PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny

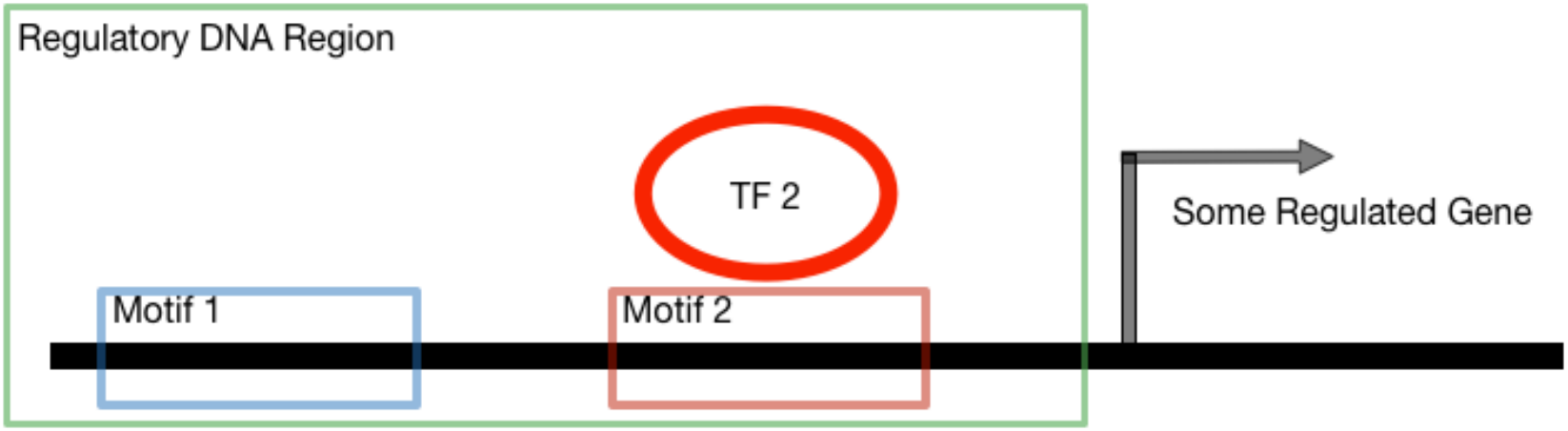
Rahul Siddharthan, Eric D Siggia, Erik van Nimwegen

http://www.imsc.res.in/~rsidd/phylogibbs/

Presentation
by
Bryan Lunt

What is a "Transcription Factor Binding Site"?

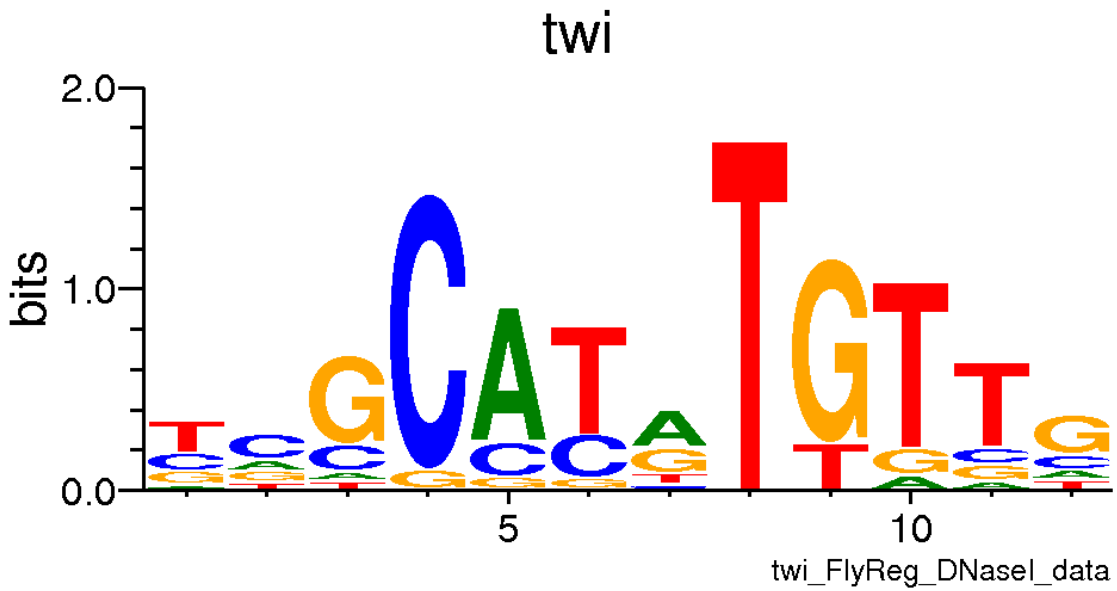
(And why do we care?)



