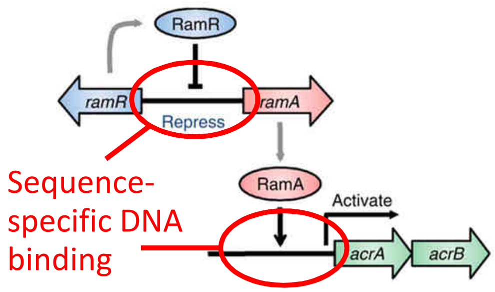
Motifs

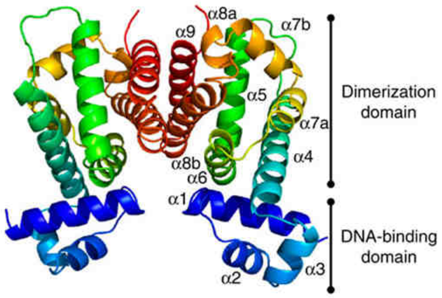
BCH339N Systems Biology / Bioinformatics – Spring 2016

Edward Marcotte, Univ of Texas at Austin

An example transcriptional regulatory cascade
Here, controlling *Salmonella* bacteria multidrug resistance



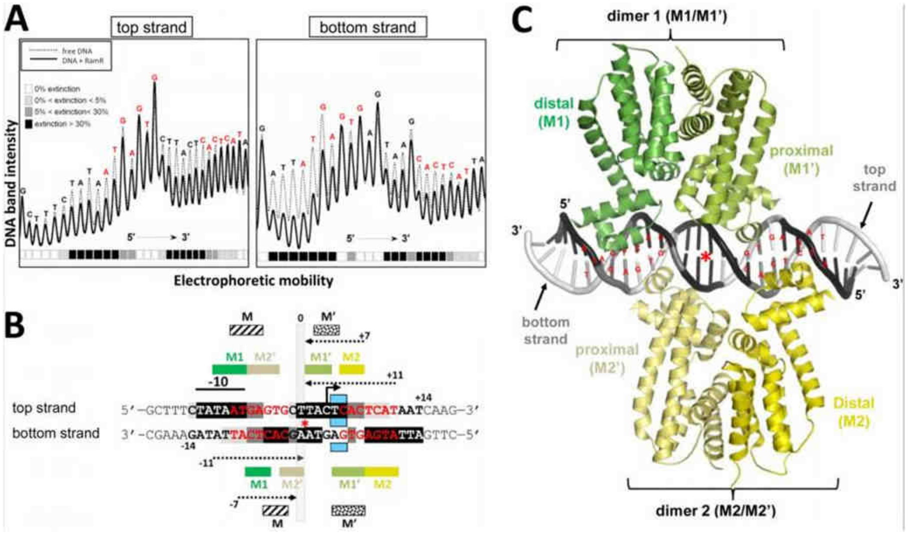
RamR represses the *ramA* gene, which encodes the activator protein for the *acrAB* drug efflux pump genes.



RamR dimer

Nature Communications 4, Article number: 2078 doi:10.1038/ncomms3078

Historically, DNA and RNA binding sites were defined biochemically (DNAse footprinting, gel shift assays, etc.)

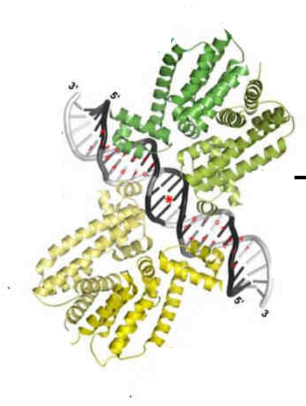


Hydroxyl radical footprinting of *ramR-ramA* intergenic region with RamR

Antimicrob Agents Chemother. Feb 2012; 56(2): 942–948.

Historically, DNA and RNA binding sites were defined biochemically (DNAse footprinting, gel shift assays, etc.)

