

Sequence Alignment and Recognition of DNA Motifs

A Bioinformatics Approach

Ani Kesanapally

Background

What are DNA sequence motifs?

A Broad Biological Phenomena

“Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins... others are involved in important processes at the RNA level...”

- Patrik D’haeseleer

Motif Recognition and Representation

A Visual Pathway

1. HEM13 CCCATTGTTCTC
HEM13 TTTCTGGTTCTC
HEM13 TCAATTGTTTAG
ANB1 CTCATTGTTGTC
ANB1 TCCATTGTTCTC
ANB1 CCTATTGTTCTC
ANB1 TCCATTGTTCGT
ROX1 CCAATTGTTTGG

2. YCHATTGTTCTC

3.

A	002700000010
C	464100000505
G	000001800112
T	422087088261



Inquiry and Approach

Research Inquiry

Discovering DNA Motifs

“If a data set of sequences is believed to possess a motif, how can we determine the composition and location of this motif?”

Addressing the Inquiry

A General Approach

1. Find a data set known to have the same, conserved motif in each sequence.
2. Select one sequence to be a reference.
3. Align all sequences to that reference at the motif location.
4. Determine the frequency and relative entropy of nucleotides at each position.

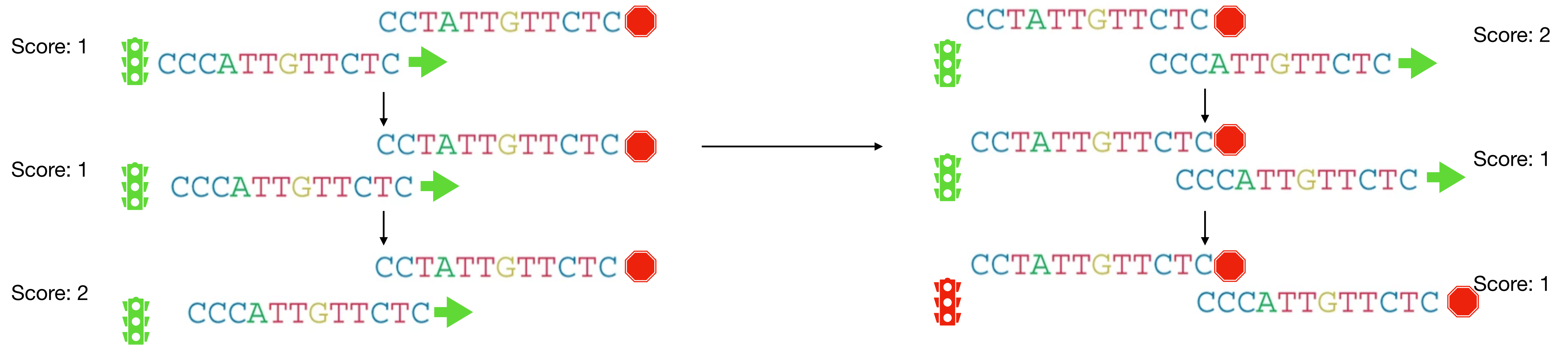
Addressing the Dilemma

Computational Concerns

- Without knowing the motifs beforehand, aligning sequences together at the motif becomes an algorithmically and computationally intensive affair.
 - The approach used for this pipeline involves finding a **“sliding” score between a reference sequence and another.**
 - The value for this score signifies at that specific shift of one sequence to the reference, **how many nucleotides are similar at each position.**
 - The higher the score, the more ideal alignment has been reached, but this may not be motif alignment.

Addressing the Dilemma

Computational Concerns



CCTATTGTTCTC
CCCATTGTTCTC
Score: 11
This is the maximum score.

Addressing the Dilemma

Computational Concerns

- Modeling and recognition of DNA sequence motifs often involves **matrix encodement** and **mathematics**.
 - It is simply quite **difficult** to run predictive complex models where the four nucleotides are represented as characters such as “**A**”, “**C**”, “**G**”, and “**T**.”
 - The approach for this pipeline uses an **encoding method** where each of the four nucleotides are represented in **3-dimensional space** where the position of each nucleotide is equidistant to zero.

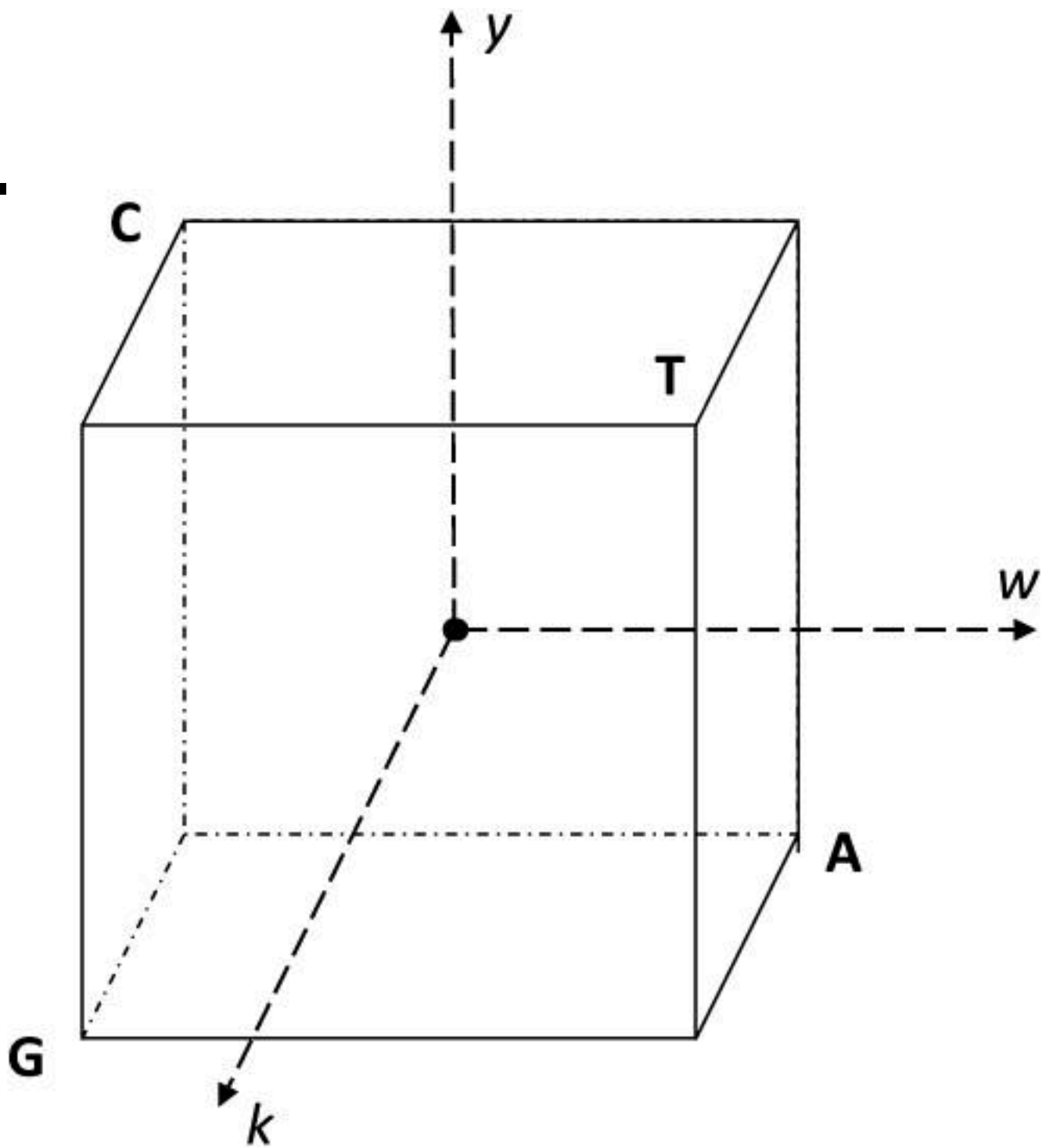
Addressing the Dilemma

Computational Concerns

1.

