYQSAINK	AIHAGRKIEL	32	A03577 (RCBPL)
P22 Cro	12	YKRDVIDHFG	TQRVAVKALGISDAVASQ	WKEVTPKDA	29	A25867 (RQBPL2)
AraC	196	ISDHLADSNF	DIAAVAGHVCLSPRLSH	LFRQQLGISV	213	A03554 (RGECA)
Fur	196	FSPRFRLTM	TRGDIQNYLGLTVETISR	LLGRFORSKM	213	A03552 (RGECT)
HcpR	252	ARWLEDEWKS	TLQELADRYGVSARVRQ	LEHNAKKLR	269	A00700 (RGECK)
NrcC (M.a.)	444	LTTALRHTQG	HKQEAARLLGWRHTLTK	KLKELGME	461	A03564 (RGRBCP)
CyR	11	MKAKEQETA	TKMDVALKAVSTATVSR	ALMNPDRVSO	28	A24963 (RPECTT)
DeoR	23	LQELRSDRL	HLKDAALLGVSQEMTIR	DLNNSAPVW	40	A24076 (RPECD0)
GalR	3	MA	TKRDVAVLAGVSVATVSR	VINNSPKASE	20	A03559 (RPEC9)
LacI	5	MKPV	TLYDVAEYAGVSYQTVSR	VVQASHVSA	22	A03558 (RPECL)
TetR	26	LLNEVGEGL	TKRLAQKLGVEQPTLYR	HVNKHALLD	43	A03576 (RPECTN)
TspR	67	IVBEELRDEM	SQRELNELGAGIATITR	QSNLKAAPV	94	A03560 (RPECM)
NLCA	495	LTAALERAGW	VQAARLLGHTPEQVAY	RIGIMDITMP	512	S02513
SpoIIG	205	RPLGVGEEK	TQKDVADMMGISQYSIR	LEKRIERLR	222	S07337
Fin	160	QAGRLLAAGT	PRKRVATIDVGVSTLYK	TFPAGDK	177	S07950
PurR	3	MA	TKEDVARRANVSTTVSH	VINKTPEVAE	20	S08477
EagR	3	MA	TKDIAIEAGVSLATVSR	VLNDDPTLAW	20	S09205
LocA	27	DHISQGMDF	TRAEIAQRLGFRSNAAB	RIHLKALRHQ	44	S11945
P22 cI	25	SSILNRIRIR	GQRKVADALGINESQISR	WEGDPIPKWG	42	B25867 (Z1BPC2)

## Pattern Probability Model

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Arg	94	222	265	137	9	9	137	137	9	9	52	222	94	94	9	265	606	
Lys	9	133	442	380	9	71	180	194	9	133	9	9	71	9	9	71	256	
Glu	53	9	96	401	9	9	140	140	9	9	53	140	140	9	9	9	53	
Asp	67	9	9	473	9	9	299	125	9	67	9	67	67	9	9	9	67	
Gln	9	600	224	9	9	9	224	9	9	9	9	9	9	278	63	278	9	170
His	240	9	9	9	9	9	125	125	9	9	9	125	125	125	9	9	240	
Asn	168	9	9	9	9	9	168	89	9	49	9	248	9	168	89	9	89	89
Ser	117	9	117	117	9	9	9	9	9	9	819	63	387	63	9	819	9	
Gly	151	9	56	9	9	151	9	9	9	1141	9	151	9	56	9	56	9	
Ala	9	9	112	43	181	901	43	181	215	9	43	9	43	181	112	43	78	9
Thr	915	130	130	9	251	9	9	9	9	9	311	130	70	855	9	130	9	
Pro	76	9	9	9	9	9	9	9	9	9	9	210	210	9	9	9	9	
Cys	9	9	9	9	9	9	9	9	295	581	295	9	9	9	9	9	9	
Val	58	107	9	9	500	9	9	156	9	598	9	205	58	9	746	9	58	
Leu	9	121	9	9	149	9	93	149	458	9	149	9	37	37	9	177	9	
Ile	9	166	114	61	323	9	114	166	9	9	427	9	61	9	61	427	9	61
Met	9	104	9	9	9	9	9	198	198	9	104	9	9	198	9	9	9	
Tyr	9	9	136	9	9	9	9	252	262	9	9	136	136	9	262	9	262	136
Phe	9	9	9	9	9	9	9	9	9	9	108	9	9	9	9	9	9	9
Trp	9	9	9	9	9	9	9	9	9	9	366	9	9	9	9	9	9	366

- 3 independent runs
- a run on sequences without signal

- Upper Points: original sequences
- Middle: original date minus pattern
- Lower: Shuffled sequences



# Natural Extensions to Basic Model I

## Multiple Pattern Occurances in the same sequences:

Liu, J. "The collapsed Gibbs sampler with applications to a gene regulation problem," *Journal of the American Statistical Association* **89** 958-966.