Now, many binding motifs are discovered bioinformatically



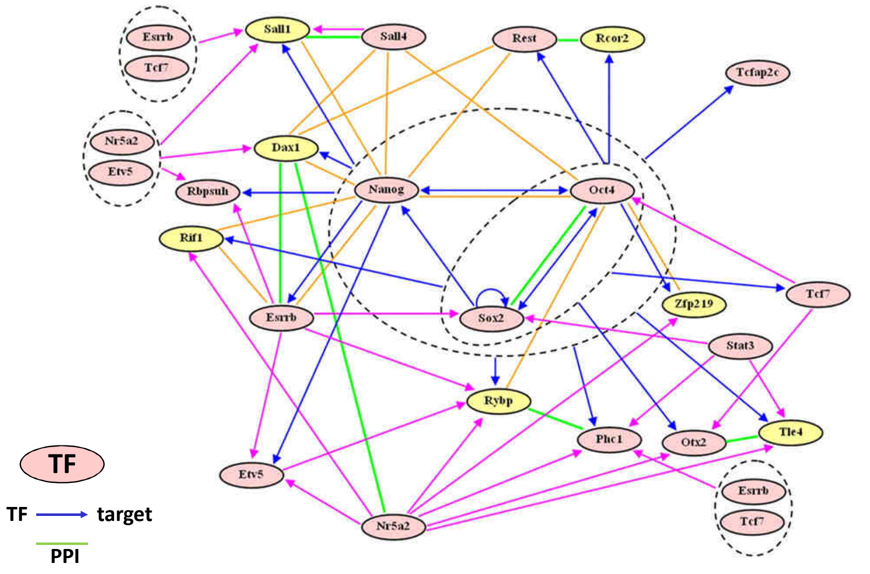
→ Isolate different nucleic acid segments bound by copies of the protein (e.g. all sites bound across a genome) → Sequence

↓

Search computationally for recurring ***motifs***

Image: Antimicrob Agents Chemother. Feb 2012; 56(2): 942–948.

Transcription factor regulatory networks can be highly complex, e.g. as for embryonic stem cell regulators



http://www.pnas.org/content/104/42/16438

MOTIFS

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

Binding sites of the transcription factor ROX1

YCHATTGTTCTC

consensus

A 002700000010
C 464100000505
G 000001800112
T 422087088261

frequencies

$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

frequency of nuc b at position i

freq of nuc b in genome

scaled by information content

NATURE BIOTECHNOLOGY VOLUME 24 NUMBER 4 APRIL 2006

So, here's the challenge:

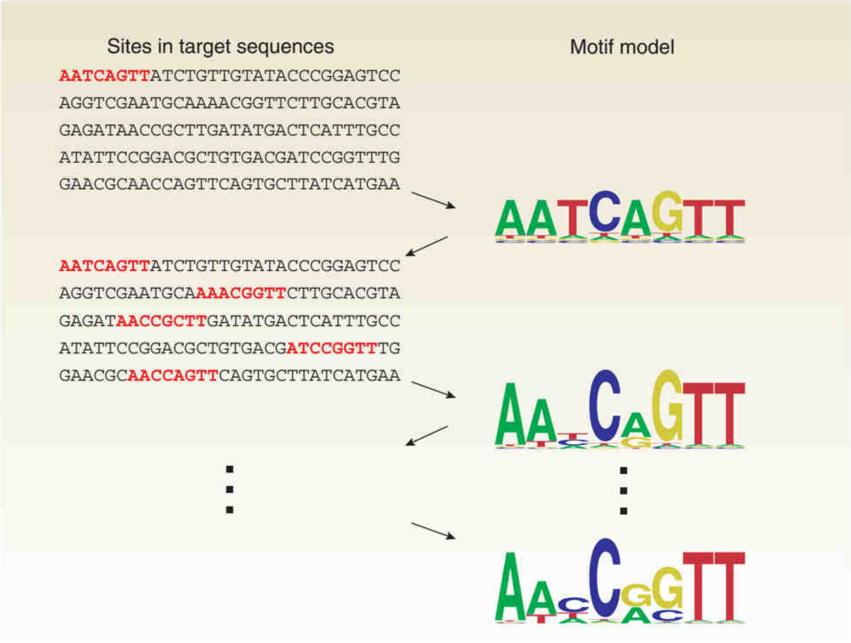
Given a set of DNA sequences that contain a motif (e.g., promoters of co-expressed genes), how do we discover it computationally?