	A	C	G	T
<i>w</i>	1	-1	-1	1
<i>y</i>	-1	1	-1	1
<i>k</i>	-1	-1	1	1

2.



Pipeline

Pipeline Steps

Reproducible Computation

1. **Encode** all the desired **data** (as *.fasta* files) into **wyk** format.
 - encoderWKY.sh runs encodeWYK.py on all files of a directory.
2. **Determine the shift value** by calculating scores between an established reference *wyk* file and all remaining *wyk* files in the data set.
3. **Shift the non-reference *wyk* files to align with the reference** by adding 0's at the beginning or end of the file with the number of 0's obtained from the shift value.
 - shifter.sh runs corrshift.py (which does steps 2 and 3) on all *wyk* files of a directory.

Pipeline Steps

Reproducible Computation

1.

main

pipeline-project-evokes / data / coli /

Go to file

Add file

...

evokes all code outputs

ee7952c 7 days ago

History

..

ecoblgr1.fasta

all code outputs

7 days ago

ecocya.fasta

all code outputs

7 days ago

ecomalba.fasta

all code outputs

7 days ago

ecomalba2.fasta

all code outputs

7 days ago

ecoompa.fasta

all code outputs

7 days ago

ecotnaa.fasta

all code outputs

7 days ago

tdc.fasta

all code outputs

7 days ago

2.

main

pipeline-project-evokes / output / coli_wyk /

Go to file

Add file

...

evokes all code outputs

ee7952c 7 days ago

History

..

wyk_ecoblgr1.txt

all code outputs

7 days ago

wyk_ecocya.txt

all code outputs

7 days ago

wyk_ecomalba.txt

all code outputs

7 days ago

wyk_ecomalba2.txt

all code outputs

7 days ago

wyk_ecoompa.txt

all code outputs

7 days ago

wyk_ecotnaa.txt

all code outputs

7 days ago

wyk_tdc.txt

all code outputs

7 days ago

3.

main

pipeline-project-evokes / output / coli_shift_wyk /

Go to file

Add file

...

evokes all code outputs

ee7952c 7 days ago

History

..

shifted_wyk_ecocya.txt

all code outputs

7 days ago

shifted_wyk_ecomalba.txt

all code outputs

7 days ago

shifted_wyk_ecomalba2.txt

all code outputs

7 days ago

shifted_wyk_ecoompa.txt

all code outputs

7 days ago

shifted_wyk_ecotnaa.txt

all code outputs

7 days ago

shifted_wyk_tdc.txt

all code outputs

7 days ago

Pipeline Steps

Reproducible Computation

4. **Add** by each index position (or line) the **contents of the reference *wyk* file and all the non-reference, shifted *wyk* files** to create a final sum file.

- summer.sh runs sumcreator.py on the chosen reference wyk file and shifted non-reference wyk files.

5. **Convert the final sum *wyk* file into a nucleotide probability matrix.**

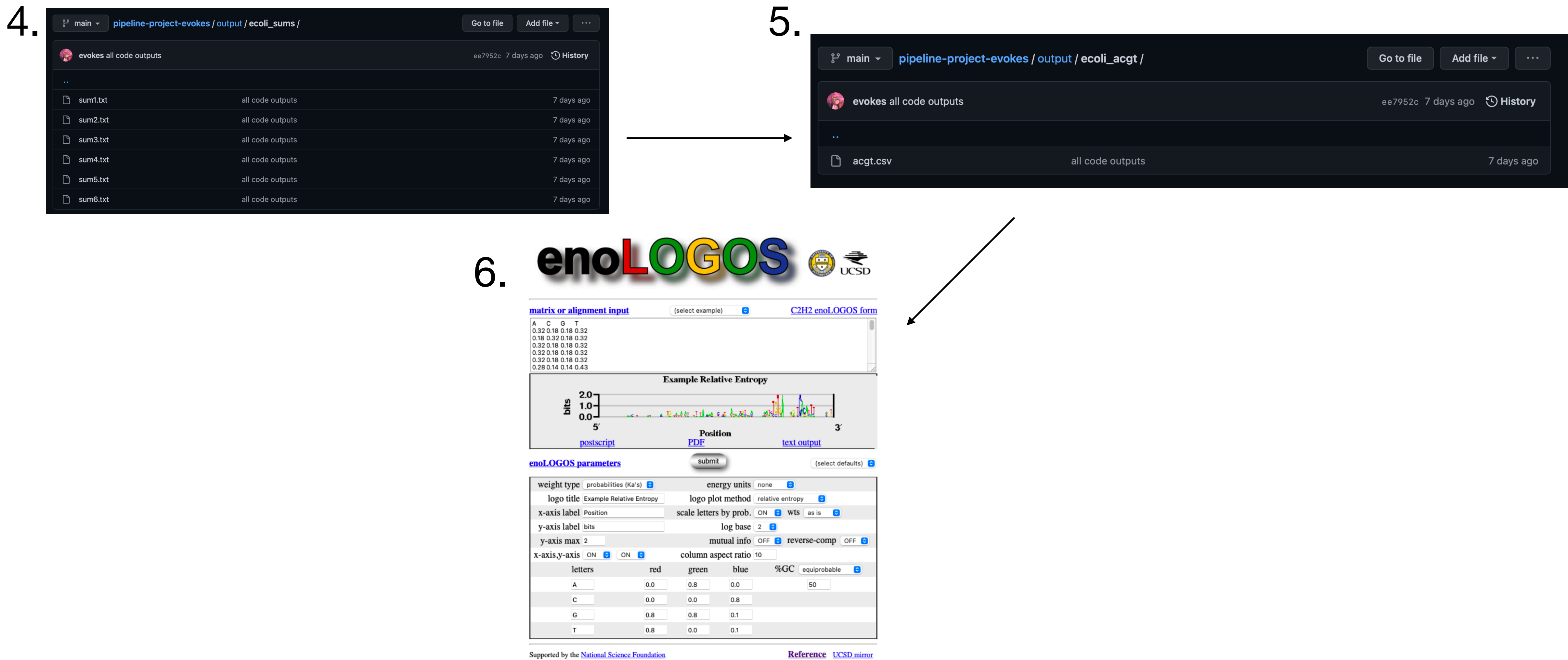
- encoderACGT.sh runs encodeACGT.py on the final sum file.

