Motif-Finding in Trypanosomatids

Kelan Wang
Eithon Cadag
CSE527, Aut04, L. Ruzzo

Presentation Agenda

- Introduction to Trypanosomatids and their genomes
- Algorithms to explore motifs in Trypanosomes
- Results and conclusion

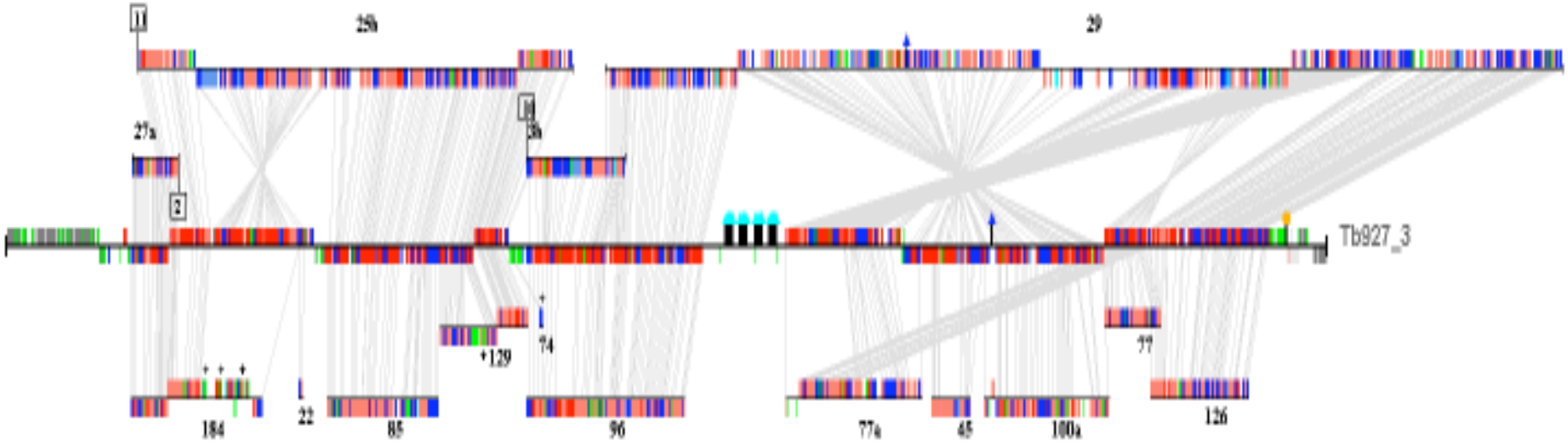
Trypanosomes in the world

- Family of parasites
- Human infective - 12 million affected by *Leishmania* species of Trypanosomes alone
- Infection can be asymptomatic to deadly
- 2 million new cases every year, estimated by World Health Organization

Genome makeup

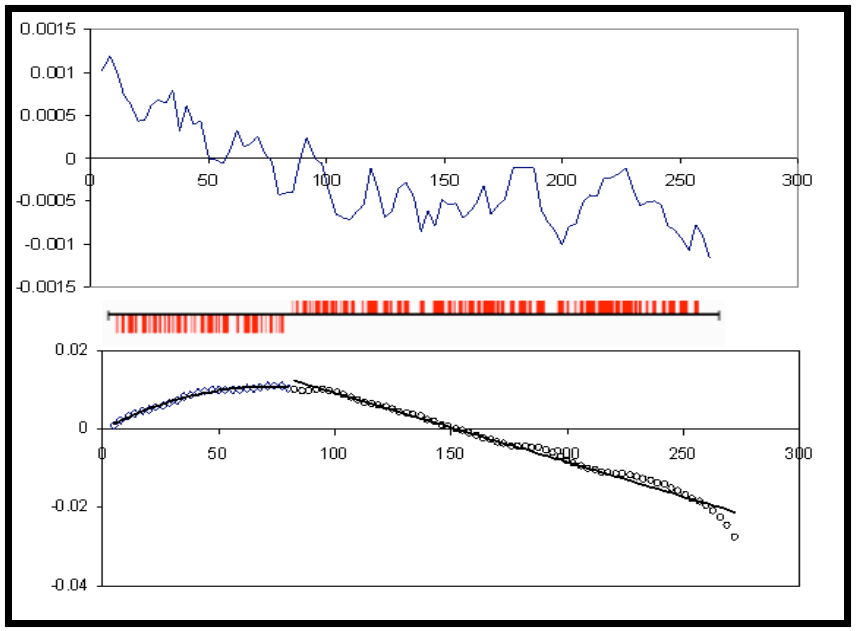
- *Leishmania major*
  - ~33.6 megabase genome
  - 36 chromosomes of sizes 300 to 2800 kilobaes
  - Chromosome 1: ~85 protein-coding ORFs
- *Trypanosoma brucei*
  - ~35 megabase genome
  - 11 chromosomes (larger than *L. major*)
  - Chromosome 1: ~145 protein-coding ORFs
- Overwhelming majority of genomes have been annotated *in silico*

