

Finding Regulatory Signals in Genomes

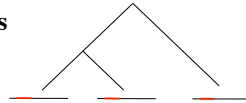
Searching for known signal in 1 sequence



Searching for unknown signal common to set of unrelated sequences



Searching for conserved segments in homologous



Challenges

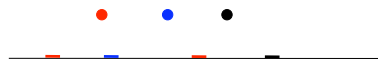
Combining homologous and non-homologous analysis



Merging Annotations

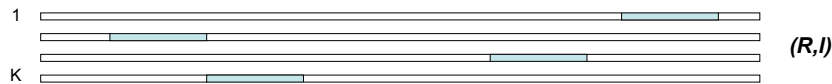


Predicting signal-regulatory protein relationships



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

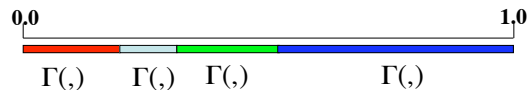
$A = (a_1, \dots, a_k)$ – positions of the windows

$\theta_0 = (\theta_{A_1}, \dots, \theta_T)$ – background frequencies of nucleotides.

$$p(R | \theta_0, \Theta, A) = \theta_0^{h(R_{1,A_1^c})} \prod_{j=1}^w \theta_j^{h(R_{A+j-1})} = \theta_0^{h(R)} \prod_{j=1}^w \left(\frac{\theta_j}{\theta_0} \right)^{h(R_{A+j-1})}$$

Priors A has uniform prior

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



Weight Matrices & Sequence Logos

Set of signal sequences:

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

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

	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	T	A	A	G	G	C	A	
3	T	G	A	C	T	A	T	A	A	A	G	G	A	
4	T	G	A	C	T	A	T	A	A	A	G	G	A	
5	T	G	C	C	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	G	G	G	C	
8	C	T	C	T	A	C	A	T	G	G	G	C		

Position Frequency Matrix - PFM

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

Consensus sequence:

B R M C W A W H R W G G B M

Position Weight Matrix - PWM

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

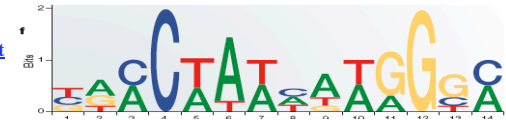
A	-1.93	.79	-.79	-1.93	.45	1.50	.79	-.45	1.07	.79	.0	-1.93	-1.93	.79
C	.45	-1.93	.79	1.66	-1.93	-1.93	-1.93	-.45	-1.93	-1.93	-1.93	-1.93	.0	.79
G	.0	.45	-1.93	-1.93	-1.93	-1.93	-1.93	.66	-1.93	1.3	1.68	1.07	-1.93	
T	.15	.66	-1.93	-1.93	1.07	.66	.79	.0	.79	-1.93	-1.93	-1.93	.66	-1.93
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

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

T	T	G	C	A	T	A	A	G	T	A	G	T	C
.45	-.66	.79	1.66	-.45	-.66	.79	.45	-.66	.79	.0	1.68	-.66	.79

Sequence Logo & Information content

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



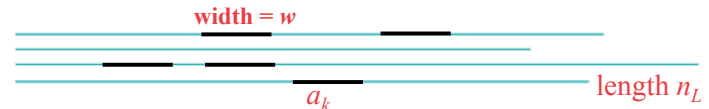
Natural Extensions to Basic Model I

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

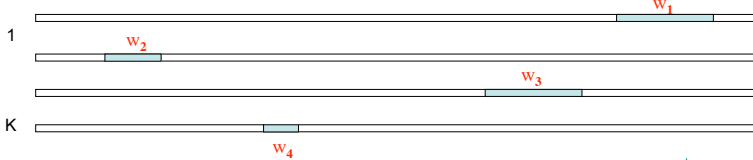
Correlated in Nucleotide Occurrence in Motif:

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



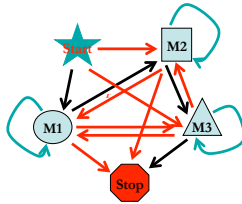
Insertion-Deletion

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 Natl Acad Sci USA*, 102, 7079-84

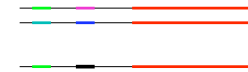


Combining Signals and other Data

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



Modified from Liu

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.

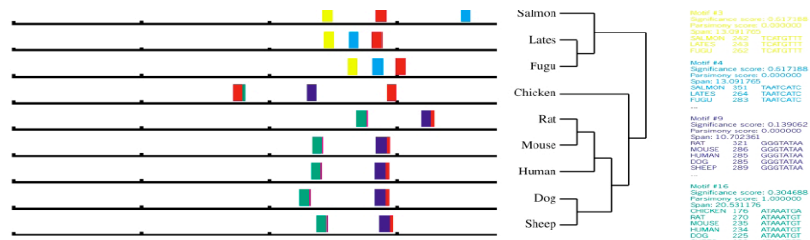
$$D_i^{begin} = \min\{D_{i,\Delta}^{begin} + d(i,\Delta)\}$$

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

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

...