6. Upload the contents of the nucleotide probability matrix to a **DNA logo generator** to visualize output.

- Benos Lab Energy Normalized Logo Generator produces this logo.

Pipeline Steps

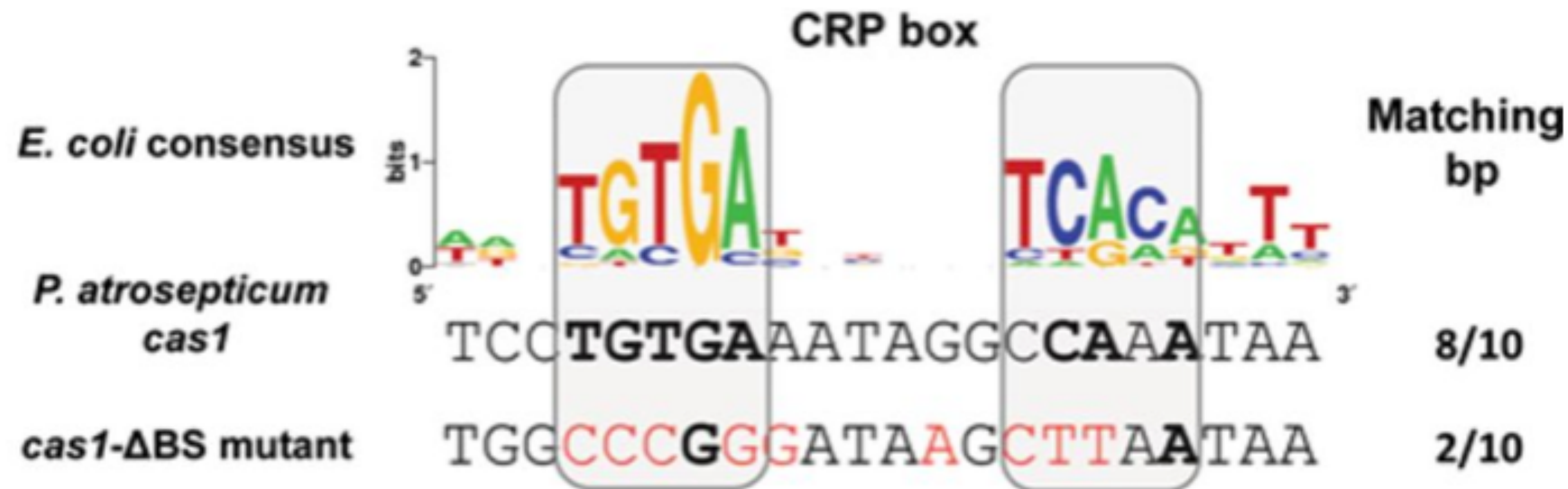
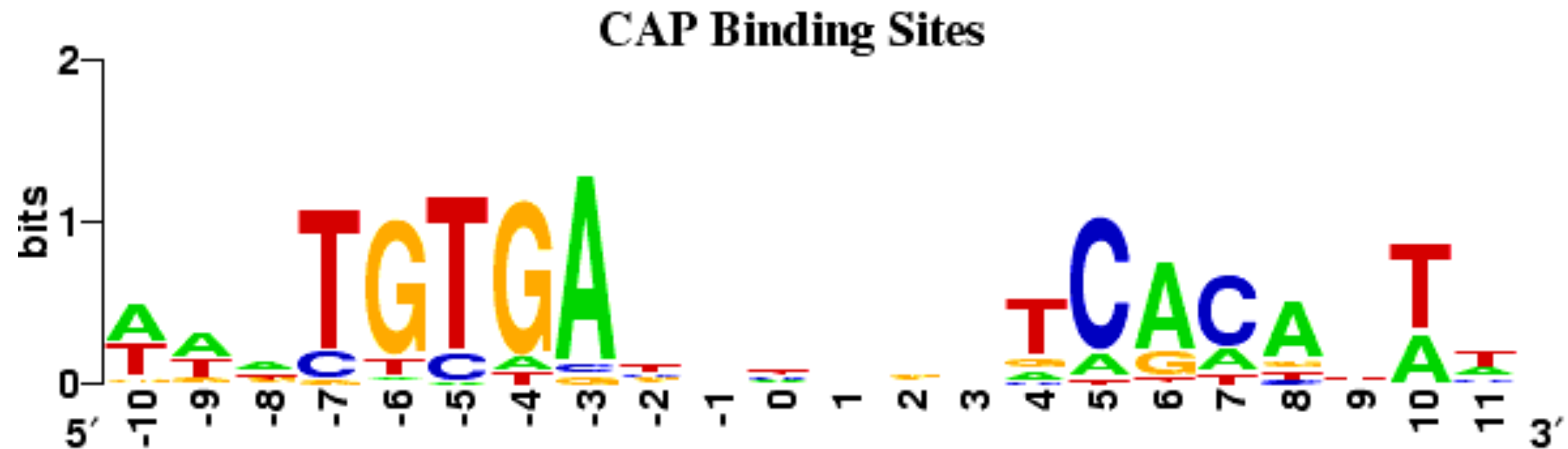
Reproducible Computation



