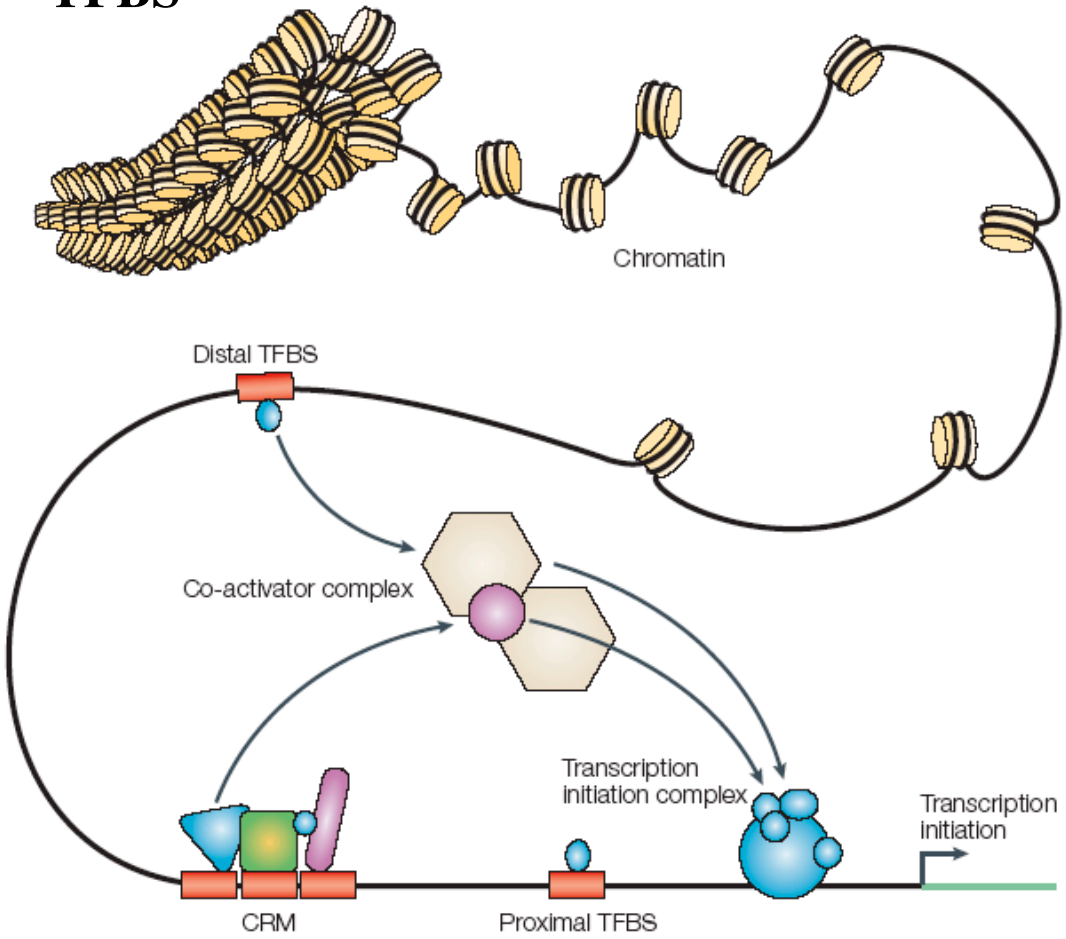
Finding Regulatory Signals in Genomes

- The Biological Problem
  - Different Kinds of Signals
    - Promotors
    - Enhancers
    - Splicing Signals
  - Different Organisms
  - Information Beyond the sequences
  - Data - known/unknown signal
    - Aligned
    - Unaligned
- The Computational Problem
- Measures of Performance Quality
- Performance of Different Methods

Regulation in Eukaryotes

- **Promotor**
- **Transcription Factors - TF**
- **Transcription Factor binding Sites - TFBS**
- **Cis-regulatory modules - CRM**
- **Transcription Start Site - TSS**
- **TATA boxes**
- **CG richness**
- **Phylogenetic Footprinting**
- **Combinatorial Interaction**
- **Enhancers**



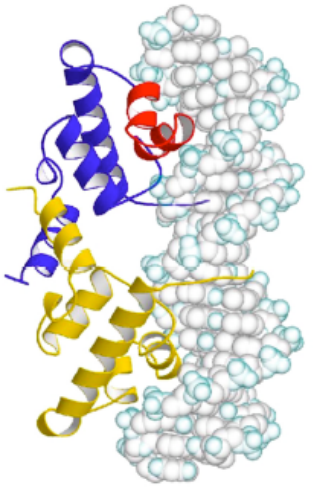
Wasserman and Sandelin (2004) 'Applied Bioinformatics for the Identification of Regulatory Elements" Nature Review Genetics 5.4.276

Regulatory Protein-DNA Complexes

1. Cro and Repressor family

1lmb*	3,4	Repressor	Phage λ	1.8	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
1li	A,B	Repressor mutant	Phage λ	2.1	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
1per	L,R	Repressor	Phage 434	2.5	AAGTACAGTTTTTCTTG-TATTATA--CAAGAAAAACTGTACT
1rpe	L,R	Repressor	Phage 434	2.5	-TATACAATGTATCTTG-TTTGACAAACAAGATACATTGTAT-
2or1	L,R	Repressor	Phage 434	2.5	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
3cro	L,R	Cro	Phage 434	2.5	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
6cro	A	Cro	Phage λ	3.0	AAGTACAAACTTTCTTG-TATTATA-CAAGAAAGTTTGTACT
3orc	A	Cro	Phage λ	3.0	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT

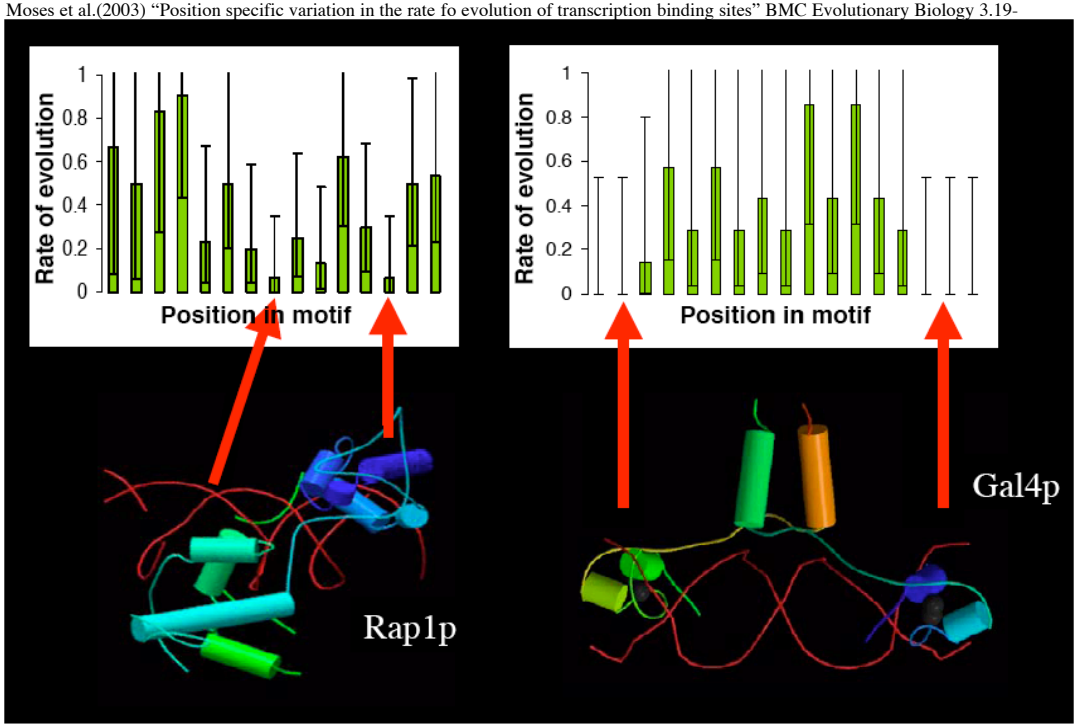
Luscome et al.(2000) An overview of the structure of protein-DNA complexes Genome Biology 1.1.1-37



1. Cro and Repressor (1lmb)

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

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



Weight Matrices,Sequence Logos

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation: $$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion: $$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

$p(b)$ = background probability of base b; $p(b,i)$ = corrected probability of base b in position i; $W_{b,i}$ = PWM vaue of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3)

Evaluation of sequences: $$S = \sum_{i=1}^{w} W_{l_i,i} \quad (3)$$

l_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

Information content calculation: $$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \quad (4)$$

D_i = information content in position i; $p(b,i)$ = corrected probability of base b in position i

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
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	2	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

d Position weight matrix (PWM)

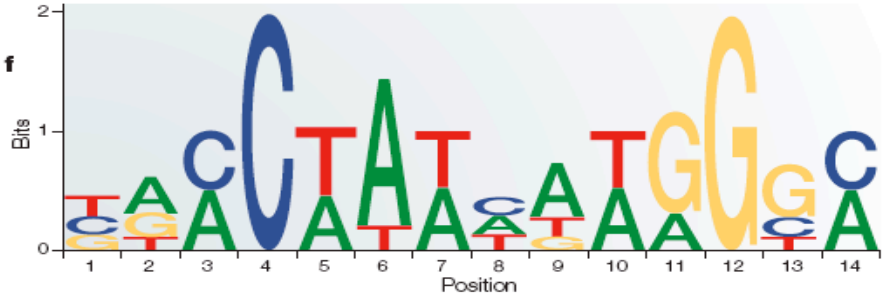
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

f



Very high frequency of false positives. A model for binding of MyoD will yield 10^6 binding sites, while only 10^3 might be real.

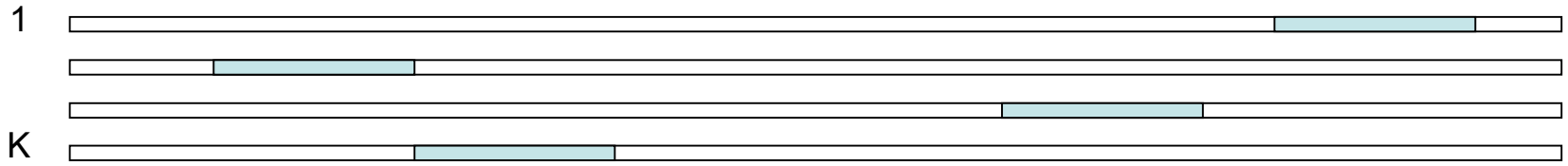
Wasserman and Sandelin (2004) 'Applied Bioinformatics for the Identification of Regulatory Elements" Nature Review Genetics 5.4.276

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.



