Ahmedabad University

MAT277 : Probability and Stochastic Processes

(Room:hn_bio_08)

Domain: BioInformatics

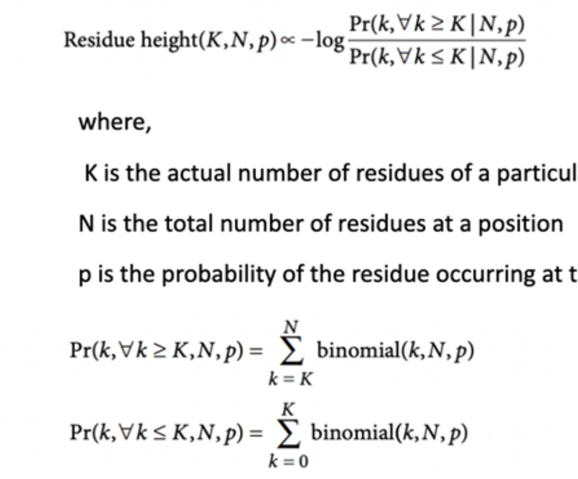
Milee Bajpai	AU2140158
Krishna Patel	AU2140170
Harsh Pandya	AU2140171
Yax Prajapati	AU2140230

Work 1:

- Title: pLogo: a probabilistic approach to visualising sequence motifs
- URL: https://arep.med.harvard.edu/pdf/Oshea_nmeth_13.pdf
- Summary:

The study introduces pLogo, a new method for visualising protein or nucleic acid motifs by stacking residues based on their statistical significance. It allows for real-time adjustments and can be used for genomic and proteomic sequencing data. It also has potential applications beyond visualisation due to its probability-based framework.

The pLogo is a graphical representation of a position weight matrix (PWM), with a value assigned to each residue at each position.These values are visually represented by pLogo as the height of the residues above or below the x-axis.In the pLogo, the height of the residues is proportional to the log odds value.



$$\text{Residue height}(K,N,p) \propto -\log \frac{\Pr(k, \forall k \geq K \mid N,p)}{\Pr(k, \forall k \leq K \mid N,p)}$$

where,

K is the actual number of residues of a particular type at a given position

N is the total number of residues at a position

p is the probability of the residue occurring at that position

$$\Pr(k, \forall k \geq K, N, p) = \sum_{k=K}^{N} \text{binomial}(k,N,p)$$

$$\Pr(k, \forall k \leq K, N, p) = \sum_{k=0}^{K} \text{binomial}(k,N,p)$$

The pLogo uses a log-odds binomial probability approximation to calculate residue heights, which has desirable properties such as being intuitive and well-defined. The computation of binomial probabilities and PWM scores is similar to that for full data sets, and logs are computed in base ten. Fixing residues in motifs involves selecting a subset of the data and calculating binomial probabilities and PWM scores. Fixed residues are depicted in pLogo with full height on a grey background.

The position of the red horizontal pLogo bar is used to determine the importance of overrepresented or underrepresented residues in a motif.

$$\text{Number of binomial calculations} = \sum_{i \in R} C_i$$

The correction formula determines the position of the bar on a log scale based on the number of statistical tests and the expected alpha value.

$$\alpha' = \frac{0.05}{\sum_{i \in R} C_i}$$

$$\text{Statistical significance bar position} = \pm\log\left(\frac{\alpha'}{1-\alpha'}\right)$$

The study used single-cell sequencing to examine the dynamics of the immune response in mice. The results showed that this technique offers a more detailed understanding of the immune response and could provide new insights into immunity and aid in therapy development.

Work 2:

- Title: A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length
- URL: https://academic.oup.com/bioinformatics/article/21/10/2240/206744
- Summary:

Introduction:

Extraction of DNA motifs from a set of unaligned sequence fragments (also known as the multiple local alignment or MLA) is often applied to identify DNA sites that are recognized by transcription factors.

System and methods:

Different approaches to this problem were reviewed and a probabilistic model based on Gibbs sampling appeared to be the most efficient method. The algorithm will be used as a modified version of Lawrence's algorithm. Two probabilistic models - foreground and background will be formulated. The algorithm, called a Gibbs sampler, uses a Markov Chain Monte Carlo (MCMC) method to search for the most likely DNA motif.