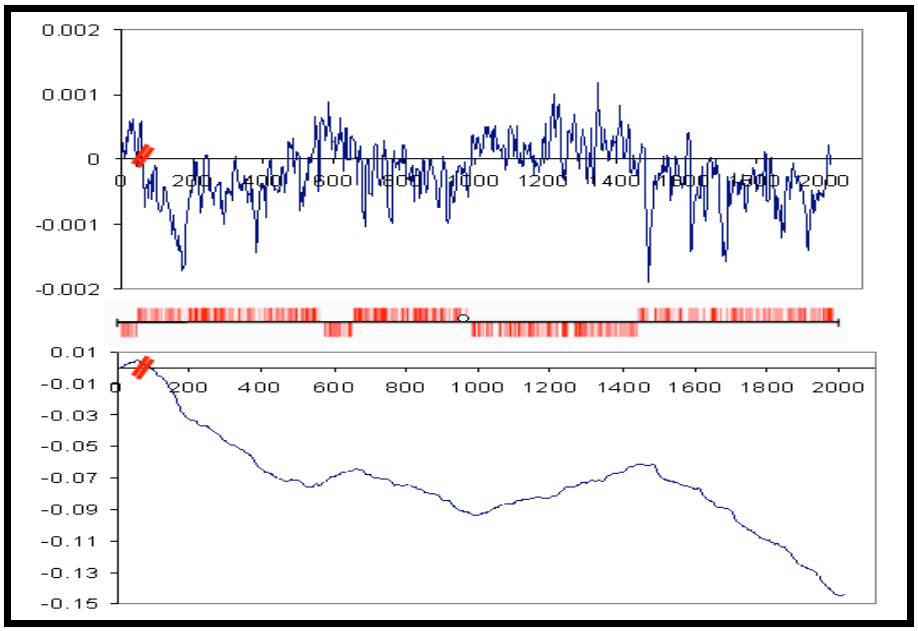
Genome structure



- High conservation between related species
- Very syntenous despite divergence
- What else is shared?

Genome characteristics





- Gene organization follows a polycistronic structure
- Predictable via GC-content skew

Algorithms

- Gibbs Sampling (Lawrence, 1993)
- Variations of Gibbs Sampling
  - AlignACE
  - GLAM (gapless local alignment of multiple sequences)
- Mismatch Tree Algorithm (MITRA)

Gibbs Sampling (review)

- Goal: locate the alignment that maximizes the ratio of the pattern probability to background probability

$$F = \sum_{i=1}^{W} \sum_{j=1}^{4} c_{i,j} \log \frac{q_{i,j}}{p_j}$$

Gibbs Sampling – basic algorithm

- Predictive update step:
  - Choose one random sequence z, and random starting positions within the various sequences.
  - Calculate pattern probability and background probability at current positions
- Sampling step:
  - Calculate probabilities of generating every possible segment of width W within sequence z according to the current pattern probability (Q), and the background probability (P).
  - The weight A = Q/P is assigned to each segment and a random one is selected for the next iteration.