Could we just count all instances of each *k*-mer?

Why or why not?

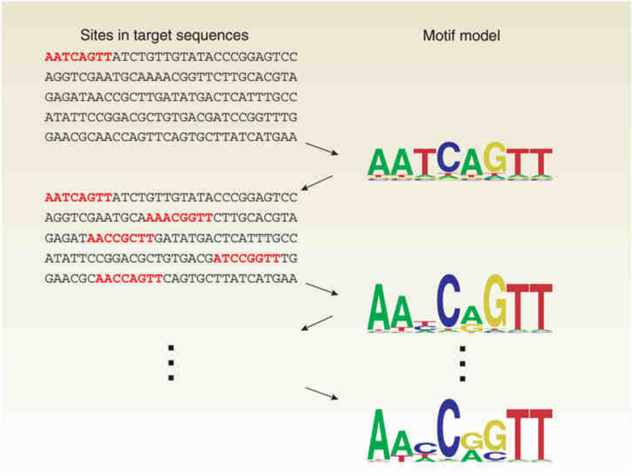
→ promoters and DNA binding sites are not well conserved

How does motif discovery work?



NATURE BIOTECHNOLOGY VOLUME 24 NUMBER 8 AUGUST 2006

How does motif discovery work?



Assign sites to motif

Update the motif model

Assign sites to motif

Update the motif model

Assign sites to motif

Update the motif model

etc.

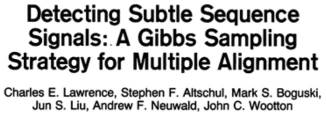
NATURE BIOTECHNOLOGY VOLUME 24 NUMBER 8 AUGUST 2006

How does motif discovery work?

Motif finding often uses *__expectation-maximization__*
***i.e.* alternating between building/updating a motif model and assigning sequences to that motif model.**

Searches the space of possible motifs for optimal solutions __without testing everything.__

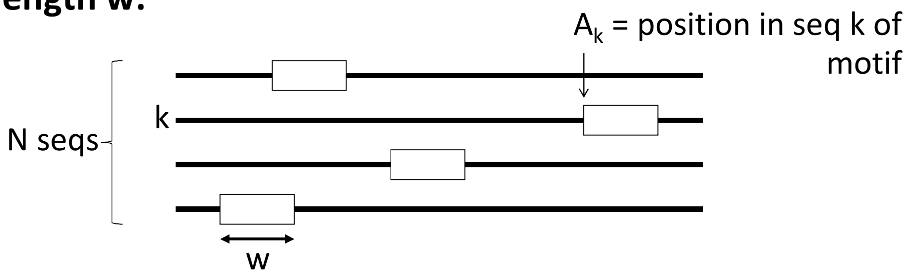
Most common approach = *Gibbs sampling*



Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, John C. Wootton

We will consider N sequences, each with a motif of length w:



q_{ij} = probability of finding nucleotide (or aa) j at position i in motif
i ranges from 1 to w
j ranges across the nucleotides (or aa)
p_j = background probability of finding nucleotide (or aa) j

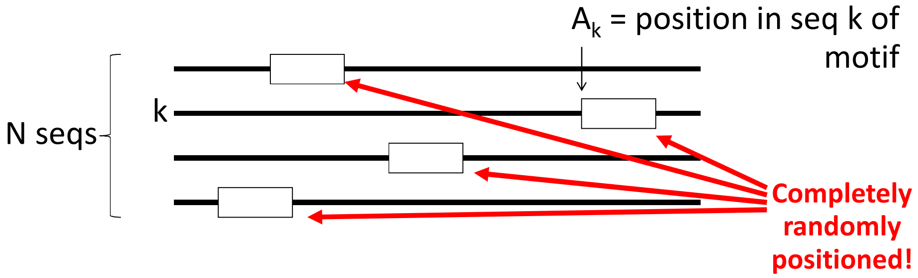
SCIENCE • VOL. 262 • 8 OCTOBER 1993

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, John C. Wootton

NOTE: You won't give any information at all about what or where the motif should be!