Algorithm:

The probabilities q(i, r) for the occurrence of nucleotide r at site position i, i = 1..s, where s is the site length and the background nucleotide probabilities f (r) are estimated from the in-site and the background counters denoted by c(i, r) and g(r):

$$q(i,r) = \frac{c(i,r) + b(r)}{M + B} \qquad (1)$$

$$f(r) = \frac{g(r) + b(r)}{K + B}, \qquad (2)$$

where M is the number of sites in the set, from which the statistics are derived, K is the number of all non-site positions in the data. Pseudocounts b(r) are proportional to the frequencies of nucleotides in the full dataset, while their sum

$$B = \sum_r b(r) \sim \sqrt{N},$$

Where N is the number of data sequences.

To select the set of similar sites, start with randomly scattered sites of a definite length, one per sequence. Then, organise a cycle of one-by-one updates of site positions. At each step, only one sequence will be selected. The site absence will be treated as a position of specific type (“null”). We estimate the positional nucleotide probabilities within the motif. For each selected sequence R = r1 r2 · · · rl−1 rl , the probability (likelihood) to obtain this sequence from a Bernoulli process (i.e. the site position likelihood) given the site position k is:

$$\begin{aligned} P(\mathrm{R}|[k], q, f) &= \prod_{i=1}^{k-1} f(r_i) \prod_{i=k}^{k+s-1} q(i-k+1, r_i) \\ &\quad \times \prod_{i=k+s}^{L-s+1} f(r_i) \quad k \neq 0 \\ P(\mathrm{R}|[0]) &= \prod_{i=1}^{L-s+1} f(r_i), \end{aligned}$$

where ri is the i-th nucleotide in sequence R and [k], k = 1..(L −s + 1) denotes the event ‘the site starts at position k’, [0] corresponds to the case where the site is absent (‘null position’).

The marginal probability of the sequence itself (evidence) is:

$$P(\mathrm{R}_{|q,f}) = \sum_{k=0}^{L-s+1} P(\mathrm{R}|[k], q, f) \cdot P([k]).$$

The probability (posterior) for a site to start at k is:

$$\begin{aligned} P([k]|\mathrm{R}, q, f) &= \frac{P(\mathrm{R}|[k], q, f) P([k])}{P(\mathrm{R}_{|q,f})} \\ &= \frac{P(\mathrm{R}|[k], q, f) P([k])}{P(\mathrm{R}_{|q,f})}. \end{aligned}$$

Combining the priors with the likelihoods in the usual Bayesian way, we obtain the posterior distribution for a site position in the current sequence and sample the new site position (possibly the 'null'one) from this distribution.The process is iterated until the chain comprising sets of site positions converges.

Conclusion:

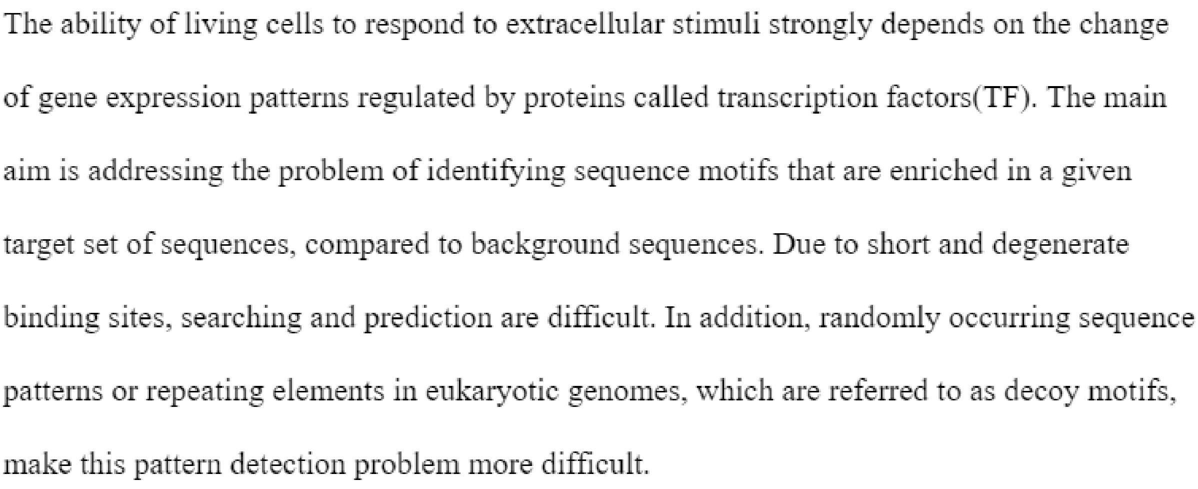
The algorithm improves the estimation of the signal length by using an estimation method that takes into account the symmetrical structure of the motif. The algorithm takes into account the symmetrical structure and spacing of the DNA motif in the DNA sequences to improve its accuracy of identifying the DNA motif. The algorithm was tested on simulated and real DNA sequences and was shown to perform well compared to other existing methods.

