# CS 412 Introduction of Data Mining

## Course Project:

## Motif Finding Problem

Instructed by Dr. Qi Li

Yining Liu,  Zhongyi Liu,  Hongjae Jeon

May 1st 2019

# Part I: Dataset Generation

In the first part, the datasets were generated, severed as the unknown DNA sequence. During this process, several decisions were made.

## A. DNA sequences generation:

Since we assumed that the probability (Background Probability) for each character at any location should be equal when there are no motif strings, the background probabilities should all be a quarter (0.25) Then, we generated the original sequences by sampling according to background probability, so that we could use this assumption in the next part.

## B. Motif table generation:

Motif is in definition, a DNA pattern that has biological importance and they are embedded in the DNA, so it is important to isolate, detect, or find them. Total probability at any position should be 1 and the constraint of each ICPC condition, we only need to choose two probability and other probability could be derived. However, the first two probabilities cannot be arbitrarily determined, since not all values could achieve given ICPC. In order to randomly generate the motif table for each given ICPC, we determined the feasible range of the first number, by calculating the maximum ICPC and minimum ICPC of any given number.

1. Iterate from 0 to 1 with increment 0.0001
2. For any number $n$, the combination of maximum ICPC is $\{n,\ 1-n,\ 0,\ 0\}$ and the combination of minimum ICPC is $\{n,\ (1-n)/3,\ (1-n)/3,\ (1-n)/3\}$
3. If max $ICPC \geq$ Given $ICPC$ and min $ICPC \leq$ Given $ICPC$, incorporate this number into feasible range
4. When iteration is finished, randomly choose one number from the obtained range as the first probability in this position

After determining the first number, we extract the ICPC by the information content of the first probability, and it will be considered in obtaining the next feasible range for second probability with the similar process using new ICPC updated using first probability. The second probability will be chosen from it. Given these two probabilities, another two probabilities are obtained with equations.

After following the instruction, 10 datasets for each combination were obtained, which contents include:

1. The Motif table shows the probability of “A”, “C”, “G”, “T” at each position
2. Site is generated according to Motif table
3. The Sites file indicates the right location and right content of each inserted strings
4. Motif Length file is used to predict the Motif with the appropriate algorithm

## Part 2: Motif Prediction

In order to choose appropriate algorithm and predict the Motif table precisely, we first analyzed the Expectation Maximization algorithm and Gibbs Sampling Algorithm. Although the Expectation Maximization converges very quickly, it will easily get trapped in the local optimum. In addition, it is very susceptible to the initial condition. On the contrary, the Gibbs sampling does not only move toward the improving direction. By tolerating some bad movements, the sampling process has a better chance to approach the true optimum. Thus, we applied the Gibbs Sampling algorithm, combined with some improvements.

Gibbs sampling algorithm was first suggested by Lawrence et. al. The algorithm in this project implemented the Gibbs sampling algorithm and add some modifications to improve the accuracy of predictions.

1. Initiate a best solution, with all starting locations to be -1 and its information content $Best\_IC = 0$, and set $Extra\_Required\_Iterations$ to be zero;
2. Randomly choose motifs starting points for each sequence and record the starting points in a list named as “pst”;
3. Randomly choose one sequence from a set as the **selected sequence**, whose starting point will be modified in the following steps;
4. Formulate position weight matrix (PWM), according to motif length (ML) and the motif starting points in all sequences except the **sequence selected** in step 2, during which 0.1 was added in the count of each position to avoid the zero probability;
5. Compute the probability for each location in the **selected sequence** to be the true motif starting points. These probabilities are calculated by normalizing the weights of each location, which are stored in a list denoted as $A_x$. It can be calculated as

   $$A_x = \frac{Q_x}{P_x}$$

   where:

(1) the probabilities $Q_x$ is computed according to $q_{i,j}$, which is the probability for each character $i \in \{A, C, G, T\}$ in each position $j$ as shown in the PWM in step 3;

(2) the probabilities $P_x$ computed using the background probability $p_{i,j}$, which in our case is 0.25 all pair of $i, j$;

6. Randomly sample a new location of starting point in the **selected sequence** based on the probability calculated in Step 4 and update the starting points list, “pst”.
7. To jump out of local optimization, phase shifts are conducted

   After every 40 iterations, the current predicted motif starting points will be moved simultaneously by certain units toward left or right.

   First of all, the information content of current predicted motif table is calculated and compared to current best solution. If current predicted one is better, update the starting point of best solution, as well as $Best\_IC$;

   In order to determine where to put new starting points, we set a searching range, from -5 to 5, and then calculate the information content ($IC$) for all 11 new “motif tables” based on different movements. After normalizing those 11 values of $IC$, we obtain the probability for the next movement and then sample the movement.

   Since the location of the starting points must be positive while smaller than the sequence length, we check feasibility of the starting point before the movement.
8. Check if the iteration times exceed $(4000 + Extra\_Required\_Iterations)$:

- If false, continue to iterate with Step 2~7
- If true:
  - Check if this is the first time to enter this block:
    - If false, go to Step 9 to check convergence
    - If true:
      - Determine the $Extra\_Required\_Iterations$:

        Set $Extra\_Required\_Iteration = 1000 * (\frac{2*ML-Best\_IC}{2})^2$
      - Go to Step 2~7

The reason why we devise this step is that we first observed the results with smaller number of iterations were always not satisfying, since it may get trapped in local optimum in early stage.