(R,I)

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

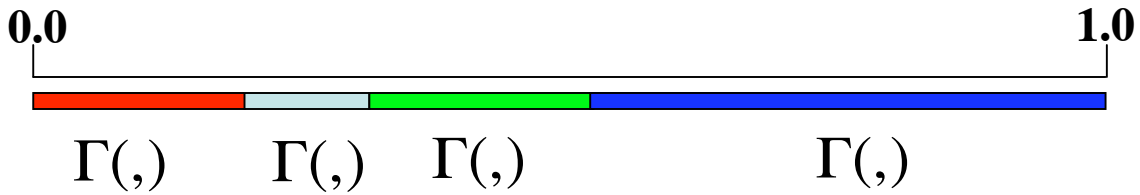
$A=(a_1,..,a_K)$ – positions of the windows

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

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

Priors A has uniform prior

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



The Gibbs Sampler

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

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

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

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

An example:

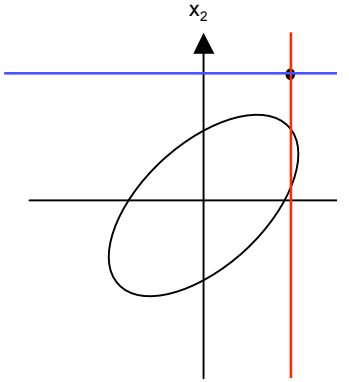
Target Distribution is $x = (x_1, x_2)$ is $N\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right\}$ distributed.

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

$x_1^{t+1} | x_2^{t+1} \sim N\{\rho x_2^t, (1-\rho)^2\}$,

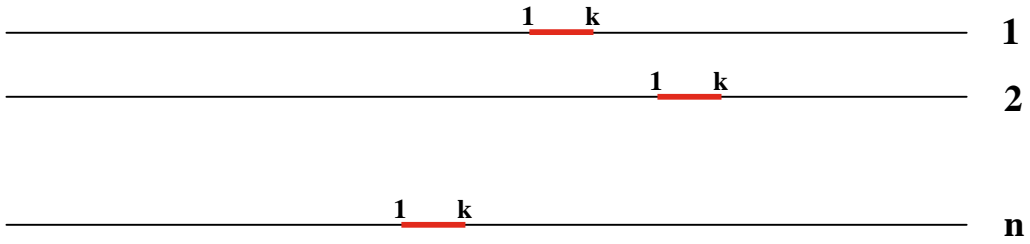
The approximating distribution after t steps of a systematic GS will be:

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



The Gibbs sampler

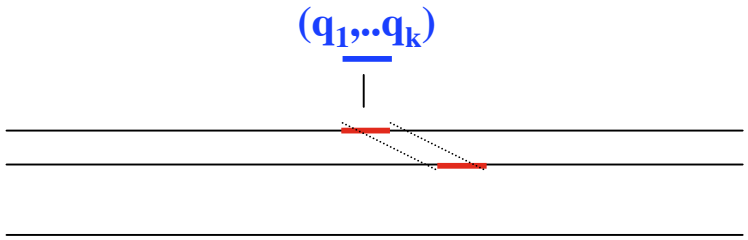
***Objective*: Find conserved segment of length *k* in *n* unrelated sequences**



Gibbs iteration:

Remove one at random - s_j

Form profile of remaining *n*-1



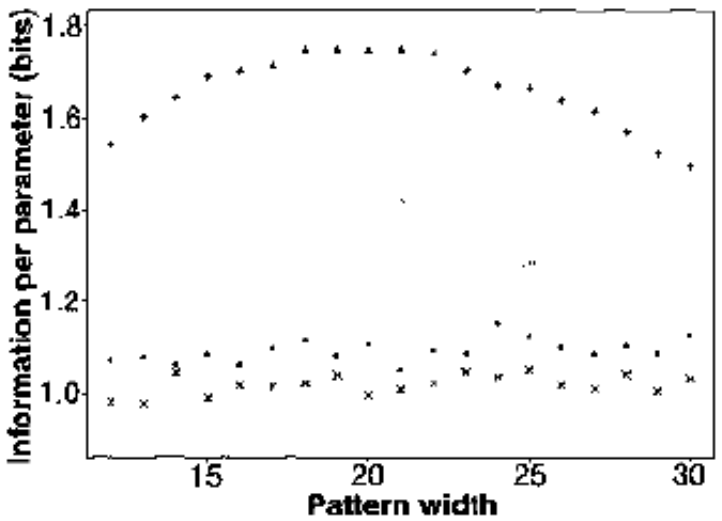
Let p_i be the probability with which $s_j[i..i+k-1]$ fits profile. Including pseudocounts. Choose to start replacement at *i* with probability proportional to p_i

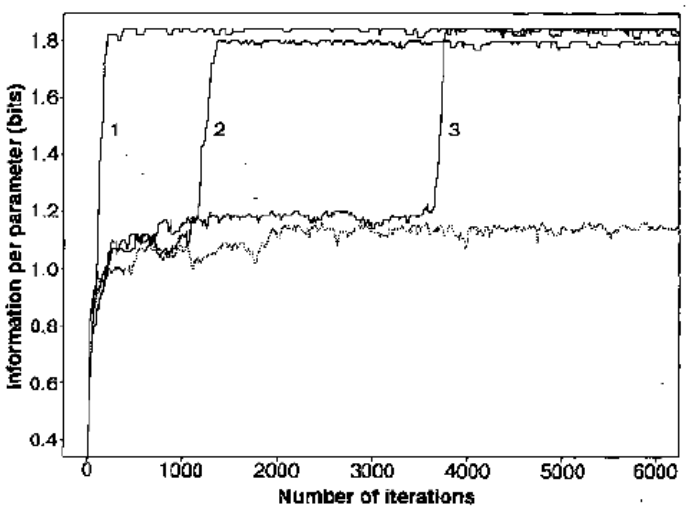
From : Lawrence, C. et al.(1993) Detecting Subtle Sequence Signals: A Geibbs Sampler approach to Multiple Alignment. Science 262.208-