**Prior:** any position  $i$  has a small probability  $p$  to start a binding site:

$$A=(a_1, \dots, a_k) \quad P(A) \approx p_0^k (1-p_0)^{N-k} \quad (\text{with nonoverlapping constraints})$$



## Composite Patterns:

BioOptimizer: the Bayesian Scoring Function Approach to Motif Discovery *Bioinformatics*



# Natural Extensions to Basic Model II

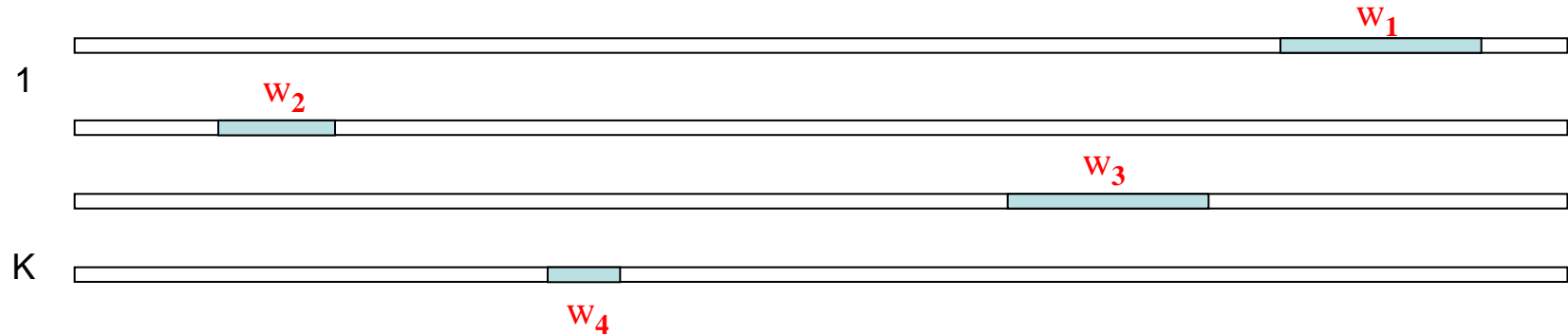
## Correlated in Nucleotide Occurrence in Motif:

Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 6, 909-916.



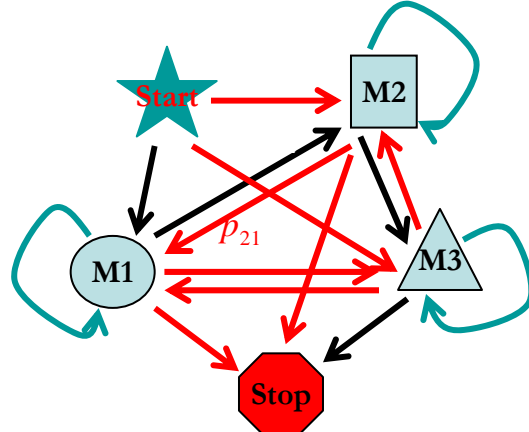
## Insertion-Deletion

BALSA: Bayesian algorithm for local sequence alignment *Nucl. Acids Res.*, 30 1268-77.



## Regulatory Modules:

De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Nat'l Acad Sci USA*, 102, 7079-84



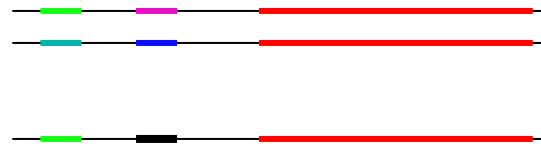
# Combining Signals and other Data

## Expression and Motif Regression:

Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci. 100.3339-44

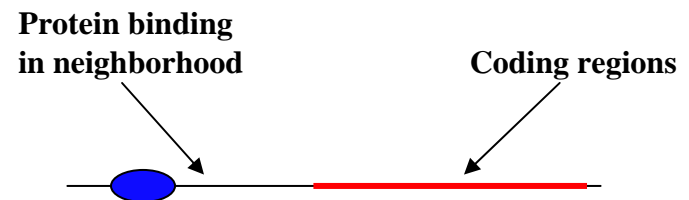


1. Rank genes by  $E = \log_2(\text{expression fold change})$
2. Find “many” (hundreds) candidate motifs
3. For each motif pattern  $m$ , compute the vector  $S_m$  of matching scores for genes with the pattern
4. Regress  $E$  on  $S_m$  
$$Y_g = \alpha + \beta_m S_{mg} + \epsilon_g$$



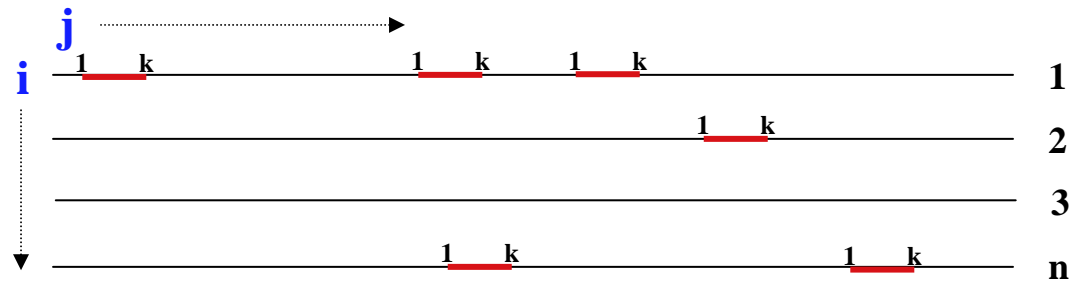
## ChIP-on-chip - 1-2 kb information on protein/DNA interaction:

An Algorithm for Finding Protein-DNA Interaction Sites with Applications to Chromatin Immunoprecipitation Microarray Experiments *Nature Biotechnology*, 20, 835-39



# MEME- Multiple EM for Motif Elicitation

$Z_{i,j} = 1$  if a motif starts at  $j$ 'th position in  $i$ 'th sequence, otherwise 0.



Motif nucleotide distribution:  $M[p,q]$ , where  $p$  - position,  $q$ -nucleotide.