Data, Results, and Analysis

Exploring An *E. coli* Motif

An Almost Palindromic Binding Site For Proteins



Dummy Data Set

Sequences With Same Motifs As *E.coli* CRP/CAP Binding Protein

>dummy1

TGTGAAAAAAATCACAAAAAAAAAAAAAAAAAAA

>dummy2

CCCCCCCCCCCCCCCCCTTGTGACCCCCCTCACA

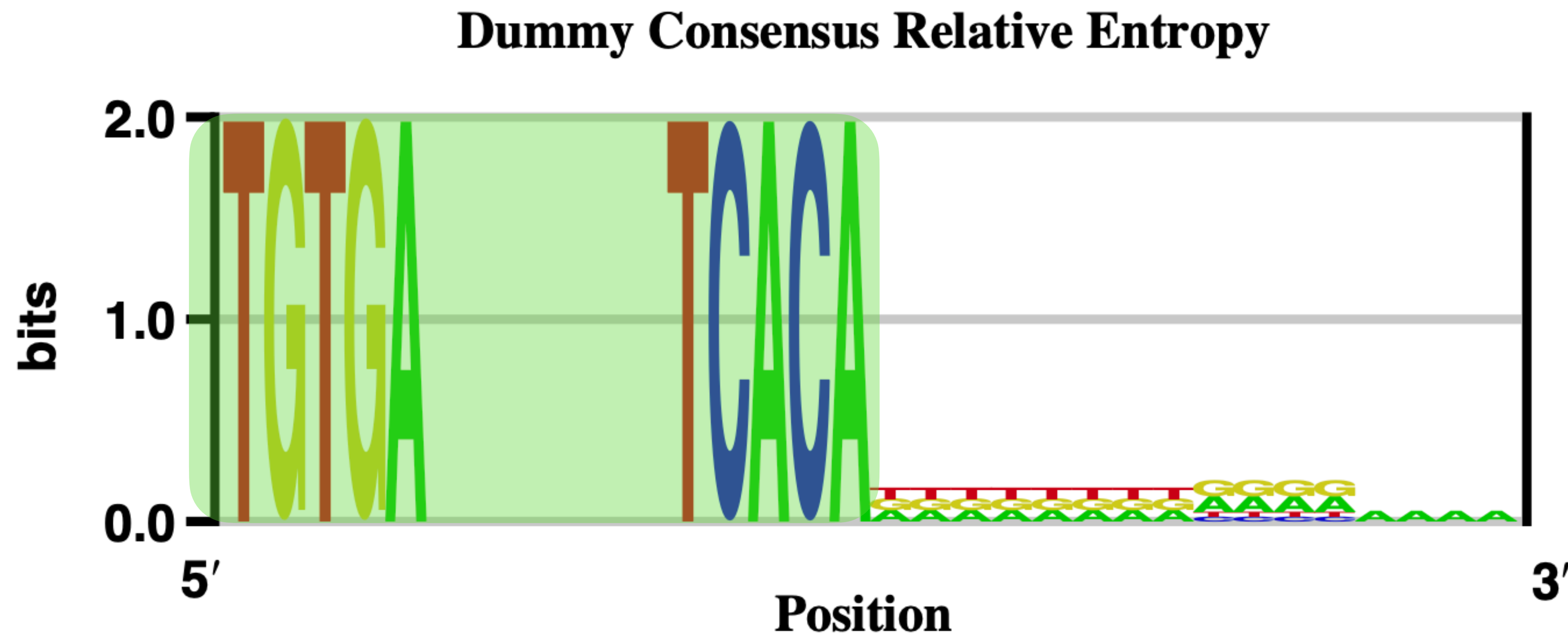
>dummy3

TTTTTTTTTTGTGATTTTTTTCACATTTTTTTT

>dummy4

GGGGTTGTGAGGGGGGGTCACAGGGGGGGGGGGGGGG

DNA Logo Results



E. coli Data Set

Seven Genes With Common Motifs

>ECOBGLR1 (REFERENCE)
ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAA
TTACACAAAGTTAATAACT**TGTGA**GCATGG**TCATA**TTTTTATCAAT

>ECOCYA
ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGTGTAAAT
TGATCACGTTTTAGACCATTTTTTCGTGCGTGAAACTAAAAAACC

>ECOTNAA
TTTTTTAAACATTAAAATTCTTACGTAATTTATAATCTTTAAAAAAGCATTTAATATTG
CTCCCCGAACGATTGTGATTGATTACATTAAACAATTCAGA

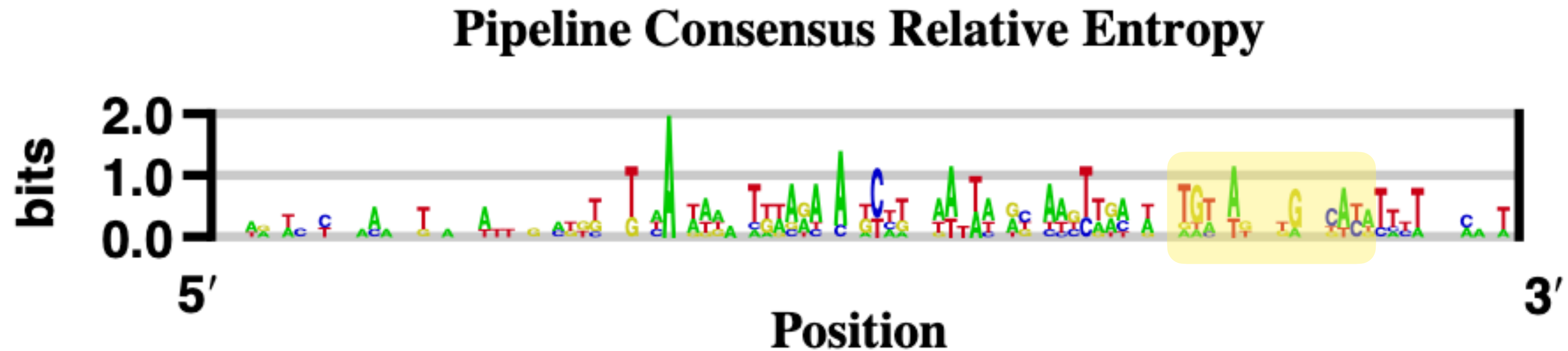
>TDC
GATTTTATACTTTAACTTGTTGATATTTAAAGGTATTTAATTGTAATAACGATACTCTG
GAAAGTATTGAAAGTTAATTTGTGAGTGGTCGCACATATCCTGTT

>ECOOMPA
GCTGACAAAAAAGATTAAACATACCTTATACAAGACTTTTTTTTCATATGCCTGACGGAG
TTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTACATCGCC

>ECOMALBA
ACATTACCGCCAATTCTGTAACAGAGATCACACAAAGCGACGGTGGGGCGTAGGGGCAAG
GAGGATGGAAAGAGGTTGCCGTATAAAGAACTAGAGTCCGTTTA

>ECOMALBA2
GGAGGAGGCGGGAGGATGAGAACACGGCTTCTGTGAACTAAACCGAGGTCATGTAAGGAA
TTTCGTGATGTTGCTTGCAAAAATCGTGGCGATTTTATGTGCGCAA

Pipeline DNA Logo Results



E. coli Data Set

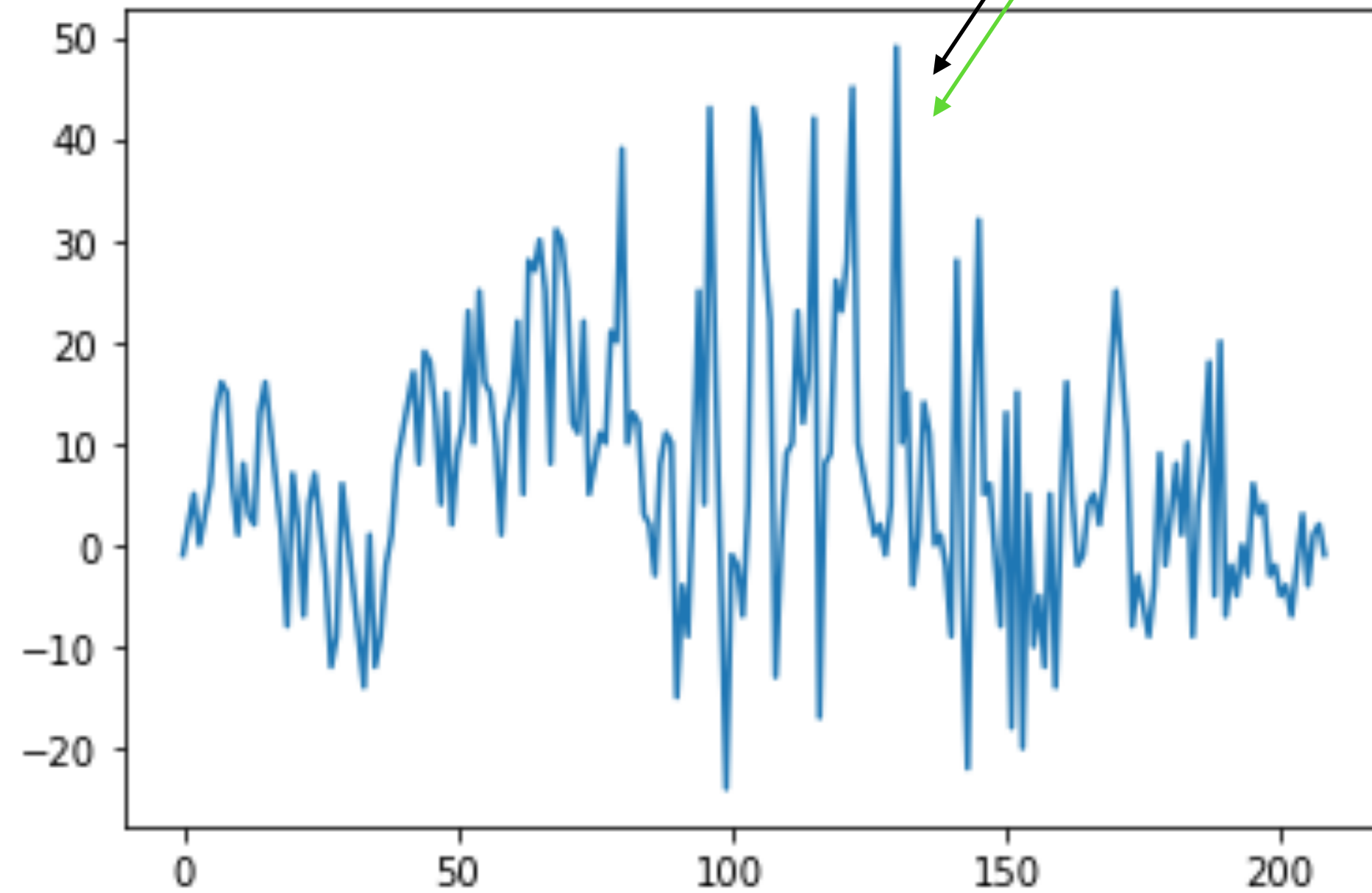
Revisiting the Data Set

Sequence	Reference	Hand-Counted Distance from One End to Motif	Relative Distance from Reference
ECOBLGR1	Yes	8	N/A
ECOCYA	No	34	26
ECOTNAA	No	13	5
TDC	No	6	-2
ECOOMPA	No	36	28
ECOMALBA	No	62	54
ECOMALBA2	No	47	39

