IITM-CS6024 : Algorithmic Approaches to Computational Biology Release Date: Aug 23

Assignment 2 Due Date : Aug 31, 23:55

Roll No: BE16B002 Name: Anoubhav Agarwaal

1. (5 points) [Surprising surprises!] When trying to find where in the genome DNA replication begins, every time we find a DnaA box using the most frequent 9-mer, we seem to find some other surprisingly frequent 9-mers. Why do you think this happens? List specific reasons.

Solution: In prokaryotic replication origins, the most abundant repeated sequences are DnaA boxes, which are the binding sites for **chromosomal replication initiation protein DnaA**.

A DnaA box can be thought of as a ‘message’ within the DNA sequence which tells the DnaA protein where to bind. On finding a DnaA box, we seem to find other frequent 9-mers. These 9-mers can be the reverse complements of the DnaA box as the proteins bind to any one of the DNA strands without any specificity. Or, these 9-mers can be hidden* messages for some other proteins to initiate their respective functions for the survival and proper functioning of the organism. Also, it has been observed that these hidden messages tend to cluster within a genome.

Hence, it is because of their **biological significance**, that they appear more surprisingly frequent to us (if assumed that the genome sequence originates from a uniform distribution). However, this is untrue as even prokaryotes have come to existence only after millions of years of evolution. Due to natural selection, i.e., conservation** of what works (about genes) and extinction of what does not, the genomic sequences have distinct evolutionary biases which cannot be explained by merely sampling from a uniform distribution of A, T, G, and Cs. Also, proteins are known to bind to these sequences in more than a single location, which explains the statistical surprise of the higher frequency of occurrence ff these specific 9-mers. [1, 2]

* Hidden because of our limited understanding of the role of the 9-mer.

** To ensure the conservation of genes and to maintain their proper function, it would be essential to have repeats of these sequences for accurate transmission to progeny cells and also to combat deleterious mutations.

2. (10 points) [Expectations in reality] Given a random length-n **circular** genome sequence $Text$ and a fixed k-mer $Pattern$, let the random variable X denote the number of exact occurrences of $Pattern$ in $Text$.

   (a) (4 points) Compute $E[X]$, and use it to bound the statistical significance $Pr[X \geq a]$ (aka p-value) of a 5-mer ACTAT occurring $a = 3$ times in a 30-length string? What is the p-value if the 5-mer is AAAAA?
   (Note: Use Markov’s inequality, $Pr[X \geq a] \leq E[X]/a$, for any non-negative r.v. X; compute $E[X]$ using “linearity of expectation” on $X = \sum_{i=1}^{n-k+1} X_i$, where X_i is

an indicator r.v. that is 1 if *Pattern* occurs in position i in *Text* and 0 otherwise.)

Solution: Expectation is a **linear** operator and does **not** require independence of X_i's. According to the linearity of expectation:

$$E[X] = \sum_{i=1}^{n} E[X_i]$$

The genome is **circular** and thus $E[X_i]$ for $i = n - k + 1, ..., n$ is also non-zero. This is not the case for a linear genome. Here, $n = 30$ and $k = 5$. Also, by symmetry the expected value of the occurrence of a pattern at all positions i is the same. This yields:

$$E[X] = 30 * E[X_1]$$

$E[X_1] = \dfrac{1}{4^k}$, where k = 5 for ACTAT and AAAAA. Thus, $E[X] = \dfrac{30}{4^5}$. This is the expected number of occurrences for both the patterns as the statements made above are general and thus apply to both.
An upper bound for the p-value can be obtained for both the patterns ACTAT and AAAAA using markov's inequality, which is given by:

$$Pr[X \geq a] \leq E[X]/a$$

Here, a = 3. Thus, p-value = $Pr[X \geq 3] \leq E[X]/3$. The upper bound for the p-value is given by:

$$p - value \leq \frac{30}{3 * 4^5} \approx \mathbf{0.009766}$$

(b) (6 points) Compute $E[X^2]$ and hence variance σ^2 for the 5-mer ACTAT and AAAAA, and use it to obtain a tighter bound for the above two p-values.
(Note: Use Chebyshev's inequality, $Pr[|X - E[X]| \geq k\sigma] \leq 1/k^2$; again, compute $E[X^2]$ using linearity of expectation).
(Bonus: Can you provide a formula for $E[X^2]$ for any k-mer *Pattern* using its "self-overlap" or "auto-correlation" structure?)