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

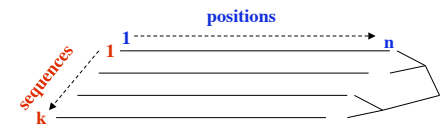


Bioinformatics and Bioprocess Engineering, (2016) 11(11): 113-121

The Basics of Footprinting

- Many aligned sequences related by a known phylogeny:

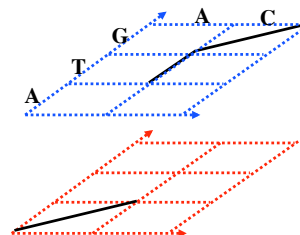
HMM:



HMM:



- Two un-aligned sequences:

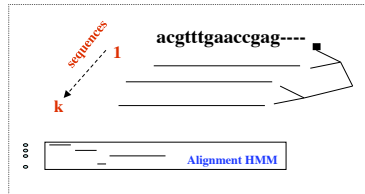


ATG
A-C

Statistical Alignment and Footprinting.

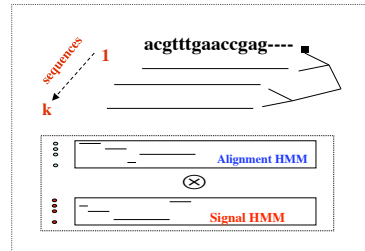
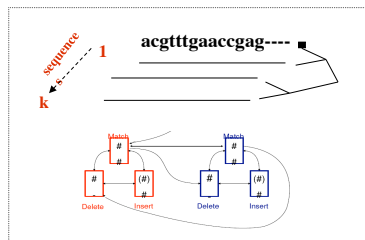
Many un-aligned sequences related by a known phylogeny:

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

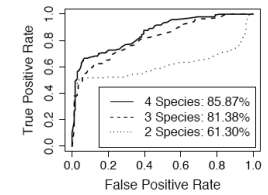
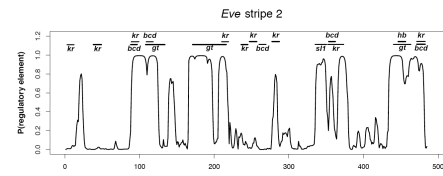
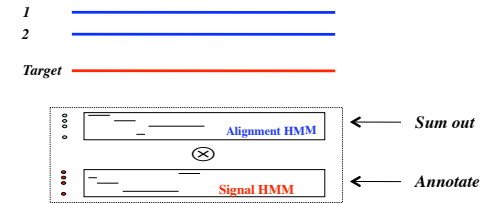
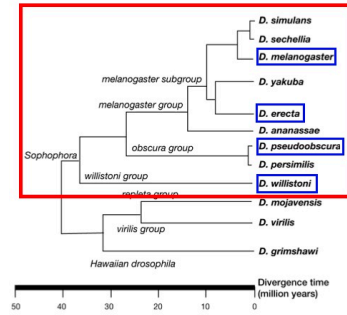


Solution:

Cartesian Product of HMMs



SAPF - Statistical Alignment and Phylogenetic Footprinting



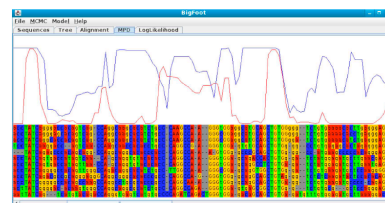
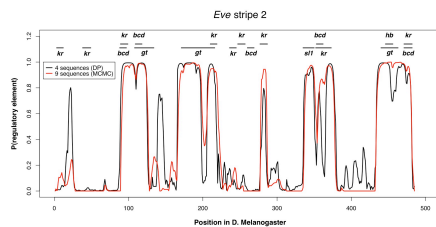
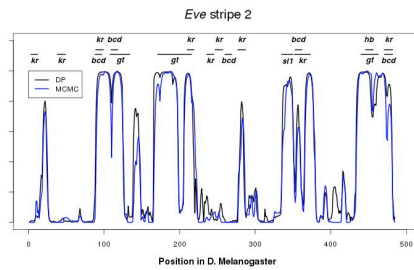
Combining Statistical Alignment and Phylogenetic Footprinting to Detect Regulatory Elements
R. Baliga¹, L. Pachter¹ and J. Henikoff¹

BigFoot: Bayesian Alignment and Phylogenetic Footprinting for Multiple Sequences with MCMC
Rahul Sanjiv¹, Adam Nevins¹, Isaac Misael², Ross Longplé¹, Jesse Hein¹

BigFoot

<http://www.stats.ox.ac.uk/research/genome/software>

- Dynamical programming is too slow for more than 4-6 sequences
- MCMC integration is used instead – works until 10-15 sequences
- For more sequences other methods are needed.

