CSE 527
Lecture 15

More on the Gibbs Sampler

Projects

- Individual or small group
- Literature: pick 3-5 papers on a coherent topic & give me a report on them, OR
- Implementation: 1-2 background papers + implement & test

Deliverables

- send me a paragraph per group outlining topic, initial paper picks, implementation & test data (if any), preferably before Thanksgiving
- Use class email if desired to brainstorm, form groups
- give me oral presentation (20-30 minutes) + written report (~5 pages) sometime during finals week.

Half-baked Ideas

- Gibbs vs MEME
- Gibbs greedy vs sampling
- Rule-based or other approach instead of k-NN for functional classification
- Microarray Normalization
- Evaluation of Microarray Normalization
- “FOM” alternative in Datta^2 (HW2)
- Try favorite motif finder on favorite organism
- ... ...

AlignAce (Roth, et al. 1998)

- Lawrence et al.: protein motifs
- Roth et al.: DNA regulatory motifs
- Differences:
  - Genomic background model,
    e.g. yeast Saccharomyces cerevisiae is 62% A-T
  - both strands used
  - overlapping sites prohibited
  - Multiple motifs: find best & mask
  - “MAP” scoring

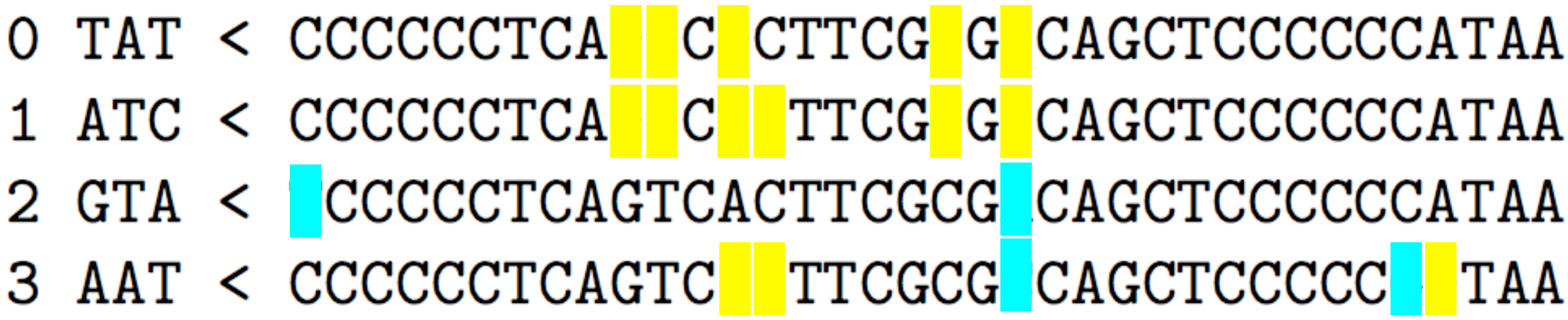
Rocke & Tompa (Recomb ‘98)

- Gibbs, adapted for gapped motifs in DNA

Why Gaps

- Biology often tolerates diversity
- 2 similar TFs bind 2 similar sites
- Same TF binds 2 sites (perhaps one better than the other)
- Dimeric TFs often “don’t care” in middle & flexible
- TF and/or DNA may twist/bulge

A Gapped Motif



```
0 TAT < CCCCCCTCA  C CTTCG G CAGCTCCCCCCATAA
1 ATC < CCCCCCTCA  C  TTCG G CAGCTCCCCCCATAA
2 GTA <  CCCCCTCAGTCACTTCGCG CAGCTCCCCCCATAA
3 AAT < CCCCCCTCAGTC  TTCGCG CAGCTCCCCC  TAA
```

Why gaps are hard

- Alignment
  - Pairwise -- O(n^2)
  - Multiple -- O(n^k)

  dynamic programming

  - Gibbs/MEME/... require *many* alignments
- Scoring

R/T Approach - Scores

- WMM
- Relative entropy, aka expected LLR
- Score gaps like background, “minus a small penalty”

R/T Approach - Alignment

- Gibbs replaces 1 string per iteration
- Use pairwise alignment between new string and previously computed alignment of remaining k-1
- Actually align motif against whole genome - Time O(genome length x motif width)