Background distribution  $B[q]$ ,  $\lambda$  is probability that a  $Z_{i,j} = 1$

Find  $M, B, \lambda, Z$  that maximize  $\Pr(X, Z | M, B, \lambda)$

Expectation Maximization to find a local maximum

Iteration  $t$ :

Expectation-step:  $Z^{(t)} = E(Z | X, (M, B, \lambda)^{(t)})$

Maximization-step: Find  $(M, B, \lambda)^{(t+1)}$  that maximizes  $\Pr(X, Z^{(t)} | (M, B, \lambda)^{(t+1)})$

# Phylogenetic Footprinting (homologous detection)

Term originated in 1988 in Tagle et al. **Blanchette et al.:** For unaligned sequences related by phylogenetic tree, find all segments of length **k** with a history costing less than **d**. Motif loss an option.

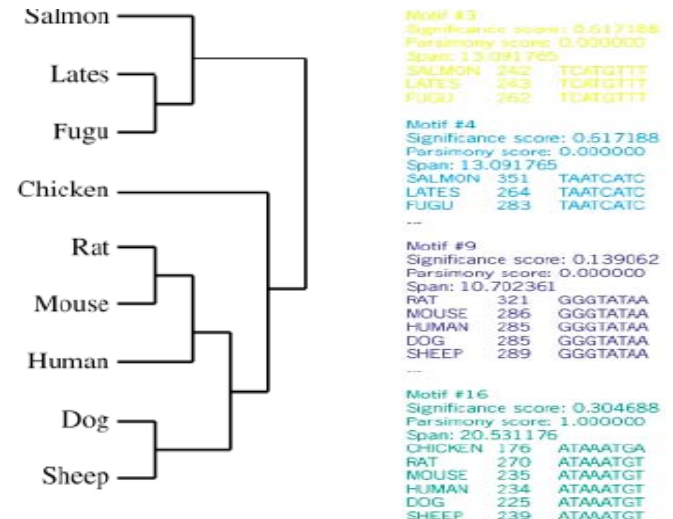
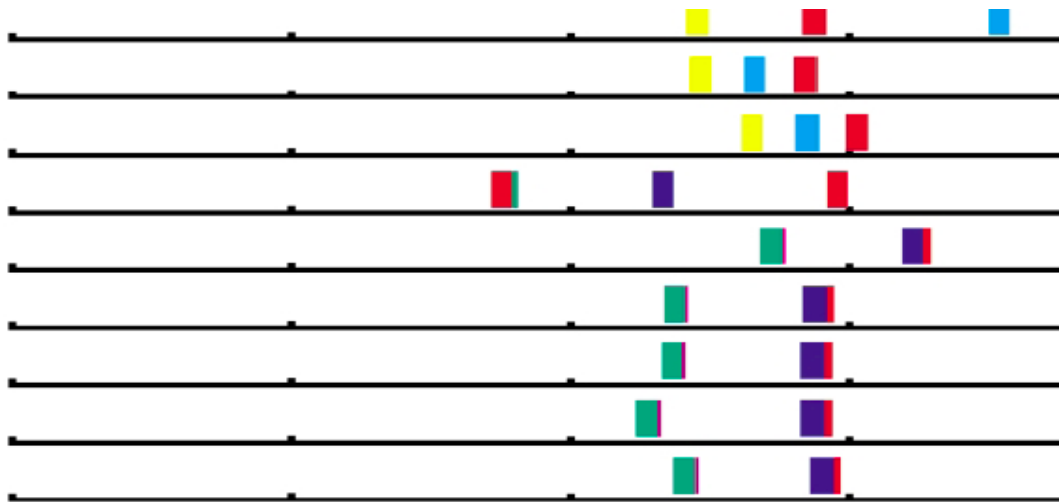
$$D_i^{begin} = \min\{ D_{i,\Delta}^{begin} + d(i,\Delta) \}$$

$$D_i^{signal,1} = \min\{ D_{i,\Delta}^{begin} + d(i,\Delta) \}$$

$$D_i^{signal,j} = \min\{ D_{i,\Delta}^{signal,j-1} + d(i,\Delta) \}$$

...

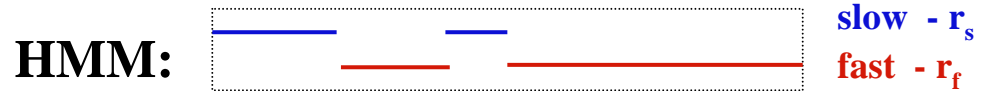
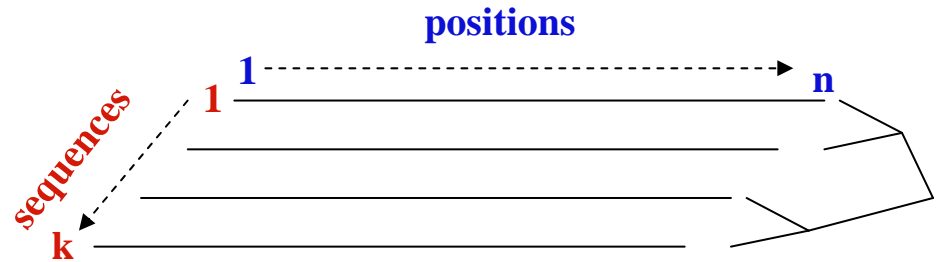
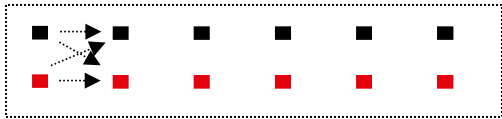
$$D_i^{end} = \min\{ D_{i,\Delta}^{end} + d(i,\Delta) \}$$