Solution: Applying linearity of expectation to calculate $E[X^2]$ we get:

$$X = \sum_{i=1}^{n} X_i$$

$$E[X^2] = E\Big[\sum_{i=1}^{n} X_i\Big]^2 = E\Big[\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i * X_j\Big]$$

$$E[X^2] = E\Big[\sum_{i=1}^{n} X_i^2\Big] + E\Big[\sum_{i\neq j} X_i * X_j\Big]; \qquad (1)$$

Now,

$$X_i^2 = X_i * X_i = \begin{Bmatrix} 1, X_i = 1 \\ 0, X_i = 0 \end{Bmatrix} = X_i$$

$$E[X_i^2] = E[X_i] \implies E\Big[\sum_{i=1}^{n} X_i^2\Big] = \sum_{i=1}^{n} E[X_i^2] = \sum_{i=1}^{n} E[X_i] = E[X] = \frac{30}{4^5}$$

Subsitituting this value back in (1),

$$E[X^2] = \frac{30}{4^5} + E\Big[\sum_{i\neq j} X_i * X_j\Big]$$

$$E[X^2] = \frac{30}{4^5} + E\Big[\sum_{i>j} X_i * X_j\Big] + E\Big[\sum_{i<j} X_i * X_j\Big]$$

$$E[X^2] = \frac{30}{4^5} + 2 * E\Big[\sum_{i<j} X_i * X_j\Big] \qquad (2)$$

Upto this point, the calculation for both patterns is the same, the second term in equation (2) gives the difference.

(i) For pattern ACTAT,

$\sum_{i<j} X_i * X_j$ has $\binom{30}{2} = 435$ terms. The pattern is **non-overlapping**. Hence, the expected value of the pattern occurring at position i AND at position $j = i+1, i+2, i+3, i+4$ is zero.

$$E\Big[\sum_{i<j<i+5} X_i * X_j\Big] = 0$$

Note: As the genome is circular it is valid for all i from 1 to 30 as after position 30 it loops back to position 1. Also, the mathematical formulation is to give the general idea. It is not accurate for $i > 26$.

Thus, out of the 435 terms in the second summation in (2), 120 of them are zero. The remaining $X_i * X_j$ pairs are **independent**. If X and Y are two independent random variables, we know that their expectation is given by:

$$E\Big[X * Y\Big] = E\Big[X\Big] * E\Big[Y\Big]$$

By symmetry of our random variables X_i for i = 1,...,30, we obtain the following for the remaining $435 - 120 = 315$ terms:

$$E\Big[\sum_{j>i+5} X_i * X_j] = \sum_{j>i+5} E[X_i] * E[X_j] = \sum_{j>i+5} E[X_i]^2 = 315 * E[X_1]^2$$

Substituting back in (2),

$$E\Big[X^2\Big] = \frac{30}{4^5} + 2*(E\Big[\sum_{i<j<i+5} X_i * X_j\Big] + E\Big[\sum_{j>i+5} X_i * X_j\Big])$$

$$E\Big[X^2\Big] = \frac{30}{4^5} + 2*(0 + 315 * E[X_1]^2) = \frac{30}{4^5} + \frac{630}{4^{10}} \approx \mathbf{0.0299}$$