R/T Approach - Gibbs

- discard 0-2 random strings at each iteration
- pick replacement greedily, not by sampling; avoid local max by random restarts (see Rocke's thesis for more on this)

Test Data

- Haemophilus influenzae
- ~1.8 megabases
- Delete all protein-coding, leaves ~ 350 kb
- Concatenate, separated with markers
- Plus reverse complement, total ~ 700 kb

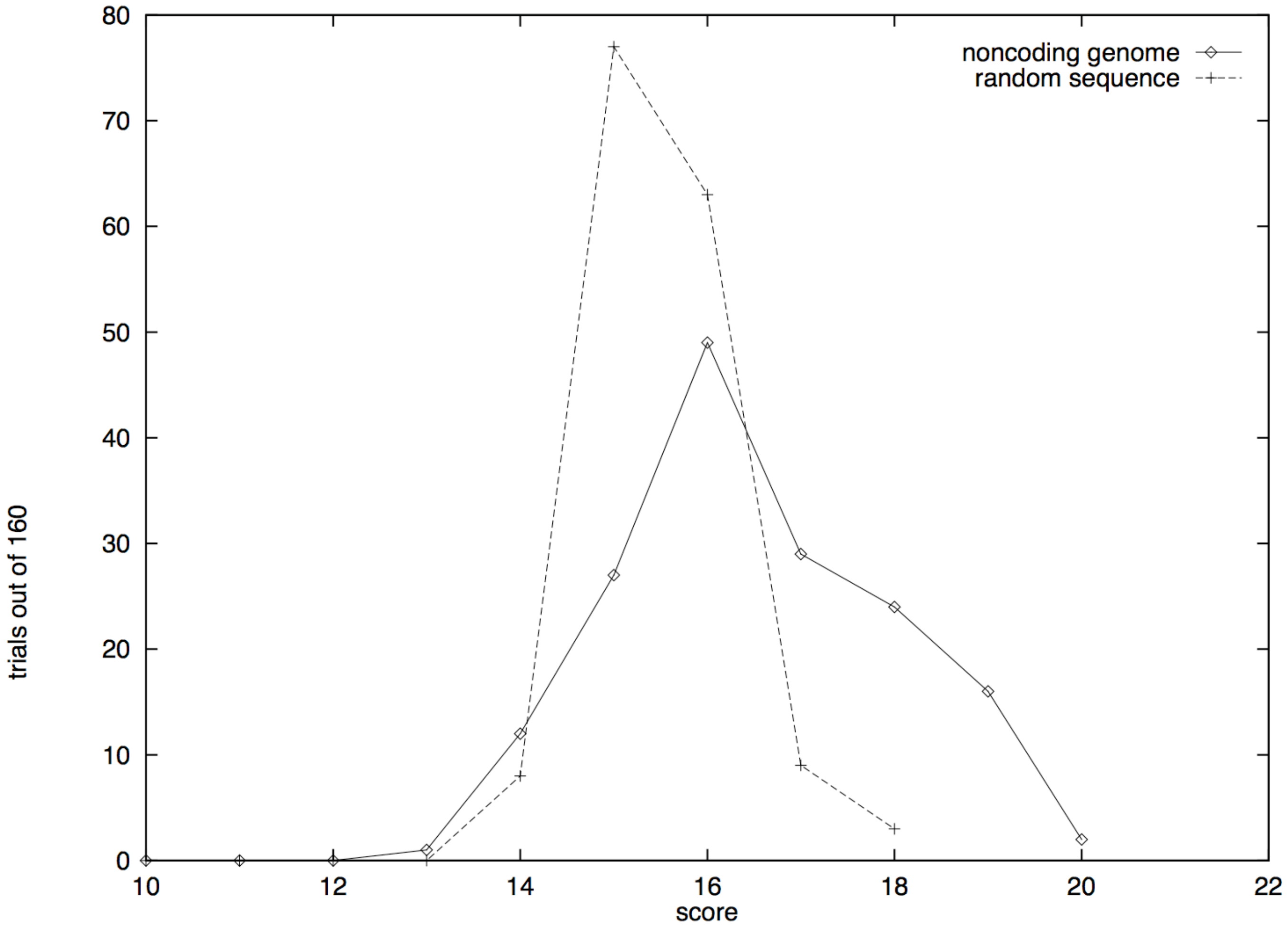
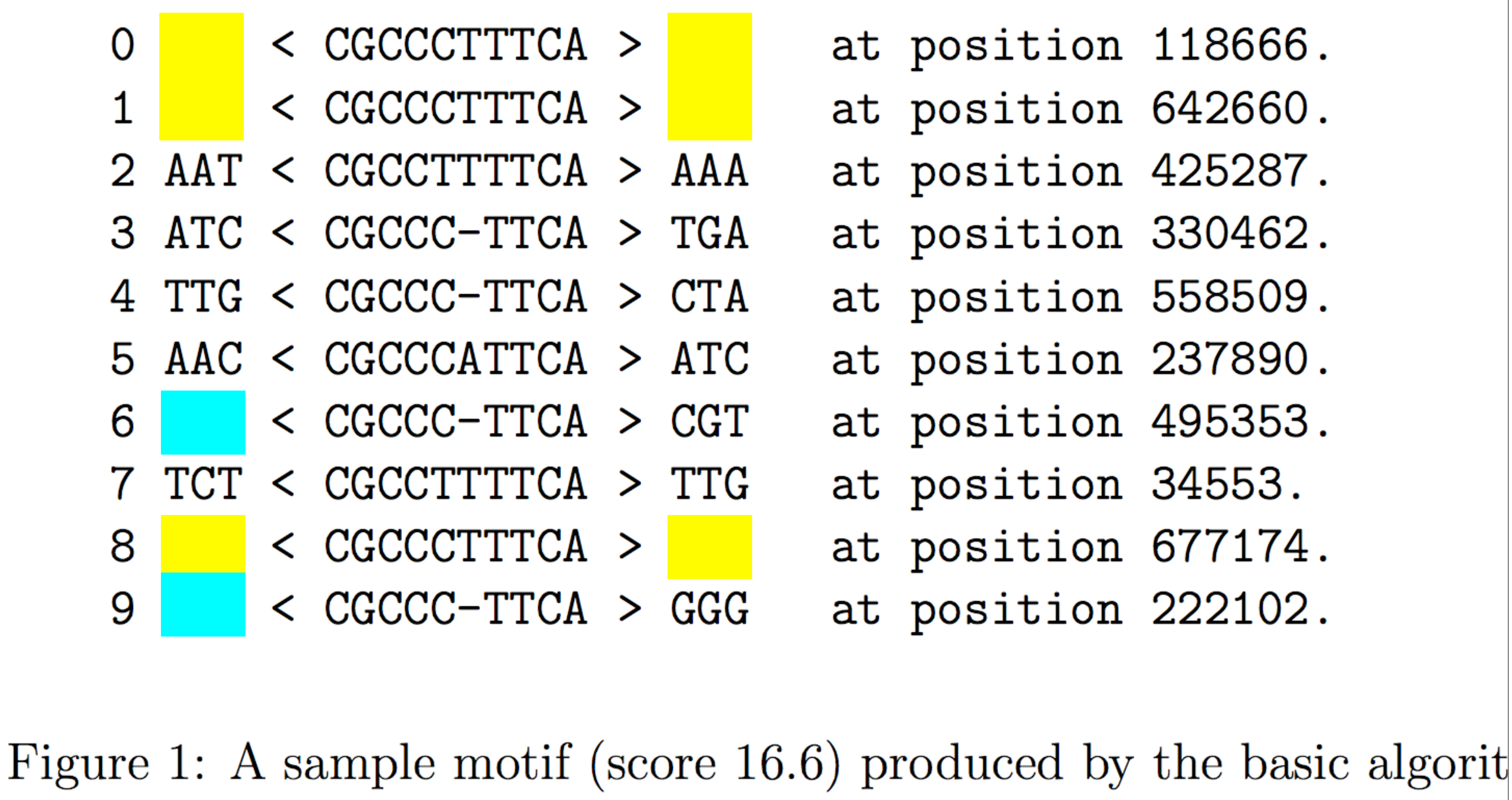


