CS 466: Mini Project

Fall 2016

Kamal Chaya

Matthew Romano

Jashn Chhaya

David Guan

Part 1: Background and Experimental Setup (Jashn Chhaya & David Guan)

In order to apply the algorithm, we first created the benchmark by creating 70 data sets with different parameters for ICPC, ML, and SC. Each dataset has the motif, motiflength, sequences, and planted sites stored in separate files. We generated all these random sequences and motifs and planted the sites randomly as well with the different parameters, repeated up to 10 times with 7 different parameter combinations.

Part 2: Our Algorithm (Matthew Romano)

We are using a Greedy Motif Search. First we are finding the best alignment between 2 sequences by trying every possible alignment and maximizing the score. Next, we are incrementally adding another sequence and trying every possible alignment with the new sequence against our current alignments to again maximize the score.

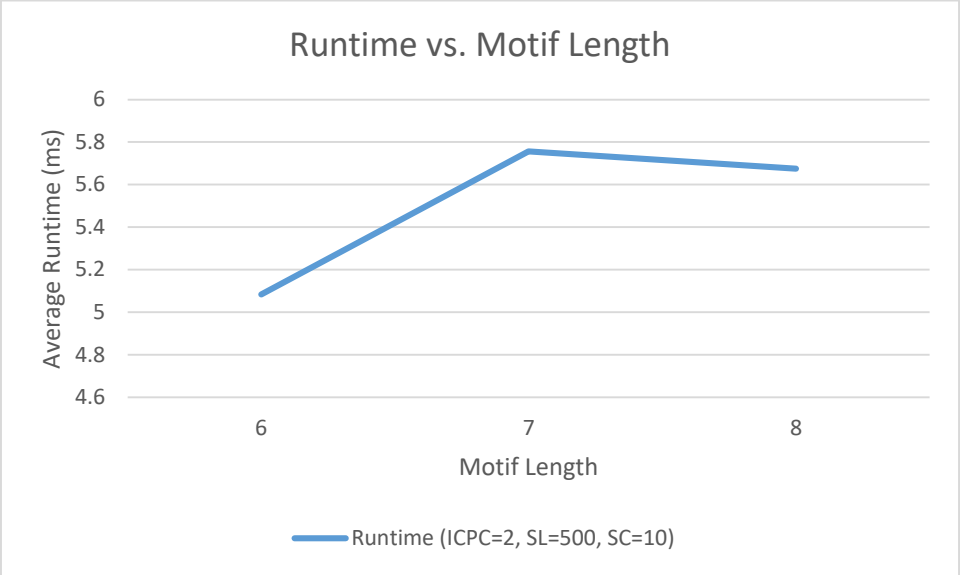
This results in a running time of $O(L^2 + (t-2)L)$ where t is the number of sequences and L is the total number of starting positions in each sequence. This does not by any means guarantee an optimal solution, however it does give a fast solution.

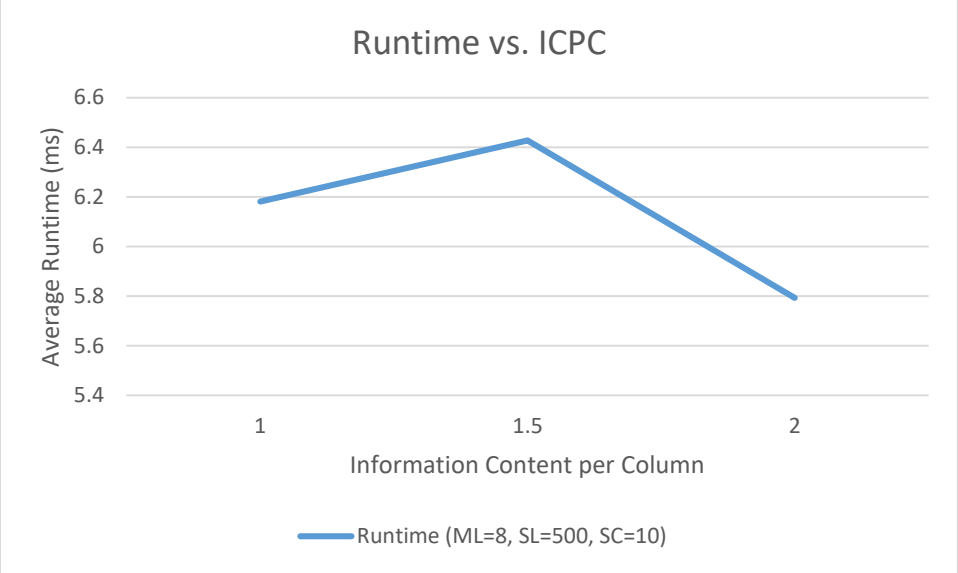
Part 3: Our Results (Kamal Chaya)