(i) For pattern AAAAA, Equation (2),

$$E\Big[X^2\Big] = \frac{30}{4^5} + 2*E\Big[\sum_{i<j} X_i * X_j\Big] \qquad (2)$$

$$E\Big[X^2\Big] = \frac{30}{4^5} + 2*(E\Big[\sum_{i<j<i+5} X_i * X_j\Big] + E\Big[\sum_{j>i+5} X_i * X_j\Big]) \qquad (3)$$

The second expectation term in (3) will be the same as the one for the previous pattern as the random variables X_i and X_j are independent for all $j > i+5$.

$$E\Big[\sum_{j>i+5} X_i * X_j\Big] = 315 * E\Big[X_1\Big]^2 = \frac{315}{4^{10}}$$

The first expectation term in (3) will be **non-zero** as the pattern is **completely overlapping.** The expectation of the occurrence of pattern AAAAA at position i AND position i+k (where k =1, 2, 3, 4) is equivalent to finding the probability of a stream of $(5+k)$ AA...AAs starting at position i. This is $\frac{1}{4^{5+k}}$. Hence, the first term in (3) is given by:

$$E\Big[\sum_{i<j<i+5} X_i * X_j\Big] = E\Big[\sum_{k=1}^{4}\sum_{i} X_i * X_{i+k}\Big] = \sum_{k=1}^{4}\sum_{i} \frac{1}{4^{5+k}} = \sum_{k=1}^{4} \frac{30}{4^{5+k}}$$

Substituting the terms back in (3) we get,

$$E\Big[X^2\Big] = \frac{30}{4^5} + 2*\left(\sum_{k=1}^{4} \frac{30}{4^{5+k}} + \frac{315}{4^{10}}\right) \approx \mathbf{0.0494}$$

Variance σ^2 for ACTAT $= E[X^2] - E[X]^2 \approx 0.0299 - 0.000858 = \mathbf{0.0290}$
Variance σ^2 for AAAAA $= E[X^2] - E[X]^2 \approx 0.0494 - 0.000858 = \mathbf{0.0485}$

Chebyshev's inequality: $Pr[|X - E[X]| \geq k\sigma] \leq 1/k^2$.
For pattern ATCAT: Here, $E[X] = 0.293$ and $\sigma = 0.029^{0.5} = 0.170$
We want to determine a tighter bound on the p-value i.e., $Pr[X \geq 3]$ using the Chebyshev's inequality as we now have information about the variance of X.

$|X - E[X]| = |X - 0.293| = X - 0.293$, as we are want to obtain $P[X \geq 3]$.

Substituting $E[X]$ and σ it into the inequality we get,

$Pr[X \geq k * 0.170 + 0.293] \leq 1/k^2$ .On comparing this with $Pr[X \geq 3]$.

$$k * 0.170 + 0.293 = 3 \implies k = \frac{3 - 0.293}{0.17} = 15.924$$

$$p - value \quad for \quad ATCAT = P[X \geq 3] \leq \frac{1}{15.924^2} = \mathbf{0.00394}$$

For pattern AAAAA: Here, $E[X] = 0.293$ and $\sigma = 0.0485^{0.5} = 0.220$
We want to determine a tighter bound on the p-value i.e., $Pr[X \geq 3]$ using the Chebyshev's inequality as we now have information about the variance of X. $|X - E[X]| = |X - 0.293| = X - 0.293$, as we are want to obtain $P[X \geq 3]$.

Substituting $E[X]$ and σ it into the inequality we get,

$Pr[X \geq k * 0.220 + 0.293] \leq 1/k^2$ .On comparing this with $Pr[X \geq 3]$.

$$k * 0.220 + 0.293 = 3 \implies k = \frac{3 - 0.293}{0.22} = 12.304$$

$$p - value \quad for \quad AAAAA = P[X \geq 3] \leq \frac{1}{12.304^2} = \mathbf{0.00660}$$

3. (10 points) [Random coding] You are about to implement a variant (Gibbs sampling) of one of the top 10 algorithms of the 20th century (Metropolis algorithm for Monte Carlo).