The Gibbs sampler: example

```
Sigma-37       223 IIDLTYIQNK SQRETGDILGISQMHVSR LQRKAVKKLR  240   A25944
SpoIIIC         94 RPGLDLKKEK TQREIAKELGISRSYVSR IEKRALMKNF  111   A28627
NahR            22 VVFNQLLVDR RVSITAENLGLTQPAVSN ALKRLRTSLQ   38   A32837
Antennapedia   326 PHENEYLTRR RRIEIAHALCLTERQIKI WFQNRRMKWK  343   A23450
NtrC (Brady.)  449 LTAALAATRG NQIRAADLLGINRNTLRK KIRDLDIQVY  466   B26499
DicA            22 IRYRRKNLKH TQRSLAKALKISHVSVSQ WERGDSEPTG   39   B24328 (BVECDA)
MerD             5        MNAY TVSRLALDAGVSVHIVRD YLLRGLLRFV   22   C29010
Fis             73 LDMVNQYTRG NQIRAALMMGINRGTLRK KLKKYGMN    90   A32142 (DNECFS)
MAT a1          99 FRRRQSLNSK EKSEVAKRCGITPLQVRV WFINKRMRSK  116   A90983 (JEBV1)
Lambda cII      25 SALLNKIAML GTEKTAEAVGVDKSQISR WKRDWIPKFS   42   A03579 (QCBP2L)
Crp (CAP)      169 THPDGMQIKI TRQEIGQIVGCSRETVGR ILKMLEDQNL  186   A03553 (QBECC)
Lambda Cro      15 ITLKDYAKRF GQTKTAKDLGVYQSAINK AIHAGRKIFL   32   A03577 (RCBPL)
P22 Cro         12 YKKDVIDKFG TQRAVAKALGISDRAVSQ WKEVIPERDA   29   A25867 (RGBP22)
AraC           196 ISDHLADSNF DIASVAQHVCLSPSRLSH LFRQQLGISV  213   A03554 (RGECA)
Fnr            196 FSPREFRLTM TRGDIGNYLGLTVETISR LLGRFQKSGM  213   A03552 (RGECF)
HtpR           252 ARWLDEDNKS TLQELADRYGVSAERVRQ LEKNAMKKLR  269   A00700 (RGECH)
NtrC (K.a.)    444 LTTALRHTQG HKQEAARLLGWGRNTLCR KLKELGME   461   A03564 (RGKBCP)
CytR            11 MKAKKQETAA TMKDVALKAKVSTATVSR ALMNPDKVSQ  28   A24963 (RPECCT)
DeoR            23 LQELKRSDRL HLRDAAALLGVSEMTIRR DLNNHSAPVV  40   A24076 (RPECDO)
GalR             3          MA TIKDVARLAGVSVATVSR VINNSPKASE  20   A03559 (RPECG)
LacI             5        MKPV TLYDVAEYAGVSYQTVSR VVNQASHVSA  22   A03558 (RPECL)
TetR            26 LLNEVGIEGL TTRKLAQKLGVEQPTLYW HVKNKRALLD  43   A03576 (RPECTN)
TrpR            67 IVEELLRGEM SQRELKNELGAGIATITR GSNSLKAAPV  84   A03560 (RPECM)
NifA           495 LIAALERAGW VQAKAARLLGMTPRQVAY RIQIMDITMP 512   S02513
SpoIIG         205 RPGLVGEEER TQKDVADMMGISQSYISR LEKRIIKRLR 222   S07337
Pin            160 QAGRLIAAGT PRQRVAIIYDVGVSTLYK TFPAGEK    177   S07950
PurR             3          MA TIKDVAKRANVSTTTVSH VINKTRFVAE  20   S08477
EbgR             3          MA TLKDIAIEAGVSLATVSR VLNDDPTLNV  20   S09205
LexA            27 DHILSQTGNDP TRAEIAQRLGFRSPNAAE EHLKALARKG  44   S11945
P22 cI          25 SSILNRIAIR GQRKVADALGINESQISR WKGDFIPKMG   42   B25867 (Z1BPC2)
                                ********** ***
```

	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	9	52	222	94	94	9	265	606
Lys	9	133	442	380	9	71	380	194	9	133	9	9	71	9	9	9	71	256
Glu	53	9	96	401	9	9	140	140	9	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	9	67
Gln	9	600	224	9	9	9	224	9	9	9	9	9	278	63	278	9	9	170
His	240	9	9	9	9	9	125	125	9	9	9	9	125	125	125	9	9	240
Asn	168	9	9	9	9	9	168	89	9	89	9	248	9	168	89	9	89	89
Ser	117	9	117	117	9	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	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	9	311	130	70	855	9	130	9
Pro	75	9	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	9
Val	58	107	9	9	500	9	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	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	9
Tyr	9	9	136	9	9	9	9	262	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





From : Lawrence, C. et al.(1993) Detecting Subtle Sequence Signals: A Gibbs Sampler approach to Multiple Alignment. Science 262.208-

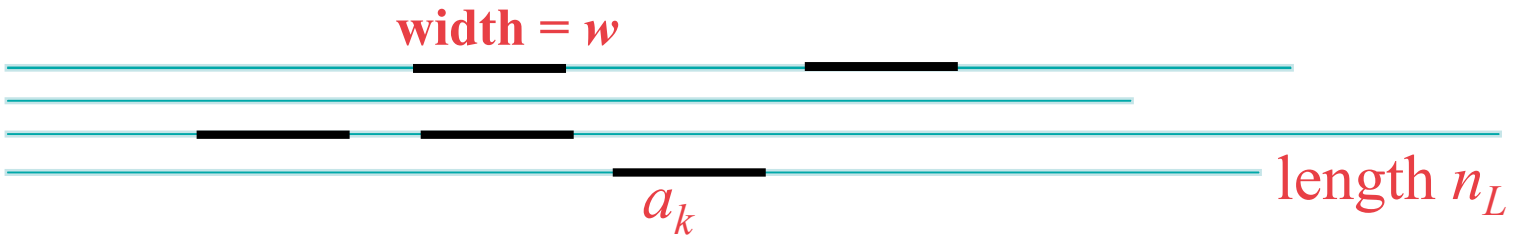
Natural Extensions to Basic Model I

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:

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



Composite Patterns:

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

Modified from Liu

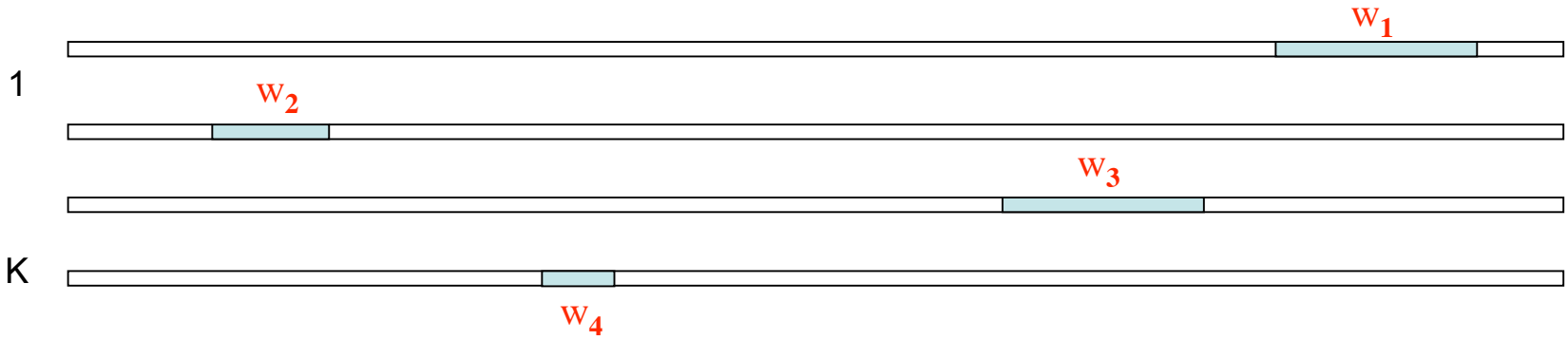
Natural Extensions to Basic Model II

Correlated in Nucleotide Occurrence in Motif:

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

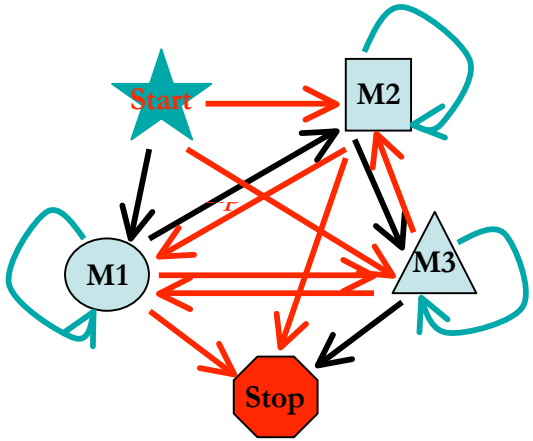
Insertion-Deletion

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

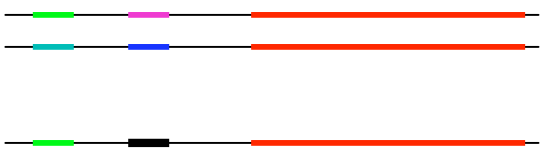


Combining Signals and other Data

Expresssion and Motif Regression:

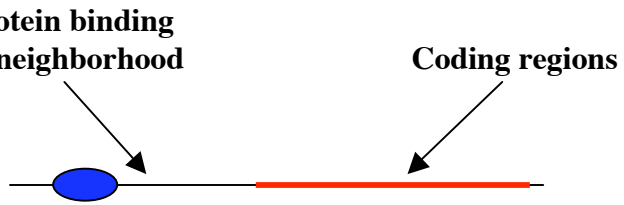
Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci. 100.3339-44

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



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



Modified from Liu

The Expectation-Maximization Algorithm (EM)

Aim: Maximizing Likelihood function in presence of missing data.

$\log P_{\Theta}(x, y)$, x is observed, y is missing and Θ is the parameter.

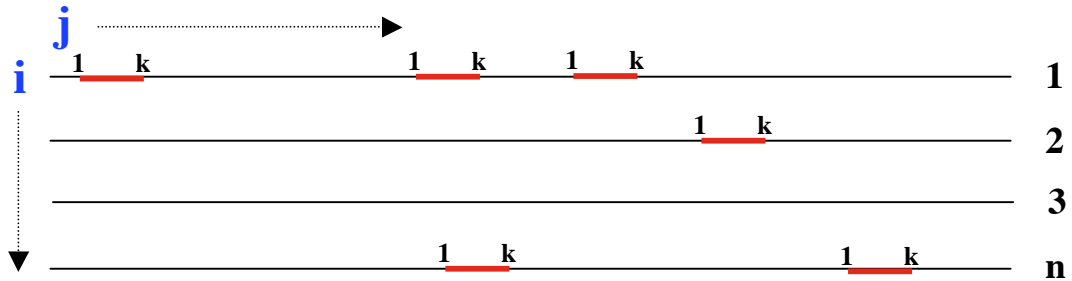
E – step : calculates expected loglikelihood averaging over the unobserved data $E[\log P_{\Theta}(x|y)]$

M – step : Maximize $E[\log P_{\Theta}(x|y)]$ as a function of Θ.