Work 3:

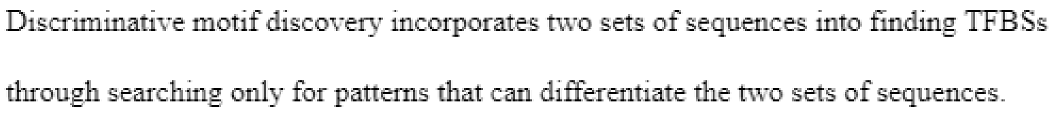
- Title: Probabilistic Models for Semisupervised Discriminative Motif Discovery in DNA Sequences
- URL: https://ieeexplore.ieee.org/document/5557858
- Summary:

1. **INTRODUCTION**

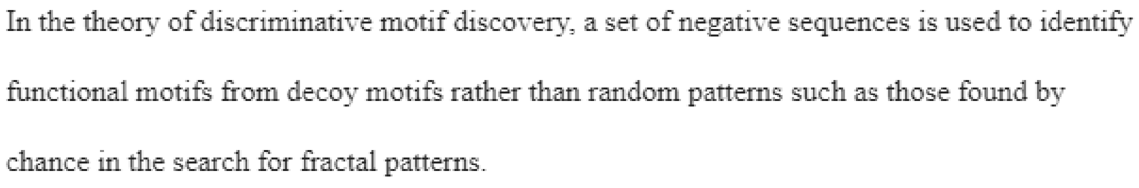
This paper describes a probabilistic model for discovering discriminatory motifs in DNA sequences with semi-supervised learning. The model uses a Bayesian framework to incorporate labelled and unlabeled data to improve the discovery of motifs that differentiate between different classes of DNA sequences.



The ability of living cells to respond to extracellular stimuli strongly depends on the change of gene expression patterns regulated by proteins called transcription factors(TF). The main aim is addressing the problem of identifying sequence motifs that are enriched in a given target set of sequences, compared to background sequences. Due to short and degenerate binding sites, searching and prediction are difficult. In addition, randomly occurring sequence patterns or repeating elements in eukaryotic genomes, which are referred to as decoy motifs, make this pattern detection problem more difficult.



Discriminative motif discovery incorporates two sets of sequences into finding TFBSs through searching only for patterns that can differentiate the two sets of sequences.



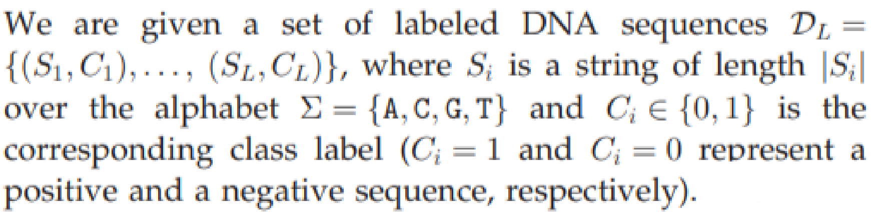
In the theory of discriminative motif discovery, a set of negative sequences is used to identify functional motifs from decoy motifs rather than random patterns such as those found by chance in the search for fractal patterns.

Generative models can easily exploit unlabeled sequences in addition to label sequences to better understand functional motifs. For this,semi-supervised learning is preferred for the set of negative sequences.

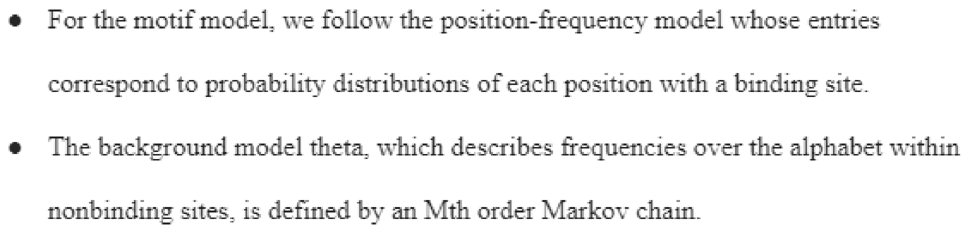
This hybrid model makes use of informative unlabeled sequences to learn motifs and at the same time exploits the set of negative sequences to remove decoy motifs.