   (a) (7 points) Solve the HackerRank Challenge `https://www.hackerrank.com/assignment-2-aacb` that asks you to apply Gibbs sampling technique to find motifs in a set of sequences. Please follow random seed, pseudocount, and other instructions carefully to produce the exact same output as the test cases.

   Solution: The link contains two code files. **gibbs sampler.py** is the gibbs sampler code I have used to find better motifs than sample output in Q3 (b). Also, it has been used in Q4 (b) for comparison with MEME output. This python file does not follow the snippets provided for pseudo random number generation. **hackerrank submission replication.py** is my attempt on replicating the snippets provided. However, I was not sucessful in reproducing the expected outputs for the test cases. This file follows (to my knowledge) all the instructions given to us. https://drive.google.com/open?id=1wEP9K1Kfv9KSZTkB09l4OtgD4xWafnLV

   (b) (3 points) Can you find a better-scoring motif for "Sample Input 1" than the one that your current algorithm outputs? What parameters of your algorithm did you tune to achieve this better-scoring motif? Feel free to run your code outside of HackerRank to answer this subquestion.

Solution: Yes, I found a better-scoring motif for 'Sample Input 1' using the same given values for the parameters a, b, m, t, N and num_random_starts. I have initialized $Xo = 1$ which gives a **motif score of 73**. The given sample 1 output has a motif score of 104. Two other sets of parameters were also tried by increasing num_random_starts and N as these parameters can be viewed as going deeper and wider (exploitation vs. exploration) in the possible search space, respectively.
Results obtained for Sample 1 (given sample output has score of 104):

(k, t, N, num_random_starts, Score) : (15, 20, 2000, 2, **73**)
(k, t, N, num_random_starts, Score) : (15, 20, 2000, 20, 63)
(k, t, N, num_random_starts, Score) : (15, 20, 4000, 60, 63)

Results obtained for Sample 0 (given sample output has score of 92):

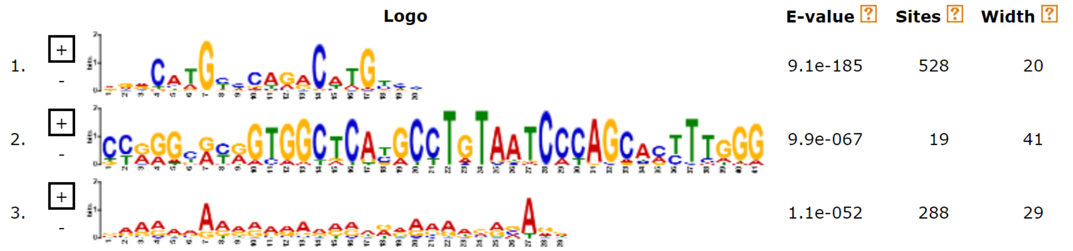
(k, t, N, num_random_starts, Score) : (15, 20, 2000, 2, 74)
(k, t, N, num_random_starts, Score) : (15, 20, 2000, 20, 64)
(k, t, N, num_random_starts, Score) : (15, 20, 4000, 25, 64)

4. (10 points) p53 is a transcription factor that suppresses tumor growth through regulation of dozens of target genes with diverse biological functions. This master regulator is inactivated in nearly all tumors. Let's try to identify the DNA-binding motif onto which p53 transcription factor binds using a motif-finding tool called MEME (available via Galaxy or stand-alone).

   (a) (6 points) Run MEME on this input fasta file to get 3 candidate motifs. Of these three, which one do you think actually binds to p53 and why? The input fasta file contains a sample of ∼200-500bp sequences bound by p53 in a ChIP-seq experiment[1]. MEME may also have output two other motifs - argue by giving specific reasons whether those motifs could be true binding elements of p53?