Also, we cannot set a very large compulsory iteration times for all situations. If current predicted motif table shows great information content, we could end the algorithm early. However, with weak information content, there might be better solution, so the algorithm requires more iterations.

In addition, we do not hope the extra required iteration times increase linearly with the deterioration of solution. According to this consideration, we choose a square function to determine whether it needs the extra iterations.

Since the maximum ICPC is 2, the maximum information content for any given motif table should be $2 * ML$. Thus, the goodness of the current best solution can be calculated by $(2 * ML - Best\_IC)$, which is the discrepancy between the maximum possible information content and the information content of current best solution

9. Repeat Step1~6 to until convergence. The criteria for convergence are shown as follows:
    - Sum of Squared Residue of the starting points obtained from $SC$ number of consecutive iterations should be zero ($SSR_{pst}$ = 0):

      Record the results of predicted starting points "$pst$" from latest SC (sequence count) iterations and then compute SSR by summing all the distance between any two records. When the SSR = 0, in other words, when a “pst” fixes for SC times, break iteration.

      The reason why we selected $SC$ as the number of records to determine convergence is that we want to make sure the majority of sequences have been selected and none of them change the starting points. If set this threshold is too small, the algorithm may break too early.
    - Number of iterations < 20000. This criterion is used to restrict the number of iterations at poor condition with weak planted motifs.
10. Do one more time of phase shift after convergence.

There might be a situation with the predicted motif table obtained after reaching the convergence criteria, that the predicted location overlapped with the actual location with the same error for each sequence, which is induced by background DNA noise, as shown in the Figure 1.

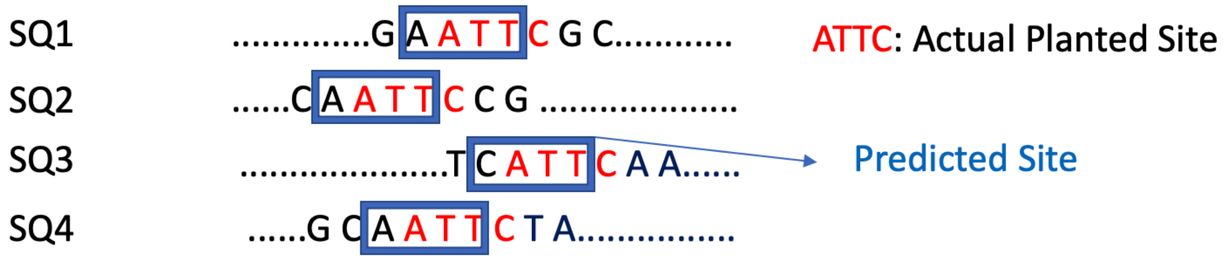

SQ1
SQ2
SQ3
SQ4
.............G A A T T C G C...........
......C A A T T C C G ...................
...................T C A T T C A A......
......G C A A T T C T A................
ATTC: Actual Planted Site
Predicted Site

*Figure 1 Illustration of Background Noise*

This situation cannot be solved by selecting single sequence to modify its starting location as from Step 2~7. Since the probabilities to sample the new location for this selected sequence is determine by other sequences, if all other sequences have the similar error, according to the PWM, the selected sequence has no incentive to change to the next step. When this situation happens. All starting points may remain the same position and then reach the convergence criteria to break out.

Thus, this situation can be improved if there will be one more step to check if moving all starting points simultaneously would be better. Also, since this step is trying to find out if there are any neighborhood with better result, we only sample the best position in this step.

## Part 3: Evaluation

To visualize the results, multiple figures are generalized to shape the influence of ICPC, ML and SC on KL-Divergence, overlapping positions, and overlapping sites. Average number and standard deviation analysis are conducted in this part. In consideration of different maximum number of overlapping sites and overlapping positions, corresponding data is normalized in box plot for accuracy analysis.