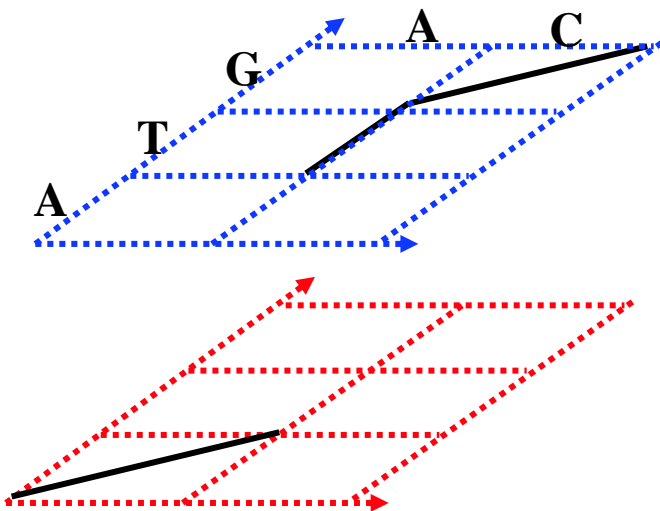
# The Basics of Footprinting I

- Many aligned sequences related by a known phylogeny:

HMM:



- Two un-aligned sequences:



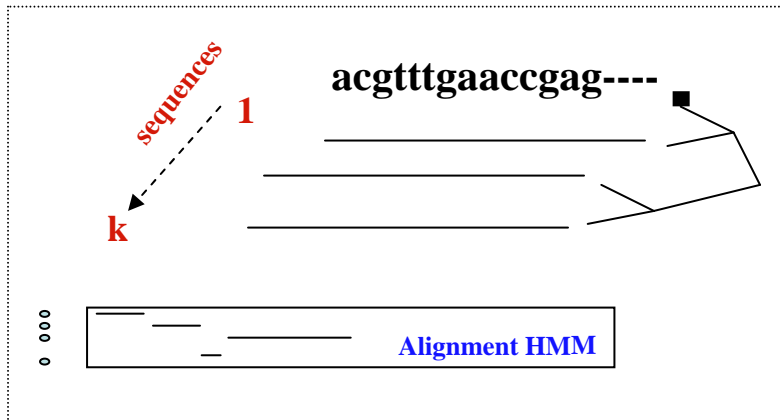
**ATG**

**A-C**

# Statistical Alignment and Footprinting.

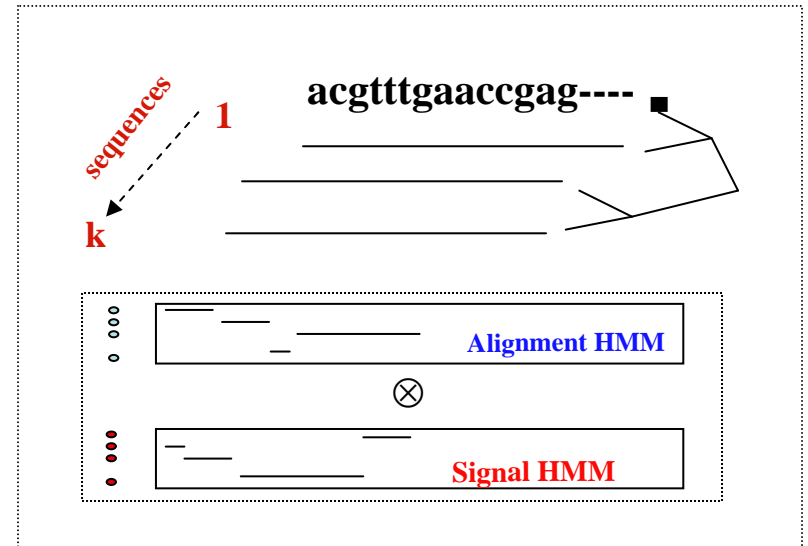
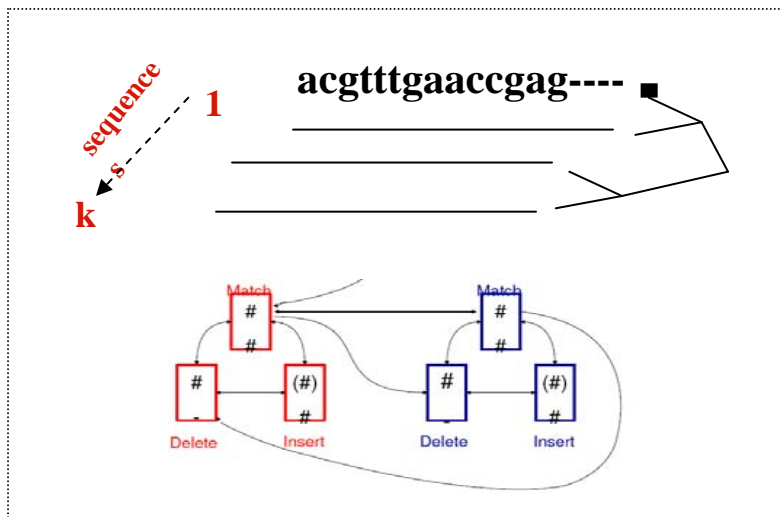
## • Many un-aligned sequences related by a known phylogeny:

- Conceptually simple, computationally hard
- Dependent on a single alignment/no measure of uncertainty

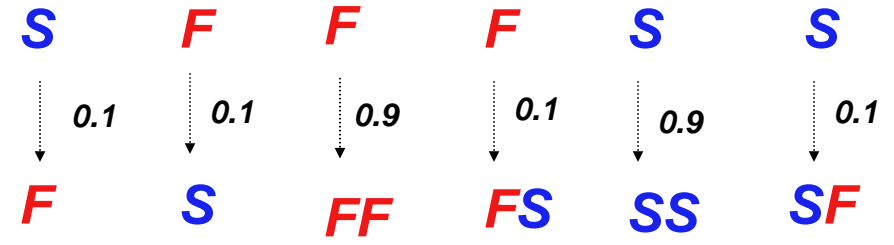
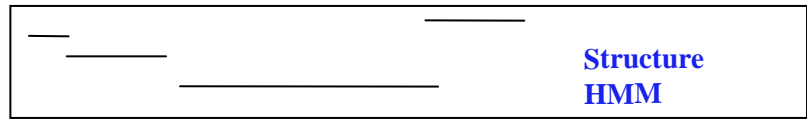