Start by choosing w and randomly positioning each motif:



A_k = position in seq k of motif

N seqs

k

Completely randomly positioned!

q_{ij} = probability of finding nucleotide (or aa) j at position i in motif
- i ranges from 1 to w
- j ranges across the nucleotides (or aa)

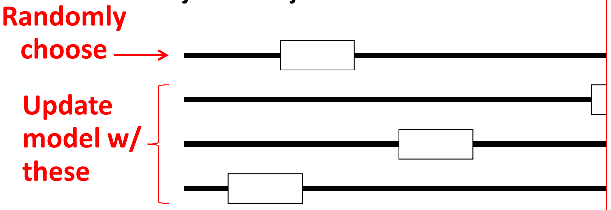
p_j = background probability of finding nucleotide (or aa) j

SCIENCE • VOL. 262 • 8 OCTOBER 1993

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, John C. Wootton

Predictive update step: Randomly choose one sequence, calculate q_{ij} and p_j from N-1 remaining sequences



Randomly choose

Update model w/ these

background frequency of symbol j

count of symbol j at position i

Σb_j

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B}$$

p_j is calculated similarly from the counts outside the motifs

q_{ij} = probability of finding nucleotide (or aa
- i ranges from 1 to w
- j ranges across the nucleotides (or a

p_j = background probability of finding nucleotide (or aa) j

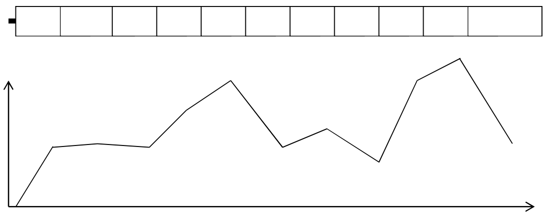
SCIENCE • VOL. 262 • 8 OCTOBER 1993

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski,
Jun S. Liu, Andrew F. Neuwald, John C. Wootton

Stochastic sampling step: For withheld sequence, slide motif down sequence & calculate agreement with model

Withheld sequence →



Odds ratio of agreement with model vs. background

Position in sequence

$$\frac{\Pi(q_{ij})^{c_{xij}}}{\Pi(p_j)^{c_{xij}}}$$

(see the paper for details)

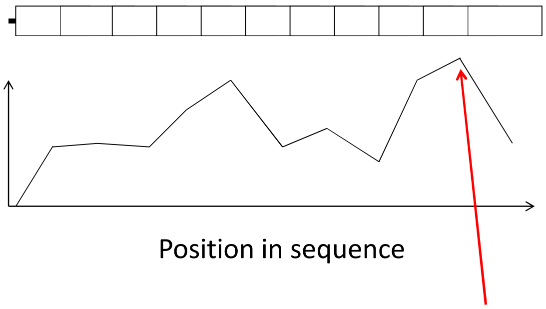
SCIENCE • VOL. 262 • 8 OCTOBER 1993

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski,
Jun S. Liu, Andrew F. Neuwald, John C. Wootton

Stochastic sampling step: For withheld sequence, slide motif down sequence & calculate agreement with model

Withheld sequence →



Odds ratio of agreement with model vs. background

Position in sequence

$$\frac{\Pi(q_{ij})^{c_{xij}}}{\Pi(p_j)^{c_{xij}}}$$

(see the paper for details)

Here's the cool part. DON'T just choose the maximum.
INSTEAD, select a new A_k position proportional to this odds ratio.

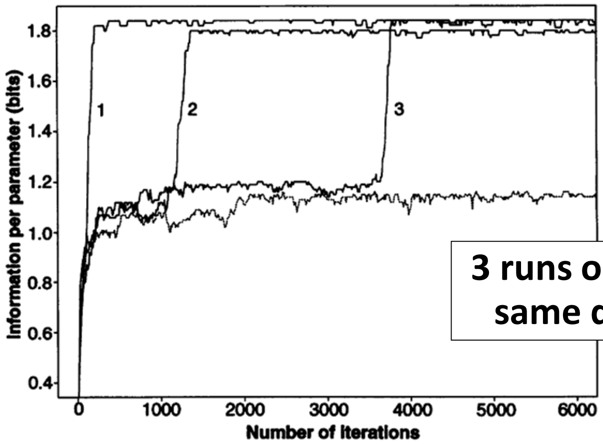
Then, choose a new sequence to withhold, and repeat everything.