Flow control for the program of life.

- Transcription factors (TFs) regulate the expression of nearby genes.
- TFs have distinct *binding motifs* that differ from the background distribution.
- Functional constraint causes motifs to be well conserved, even under great phylogenetic distance.

Motifs / Position Weight Matrics



[Fly Factor Survey]

Motif Discovery

non-phylo

- Find co-regulated genes. Cut out their upstream sequences.

  thisisatestoftheemergencybroadcastsystemthisisonlyatest

  classesatthisuniversityoftenculminateinatest

  thisisnotapipe [Magritte]

- Search this collection of sequences for common well-conserved blocks.

  thisisatestoftheemergencybroadcastsystemthisisonlyatest

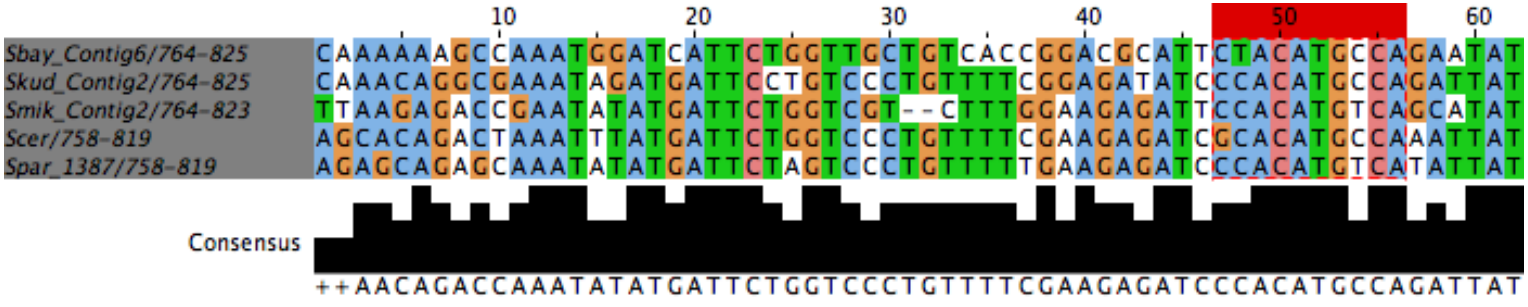
  classesatthisuniversityoftenculminateinatest

  thisisnotapipe [Magritte]

This only works if the sequences have had sufficient time to diverge so far that similarity can only be because of the outside constraint.

The problem of close phylogenetic relationships

This won't work when sequences have not had time to diverge.
Everything is well conserved.

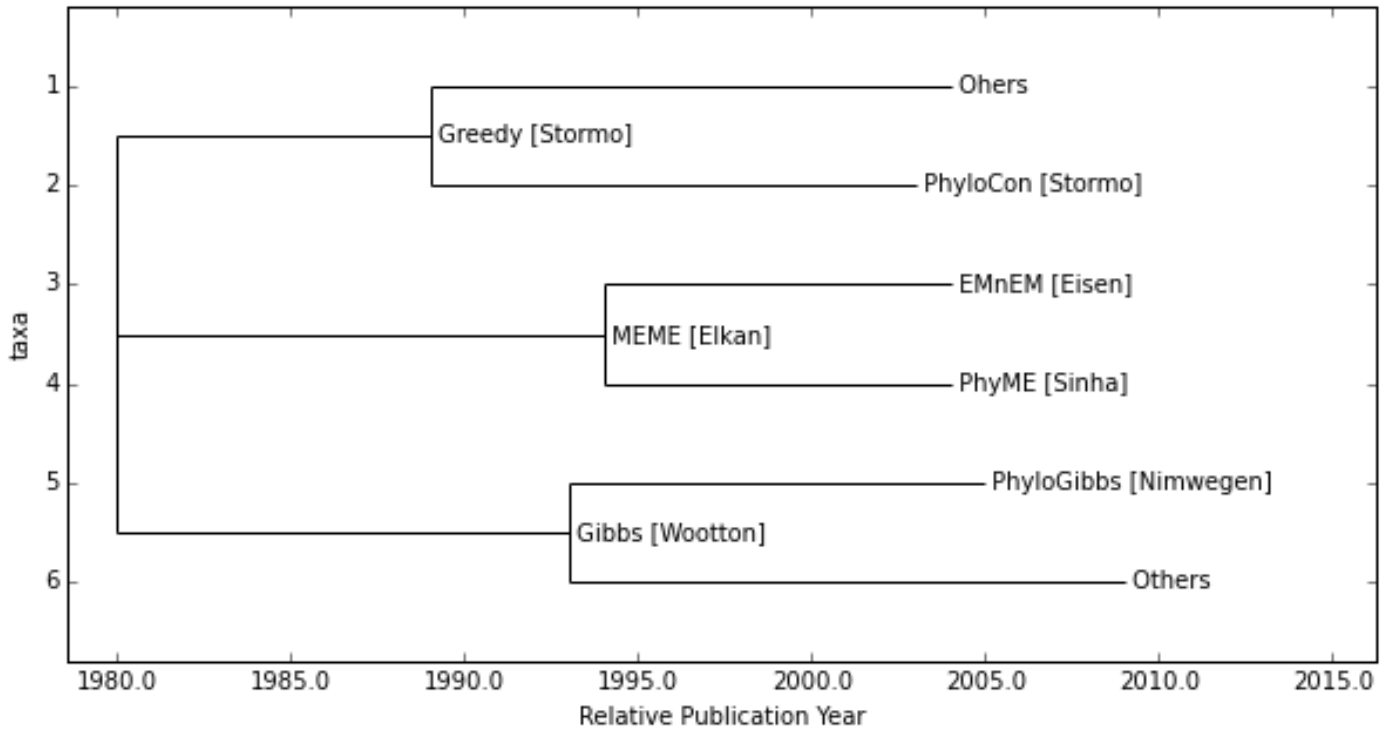


[Subset of data from YBR093C_al.fna shipped with PG code. Realigned with MUSCLE.]

Here, the highlighted part is a motif, the rest is not.

A phylogeny of phylogeny-aware motif finding algorithms.

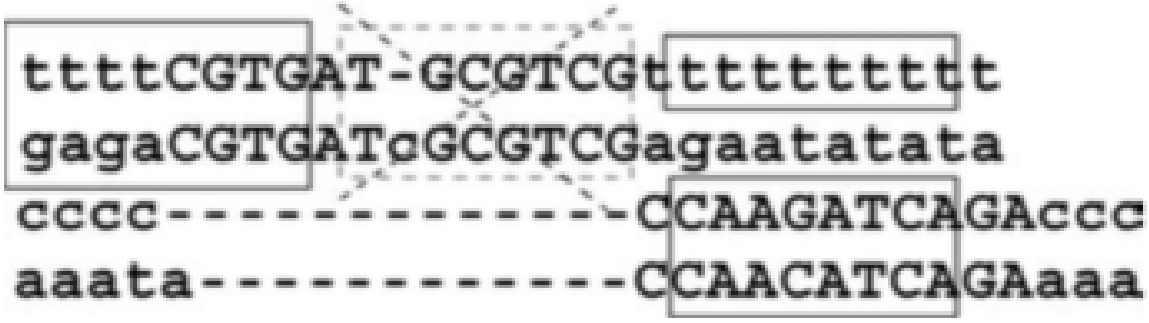
HGT not shown.



Gibbs Sampling

- We want to sample (actually maximize) the posterior $P(C|S)$ when we have S.
- We only have $P(S|C)$ and $P(C)$.
- Appeal to Bayes' rule $P(C|S) \propto P(S|C)P(C)$.
- Use this to compare the scores of different proposed moves (changes in configuration) and move around the state space probabilistically.
  Given a set of proposed moves X, choose a move $Y \in X$ according to $P_{choose}(Y) = P(Y|S)/(\sum_{x \in X} P(x|S))$
  Here, the normalization term of Bayes' rule falls out, so we can calculate this tractably for a finite set of moves.

PhyloGibbs state space



[Paper fig 2]

- Windows in aligned areas must agree.
- Windows may spill from aligned areas to unaligned areas, but the aligned parts must agree.
- Windows in unaligned areas and unaligned sequences are independent.