## Solution:

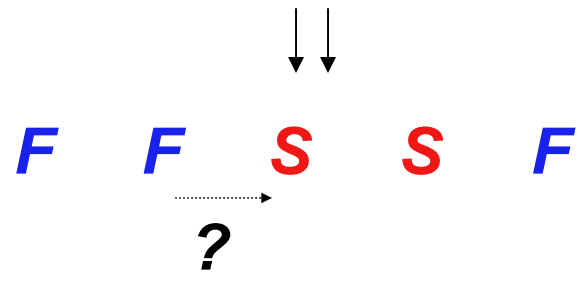
Cartesian Product of HMMs



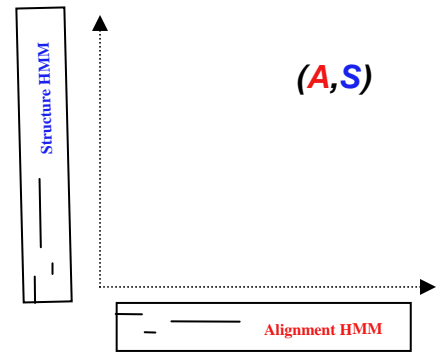
# “Structure” does not stem from an evolutionary model



• *The equilibrium annotation does not follow a Markov Chain:*



• *Each alignment in from the Alignment HMM is annotated by the Structure HMM:*

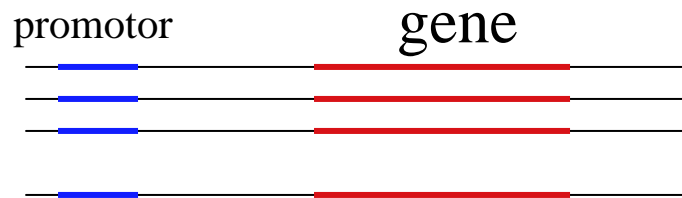


• *No ideal way of simulating:*

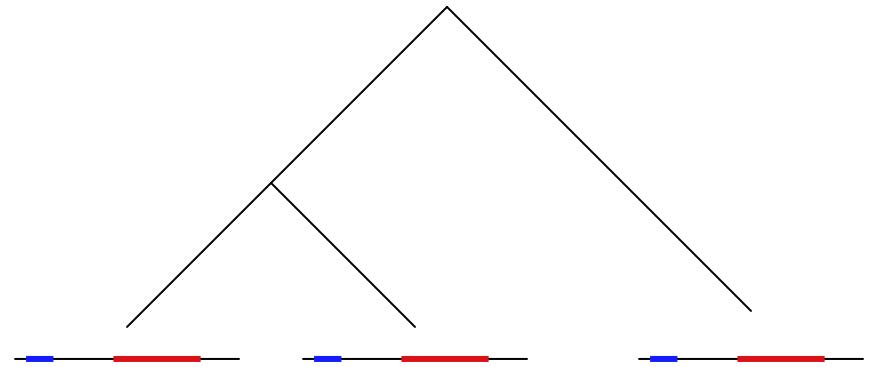
using the **HMM at the alignment** will give other distributions on the leaves  
 using the **HMM at the root** will give other distributions on the leaves

# (Homologous + Non-homologous) detection

## Unrelated genes - similar expression

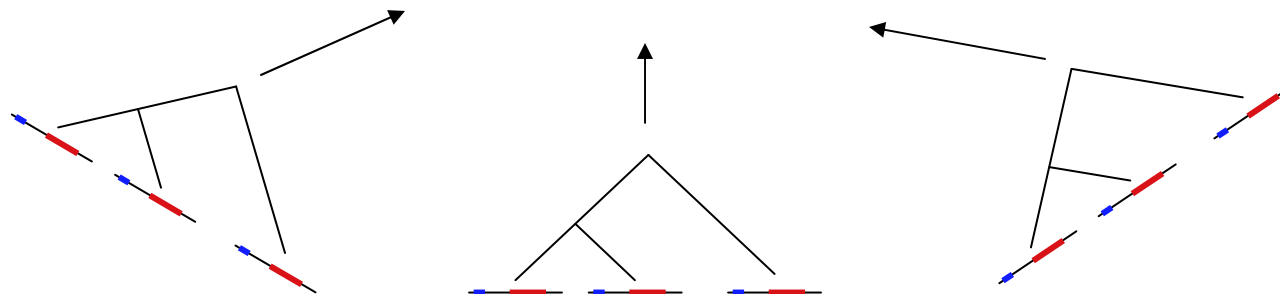


## Related genes - similar expression



## Combine above approaches: Mixed genes - similar expression

Combine "profiles"