AlignACE

- Based on Gibbs sampling
- Differences:
  - Both strands of the input sequences are considered
  - Simultaneous vs. single motif searching: masking
  - MAP score (maximum *a priori* log likelihood):
    - Degree of which a motif is over-represented relative to the expectation of the random occurrence in sequence
    - Drawbacks: ubiquitous but not relevant motifs

GLAM

- Based on Gibbs sampling
- Bayesian scoring scheme
  - Prior probability distribution
  - Dirichlet function

$$prior\{q_i\} = \frac{1}{Z}\prod_i q_i^{\alpha_i - 1}$$

GLAM – alignment score

- Scoring scheme:

$$S = \sum_{k=1}^{W} \ln \left[\frac{\dfrac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)} \dfrac{\prod_i \Gamma(c_{ki} + \alpha_i)}{\Gamma(N + A)}}{\prod_i p_i^{c_{ki}}} \right]$$

GLAM – resizing

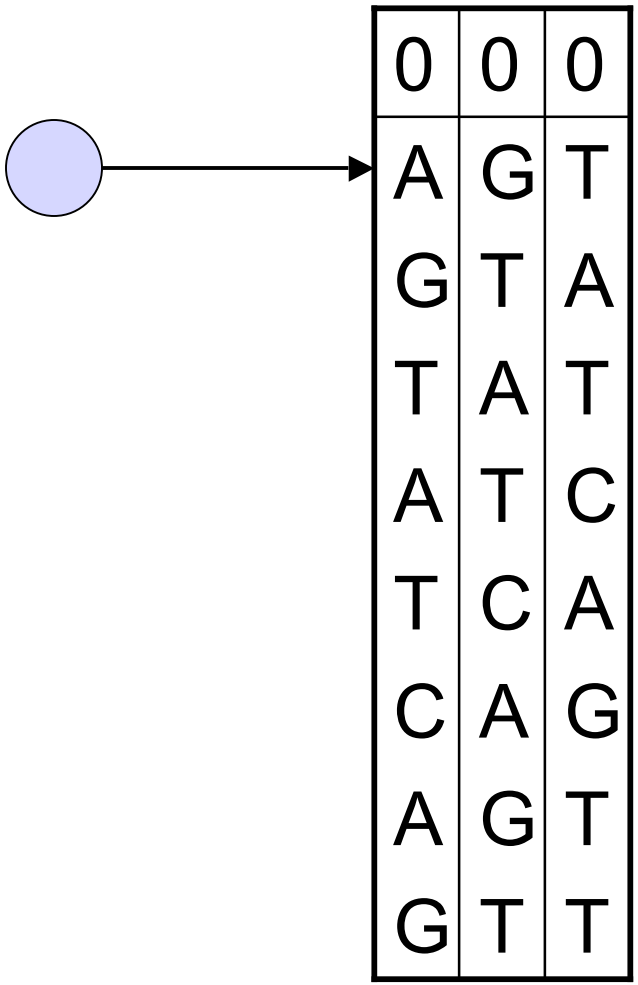
- Automatic adjustment of width of alignment
  - Fix left ends and right ends are varied
  - Fix right ends and left ends are varied
- Over the problem of fixed width algorithm where end points are shifted left or right relative to the optimal

MITRA

- Definitions:
  - search of all L-mers (a continuous string of length L) that occur with up to d mismatches in at least k sequences in the sample S.
  - Weak pattern: has less than k (L, d)-neighbors (all possible L-mers with up to d mismatches as compared to the canonical pattern) in the sample
  - Weak subspace: all patterns are weak
- Data Structure:
  - Rooted tree where each node has 4 branches {A, C, G, T}
  - Maximum depth: L

MITRA – algorithm: 1

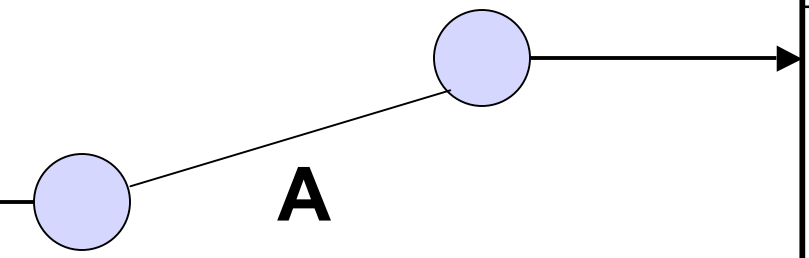
Search for (8, 1) motif for a sequence **AGTATCAGTT**