2. PROBABILISTIC MODELS FOR MOTIF DISCOVERY

2.1 Notation

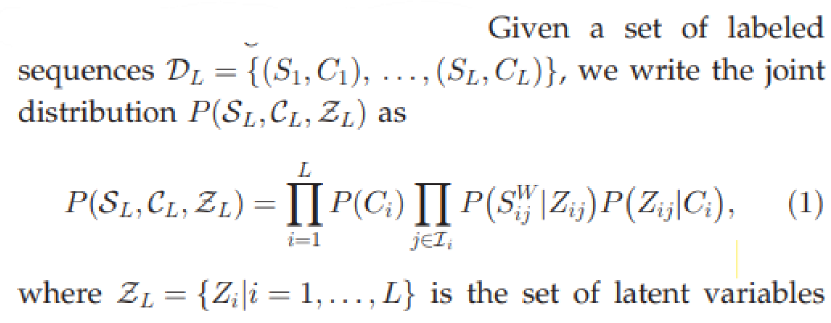


We are given a set of labeled DNA sequences $\mathcal{D}_L = \{(S_1, C_1), \ldots, (S_L, C_L)\}$, where S_i is a string of length $|S_i|$ over the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $C_i \in \{0, 1\}$ is the corresponding class label ($C_i = 1$ and $C_i = 0$ represent a positive and a negative sequence, respectively).



- For the motif model, we follow the position-frequency model whose entries correspond to probability distributions of each position with a binding site.
- The background model theta, which describes frequencies over the alphabet within nonbinding sites, is defined by an Mth order Markov chain.

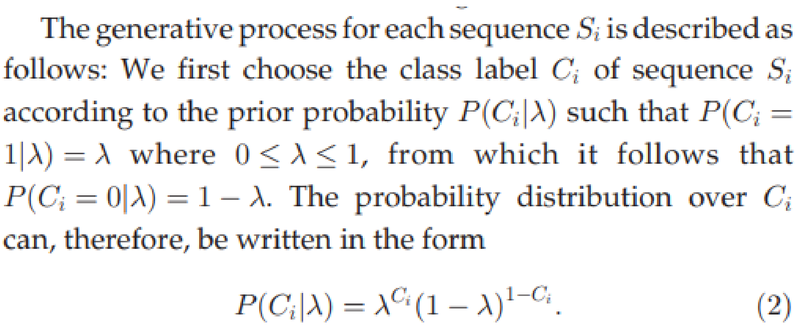
2.2 Generative Model



Given a set of labeled sequences $\mathcal{D}_L = \{(S_1, C_1), \ldots, (S_L, C_L)\}$, we write the joint distribution $P(\mathcal{S}_L, \mathcal{C}_L, \mathcal{Z}_L)$ as

$$P(\mathcal{S}_L, \mathcal{C}_L, \mathcal{Z}_L) = \prod_{i=1}^{L} P(C_i) \prod_{j \in \mathcal{I}_i} P(S_{ij}^{W} | Z_{ij}) P(Z_{ij} | C_i), \qquad (1)$$

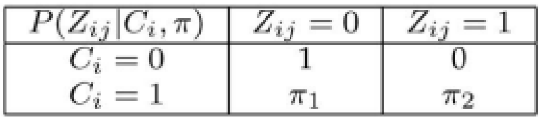
where $\mathcal{Z}_L = \{Z_i | i = 1, \ldots, L\}$ is the set of latent variables



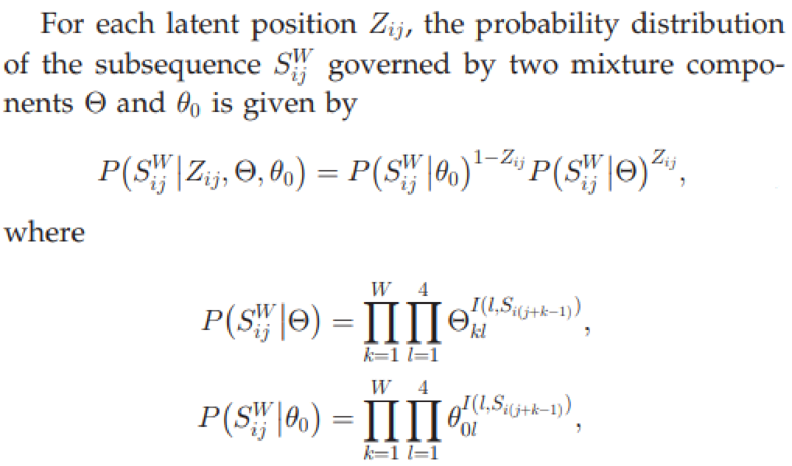
The generative process for each sequence S_i is described as follows: We first choose the class label C_i of sequence S_i according to the prior probability $P(C_i|\lambda)$ such that $P(C_i = 1|\lambda) = \lambda$ where $0 \leq \lambda \leq 1$, from which it follows that $P(C_i = 0|\lambda) = 1 - \lambda$. The probability distribution over C_i can, therefore, be written in the form