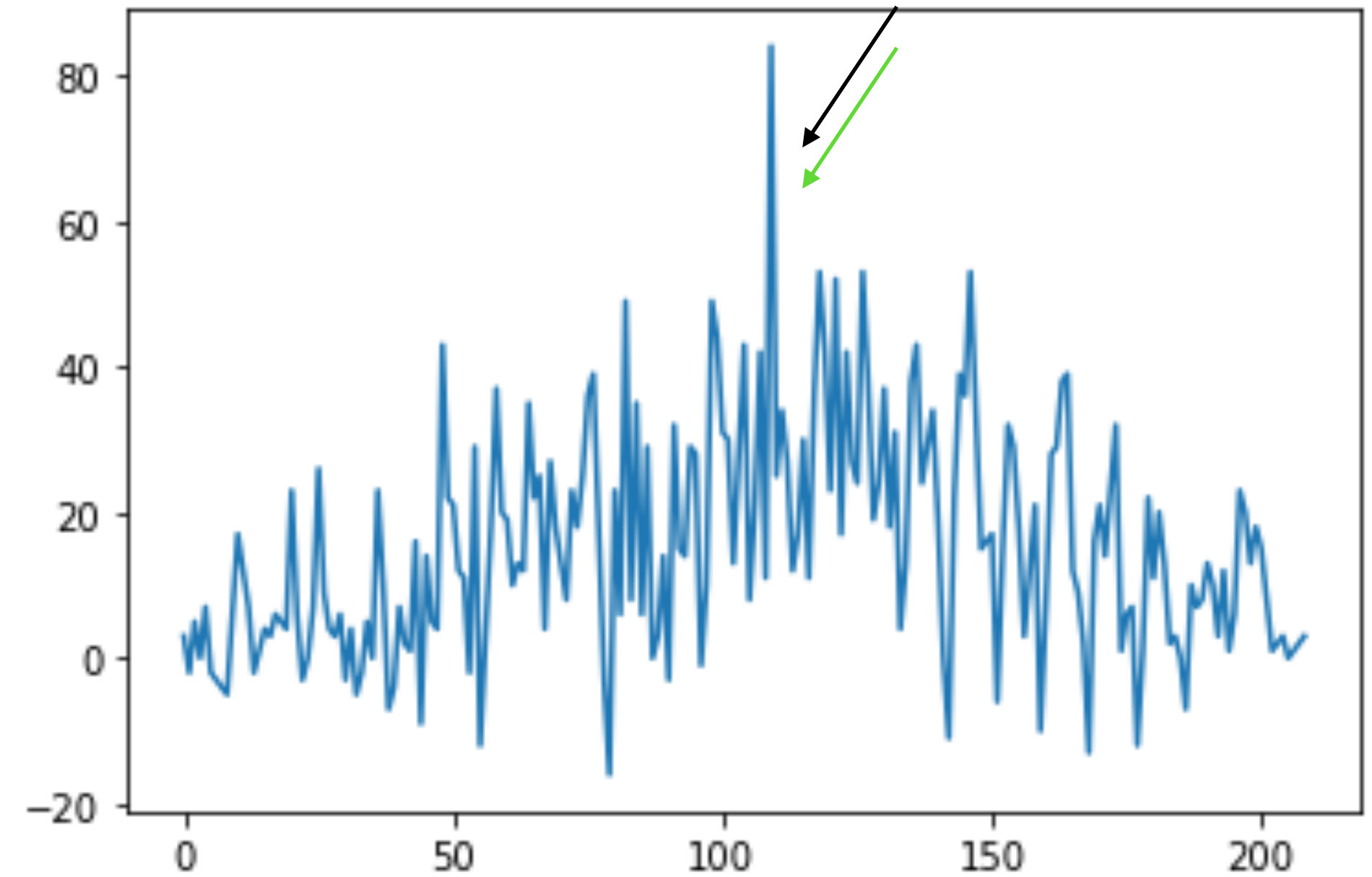
E. coli Data Set

A Graphical Troubleshoot

ECOBLGR1 vs. ECOCYA

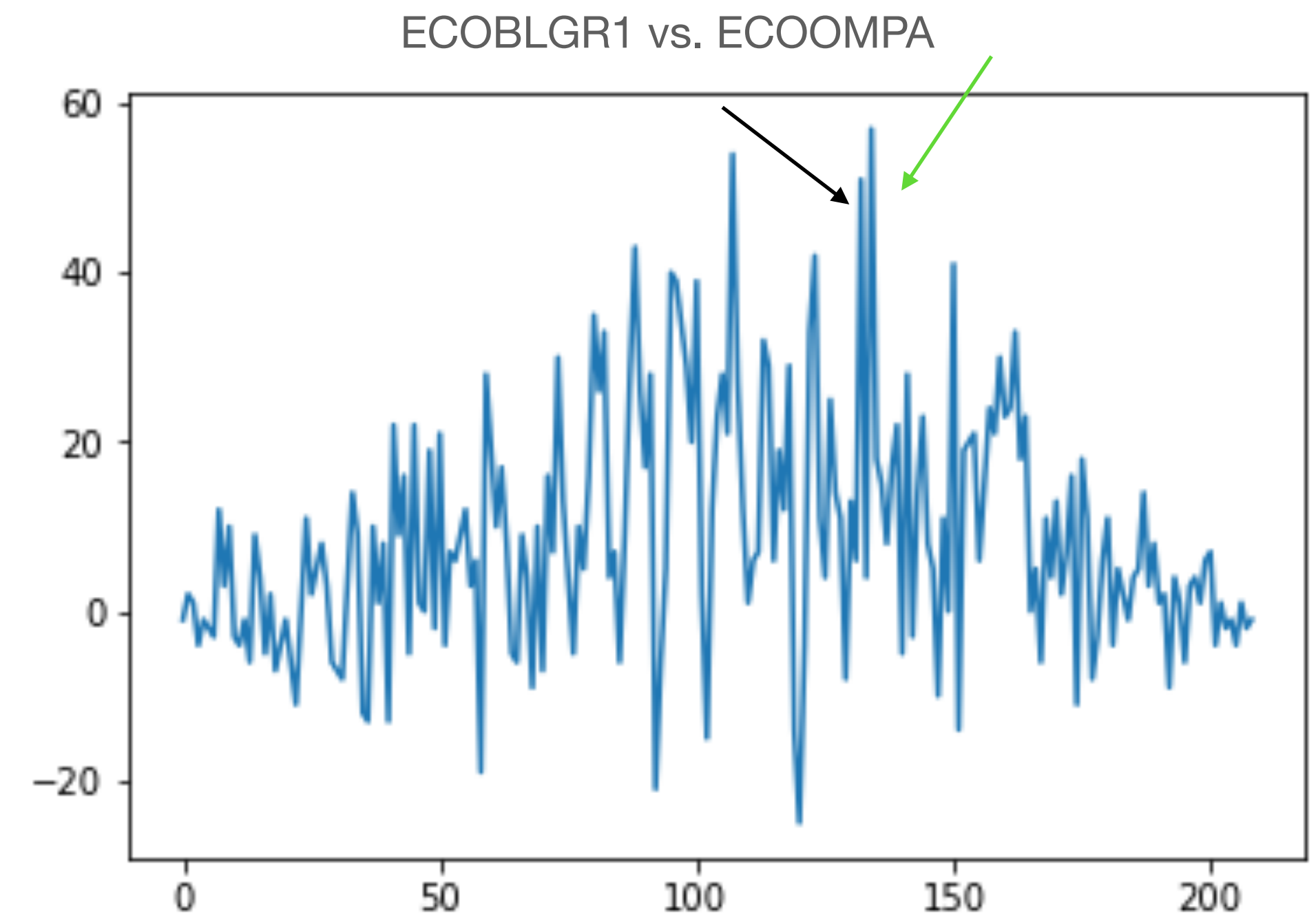
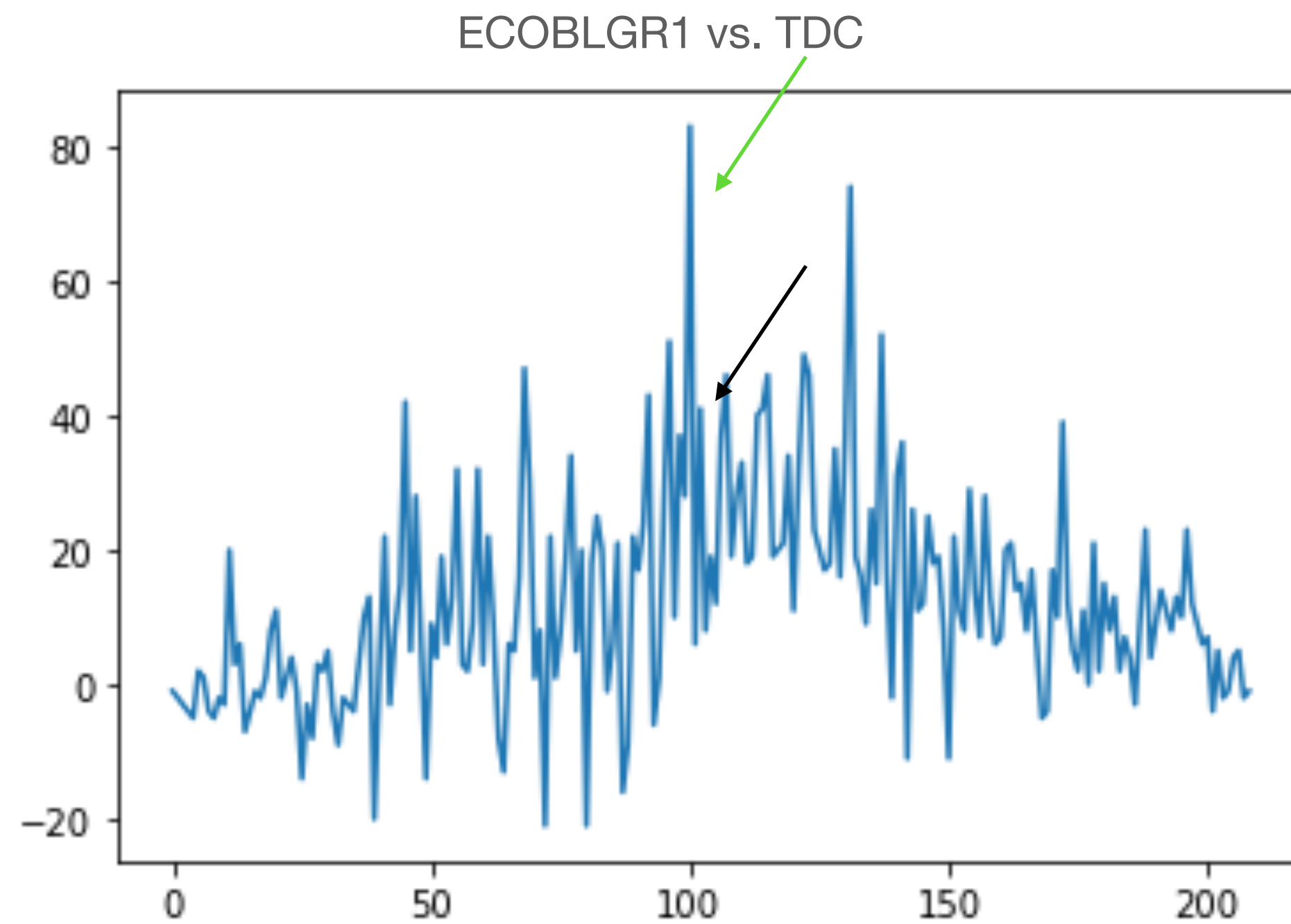