Initial State

MITRA – algorithm: 2

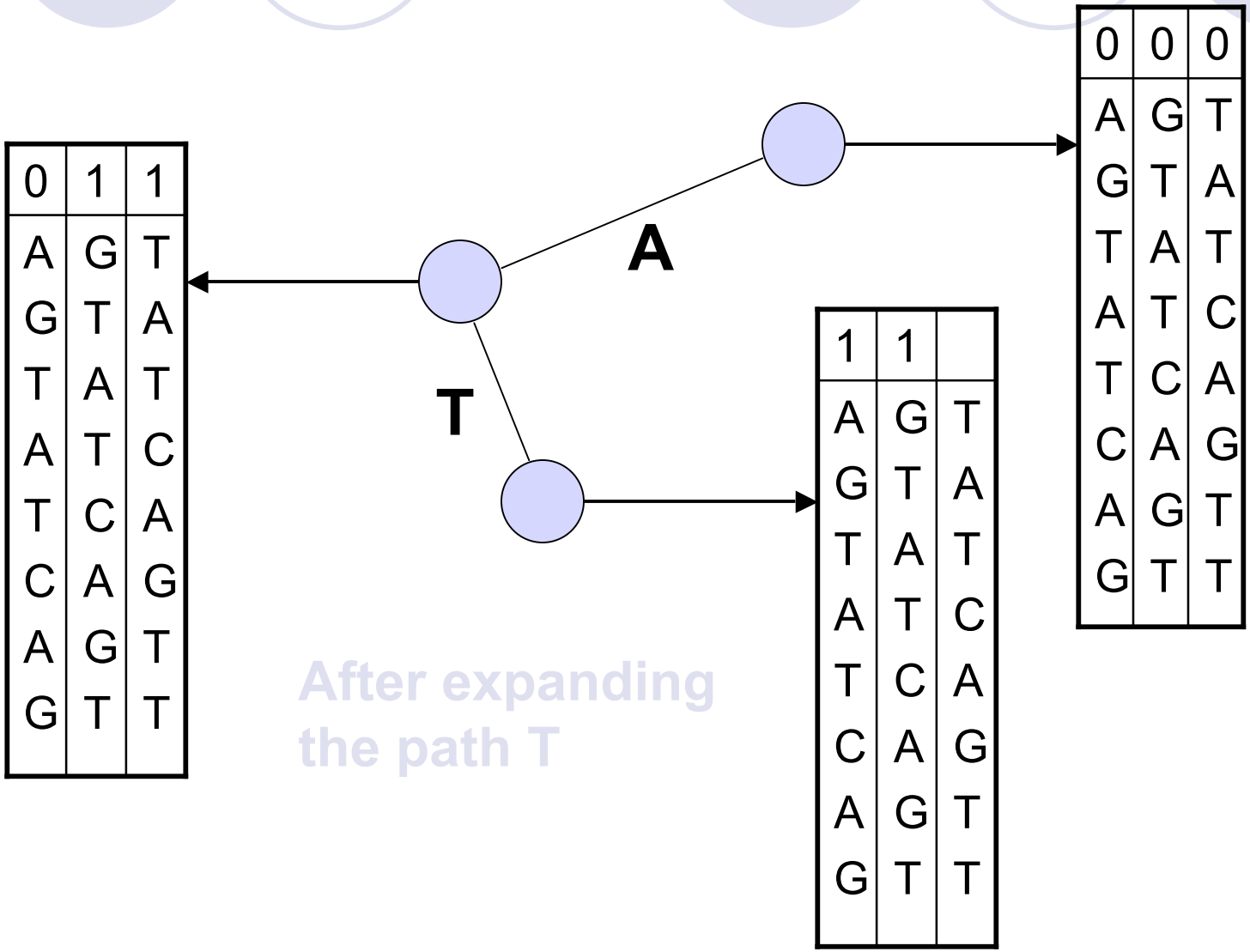
0	1	1
A	G	T
G	T	A
T	A	T
A	T	C
T	C	A
C	A	G
A	G	T
G	T	T