Figure 2: 160 trials of the basic algorithm on the noncoding genome vs. a random sequence

A Motif + Context



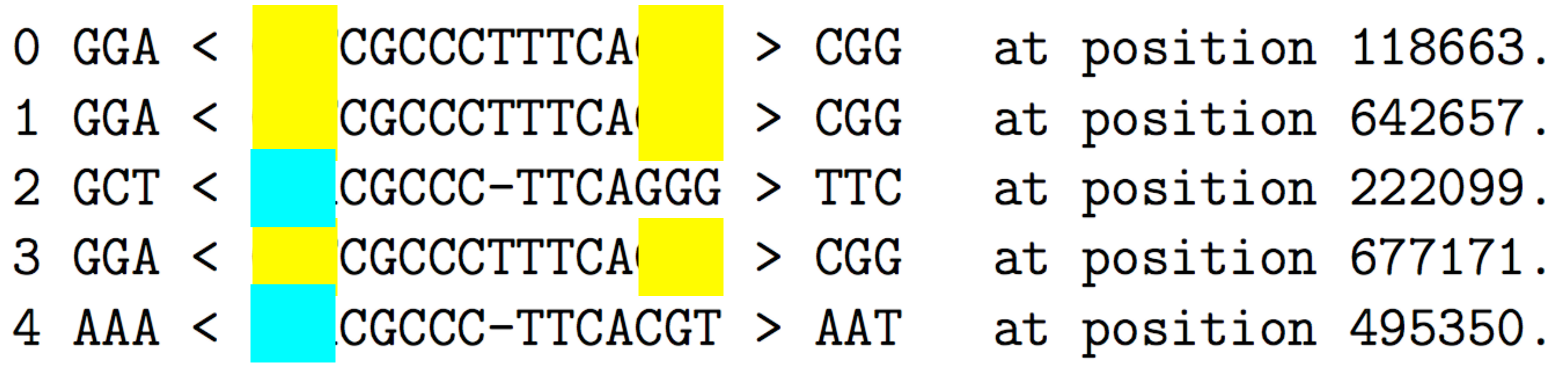
```
0 🟨  < CGCCCTTTCA > 🟨   at position 118666.
1 🟨  < CGCCCTTTCA > 🟨   at position 642660.
2 AAT < CGCCTTTTCA > AAA  at position 425287.
3 ATC < CGCCC-TTCA > TGA  at position 330462.
4 TTG < CGCCC-TTCA > CTA  at position 558509.
5 AAC < CGCCCATTCA > ATC  at position 237890.
6 🟦  < CGCCC-TTCA > CGT  at position 495353.
7 TCT < CGCCTTTTCA > TTG  at position 34553.
8 🟨  < CGCCCTTTCA > 🟨   at position 677174.
9 🟦  < CGCCC-TTCA > GGG  at position 222102.
```

Figure 1: A sample motif (score 16.6) produced by the basic algorit

Rewindowing

- After convergence, “rewindow” -- choose subset of rows and adjust left/right boundaries to maximize score.
- NP-hard? Use another greedy heuristic

Rewindowing



```
0 GGA <     CGCCCTTTCA     > CGG   at position 118663.
1 GGA <     CGCCCTTTCA     > CGG   at position 642657.
2 GCT <     CGCCC-TTCAGGG  > TTC   at position 222099.
3 GGA <     CGCCCTTTCA     > CGG   at position 677171.
4 AAA <     CGCCC-TTCACGT  > AAT   at position 495350.
```

Figure 3: The motif of Figure 1 after rewindowing (score 20.8)