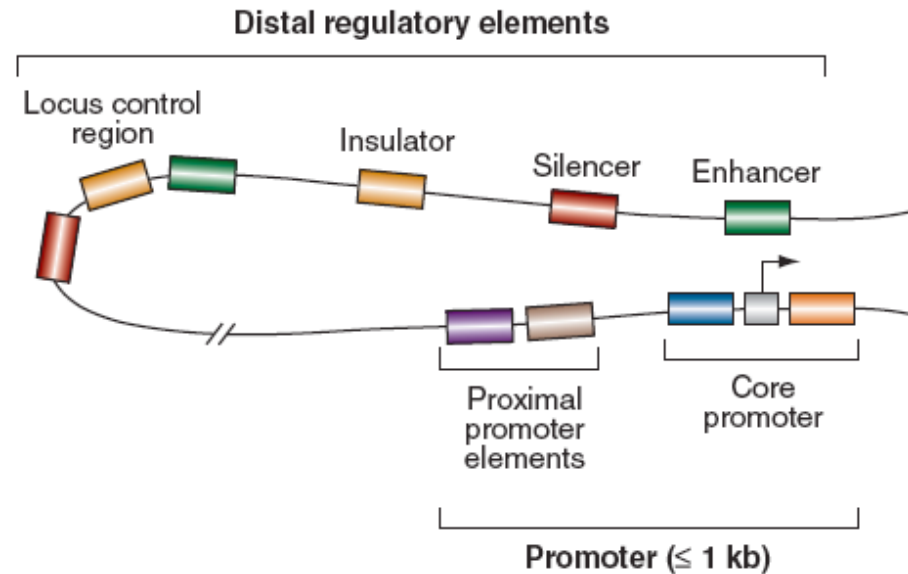


# Regulatory Signals in Humans

Transcription in Eukaryotes is done by RNA Polymerase II.

1850 DNA-binding proteins in the human genome.

- **Transcription Start Site - TSS**
- **Core Promoter - within 100 bp of TSS**
- **Proximal Promoter Elements - 1kb TSS**
- **Locus Control Region - LCR**
- **Insulator**
- **Silencer**
- **Enhancer**



# Core Promoter Elements

**TATA - box**

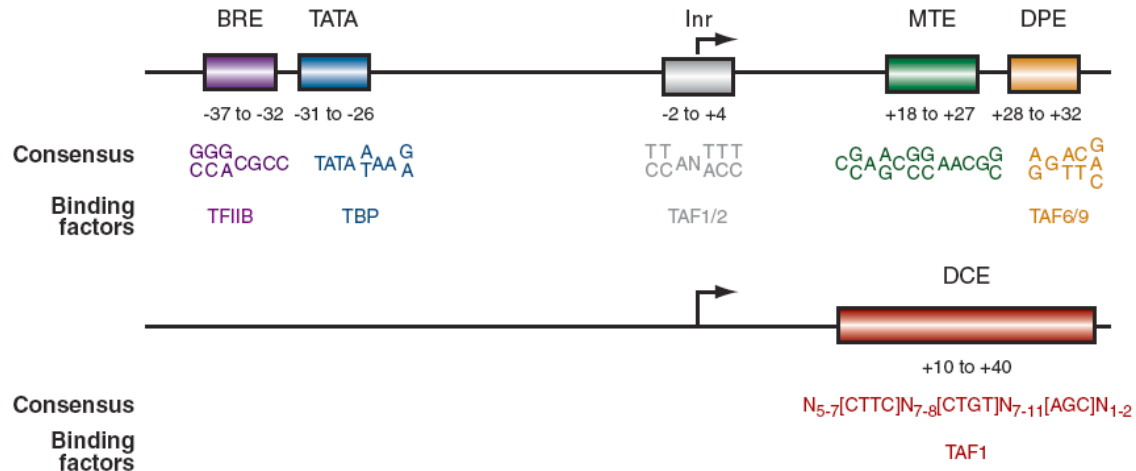
**Inhibitor element (Inr)**

**Downstream Promoter Element (DPE)**

**Downstream Core Element (DCE)**

**TFIIB-recognition element (BRE)**

**Motif Ten Element (MTE)**



## Examples of Disease Mutations:

Core promoter	$\beta$ -thalassemia	TATA-box	$\beta$ -globin
Enhancer	X-linked deafness	900kb deletion	POU3P4
Silencer	Asthma	509bp mut TSS	TFG-b
Activator	Prostate Cancer		ATBF1
Coactivator	Parkinson disease		DJ-1
Chromatin	Cancer		BRG1/BRM