Solution: Three candidate motifs were obtained after running MEME on p53.fa. The fasta file contained 721 sequences. We allowed for zero or one occurrence per sequence(zoops). The following motifs were obtained:

[1] Our initial search for motifs in the 2000-bp promoter sequences of known target genes of the p53 didn't yield expected results, so we resorted to ∼200-500bp p53-bound sequences obtained from the peaks of a genome-wide p53 ChIP-seq experiment described at the p53 BAER resource



	Logo	E-value	Sites	Width
1.		9.1e-185	528	20
2.		9.9e-067	19	41
3.		1.1e-052	288	29

Of these three, I think **the first motif logo** in the image (above) binds to p53 because of two reasons. Firstly, it has the lowest E-value of 9.1e-185. This is a statistical significance metric calculated for the motif by MEME. The E-value of a motif is based on its log likelihood ratio, width, sites, the background letter frequencies, and the size of the training set. Secondly, The number of sites(or locations in sequences) contributing to the construction of the motifis is much higher for the first motif compared to the other two. It uses 528 sites out of 721. Whereas, the other two use only 19 and 288 sites out of 721 possible sequences. The width of the motif is also a critical factor. However, that has been accounted for in the E-value.

(b) (4 points) Run your Gibbs Sampler code from last question on the sequences provided. Report the motif you found, and did it match the one output by MEME? Report the parameter values you tried and chose for motif width (k), number of iterations (N), and the number of random restarts.

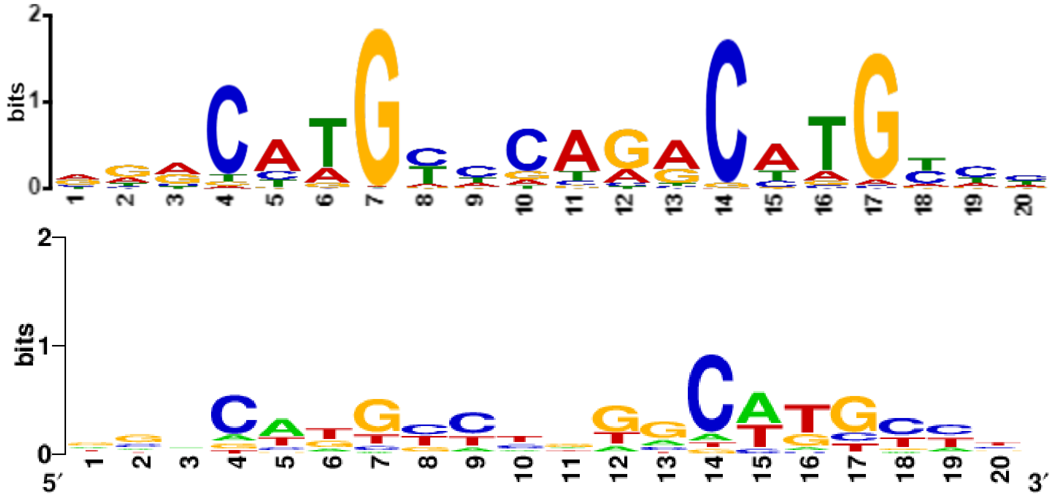
Solution: The gibbs sampler code was ran on the p53 fasta file using multiple combinations of parameters. Here, I am reporting only the best parameters(based on motif score) for k = 20, 41 and 29 and their corresponding motif logos. The following motif widths (k) were chosen in order to compare with the three candidate motifs obtained from MEME.

K	t	N	num_random_starts	Score	Max Score
20	721	8000	40	7088	10815
41	721	8000	40	17405	22170
29	721	8000	40	11819	15681