0	0	0
A	G	T
G	T	A
T	A	T
A	T	C
T	C	A
C	A	G
A	G	T
G	T	T

After expanding the path A

MITRA – algorithm:3



MITRA-graph

- Pairwise similarity match
  - Graph G(P, S)
  - Vertex: L-mer in the sample
  - Edge: if two L-mers are similar
  - Subspace is empty if clique of size k does not exist
- More efficient pruning of mismatch tree

In search of common motifs

- Ran GLAM, AlignACE and MITRA motif-finding programs on upstream non-coding regions of annotated genes
  - GLAM: http://zlab.bu.edu/glam/ (binary)
  - AlignACE: http://atlas.med.harvard.edu/ (binary)
  - MITRA: http://fluff.cs.columbia.edu:8080/domain/mitra.html (webpage)
- Parsed out sequences, generated WMMs
- Hypothesis: outstanding motifs will appear in 2-3 of the algorithms (detection via consensus/overlap)

AlignACE results

- Data profile:
  - Max = 5272.33 (LmjF)
  - 21 motifs in LmjF, 100 motifs in Tb
- Repetitious motifs frequent in short windows
- Large number of simple repeating sequences (e.g. ACACAC.., AGAGAG..)

AlignACE WMM

Motif #8:

	A	C	T	G	N		
1	0.03	0.38	0.03	0.57	0.00	G	
2	0.28	0.20	0.19	0.33	0.00	*	
3	0.22	0.26	0.25	0.27	0.00	*	
4	0.01	0.18	0.01	0.80	0.00	G	
5	0.06	0.38	0.16	0.41	0.00	*	
6	0.09	0.37	0.00	0.54	0.00	G	
7	0.25	0.21	0.29	0.24	0.00	*	
8	0.04	0.14	0.03	0.80	0.00	G	
9	0.00	0.77	0.00	0.23	0.00	C	
10	0.20	0.13	0.00	0.66	0.00	G	
11	0.00	0.52	0.12	0.35	0.00	C	
12	0.11	0.36	0.00	0.53	0.00	G	
13	0.08	0.41	0.09	0.42	0.00	C/G	

Trypanosome motifs in AlignACE

- Commonalities
  - G*G*G.. repeating pattern common to both *L. major* and *T. brucei*
  - Generally of type GAGA or GCGC
- Differences
  - *T. brucei* possessed high-scoring relatively complex repeating sequences, while *L. major* did not

Complicated reoccurring motifs in *T. brucei*

MOTIF #1:

	A	C	T	G	N		
1	1.00	0.00	0.00	0.00	0.00	A	
2	0.00	0.00	0.00	1.00	0.00	G	
3	0.00	0.00	1.00	0.00	0.00	T	
4	1.00	0.00	0.00	0.00	0.00	A	
5	0.00	1.00	0.00	0.00	0.00	C	
6	1.00	0.00	0.00	0.00	0.00	A	
7	0.45	0.48	0.00	0.07	0.00	C/A	
8	0.00	0.00	0.00	1.00	0.00	G	
9	0.00	1.00	0.00	0.00	0.00	C	
10	0.44	0.00	0.05	0.51	0.00	G/A	
11	0.51	0.04	0.01	0.44	0.00	A/G	
12	0.01	0.00	0.48	0.51	0.00	G/T	
13	0.48	0.00	0.51	0.00	0.00	T/A	
14	0.51	0.01	0.48	0.00	0.00	G/A	
15	0.48	0.51	0.00	0.00	0.00	C/A	
16	0.51	0.40	0.00	0.09	0.00	A/C	
17	0.00	0.52	0.48	0.00	0.00	T/G	
18	0.06	0.00	0.00	0.93	0.00	G	
19	0.00	0.99	0.00	0.00	0.00	C	

GLAM results

- Data profile:
  - Max = 2945.25 (*Tb)*
  - Lowest = 2072.79 (*LmjF*)
  - 30 total alignments found in *Tb* and *LmjF*
- AGAG.., ACAC.., patterns reoccur in *L. major*
- Much less variability found than from AlignACE

Trypanosome motifs in GLAM

- Commonalities
  - Very few at the sequence level
- Differences
  - *L. major* dominated by alternating bases
  - *T. brucei* dominated by repeating adenine sequences (possibly poly-A tails?)

GLAM: *L. major* vs. *T. brucei*

```
AGAGAGAGAGAGAGAGAGAG
GGAGAGAAAGACAAGCGGAG
GCAGAGAGAGAGAGAGAGAC
AGAACGAGAGACGGACACAG
ACGACGAGAGAGAGAGAGAA
AGAAAGAGAGACGGAGAGAG
GCGGAGAAAGGGAAAAGGAG
GCAGAGCGAGAGAGCGAGAG
ACAGAGAGAGAGAGAGAGAG
ACACAGAGAGAGAGAGAGAG
ACAGCGCGAGACAGAGGGAG
AGAAAGAGAGAGAAGAGGCA
AGAGAGAGAGAGAGAGAGAG
AGAGAGAGAGAGAGCGGGGG
AGAAGAAGAAGAAGGGGGAG
AGAAAGAGAGAGAGAGAGAG
AGGAAGAGAGAAAGAGAGAA
AGAGAGAGAGAGAGAGAGAG
AGAGAGAGAGAGAGAGAGAG
```

L. major

```
AAAAAAGAAAAGAAAAACAA
CAAGAAAGAAAAAAAAAAAA
GAAAAAAAAAAGAAAAAGAA
AAAAAAAGAAAGGAAAAGAA
AAGAAAAGAAAAGAAAAGAA
GAAAGAAAAAAAAAAGACAG
AAGAAAAAAAAGGAAAAGGA
GAAAGAAATAATGAAGAGAA
AAAAAAAAAAAGAGAAAGAA
AAAAAAAAAAAGAAAAAGAA
AAAGGAGGAAAAAACAAAAG
GAAAAAACAAAAGAAAACAT
AAAAGATATAAAAGAGGAAA
CAAAAAAAAAAAGTAGAGGC
GAAAAAGAAAAAGAAGGGGA
CAAAAAATAAATGGCAACAA
GAAAAAAACAAAAGAAAACA
GAAAGATATAAAAGAGGAAA
```

T. brucei

MITRA results

- Data profile:
  - Web interface returned A LOT of data
  - Max = 35.0
- Motifs mostly ATAT variants
- Some ACAC, AGAG seen as in AlignACE and GLAM
- Notable limitation - web interface had sequence size limitation

Trypanosome motifs in MITRA

- Commonalities
  - NTNT, NTTNTT patterns
- Differences
  - *T. brucei* included results with repeating adenine's (AAAANA, etc)

	A	C	T	G	N		
1	0.25	0.25	0.25	0.25	0.00	*	
2	0.50	0.25	0.25	0.00	0.00	A	
3	0.00	0.00	0.75	0.25	0.00	T	
4	0.25	0.00	0.25	0.25	0.25	*	
5	0.25	0.25	0.25	0.00	0.25	*	
6	0.25	0.00	0.75	0.00	0.00	T	
7	0.25	0.25	0.25	0.25	0.00	*	

Characterizing possible motifs across two genomes

- CACA, GAGA patterns very common in both genomes from all algorithms
  - GAGA.. possible true motif within both genomes
  - CACA.. perhaps - maybe altered poly-A tail?
- Extremely high scores for possible motifs
  - AlignACE, GLAM had scores upwards of 2000+

Altering the experiment

- Stuff to try in the future
  - Use more of the genome, once its completely annotated officially
  - Inclusion of other highly-conserved species in the Trypanosoma family
    - *L. infantum, T. cruzi*, etc.
  - Prune out possible poly-A regions to refine searching
  - Apply or alter other algorithms to increase breadth of search
  - Connect found motifs to gene function - is there a relation?

Questions?