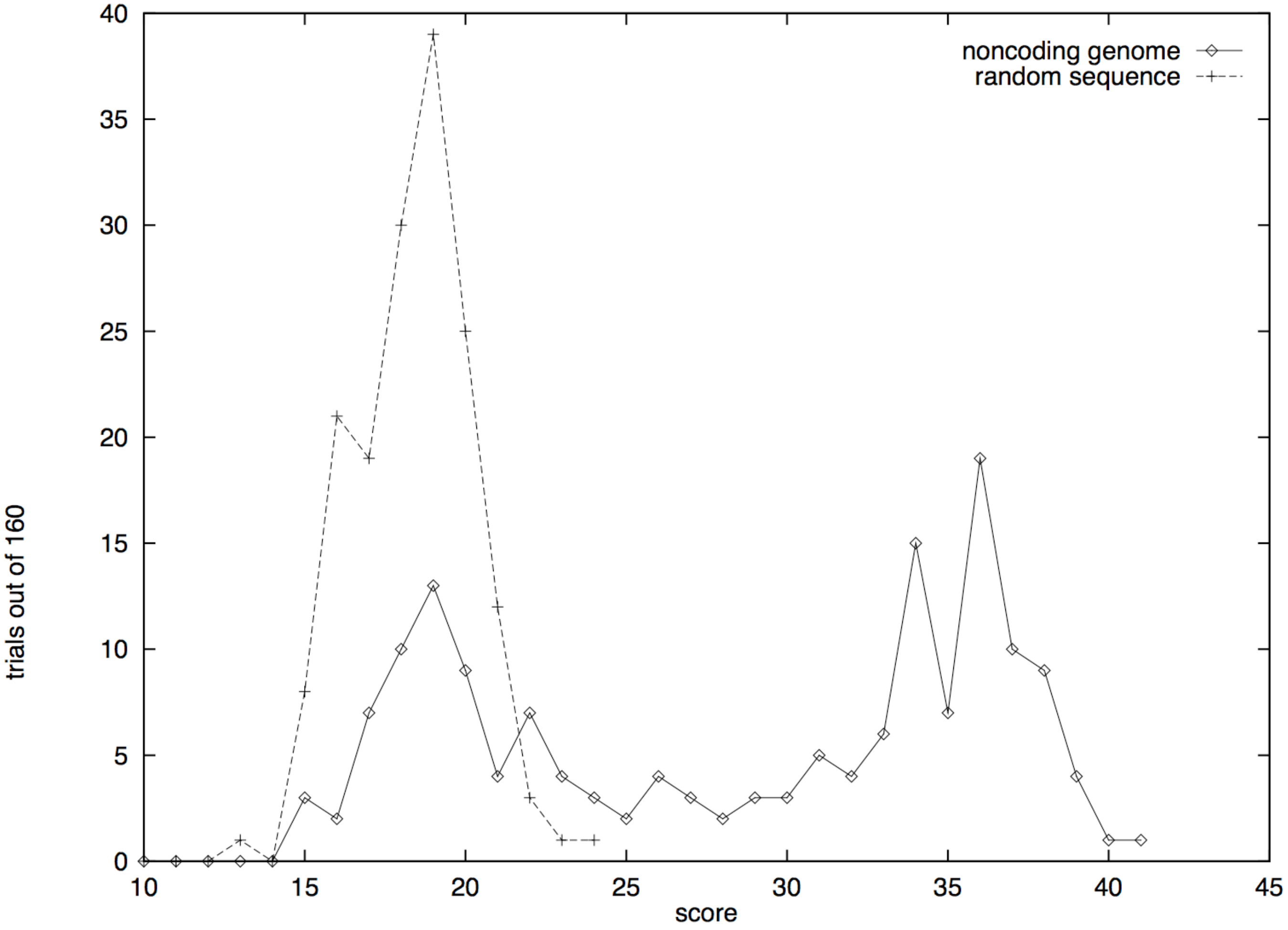


Figure 4: 160 trials of the two-phase algorithm on the noncoding genome vs. a random sequence

A closer look at 35

- 6 almost perfectly identical regions of 5.3 kb, each 3 rRNA genes plus some tRNA genes
- 9% of genome but 50% of high-scoring motifs
- removed 80kb containing them & re-ran