Workflow

- Simulated annealing to find the best configuration $C*$.
  That's just lowering the temperature until it gets stuck in one configuration.
- Tracking to see how often assignments are the same as $C*$, and to find other windows with high scores.
  Gibbs Sampling to see how often configurations make the same window assignments as does $C*$

Evolutionary model

q : Probability of no mutation from the ancestor.
If there is a mutation, fix to an appropriate equilibrium distribution.
Uses the F81 model for all sites. [Felsenstein 1981]

$$Q_{bg} = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$

$\vec{\pi}$ is the same background for all sites.

$$Q_{W_i} = \begin{pmatrix} * & \omega_{iC} & \omega_{iA} & \omega_{iG} \\ \omega_{iT} & * & \omega_{iA} & \omega_{iG} \\ \omega_{iT} & \omega_{iC} & * & \omega_{iG} \\ \omega_{iT} & \omega_{iC} & \omega_{iA} & * \end{pmatrix}$$

[LaTeX from Wikipedia]

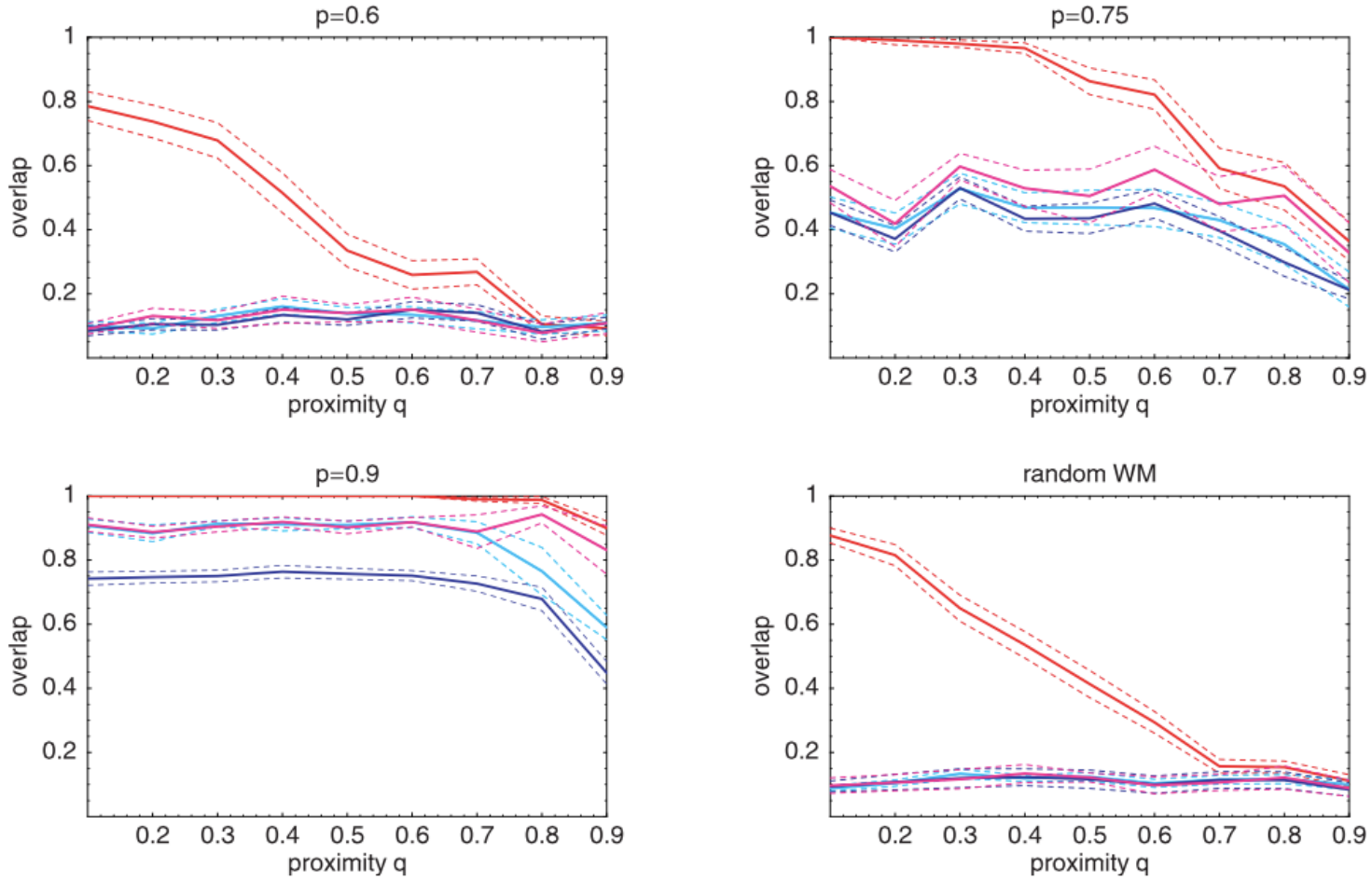
$\vec{\omega_i}$ is a column out of the appropriate PWM.

Extending the math of Saurabh Sinha, for aligned segments they form the probability

$$P(S|C)P(C) = \int_w P(S|w;C)P(w|C)P(C)$$

Various approximations are used to rearrange the phylogeny into an approximating star-phylogeny.
If w were fixed, and we were not integrating over all w, this could be exactly calculated with DP.

Small artificial data



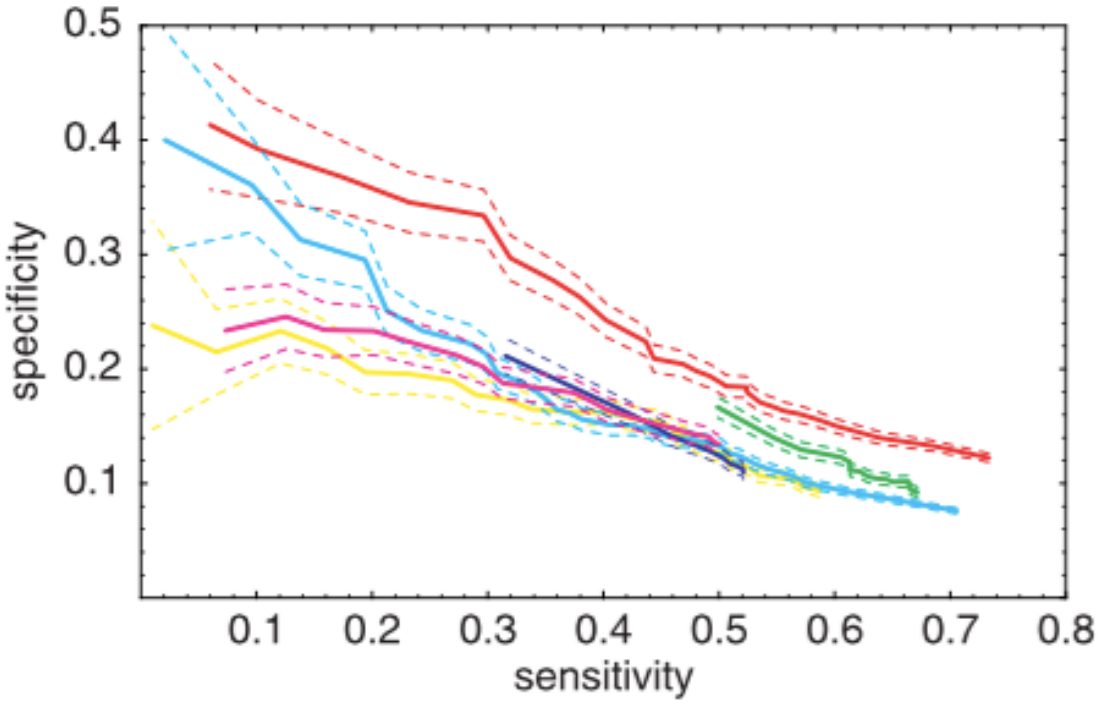
[paper fig 3]

(red) PhyloGibbs ; (light-blue) PhyloGibbs in non-phylo mode ;
(dark-blue) WGibbs ; (pink) MEME

I have trouble believing the other methods are this bad around $q = 0$.

Yeast

- 200 *S. cerevisiae* genes, and orthologs from other yeast.
- 466 experimentally verified sites from 1000 to 0 bp upstream.



[paper fig 6]

(red) PhyloGibbs ; (light-blue) PhyloGibbs in non-phylo mode ;
(dark-blue) WGibbs ; (pink) MEME ; (yellow) EMnEM; (green) PhyME

Major Criticisms

- All experiments used very small phylogenies.
- There was no measurement of the change in accuracy with change in alignment quality.
- They turned *everything* into a star phylogeny.
- Their experiments seem deliberately designed to make other programs look bad. I can't believe figure 3.

links

Code:
http://www.imsc.res.in/~rsidd/phylogibbs/
http://www.imsc.res.in/~rsidd/phylogibbs-mp/
c-REDUCE (Includes a third-party comparison of PhyloGibbs and others)
http://www.biomedcentral.com/content/pdf/1471-2105-9-506.pdf