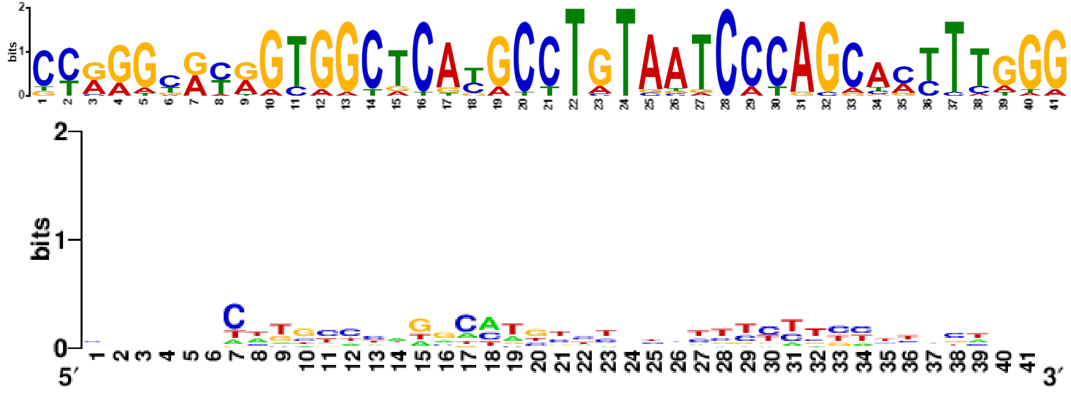
The Max Score is theoretically the maximum possible difference between a consensus string and its motifs of length k. The Gibbs sampler implementation does not provide similar results to MEME. The motif scores obtained from the Gibbs sampler are very high and close to the maximum possible score. This is mainly because the gibbs sampler considers all 721 sequences to find the motifs. Whereas, MEME considers **zero or one occurrence** of a motif in a sequence.

Comparison of motif logos between:

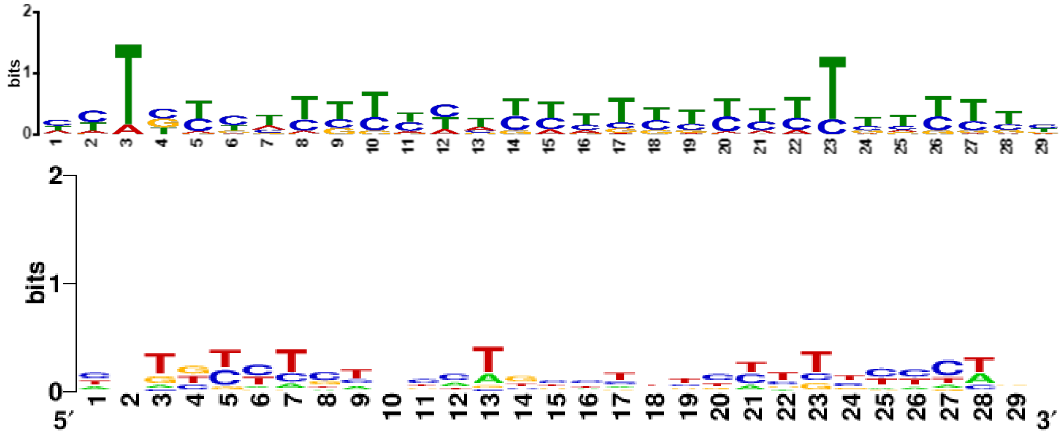
First candidate motif(top) and Gibbs sampler output(bottom)



Second candidate motif(top) and Gibbs sampler output(bottom)



Third candidate motif(top) and Gibbs sampler output(bottom)



In the first candidate motif logos (k = 20), they show some similarity at positions 4, 7, 14, 15, 16, and 17. In the second candidate motif (k=41), the Gibbs sampler performs much worse with no similarity between the motif logos. Also, the information content (in bits) is very low for the Gibbs sampler output. For the third candidate motif (k = 29), the results are again poor with minimal similarity between the motifs. The only way to improve this with the current Gibbs sampling implementation is to run the code for even more larger values of N and num_random_start.

5. (5 points) [Research exploratorium]: Provide properly-formatted references for papers in this solution.

   (a) (3 points) Read this paper on how to read a paper `https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf`. What new tip/trick did you learn from this paper that you didn't already know before?