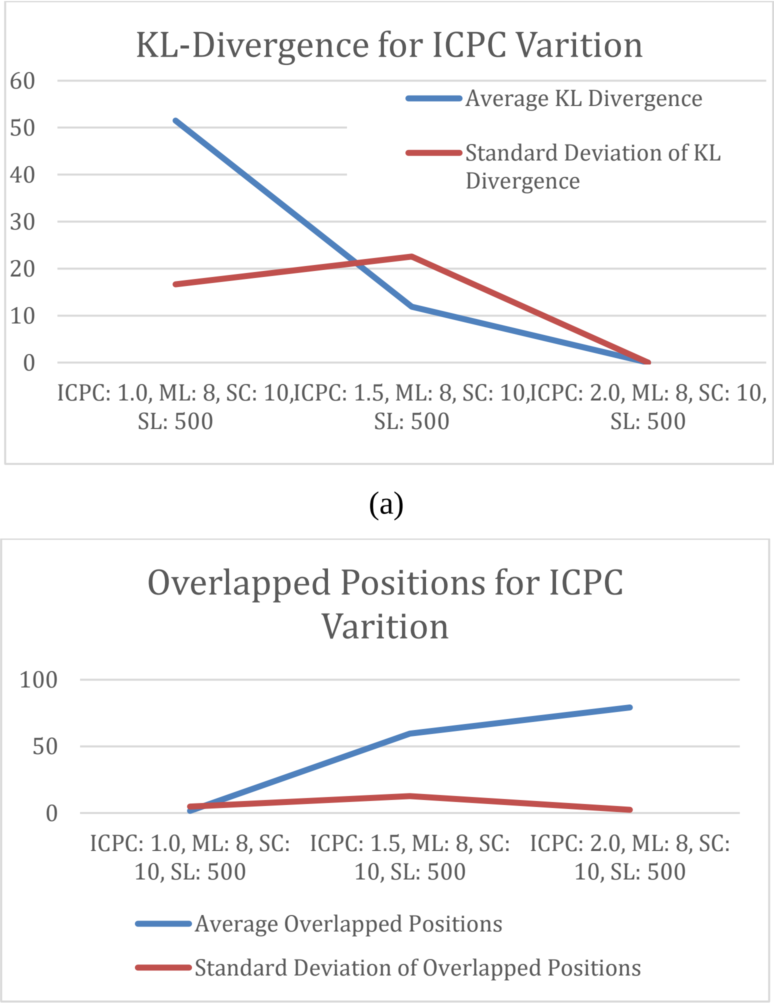
1. Effect of Variation of ICPC

Figure 2 shows the average number and standard deviation for the 10 data sets in the three cases with different ICPC assigned.

Both the average value and standard deviation of KL-Divergence is 0 in when ICPC equals to 2.0 according to the Figure 2(a), which illustrates that the predicted sites are totally consistent with original data given this strong motif, which can also be confirmed by the number of overlapped positions and sites as shown in Figure 2(b) and 2(c).

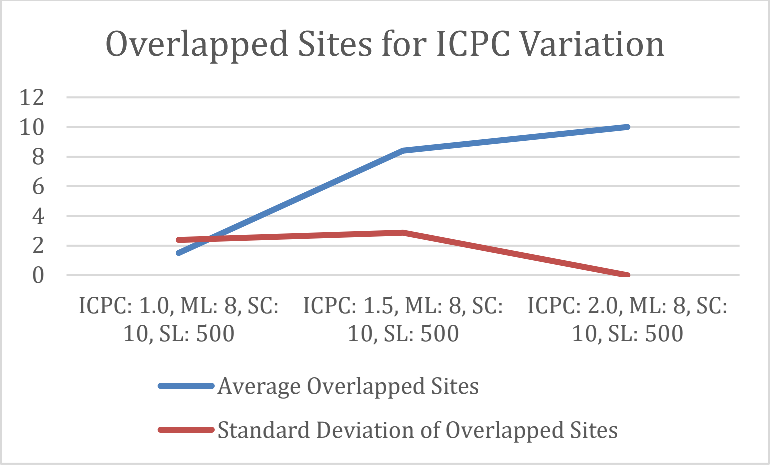
For ICPC = 1.5, the average KL-Divergence is about 10, which is acceptable, but its standard deviation is the largest one. This evidence shows that the performance of our algorithm is not very stable given this medium motif but still can obtain reasonable prediction. With large KL-Divergence but small deviation, the performance of algorithm could not handle the weak motif situation.

The number of overlapping sites and locations are all at a low level in the situation that ICPC is 1.0, so it can be concluded that the prediction results for ICPC=1.0 are not reliable. Also, according to the Figure 2(b), there are over 50 predicted positions overlapped with original one, showing acceptable prediction ability for ICPC = 1.5.



(a)

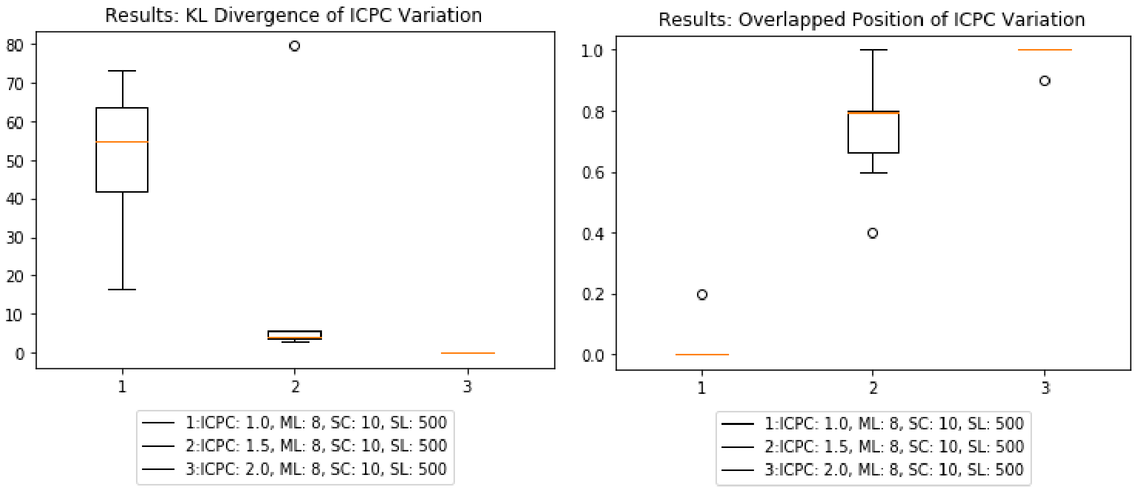
(b)