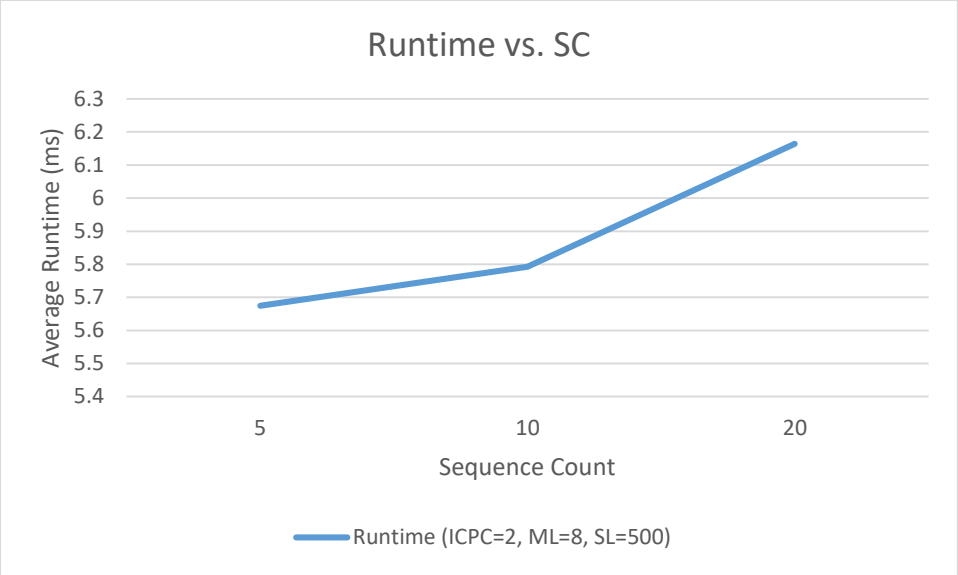
Below are all of the metrics we calculated along with the raw data in order to evaluate our motif finder. We calculated the number of overlapping sites between sites.txt and predictedsites.txt. A higher number of overlapping sites indicates better performance of the motif finder as more common subsequences were found. We also calculated the relative entropy between our motif.txt and predictedmotif.txt. A higher relative entropy indicates a higher amount of information content, (e.g. essentially how skewed the probability of each nucleotide occurring is). Lastly, we evaluated the running time of our motif finder as well, to see how performant our code was. For each of these metrics we compared against our parameter variations, namely, motif length, information content per column, and sequence count.

Please note our relative entries are above 2 since we calculate the relative entropy for each position for each nucleotide in predictedmotif.txt (using motif.txt to calculate the base probabilities of each nucleotide), and then add all of these across the entire matrix.

Runtime


Runtime vs. Motif Length
Average Runtime (ms)
6
5.8
5.6
5.4
5.2
5
4.8
4.6
6
7
8
Motif Length
Runtime (ICPC=2, SL=500, SC=10)