ECOBLGR1 vs. ECOTNAA



E. coli Data Set

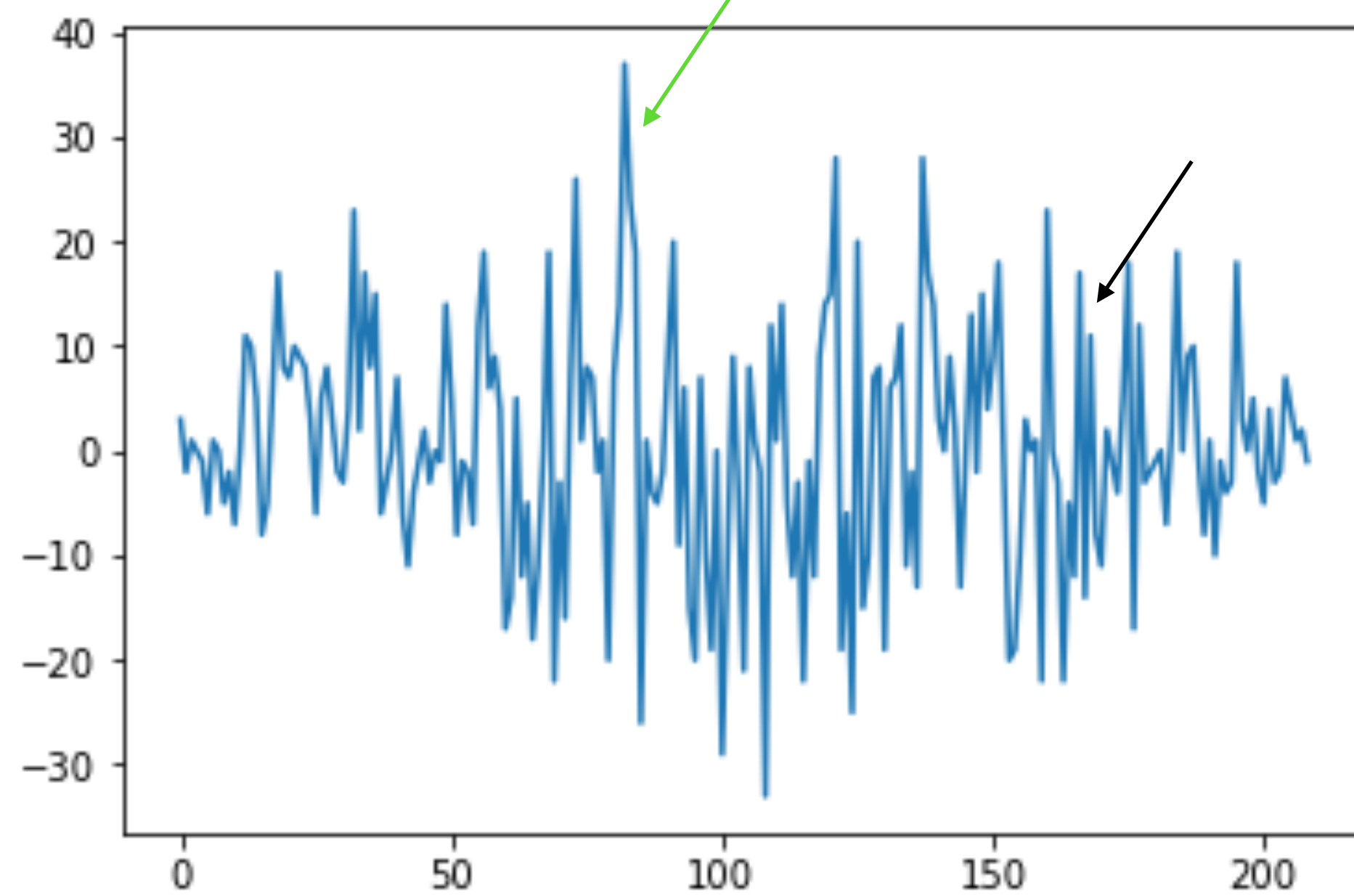
A Graphical Troubleshoot



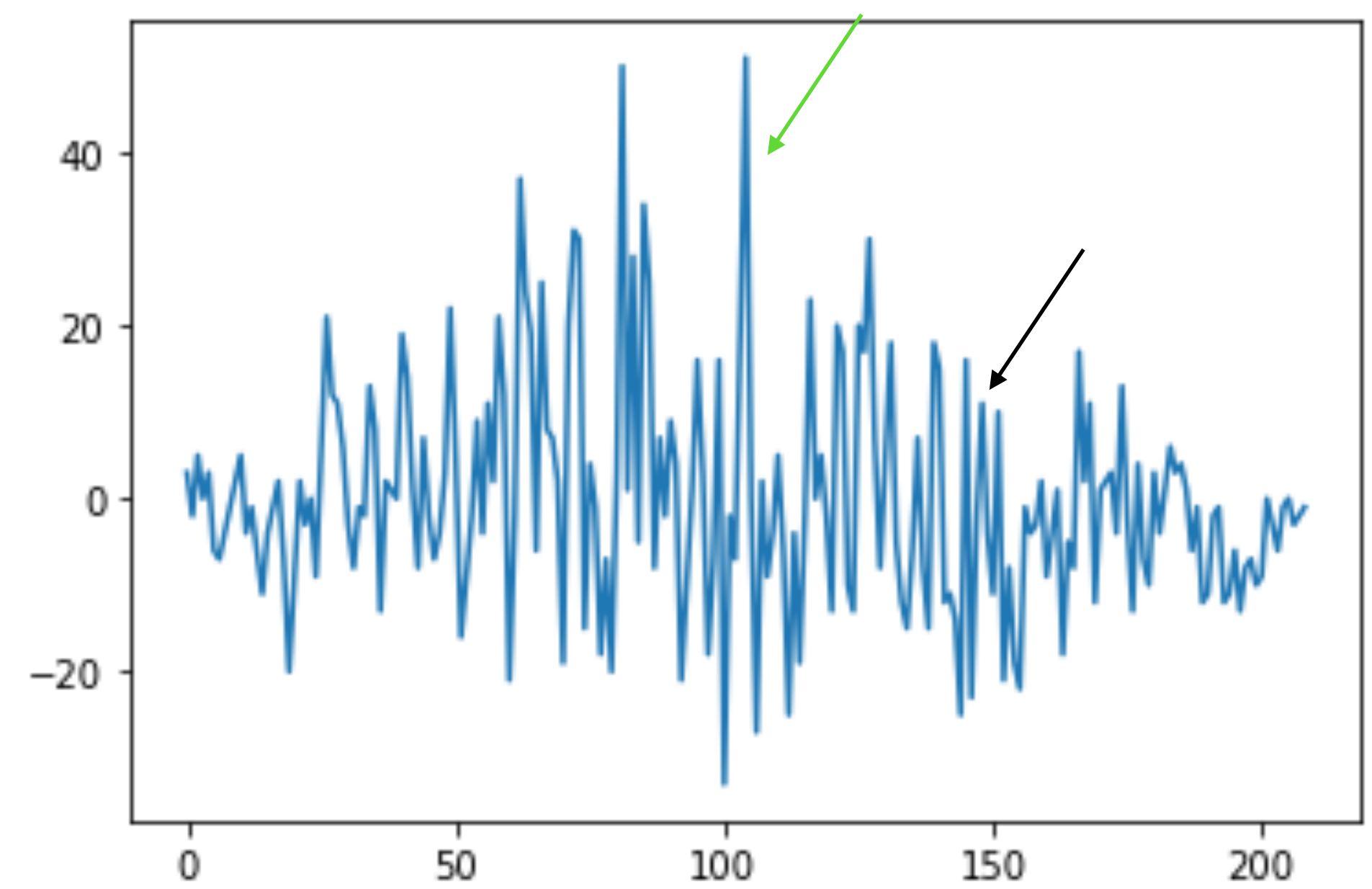
E. coli Data Set

A Graphical Troubleshoot

ECOBLGR1 vs. ECOMALBA



ECOBLGR1 vs. ECOMALBA2



E. coli Data Set

Further Data Analysis

Sequence	Reference	Hand-Counted Distance from One End to Motif	Relative Distance from Reference	Obtained Shift/ Relative Distance	Consensus
ECOBLGR1	Yes	8	N/A	N/A	N/A
ECOCYA	No	34	26	26	Good
ECOTNAA	No	13	5	5	Good
TDC	No	6	-2	-4	Close
ECOOMPA	No	36	28	30	Close
ECOMALBA	No	62	54	-22	Little to None
ECOMALBA 2	No	47	39	0	Little to None

Good Consensus DNA Logo Results

[illegible]

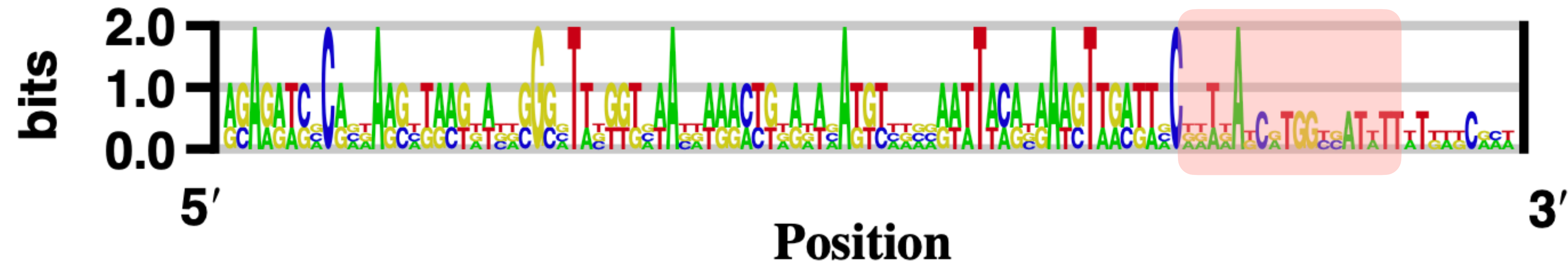
Close Consensus DNA Logo Results

[illegible]

E. coli Data Set

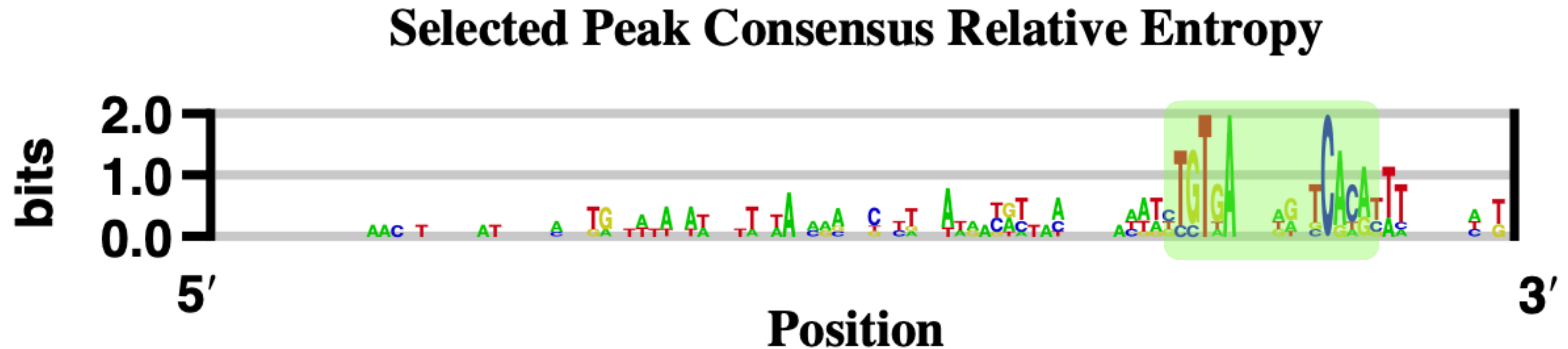
Little to No Consensus DNA Logo Results

Little to No Peak Consensus Relative Entropy



E. coli Data Set

Manually Selected Peak Consensus DNA Logo Results



Citations and Conclusion

Sources & Works Cited

Special Thanks to the Gary D. Stormo, *Ph.D* Lab

Stormo GD, Hartzell GW 3rd. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A*. 1989 Feb;86(4):1183-7. doi: 10.1073/pnas.86.4.1183. PMID: 2919167; PMCID: PMC286650.

D'haeseleer, P. What are DNA sequence motifs?. *Nat Biotechnol* **24**, 423–425 (2006). <https://doi.org/10.1038/nbt0406-423>

Stormo GD. Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics*. 2011 Apr;187(4):1219-24. doi: 10.1534/genetics.110.126052. Epub 2011 Feb 7. Erratum in: *Genetics*. 2011 Dec;189(4):1525. PMID: 21300846; PMCID: PMC3070529.

Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*. 2005 Jul 1;33(Web Server issue):W389-92.

Robison, K., McGuire, A. M., Church, G. M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K12 genome. *Journal of Molecular Biology* (1998) 284, 241-254.

Patterson, Adrian & Chang, James & Taylor, Corinda & Fineran, Peter. (2015). Regulation of the type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. *Nucleic acids research*. 43. 10.1093/nar/gkv517.