Each E+M step will not decrease the likelihood, E+M steps are continued until little change in likelihood function.

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], λ is probability that a $Z_{i,j}$ = 1

Find M,B, λ, Z that maximize Pr (X, Z | M, B, λ)

Expectation Maximization to find a local maximum

Iteration t:

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

Maximization-step: Find (M, B, λ) $^{(t+1)}$ that maximizesPr (X, $Z^{(t)}$ | (M, B, λ) $^{(t+1)}$)

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol 2: 28-36.

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.**

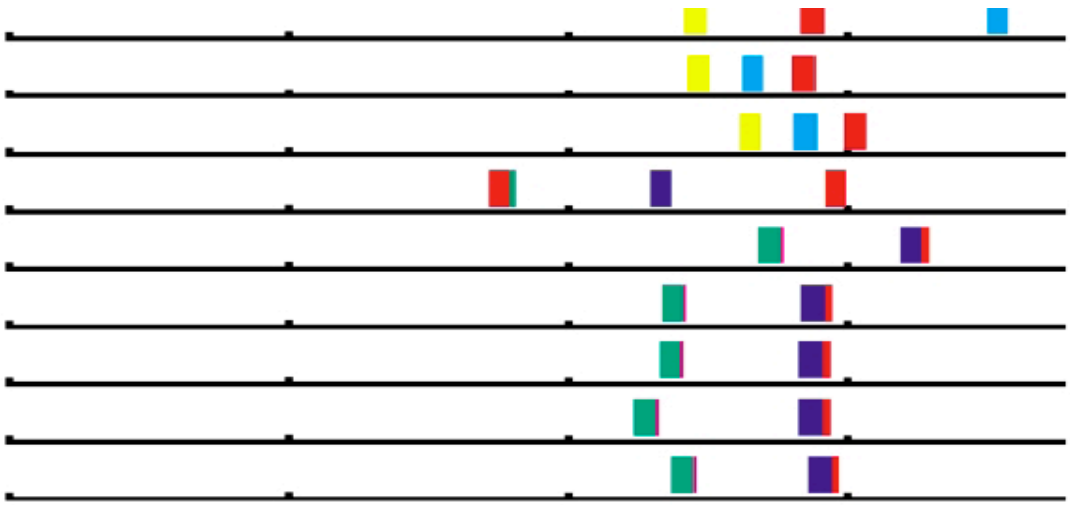
$$D_i^{begin} = \min\{D_{i,\Delta}^{begin} + d(i,\Delta)\}$$

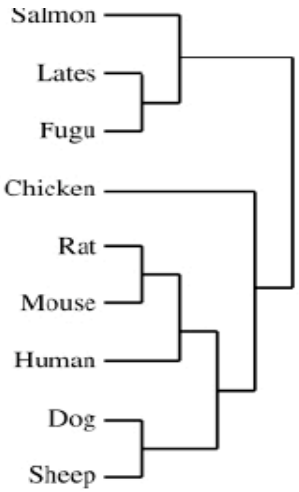
$$D_i^{signal,0} = \min\{D_{i,\Delta}^{b} + d(i,\Delta)\}$$

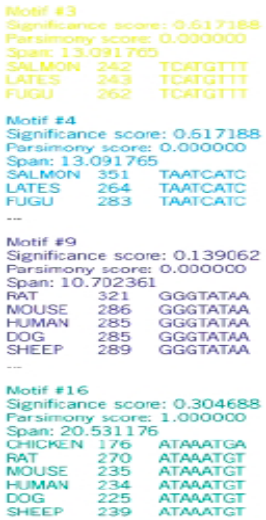
$$D_i^{signal,j+1} = \min\{D_{i,\Delta}^{signal,j} + Kd(i,\Delta)\}$$

$$\ldots$$

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





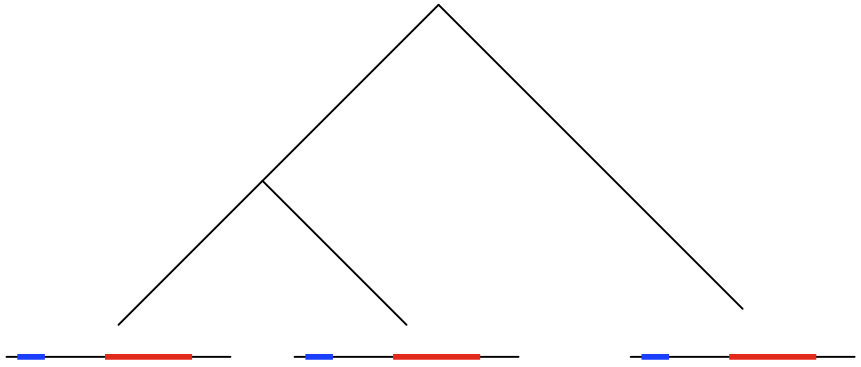


Blanchette and Tompa (2003) "FootPrinter: a program designed for phylogenetic footprinting" NAR 31.13.3840-