# $\alpha$ -globins

**Multispecies Conserved Sequences - MCSs**

**Analyzed 238kb in 22 species**

**Found 24 MCSs**

**Programs use**

**GUMBY - VISTA - MULTIPIPMAKER**

**MULTILAGAN - CLUSTALW - DIALIGN**

**TRANSFAC 6.0 - TRES -**

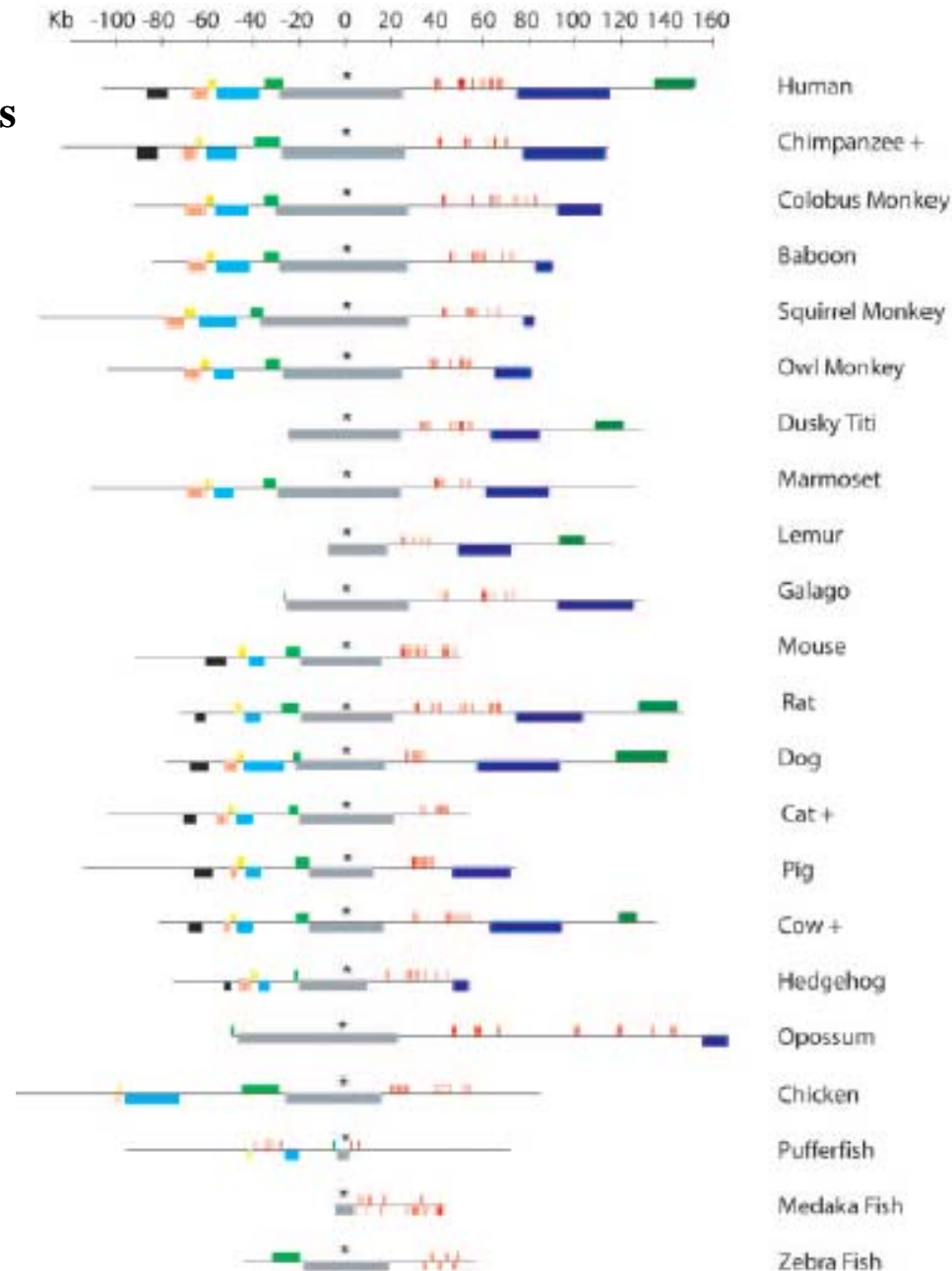
**Experimental Knowledge of the region**

**Hypersensitive sites (DHSs)**

**DNA Methylation**

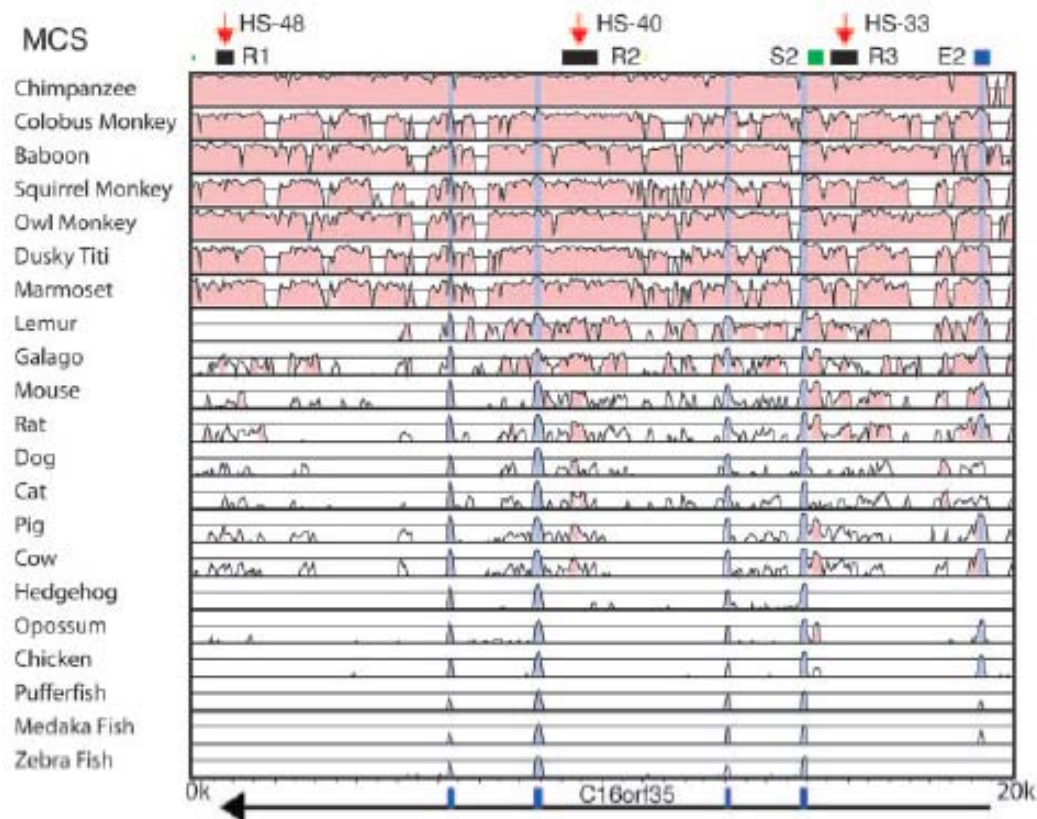
**Region lies in CG rich, gene rich region  
close to the telomeres.**