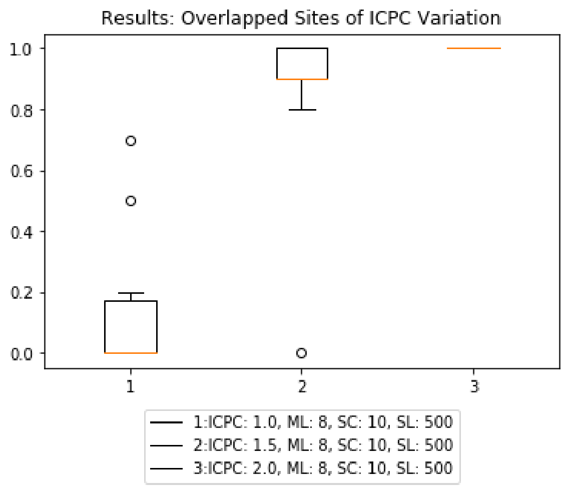
(c)

*Figure 2 Average Value and Standard Deviation of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for ICPC Variation*

By comparing the results for different ICPC, it shows that higher ICPC can help produce better results, which is also shown in the normalized box plots in Figure 3. There is one outlier for the default case in the plot of overlapping position but not existing in the overlapping sites, which can be explained that background noise influences result. As a comparison between "sites.txt" and "predictedsites.txt" of default case in Figure 4, the outlier is underlined, and it can be obviously realized that the sites are totally the same, but the locations are not overlapped.

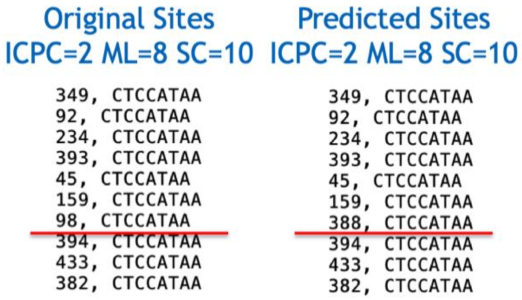


(a) (b)



(c)

*Figure 3 Box Plots of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for ICPC Variation*



Original Sites
ICPC=2 ML=8 SC=10

```
349, CTCCATAA
92, CTCCATAA
234, CTCCATAA
393, CTCCATAA
45, CTCCATAA
159, CTCCATAA
98, CTCCATAA
394, CTCCATAA
433, CTCCATAA
382, CTCCATAA
```

Predicted Sites
ICPC=2 ML=8 SC=10

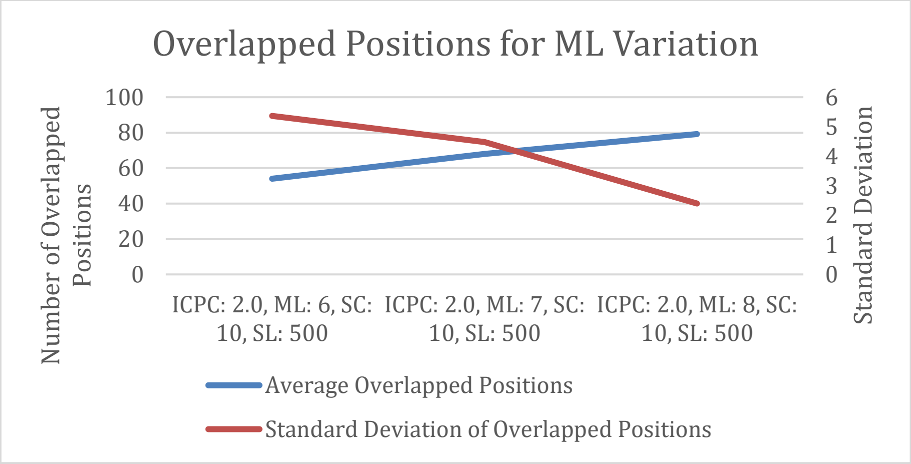
```
349, CTCCATAA
92, CTCCATAA
234, CTCCATAA
393, CTCCATAA
45, CTCCATAA
159, CTCCATAA
388, CTCCATAA
394, CTCCATAA
433, CTCCATAA
382, CTCCATAA
```

*Figure 4 Comparison one Set of Original and Predicted Data in Case of ICPC=2.0, ML=8, SC=10*