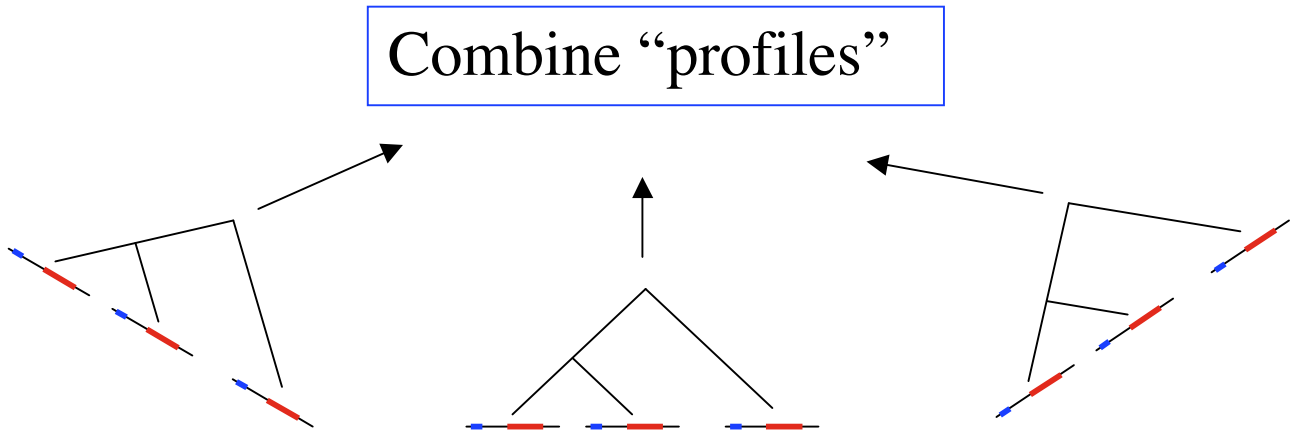
(Homologous + Non-homologous) detection

Unrelated genes - similar expression

Related genes - similar expression



Combine above approaches:Mixed genes - similar expression



Wang and Stormo (2003) “Combining phylogenetic data with co-regulated genes to identify regulatory motifs” Bioinformatics 19.18.2369-80

Rate of Molecular Evolution versus estimated Selective Deceleration

Neutral Process

	A	C	G	T
A	–	$q_{A,C}$	$q_{A,G}$	$q_{A,T}$
C	$q_{C,A}$	–	$q_{C,G}$	$q_{C,T}$
G	$q_{G,A}$	$q_{G,C}$	–	$q_{G,T}$
T	$q_{T,A}$	$q_{T,C}$	$q_{T,G}$	–

How much selection?

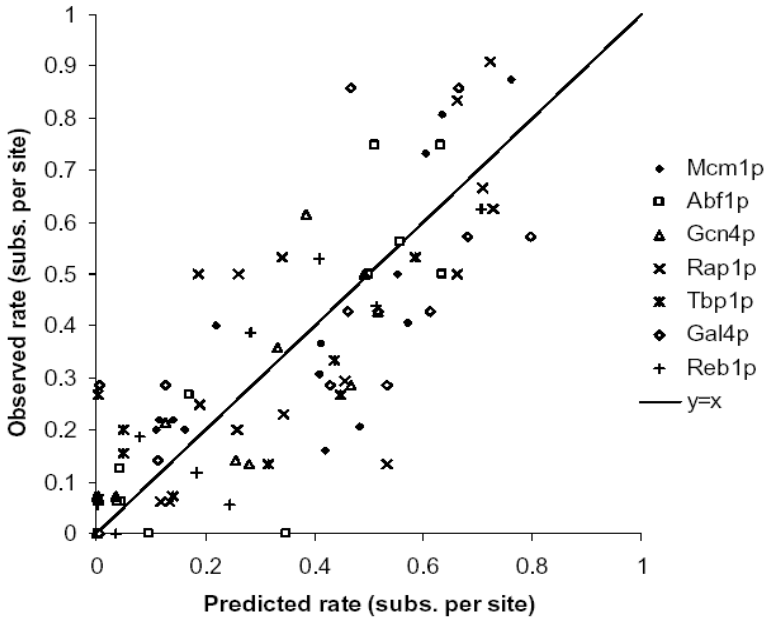
Selection => deceleration

Selected Process

	A	C	G	T
A	–	$q'_{A,C}$	$q'_{A,G}$	$q'_{A,T}$
C	$q'_{C,A}$	–	$q'_{C,G}$	$q'_{C,T}$
G	$q'_{G,A}$	$q'_{G,C}$	–	$q'_{G,T}$
T	$q'_{T,A}$	$q'_{T,C}$	$q'_{T,G}$	–

Neutral Equilibrium

$(\pi_A,\pi_C,\pi_G,\pi_T)$



Observed Equilibrium

$(\pi_A,\pi_C,\pi_G,\pi_T)'$

Halpern and Bruno (1998) "Evolutionary Distances for Protein-Coding Sequences" MBE 15.7.910- & Moses et al.(2003) "Position specific variation in the rate fo evolution of transcription binding sites" BMC Evolutionary Biology 3.19-

Summary

- The Biological Problem
  - Different Kinds of Signals
    - Promotors
    - Enhancers
    - Splicing Signals
  - Different Organisms
  - Information Beyond the sequences
  - Data - known/unknown signal
    - Aligned
    - Unaligned
- The Computational Problem
- Measures of Performance Quality
- Performance of Different Methods