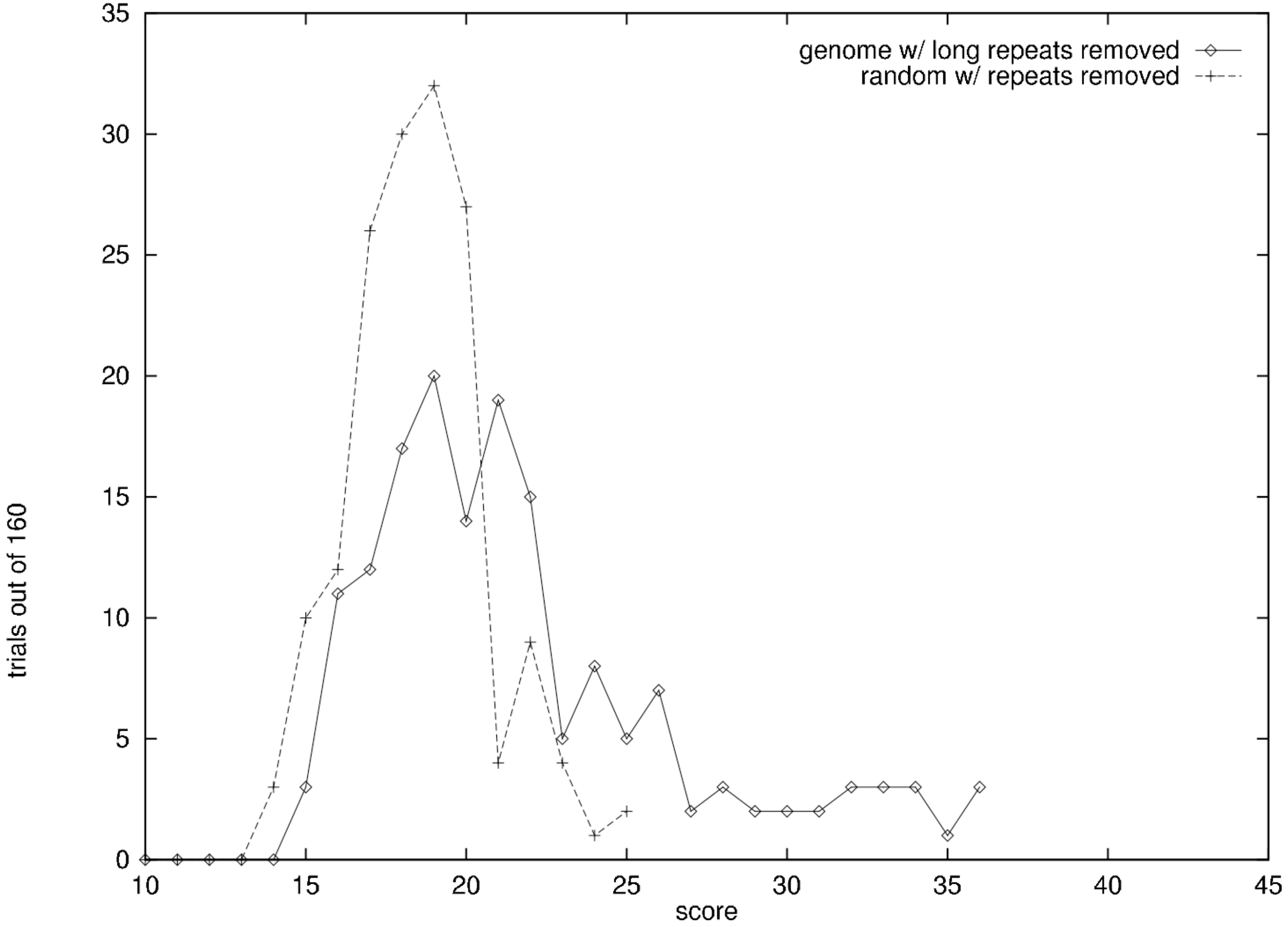
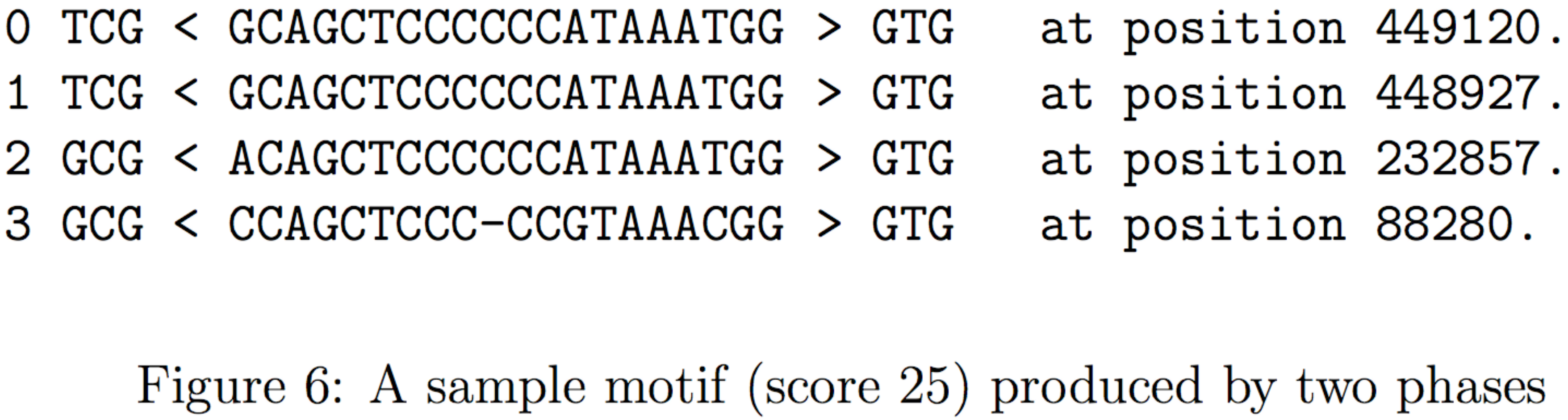