SCIENCE • VOL. 262 • 8 OCTOBER 1993

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, John C. Wootton

Over many iterations, this magically converges to the most enriched motifs. Note, it's stochastic:

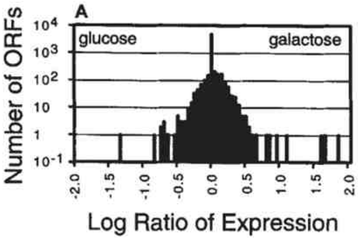


3 runs on the same data

SCIENCE • VOL. 262 • 8 OCTOBER 1993

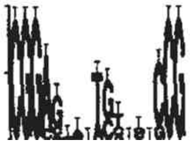
Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation

Frederick P. Roth[1,#], Jason D. Hughes[1,#], Preston W. Estep[2], and George M. Church[1,*]



galactose vs. glucose → **Measure mRNA abundances using DNA microarrays** → **Search for motifs in promoters of glucose vs galactose controlled genes**

Discovered motifs



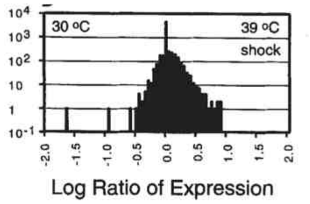
Known motif

Galactose upstream activation sequence

"AlignAce" NATURE BIOTECHNOLOGY VOLUME 16 OCTOBER 1998

Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation

Frederick P. Roth[1,*], Jason D. Hughes[1,*], Preston W. Estep[1], and George M. Church[1,*]



heat shock
vs.
30°C

→ **Measure mRNA abundances using DNA microarrays** → **Search for motifs in promoters of heat-induced and repressed genes**

Discovered motifs

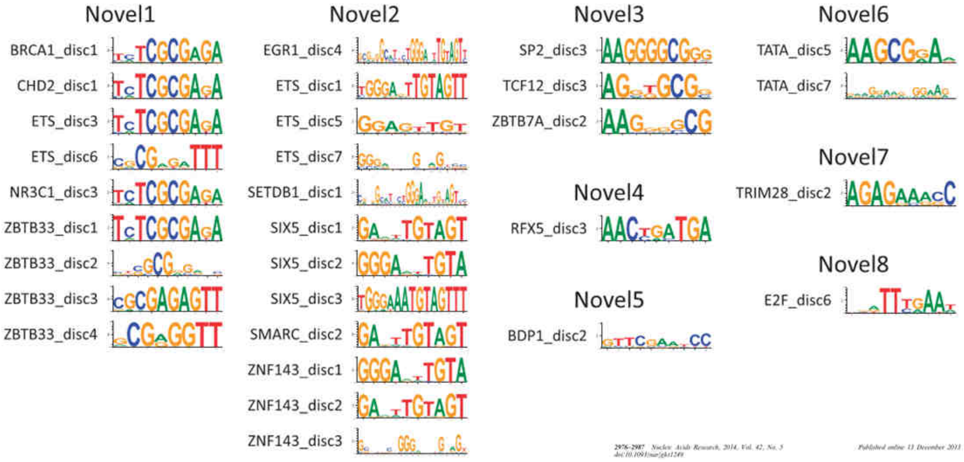
Known motif

Cell cycle activation motif, histone activator

“AlignAce” NATURE BIOTECHNOLOGY VOLUME 16 OCTOBER 1998

If you need them, we now know the binding motifs for 100’s of transcription factors at 1000’s of distinct sites in the human genome, including many new motifs.

e.g., http://compbio.mit.edu/encode-motifs/



2976–2987 Nucleic Acids Research, 2014, Vol. 42, No. 5 Published online 13 December 2013
doi:10.1093/nar/gkt1249

Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments