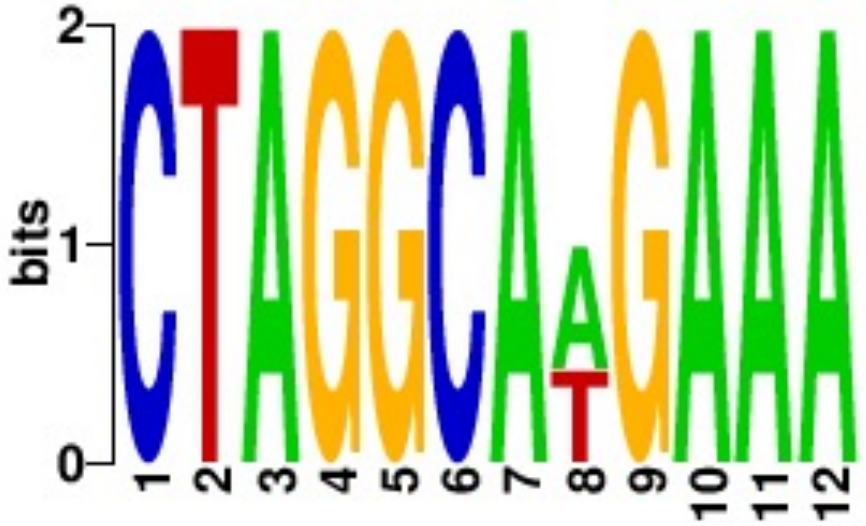
Finding Motifs in TB using Gibbs Sampling

Anirudh Narsipur

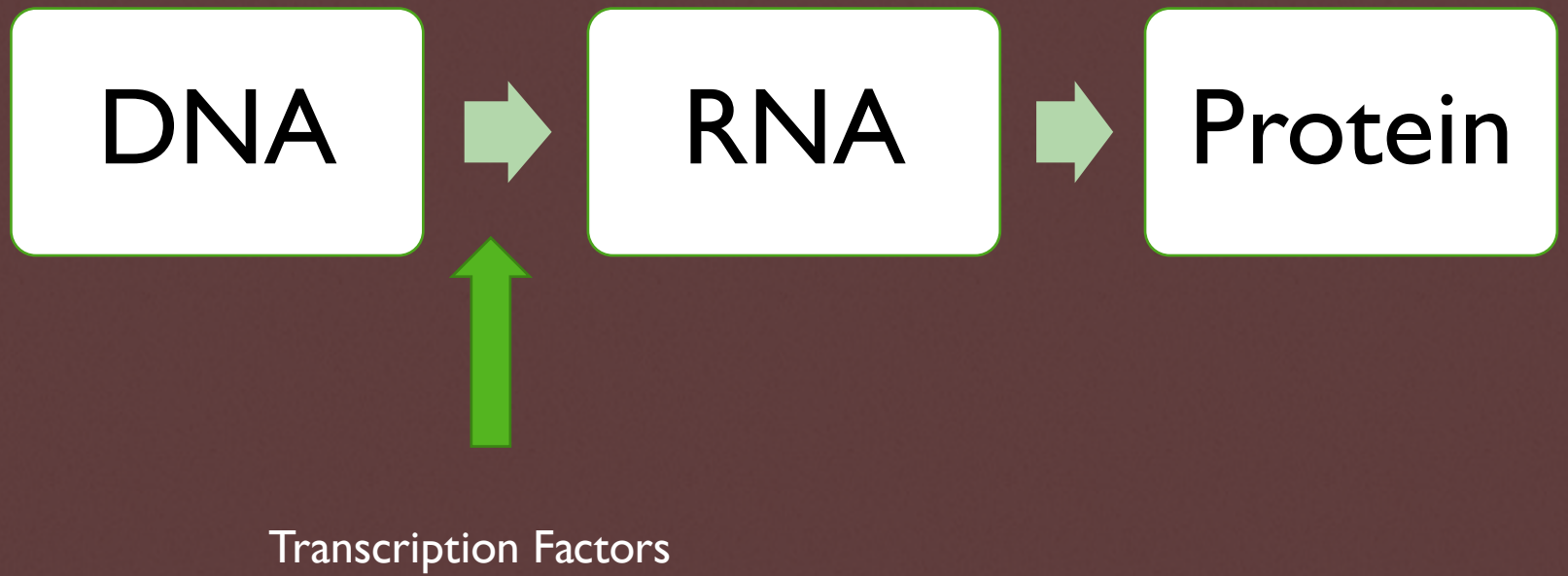


The Path

- Context sensitive HMMs

- miRNA prediction

Some Biology



Why do we care?

- Transcription factors can activate or repress the transcription of a gene.
- By turning genes on and off TFs can direct cell division, growth and death
- Understanding how they work and where they bind is critical to understanding the function and role.

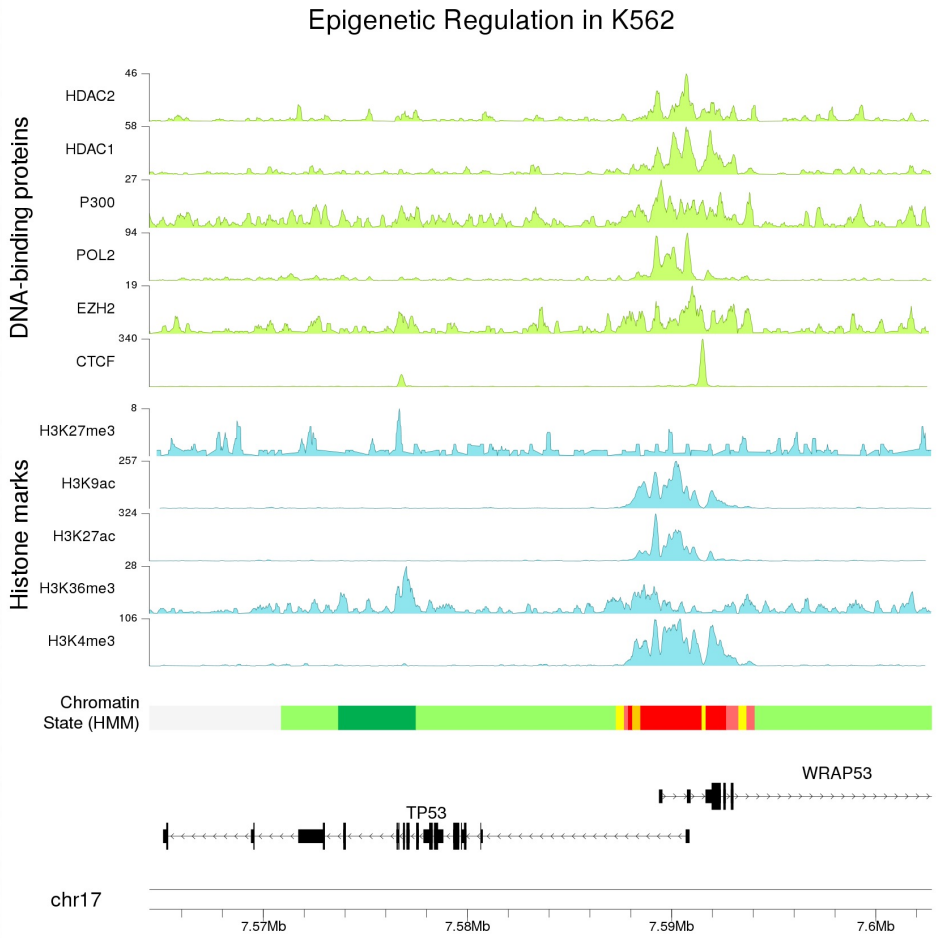
The Problem

- Transcription Factors only bind at specific sites : DNA Binding Domains (DBD).

- We want to identify these motifs to identify TF binding sites, classify TF sites

The Problem

- ChIP-Seq = chromatin immunoprecipitation + DNA Sequencing
- The ”peaks” tell us the location of DNA binding sites
- We can Sequence these peaks



Sequencing gives us….

- TGTCGCTTCTTGAGTAGGCGGGGGCTAACAGCTCAACGATGTC AATGCTCCCTGATCGAA….
- GTTCTGCCATTGCTTCCTTATTATCTGGTCCTTACGGATCTGCC GAACAAAAAGCGGCTA…..

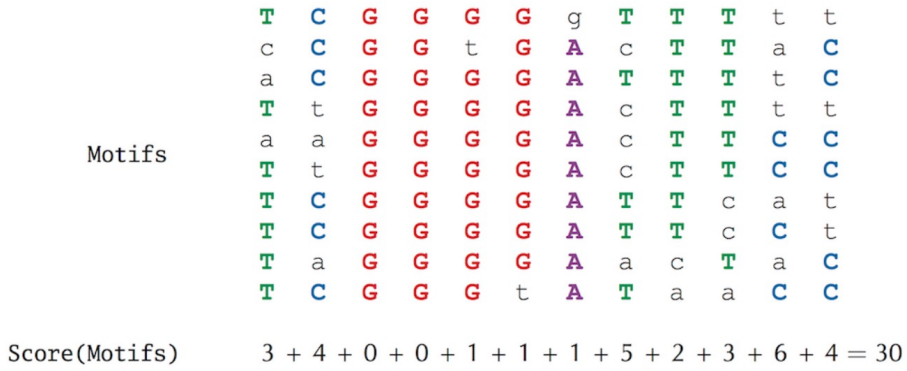
- CACAAACAGCGTACCTCGATTCGGTTTCCTCGCGGGATGTTAT TCCGGTACTGCCCAGGC

Embedded we have ..

- TGTCGCTTCTTGAGTAGGC**ATAGCTA**GGGGGCTAACAGCTCA
- AGTTCTGCATAcCTACATTGCTTCCTTAGGTCCTTACGGATCTGC
- TTTCCTCGCGGGATGTTATTCCGGTACTGCCAgAGCTACAGGC

Let's Get To Computation

- Given p strings of length n
- Find the most similar k-length sequences in the sequences s_1 ,s_2 …. s_p.
- Multiple definitions of similarity. Hamming Distance is a common definition.



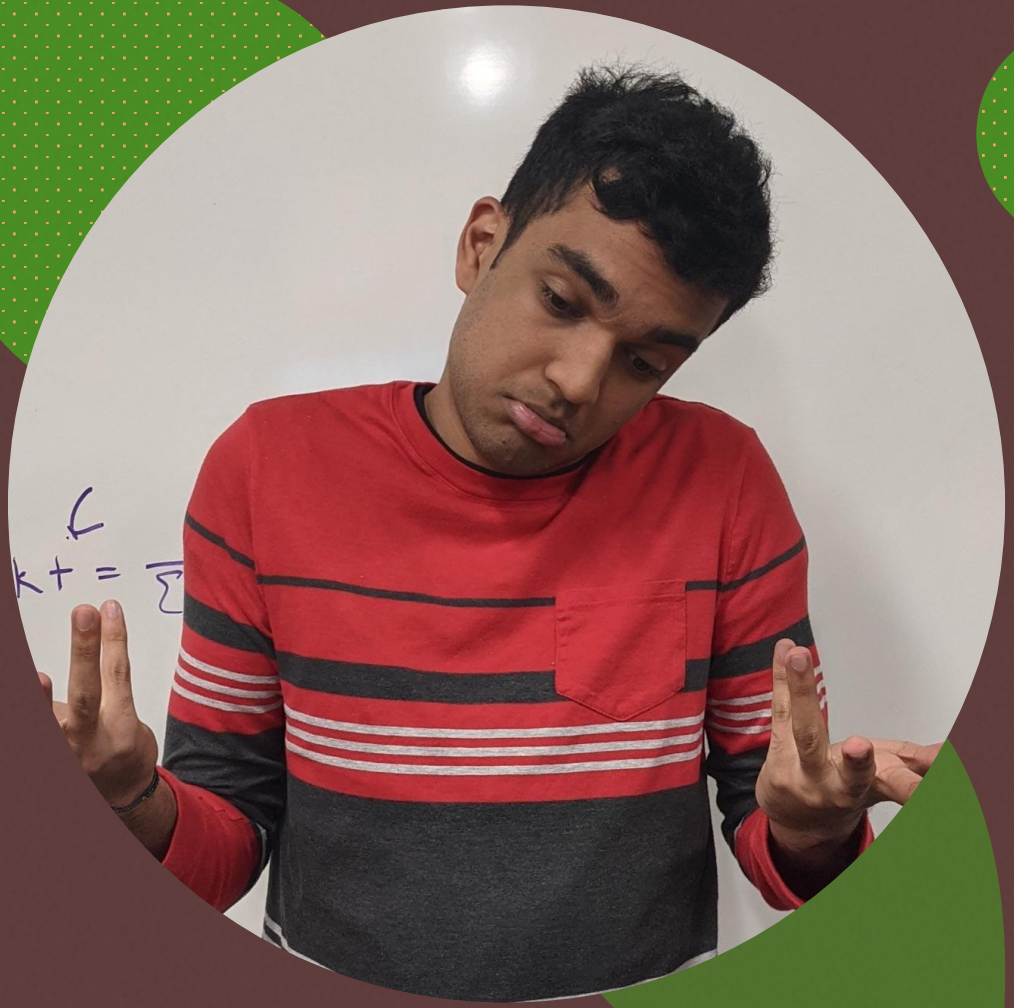
How?