Solution: I was aware of only the first pass in the three-pass approach. I had no formal methodology to perform a literature survey. This was completely new to me. The new tricks and tips I learned that are highly applicable to me are:

Three-pass approach

1. The first pass should take from 5-10 minutes. In which one should be able to determine if the paper should be read further.
2. Answering the 5 Cs, i.e., Category, Context, Correctness, Contributions, and Clarity after the first pass.
3. The second pass (takes an hour) involves reading the paper with greater emphasis on understanding the figures, graphs, and ignoring proofs. The goal is to grasp the content by the end of this pass.
4. The goal of the third pass is to re-create the work of the author. During this pass, one should jot down ideas for future work. This pass takes about 4-5 hours for a beginner. One should be able to identify the papers strong and weak points as well as its implicit assumptions.

How to perform a literature survey

1. Through an academic search engine such as google scholar or *CiteSeer* and use apt keywords to find 3-5 recent papers in the area.
2. Perform one-pass on each of the papers. Then read the related work sections.

3. Finding a recent survey paper is the best situation.
4. If there is no survey paper, find shared citations and repeated author names to identify the key papers and researchers in the field.
5. Download the key papers.
6. Go to the websites of key researchers and identify the top conferences by checking where they have published recently.
7. Go to the conference websites, scan through recent papers. One should be able to identify high-quality related work on scanning.
8. Make two passes through the key papers downloaded before and the ones obtained from the conference website.
9. Iterate if necessary, to find more influential papers based on shared citations from the bunch.

(b) (2 points) What are some latest research publications you could find on *de novo* motif finding? Report one such paper based on Gibbs-sampling-like heuristics and another based on machine learning approaches such as SVM or deep learning. Try to answer this question by looking only at the papers' title/abstract.

Solution:
Some latest research publications in de novo motif finding are:

1. Paper based on Gibbs-sampling-like heuristic:
   ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. [3]
   The researchers have developed an ultrafast and accurate motif-finding algorithm using a combination of novel discriminative heuristic seeding, Gibbs sampling and length extension methods. The tool is named ProSampler. The core algorithm has the following steps:
   (a) Generating background sequences
   (b) Indentifying significant k-mers
   (c) Constructing preliminary motifs and their position weight matrices.
   (d) Constructing the motif similarity graph
   (e) Gibbs sampling
   (f) Extending core motifs

2. **A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs.**[4]
   The paper presents ShapeMF, a Gibbs sampling algorithm that identifies the de novo shape motifs by working on DNA shape data and thus extending de novo motif discovery.

3. Paper based on SVM/ Deep learning:
   DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery[5]
   DeepFinder is a motif discovery pipeline which uses an ensemble of motif discovery tools like MEME, Bioprospector, MDSCAN, and MotifSampler for the initial prediction of candidate binding sites from a subset of input sequences. Then from these candidate binding sites features are extracted for a deep neural network. This is then used to construct a motif prediction model.

References

[1] M. Rajewska, K. Wegrzyn, and I. Konieczny, “AT-rich region and repeated sequences – the essential elements of replication origins of bacterial replicons,” *FEMS Microbiology Reviews*, vol. 36, pp. 408–434, 03 2012.

[2] P. Compeau and P. Pevzner, “Bioinformatics Algorithms: An Active Learning Approach 2nd Edition, Vol. I,”

[3] Y. Li, P. Ni, S. Zhang, G. Li, and Z. Su, “ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery,” *Bioinformatics*, 05 2019.

[4] M. A. H. Samee, B. G. Bruneau, and K. S. Pollard, “A de novo shape motif discovery algorithm reveals preferences of transcription factors for dna shape beyond sequence motifs,” *Cell Systems*, vol. 8, no. 1, pp. 27 – 42.e6, 2019.

[5] N. K. Lee, F. L. Azizan, Y. S. Wong, and N. Omar, “Deepfinder: An integration of feature-based and deep learning approach for dna motif discovery,” *Biotechnology & Biotechnological Equipment*, vol. 32, no. 3, pp. 759–768, 2018.