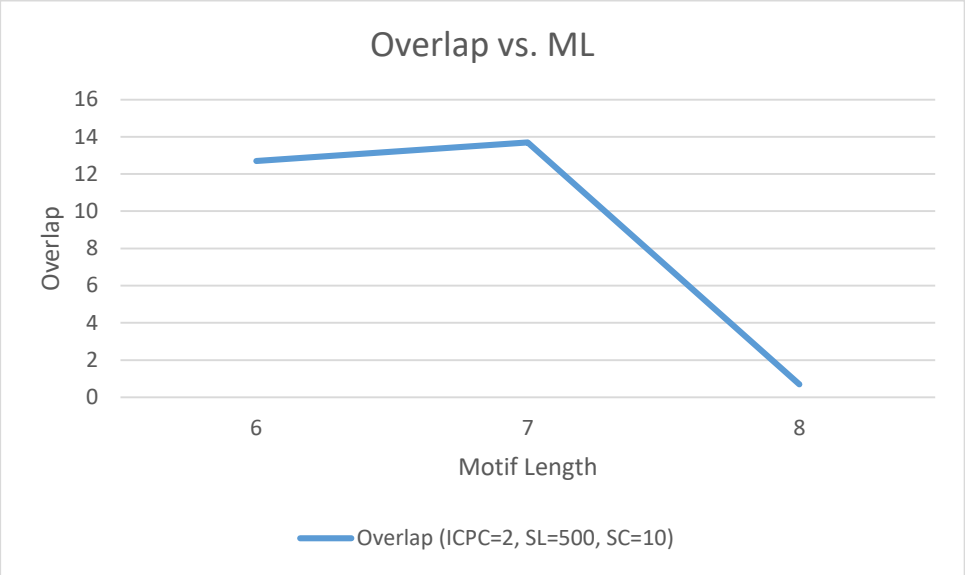

Runtime vs. ICPC
Average Runtime (ms)
6.6
6.4
6.2
6
5.8
5.6
5.4
1
1.5
2
Information Content per Column
Runtime (ML=8, SL=500, SC=10)

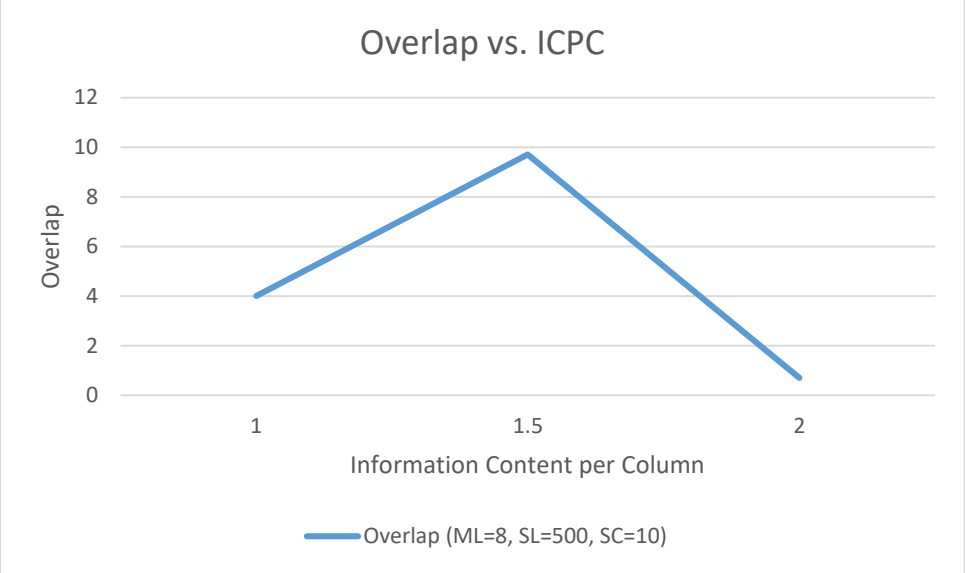


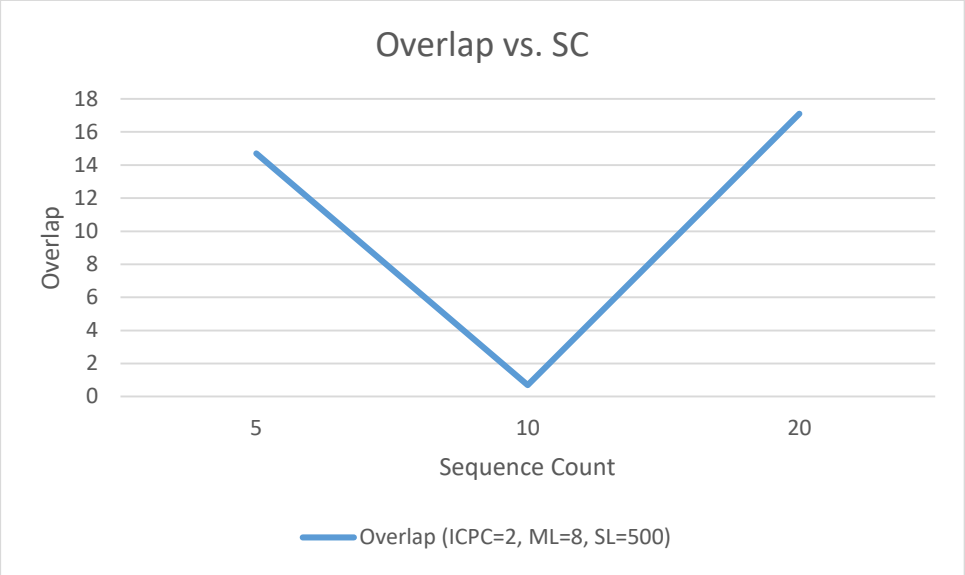
Raw Data:

	data 1	data 2	data 3	data 4	data 5	data 6	data 7	data 8	data 9	data 10
dataset 1	5.765477	5.670561	5.327091	5.306202	5.446834	5.160634	5.571666	5.211826	7.07916	7.389279
dataset 2	5.481457	6.438741	5.59228	5.640076	5.556055	5.96022	5.695537	8.786644	6.84303	5.809966
dataset 3	6.211197	5.860791	6.938735	6.321164	5.406765	6.232387	5.587406	7.607035	7.985902	6.122553
dataset 4	4.513918	5.052624	4.74889	5.217777	5.465393	4.706549	4.750098	5.117223	5.900684	5.369125
dataset 5	5.331383	8.740659	6.517275	6.213104	5.147392	4.726517	4.717885	4.687013	6.037205	5.415639
dataset 6	6.208988	5.589859	5.382144	6.247922	5.200696	6.846894	5.154531	5.719002	5.108847	5.284432
dataset 7	6.320182	5.59873	6.184986	5.558619	5.968139	6.275814	6.030023	5.4567	6.244095	7.999716

Overlap


Overlap vs. ML
16
14
12
10
8
6
4
2
0
Overlap
6
7
8
Motif Length
Overlap (ICPC=2, SL=500, SC=10)

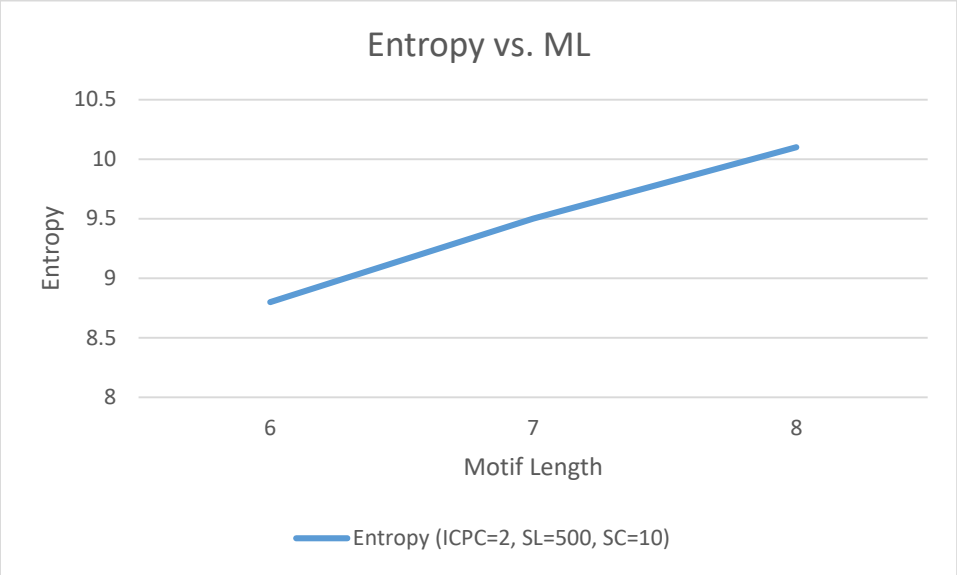

Overlap vs. ICPC
12
10
8
6
4
2
0
Overlap
1
1.5
2
Information Content per Column
Overlap (ML=8, SL=500, SC=10)