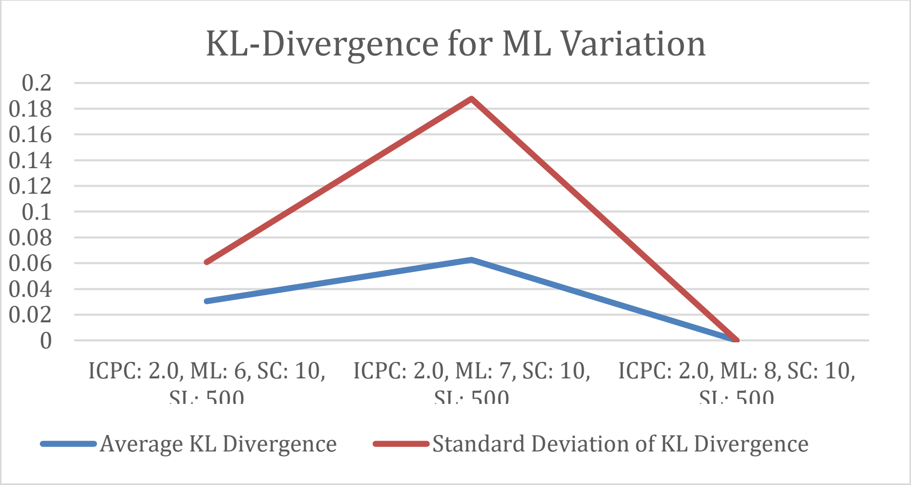
2. Effect of Variation of ML

The average values and standard deviations are shown in the Figure 5. The default case with ML=8 can also offer the best result with zero KL-Divergence and corresponding standard deviation among 10 data sets. The situation with ML=7 does not provide with the second-best result, which it was supposed to have, because of background noise and getting stuck in local optimization as shown in the Figure 6 and Figure 7. The background noise is similar to that of the default case. The labeled site in the Figure 6 is same as the corresponding original site, however, errors will be introduced into the number of overlapping location and KL-Divergence of PWM.

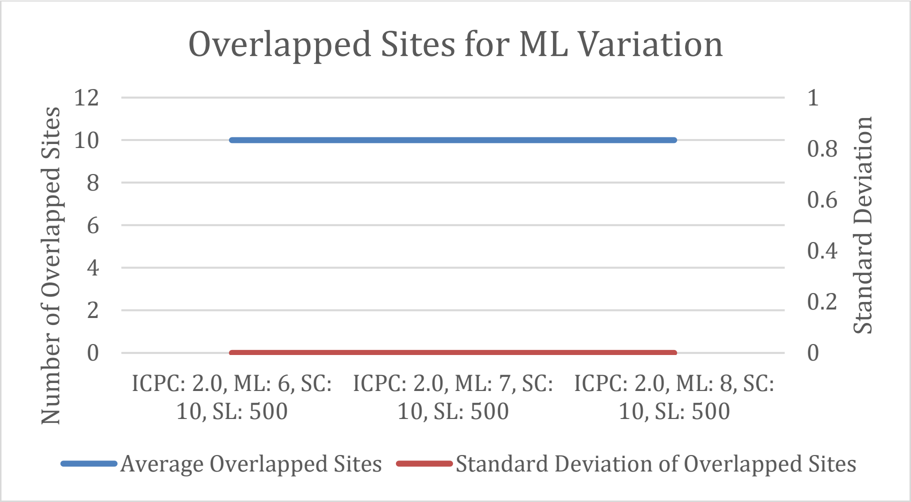
From the box plot in Figure 8, the predicted sites in all ML cases are absolutely the same as what are originally planted in Part 1, which proves that the algorithm can predict sites accurately, even though there is slightly error in the predicted PWM and locations, no matter how ML varies.



(a)

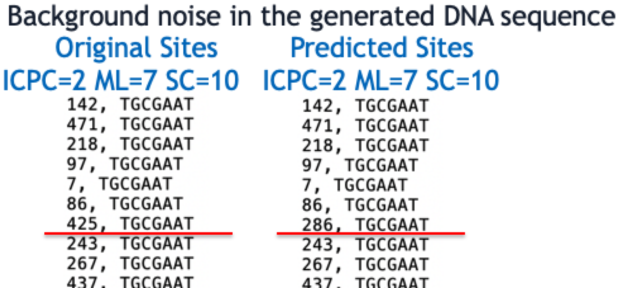


(b)



(c)

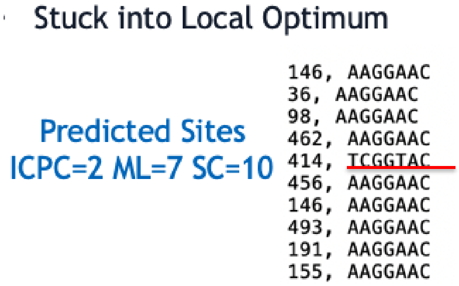
*Figure 5 Average Value and Standard Deviation of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for ML Variation*



Background noise in the generated DNA sequence

```
   Original Sites          Predicted Sites
ICPC=2 ML=7 SC=10       ICPC=2 ML=7 SC=10
  142, TGCGAAT            142, TGCGAAT
  471, TGCGAAT            471, TGCGAAT
  218, TGCGAAT            218, TGCGAAT
  97, TGCGAAT             97, TGCGAAT
  7, TGCGAAT              7, TGCGAAT
  86, TGCGAAT             86, TGCGAAT
  425, TGCGAAT            286, TGCGAAT
  243, TGCGAAT            243, TGCGAAT
  267, TGCGAAT            267, TGCGAAT
  437, TGCGAAT            437, TGCGAAT
```

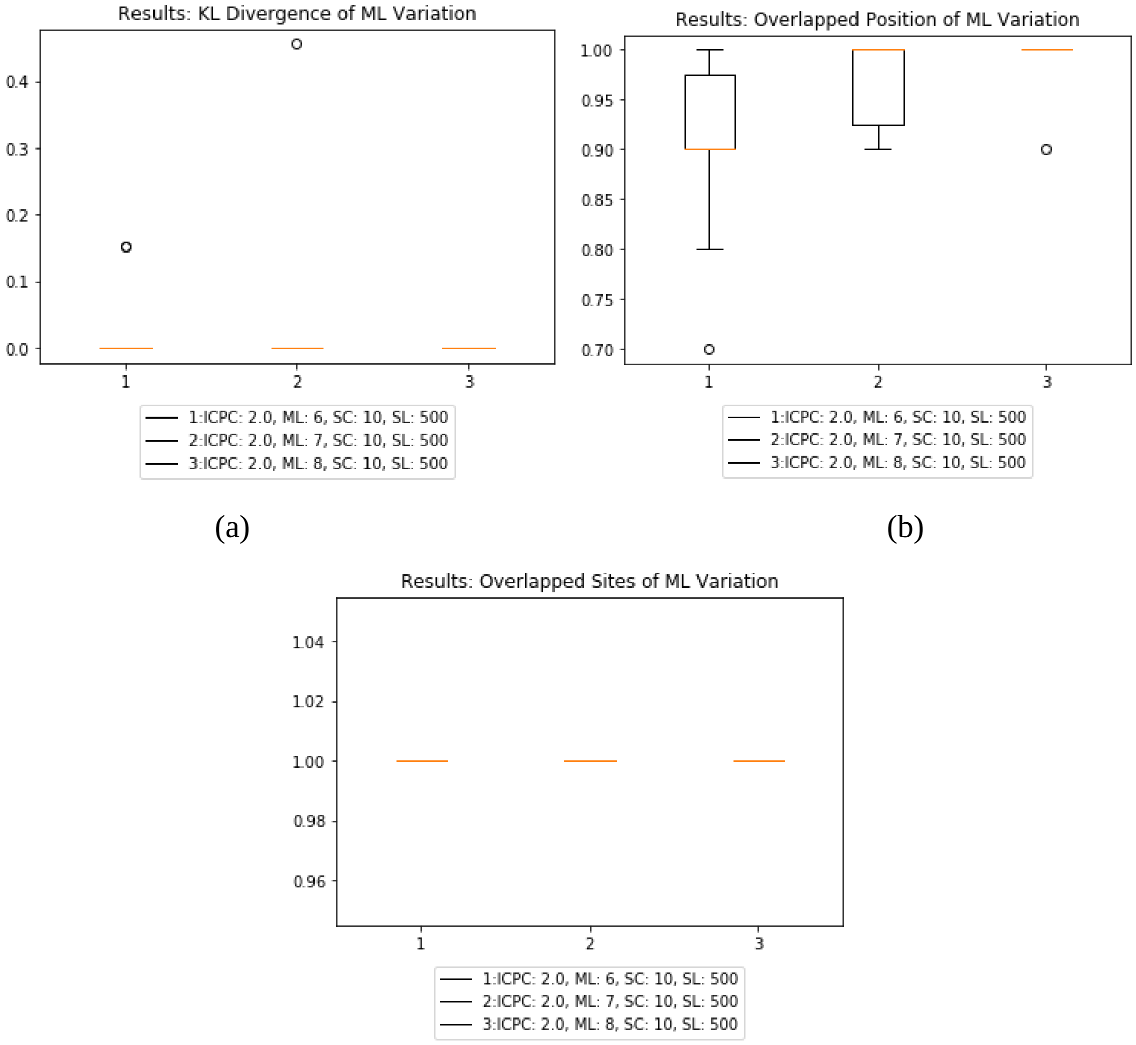
*Figure 6 Comparison one Set of Original and Predicted Data in Case of ICPC=2.0, ML=7, SC=10*



Stuck into Local Optimum

```
                     146, AAGGAAC
                     36, AAGGAAC
                     98, AAGGAAC
  Predicted Sites    462, AAGGAAC
ICPC=2 ML=7 SC=10    414, TCGGTAC
                     456, AAGGAAC
                     146, AAGGAAC
                     493, AAGGAAC
                     191, AAGGAAC
                     155, AAGGAAC
```

*Figure 7 One Set of Getting Stuck into Local Optimum in Case of ICPC=2.0, ML=7, SC=10*



(a)

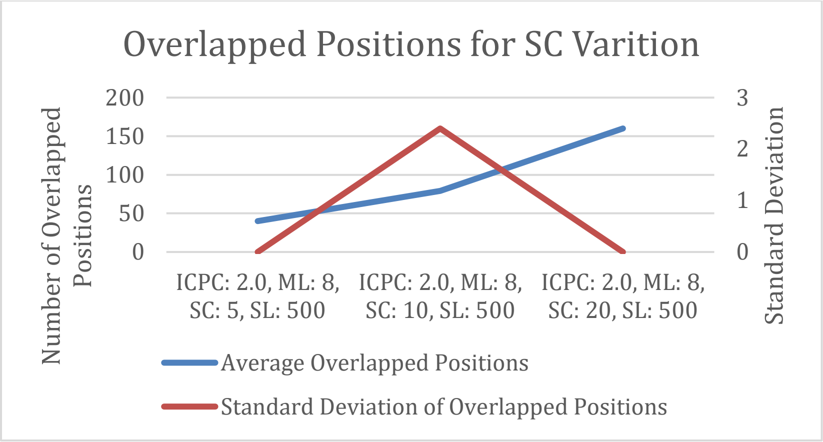
(b)

(c)

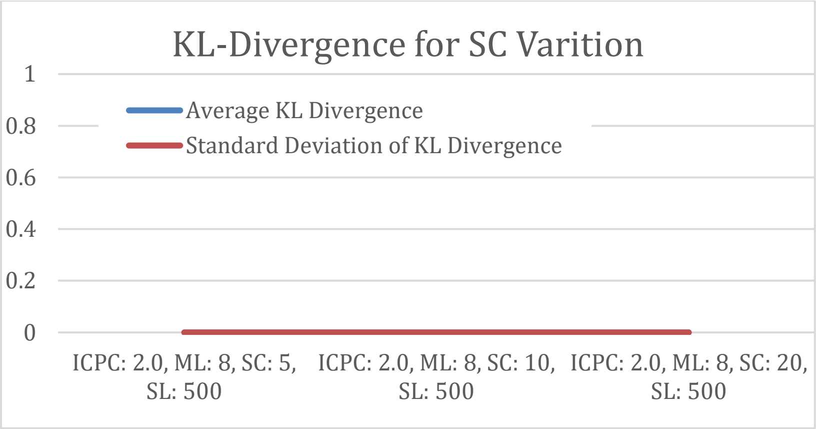
*Figure 8 Box Plots of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for ML Variation*

3. Effect of Variation of SC

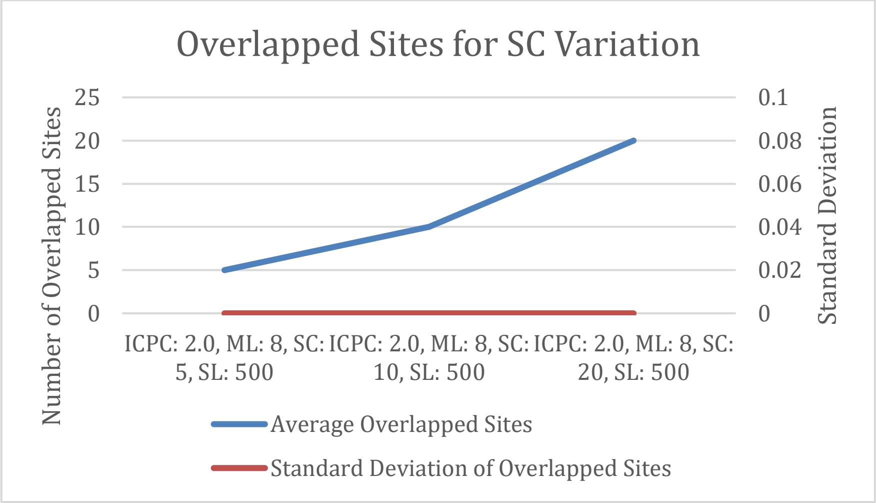
The KL-Divergence keeps 0 in all sets of data as illustrated in the Figure 9, showing that the algorithm can finalize a fully same PWM as original one without influenced by the change of SC. The box plots in Figure 10 reveal that the algorithm can make a significant prediction with SC changes on sites, locations, and PWM at ICPC = 2.0.



(a)

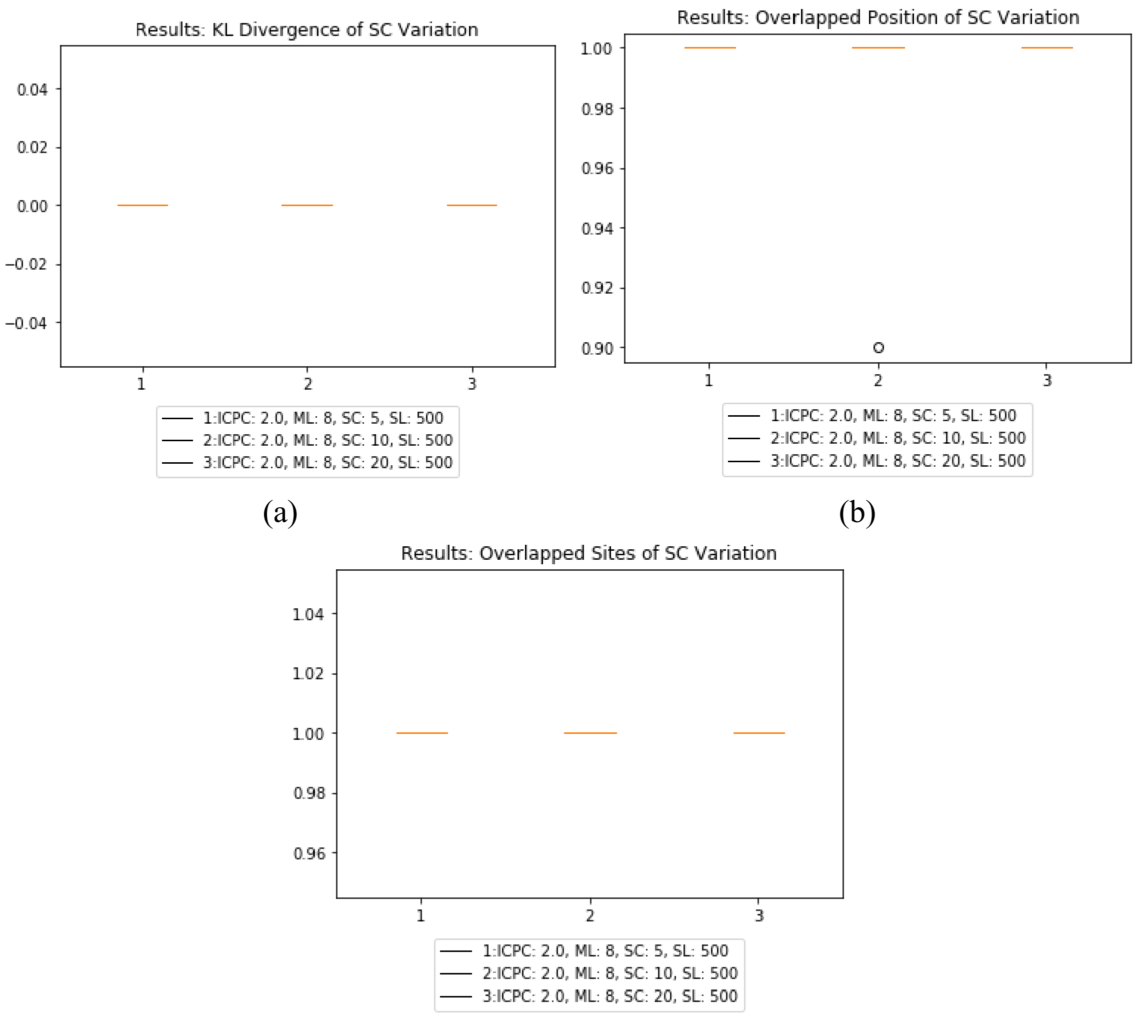


(b)



(c)

*Figure 9 Average Value and Standard Deviation of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for SC Variation*



(a)

(b)

(c)

*Figure 10 Box Plots of (a) KL-Divergence (b) Overlapped Positions (c) Overlapped Sites for SC Variation*

## Conclusion

- No matter how ML and SC change, predicted sites can have good consistency with original planted sites when ICPC = 2. This algorithm has good performances in cases of high ICPC because it has a trend to seek for the starting points with high information content. On the contrary, when ICPC=1, predictions are not reliable.
- The final results may be influenced by background DNA generated in part I.
- There is still a slight possibility that results approach local optimum but not global optimum. However, from the data sets ran in the project previously, though results get stuck into a local optimum, the predicted sites are still thought as the same with planted sites in Part I because of the definition of "same site" in the assignment of project. Therefore, it can be concluded that the algorithm will provide a perfect site prediction when ICPC is at 2.0.

## Reference:

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., & Wootton, J. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science,262(5131), 208-214. doi:10.1126/science.8211139

OpenCourseWare, M. (2015, January 20). 9. Modeling and Discovery of Sequence Motifs. Retrieved from https://www.youtube.com/watch?v=1EMonM7qAU8&t=2459s

Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423

Chapman BA and Chang JT (2000). Biopython: Python tools for computational biology. ACM SIGBIO Newsletter, 20, 15-19

## *Appendix A*

Evaluation Chart

| Parameter Setting | KL divergence | Number of overlapping positions between “sites.txt” and “predictedsites.txt”. | Number of overlapping sites between “sites.txt” and “predictedsites.txt”. | Running time.(sec) |
|---|---|---|---|---|
| ICPC = 2, ML = 8, SL = 500, SC = 10 | 0 | 79.2 | 10 | 29.21801238 |
| ICPC = 1, ML = 8, SL = 500, SC = 10 | 51.50498 | 1.6 | 1.5 | 73.77654171 |
| ICPC = 1.5, ML = 8, SL = 500, SC = 10 | 11.91772 | 59.5 | 8.4 | 58.94058328 |
| ICPC = 2, ML = 6, SL = 500, SC = 10 | 0.030397 | 54 | 10 | 37.74416091 |
| ICPC = 2, ML = 7, SL = 500, SC = 10 | 0.045596 | 67.9 | 10 | 33.9980592 |
| ICPC = 2, ML = 8, SL = 500, SC = 5 | 0 | 40 | 5 | 24.45829511 |
| ICPC = 2, ML = 8, SL = 500, SC = 20 | 0 | 160 | 20 | 49.59641387 |
| Average | 9.071241857 | 66.02857143 | 9.271428571 | 43.96172378 |
| Standard Deviation | 47.10399151 | 118.4206666 | 14.00336694 | 17.70467353 |