Figure 5: 160 trials of the two-phase algorithm on the noncoding genome with long repeats removed vs. a random sequence

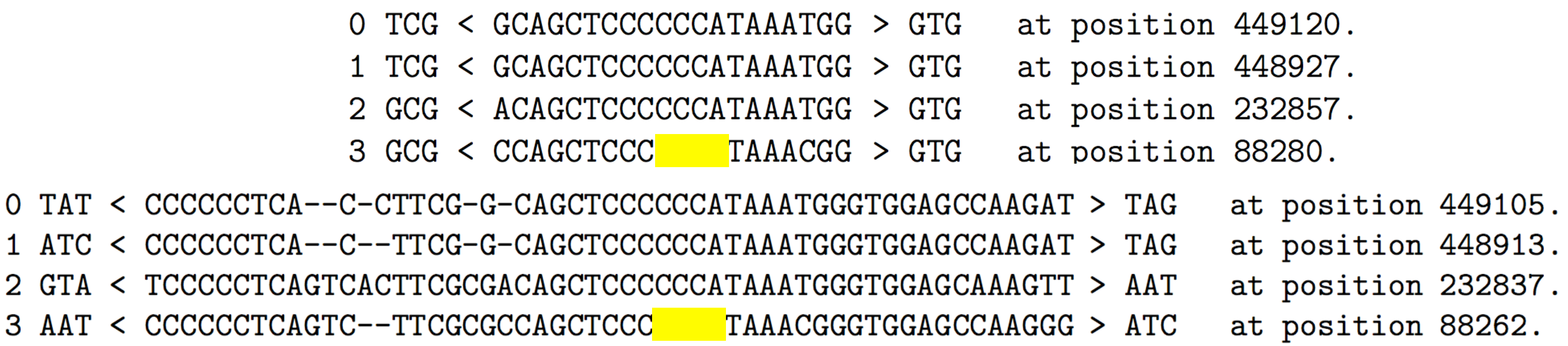
After Removal



```
0 TCG < GCAGCTCCCCCATAAATGG > GTG   at position 449120.
1 TCG < GCAGCTCCCCCATAAATGG > GTG   at position 448927.
2 GCG < ACAGCTCCCCCATAAATGG > GTG   at position 232857.
3 GCG < CCAGCTCCC-CCGTAAACGG > GTG   at position 88280.
```

Figure 6: A sample motif (score 25) produced by two phases

More rewindowing



```
                    0 TCG < GCAGCTCCCCCCATAAATGG > GTG   at position 449120.
                    1 TCG < GCAGCTCCCCCCATAAATGG > GTG   at position 448927.
                    2 GCG < ACAGCTCCCCCCATAAATGG > GTG   at position 232857.
                    3 GCG < CCAGCTCCC    TAAACGG > GTG   at position 88280.
0 TAT < CCCCCCTCA--C-CTTCG-G-CAGCTCCCCCCATAAATGGGTGGAGCCAAGAT > TAG   at position 449105.
1 ATC < CCCCCCTCA--C--TTCG-G-CAGCTCCCCCCATAAATGGGTGGAGCCAAGAT > TAG   at position 448913.
2 GTA < TCCCCCTCAGTCACTTCGCGACAGCTCCCCCCATAAATGGGTGGAGCAAAGTT > AAT   at position 232837.
3 AAT < CCCCCCTCAGTC--TTCGCGCCAGCTCCC    TAAACGGGTGGAGCCAAGGG > ATC   at position 88262.
```

Figure 7: The motif of Figure 6 after seven phases (score 62)

0 & 1 identical for another 55 bases;
5 differences in next 44.
Probably not a TFBS, but not “random”

Summary

- Handles gaps
- avoids full multiple alignment by exploiting good partial alignment
- validation - null model for comparison
- look at data -
  - rewindowing
  - rRNA cluster