References I

- J Amer "***Bayesians Models for multiple local sequence alignment***" Statist.Assoc. **90**, 1156-1170
- J Amer "***The collapsed Gibbs sampler with applications to a gene regulation problem,***" Journal of the American Statistical Association 89 958-966
- Bailey, T. L. and C. Elkan (1994). "***Fitting a mixture model by expectation maximization to discover motifs in biopolymers.***" Proc Int Conf Intell Syst Mol Biol 2: 28-36.
- Boffelli, Nobrega and Rubin (2004) "***Comparative genomics at the Vertebrate Extremes***" Nature Review Genetics 5.6.456-
- Blanchette,M, B.Schwikowski and M.Tompa (2002) "***Algorithms for Phylogenetic Footprinting***" J. Comp.Biol.9.2.211-
- Blanchette and Tompa (2003) "***FootPrinter: a program designed for phylogenetic footprinting***" NAR 31.13.3840-
- D.Che, S Jensen L.Cai "***BEST: Binding-site estimation suite of tools .***" Bioinformatics, 21, 2209-11.
- E Conlon "***Integrating Sequence Motif Discovery and Microarray Analysis*** " Proc.Natl.Acad.Sci. 100.3339-44
- Chuzhanova et al.(2002) "***The Evolution of Vertebrate b-globin promotor.***" Evolution 56.2.224-232
- Dermitzakis, E. T., A. Reymond, et al. (2003). "***Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (CNGs).***" Science.
- Fickett and Hartzegiorgiou (1997) "***Eukaryotic Promotor Recognition***" Genome Research 7.861-
- Gribskov, M., McLachlan, A.D., and Eisenberg, D., "***Profile analysis: detection of distantly related proteins*** **".** Proceedings of the National Academy of Sciences 84, 4355-4358, 1987
- Halpern and Bruno (1998) "***Evolutionary Distances for Protein-Coding Sequences***" MBE 15.7.910-
- M.Gupta "***Statistical models for biological sequence motif discovery*** " Case Studies in Bayesian Statistics VI, 2002. Springer
- M Gupta "***De novo cis-regulatory module elicitation for eukaryotic genomes.*** " Proc Nat'l Acad Sci USA, 102, 7079-84**.**
- M Gupta "***Discovery of conserved sequence patterns using a stochastic dictionary model.***" J. Amer. Statist. Assoc., 98, 55-66.
- H.Huang M.J.Kao X. Zhou WH Wong "***Identification of transcription factor binding sites using local Markov models.*** " J. Computational Biology
- P. Hong XS Liu WH Wong "***A Boosting Approach for Motif Modeling Using ChIP-chip Data.*** " Bioinformatics, 21, 2636-43.
- S. Jensen "***BioOptimizer: The Bayesian Scoring Function Approach to Motif Discovery*** " Bioinformatics
- S. Jensen L.Shen " ***Combining Phylogenetic Motif Discovery and Motif Clustering to Predict Co-Regulated Genes.*** " Bioinformatics In
- Lawrence, C. et al.(1993) " ***Detecting Subtle Sequence Signals.***" A Gibbs Sampler approach to Multiple Alignment. Science 262.208-

References II

- CE Lawrence "***Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment***" Science 262, 208-214.
- CE Lawrence *et al* "***Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective.***" Statistical Science, 19, 188-204
- JS Liu "***A Gibbs sampler for the detection of subtle motifs in multiple sequences***" Proc. 27th Hawaii International Conference on System

- JS Liu et al "***Unified Gibbs Method for Biological Sequence Analysis***"Proc. ASA Biometrics Section, 194-199.
- X Liu et al "***Bioprospector: Discovering Conserved DNA motifs in upstream regulatory regions.***" Proceedings of the Pacific Symposium on Biocomputing (PSB)
- XS Liu DL Brutlag"***An Algorithm for Finding Protein-DNA Interaction Sites with Applications to Chromatin***
- Lenhard, B., A. Sandelin, et al. (2003). "***Identification of conserved regulatory elements by comparative genome analysis.***" J Biol 2(2): 13.
- Loots, G. G., I. Ovcharenko, et al. (2002). "***Vista for comparative sequence-based discovery of functional transcription factor binding sites.***" Genome Res 12(5): 832-9.
- Luscome et al.(2000) An overview of the structure of protein-DNA complexes Genome Biology 1.1.1-37
- Marchal et al.(2003) "***Genome Specific higher order background models to improve motif detection***" Trends in Genetics 11.2.61-
- *LA McCue et al* "***Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes*** " Nucleic Acids Research, 29,774-782
- Moses et al.(2003) "***Position specific variation in the rate of evolution of transcription binding sites***" BMC Evolutionary Biology 3.19-
- Pennachio and Rubin (2001) "***Genomic Strategies in Identifying Mammalian Regulatory Sequences***" Nature Review Genetics 2.2.100-109
- Christoph D. Schmid, Viviane Praz, Mauro Delorenzi, Rouaïda Périer, and Philipp Bucher "***The Eukaryotic Promoter Database EPD: the impact of in silico primer extension***" Nucl. Acids. Res. 2004 32: D82-D85.
- Stormo, G. (2000) "***DNA binding sites: representation and discovery***" Bioinformatics 16.16-23.
- Struhl, K. (1999). "***Fundamentally different logic of gene regulation in eukaryotes and prokaryotes.***" Cell 98(1): 1-4.
- Wasserman and Sandelin (2004) "***Applied Bioinformatics for the Identification of Regulatory Elements***" Nature Review Genetics 5.4.276
- Wang and Stormo (2003) "***Combining phylogenetic data with co-regulated genes to identify regulatory motifs***" Bioinformatics 19.18.2369-80
- Wray, G. A., M. W. Hahn, et al. (2003). "***The evolution of transcriptional regulation in eukaryotes.***" Mol Biol Evol 20(9): 1377-419.

- B.M Webb C.E. Lawrence " ***BALSA: Bayesian algorithm for local sequence alignment*** " Nucl. Acids Res., 30 1268-77.
- SZ Qin et al "***Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites*** " Nature Biotechnology 21, 435-39.
- Qing Zhou et al"***Modeling within-motif dependence for transcription factor binding site predictions.***" Bioinformatics, 6, 909-916. Sciences, 245-Press
- "***Immunoprecipitation Microarray Experiments***" Nature Biotechnology, 20, 835-39