- Brute force ….. ?

- Naïve Alignment … ?

- Greedy Search

- Multiple Sequence Alignment



Just Guess!

Gibbs Sampling

- Randomly pick a k-mer in each of each string s_1 s_2…. s_P :
  - $Motif_1, Motif_2, \ldots\ldots Motif_p$
- Randomly pick s_j
- Create a profile matrix Ω from all motifs **except Motif$_j$**.

- Score all k-mers in s_j using Ω and normalize to get distribution D_j

- Pick new $Motif_j$ by sampling from D_j

- Repeat

Gibbs Sampling

GibbsSampler(*Dna*, *k*, *t*, *N*)

randomly select *k*-mers *Motifs* = ($Motif_1$, …, $Motif_t$) in each string from *Dna*

BestMotifs ← *Motifs*

for *j* ← 1 to *N*

i ← *Random*(*t*)

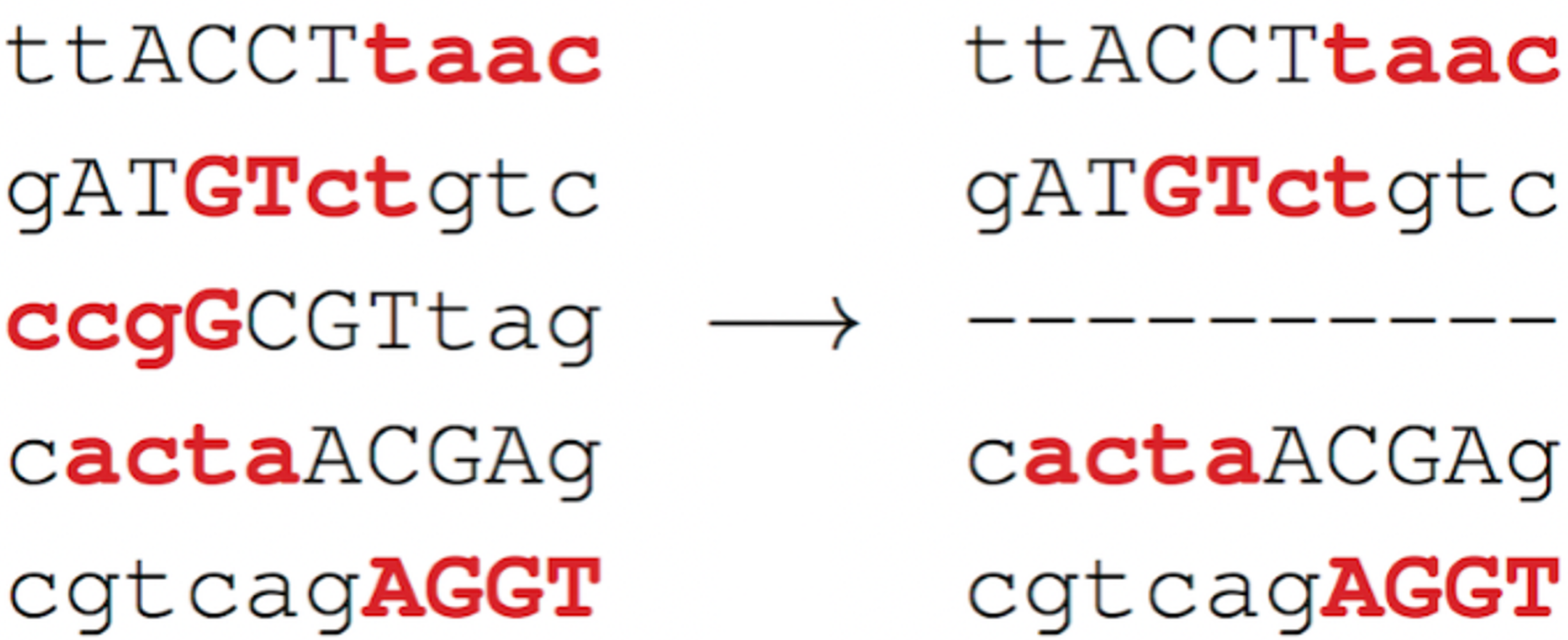
Profile ← profile matrix constructed from all strings in *Motifs* except for $Motif_i$

$Motif_i$ ← *Profile*-randomly generated *k*-mer in the *i*-th sequence

if *Score*(*Motifs*) < *Score*(*BestMotifs*)

BestMotifs ← *Motifs*

return *BestMotifs*


ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT
ttACCTtaac
gATGTctgtc
----
cactaACGAg
cgtcagAGGT

Motifs

t a a c
G T c t
a c t a
A G G T

otifs)

A: 2 1 1 1
C: 0 1 1 1
G: 1 1 1 0
T: 1 1 1 2

Profile(Motifs)

A: 2/4 1/4 1/4
C: 0 1/4 1/4
G: 1/4 1/4 1/4
T: 1/4 1/4 1/4

Count(Motifs)				
A:	3	2	2	2
C:	1	2	2	2
G:	2	2	2	1
T:	2	2	2	3

Profile(Motifs)				
A:	3/8	2/8	2/8	2/8
C:	1/8	2/8	2/8	2/8
G:	2/8	2/8	2/8	1/8
T:	2/8	2/8	2/8	3/8

- Score the string **ccgGCGTtag**
- **Get the distribution**
- **Get a new k-mer from distribution**
- ccgGCGTtag

ccgG	cgGC	gGCG	GCGT	CGTt	GTta	Ttag
$4/8^4$	$8/8^4$	$8/8^4$	$24/8^4$	$12/8^4$	$16/8^4$	$8/8^4$

Random (4/80, 8/80, 8/80, 24/80, 12/80, 16/80, 8/80)

Why does this work? It's random!

Why does this work?

- Consider if everything was random
- But motifs are not random
- Have a statistical bias towards motif
- Through iteration and repitition we get to the motif

Application

Mycobacterium tuberculosis

- M.tb is a species of pathogenic bacteria that causes tuberculosis
- 25% of the world has latent infection
- 10 million fall sick every year
- 1.5 million die every year
- Antibiotic Resistance

Let's Find Motifs

- Chip-Seq dataset from Minch et al.
- Have sequenced peaks and associated TF
- Task is to find Motifs
- Take all TFs with > 10 peaks
- Find motifs through Gibbs Sampling

Rv0022c

Rv0023

We're Done!

No

What is our Background?

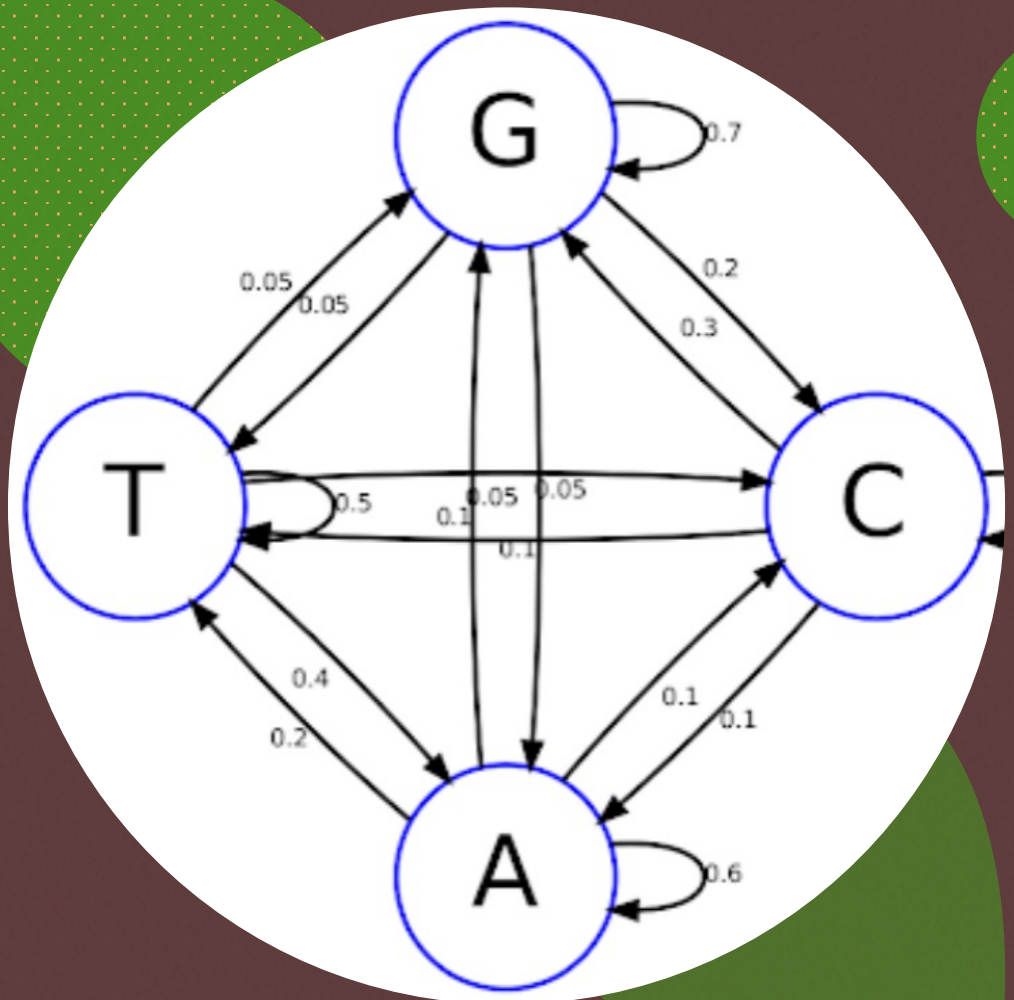
- Nucleotides are not independent
- They are not distributed uniformly
- You have repetitions
- Mtb is heavily imbalanced -> GC content is 65.6%!

The Solution

- Build a Background model B_m of the probability of observing a nucleotide b given the m previous nucleotides

- Can be built from an independent data set (has to be intergenic!) or from the sequences itself

- This is a Markov Chain!
- The value of m gives you the "order" of the model
- You only need to calculate B_m only once.



Integrating Into Gibbs Sampling

- When scoring motifs in the excluded sequence:
- Get the score Q from the profile matrix P_x
- Get score C from background model B_m

- Divide to get A = Q/C

- Sample new motif from $\{A_1, A_2 \ldots A_{n-k}\}$ where n is sequence length, for motif of length k

Profile(Motifs)				
	A:	3/8	2/8	2/8	2/8
	C:	1/8	2/8	2/8	2/8
	G:	2/8	2/8	2/8	1/8
	T:	2/8	2/8	2/8	3/8

Also

- Statistics on information content
- Comparison to existing Results
- Effect of using background model

Thank You!