**It is not easy to align CG-islands.**



# Promoters in $\alpha$ -globins

- 94.273-114.273 vista illus.
- 5 MCSs
- Divergence relative to human



1. Promoters MCSs - 11
2. Regulatory MCSs - 4
3. Intronic MCSs - 2
4. Exonic MCSs - 4
5. Unknown - 3

	MCS-P1	MCS-P2	MCS-P3	MCS-P4	MCS-P5	MCS-P6	MCS-P7	MCS-P8	MCS-P9	MCS-P10	MCS-P11	MCS-R1	MCS-R2	MCS-R3	MCS-R4	MCS-S1	MCS-S2	MCS-E1	MCS-E2	MCS-E3	MCS-E4	MCS-U1	MCS-U2	MCS-U3
Length (bp)	103	155	-	151	292	-	367	-	129	63	91	68	256	268	181	158	158	135	251	232	665	222	42	103
HS Human	-80	-80	-77	-77	-77	-14	-	-	-	-	-	-48	-40	-33	-10	-	-	-	-	-	-	-	-	-
HS Mouse	-58.4	-58.4	-50	-50	-50	-9.7	-	-	-	-	-	-31	-26	-21	-8	-	-	-	-	-	-	-	-	-
EST	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-
Human	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Chimp	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
C. Monkey	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Baboon	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
S. Monkey	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
O. Monkey	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Dusky Titi	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Marmoset	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Lemur	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Galago	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Mouse	+	+	N	+	+	N	+	+	+	+	T	+	+	+	+	+	+	+	+	N	T	+	+	+
Rat	+	+	N	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	N	+	+	+	+
Dog	NS	NS	+	+	+	NS	+	+	+	NS	+	+	+	+	+	+	+	+	+	+	NS	NS	+	+
Cat	+	+	+	+	+	+	+	+	+	NS	NS	+	+	+	+	+	+	+	+	NS	NS	+	+	+
Pig	+	+	+	+	+	+	+	+	+	NS	NS	+	+	+	+	+	+	+	+	N	+	+	+	+
Cow	+	+	+	+	+	+	+	+	+	NS	NS	+	+	+	+	+	+	+	+	N	+	+	+	+
Hedgehog	+	+	+	+	+	+	+	+	+	NS	NS	+	+	+	+	+	+	+	+	N	+	+	+	+
Opossum	N	N	N	N	N	N	N	N	N	(-)	NS	N	+	N	N	N	N	N	N	NS	NS	NS	N	N
Chicken	N	N	N	N	N	N	N	N	N	(-)	T	N	(+)	N	N	N	N	N	N	NS	T	N	N	N
Pufferfish	N	N	N	N	N	N	N	N	N	(-)	T	N	(+)	N	N	N	N	N	N	NS	T	N	N	N
Medaka Fish	NS	NS	NS	NS	NS	NS	NS	NS	NS	(-)	NS	N	(+)	N	N	NS	NS	NS	NS	NS	NS	NS	NS	NS
Zebra Fish	NS	NS	N	N	N	N	N	N	N	(-)	NS	N	(+)	N	N	NS	NS	NS	NS	NS	NS	NS	NS	NS
% HMR	74	74	-	68	52	-	64	-	48	48	69	57	69 (80)	58	46	65	65	82	86	-	92	69	90	63
% all	41	55	-	36	21	-	20*	-	22	27	44	30	13	25	19	16	40	24	24	-	59	40	71	28
MCS 95	+	+					+				+		+	+		+		+	+	+	+	+	+	1
MCS 94	+	+					+				+		+	+		+		+	+	+	+	+	+	16
MCS 93	+	+	+				+	+	+	+	+		+	+		+		+	+	+	+	+	+	37
MCS 90	+	+	+	+			+	+	+	+	+		+	+		+		+	+	+	+	+	+	137

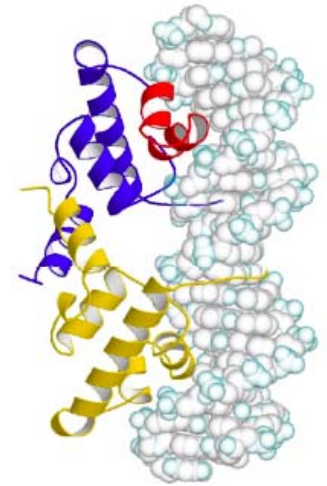
Source: Hughes et al. (2005) Annotation of cis-regulatory elements by identification

# Regulatory Protein-DNA Complexes

## 1. Cro and Repressor family

11mb*	3,4	Repressor	Phage $\lambda$	1.8	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
11li	A,B	Repressor mutant	Phage $\lambda$	2.1	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
1per	L,R	Repressor	Phage 434	2.5	AAGTACAGTTTTTCTTG-TATTATA--CAAGAAAACTGTACT
1rpe	L,R	Repressor	Phage 434	2.5	-TATACAATGTATCTTG-TTTGACAAACAAGATACATTGTAT-
2or1	L,R	Repressor	Phage 434	2.5	AAGTACAAACTTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
3cro	L,R	Cro	Phage 434	2.5	AAGTACAAACTTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
6cro	A	Cro	Phage $\lambda$	3.0	AAGTACAAACTTTTCTTG-TATTATA-CAAGAAAGTTTGTACT
3orc	A	Cro	Phage $\lambda$	3.0	AAGTACAAACTTTTCTTG-TATTATA--CAAGAAAGTTTGTACT

Luscombe et al.(2000) An overview of the structure of protein-DNA complexes Genome Biology 1.1.1-37



1. Cro and Repressor (11mb)

Moses et al.(2003) "Position specific variation in the rate fo evolution of transcription binding sites" BMC Evolutionary Biology 3.19-

- Databases with the 3-D structure of combined DNA -Protein
- Data bases with known promoters