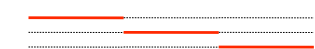


FSA - Fast Statistical Alignment Pachter, Holmes & Co

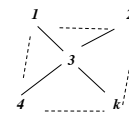
Data – k genomes/sequences:



Iterative addition of homology statements to shrinking alignment:



Spanning tree Additional edges

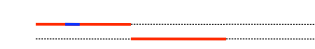


An edge – a pairwise alignment



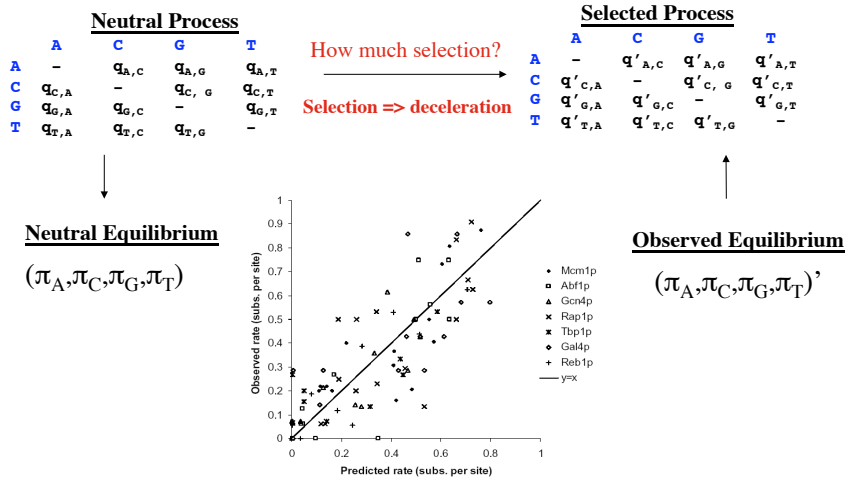
- 1,3 2,3 3,4 3,k
- - - 1,2 2,k 1,4 4,k

Add most certain homology statement from pairwise alignment compatible with present multiple alignment



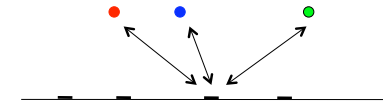
- i. Conflicting homology statements cannot be added
- ii. Some scoring on multiple sequence homology statements is used.

Rate of Molecular Evolution versus estimated Selective Deceleration



Halpern and Bruno (1998) "Evolutionary Distances for Protein-Coding Sequences" MBE 15:7:910. & Moses et al.(2003) "Position specific variation in the rate of evolution of transcription binding sites" BMC Evolutionary Biology 3:19.

Signal Factor Prediction

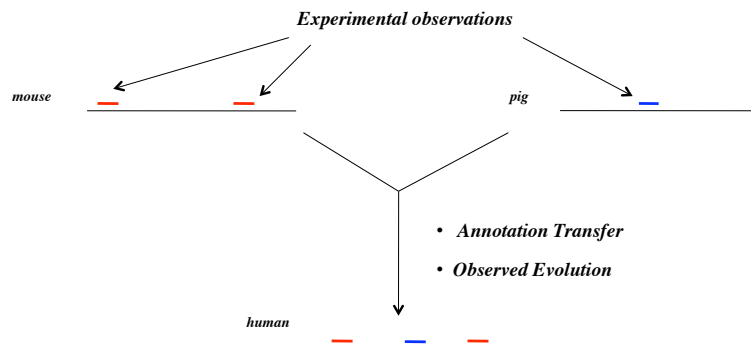


Given set of homologous sequences and set of transcription factors (TFs), find signals and which TFs they bind to.

- Use PWM and Bruno-Halpern (BH) method to make TF specific evolutionary models
- Drawback BH only uses rates and equilibrium distribution
- Superior method: Infer TF Specific Position Specific evolutionary model
- Drawback: cannot be done without large scale data on TF-signal binding.

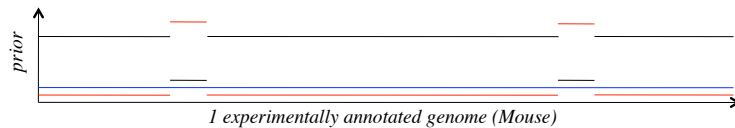
<http://jaspar.cgb.ki.se/> <http://www.gene-regulation.com/>

Knowledge Transfer and Combining Annotations



Must be solvable by Bayesian Priors

Each position p_i probability of being j 'th position in k 'th TFBS
 If no experiment, low probability for being in TFBS

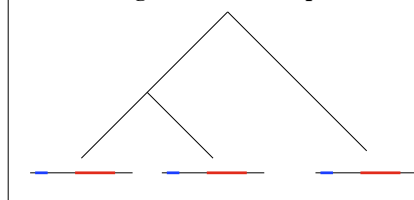


(Homologous + Non-homologous) detection

Unrelated genes - similar expression

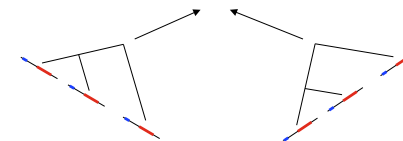


Related genes - similar expression



Combine above approaches

Combine "profiles"



Wang and Stormo (2003) "Combining phylogenetic data with co-regulated genes to identify regulatory motifs" Bioinformatics 19:18:2369-80
 Zhou and Wong (2007) Coupling Hidden Markov Models for discovery of cis-regulatory signals in multiple species. Annals Statistics 1.1:36-65