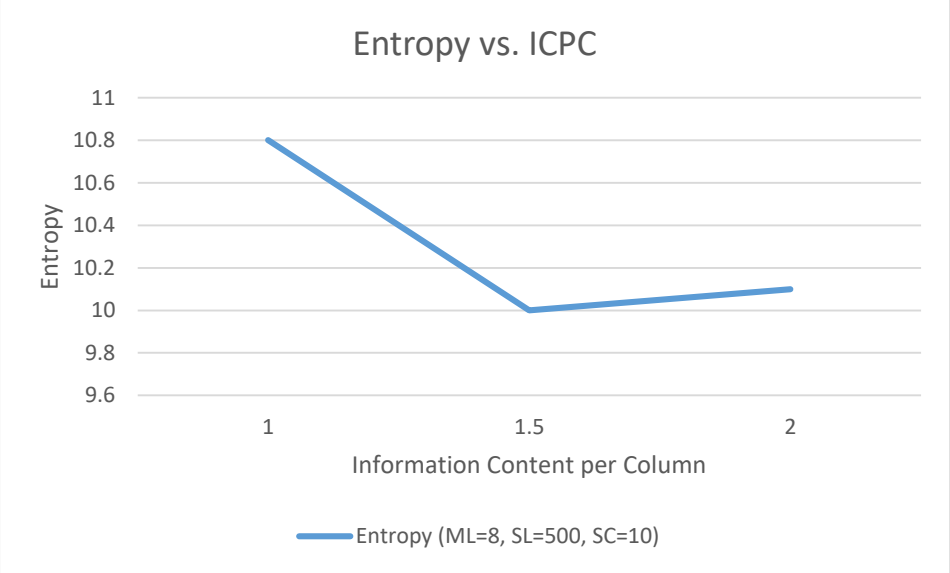


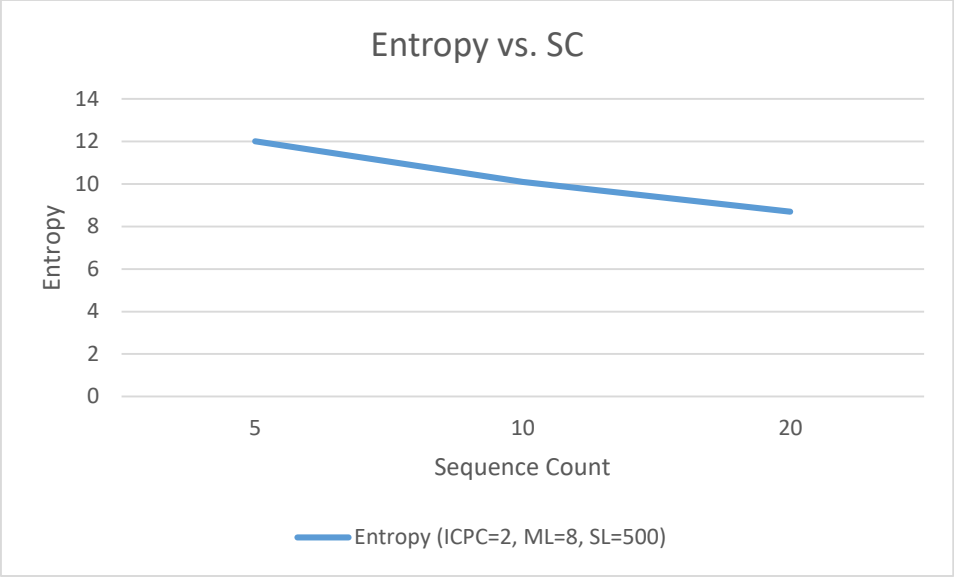
Raw Data:

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
Dataset 1	0	0	0	1	1	1	1	1	1	1
Dataset 2	2	2	2	3	4	5	5	5	6	6
Dataset 3	7	8	9	9	9	11	11	11	11	11
Dataset 4	12	12	12	13	13	13	13	13	13	13
Dataset 5	13	13	13	14	14	14	14	14	14	14
Dataset 6	14	14	14	14	14	14	15	15	16	16
Dataset 7	16	16	16	17	17	17	18	18	18	18

Entropy


Entropy vs. ML
10.5
10
9.5
9
8.5
8
Entropy
6
7
8
Motif Length
Entropy (ICPC=2, SL=500, SC=10)


Entropy vs. ICPC
11
10.8
10.6
10.4
10.2
10
9.8
9.6
Entropy
1
1.5
2
Information Content per Column
Entropy (ML=8, SL=500, SC=10)



Raw Data:

	data 1	data 2	data 3	data 4	data 5	data 6	data 7	data 8	data 9	data 10
Dataset 1	11	11	10	9	11	11	8	10	10	10
Dataset 2	10	11	12	10	11	12	10	10	11	11
Dataset 3	10	9	11	9	10	10	10	10	11	10
Dataset 4	8	8	8	10	8	9	8	9	10	10
Dataset 5	8	10	10	9	8	10	11	9	9	11
Dataset 6	11	11	11	12	11	12	16	11	13	12
Dataset 7	9	9	9	9	9	8	9	8	8	9