$$P(C_i|\lambda) = \lambda^{C_i}(1-\lambda)^{1-C_i}. \qquad (2)$$

The Table of Probabilities $P(Z_{ij}|C_i, \pi)$



$P(Z_{ij}\|C_i, \pi)$	$Z_{ij} = 0$	$Z_{ij} = 1$
$C_i = 0$	1	0
$C_i = 1$	π_1	π_2



For each latent position Z_{ij}, the probability distribution of the subsequence S_{ij}^W governed by two mixture components Θ and θ_0 is given by

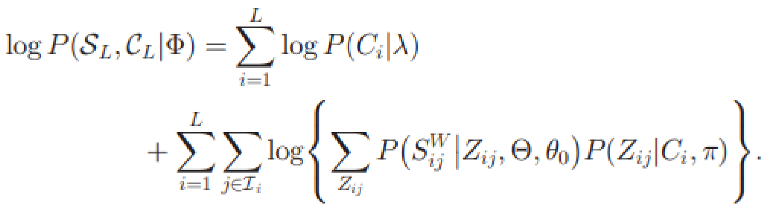
$$P(S_{ij}^W|Z_{ij}, \Theta, \theta_0) = P(S_{ij}^W|\theta_0)^{1-Z_{ij}} P(S_{ij}^W|\Theta)^{Z_{ij}},$$

where

$$P(S_{ij}^W|\Theta) = \prod_{k=1}^{W} \prod_{l=1}^{4} \Theta_{kl}^{I(l, S_{i(j+k-1)})},$$

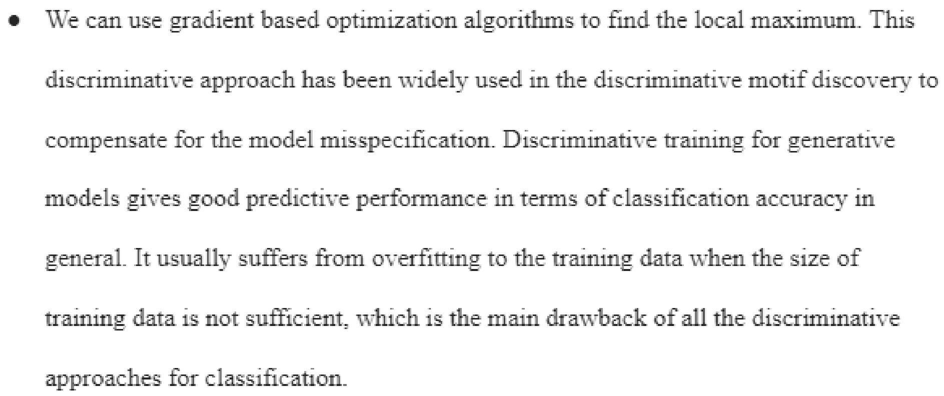
$$P(S_{ij}^W|\theta_0) = \prod_{k=1}^{W} \prod_{l=1}^{4} \theta_{0l}^{I(l, S_{i(j+k-1)})},$$

2.3 Learning: EM and Discriminative Training



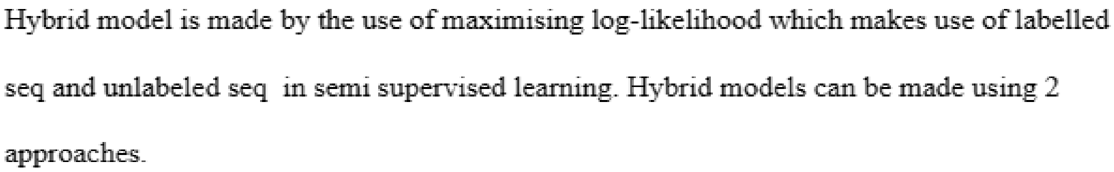
$$\log P(\mathcal{S}_L, \mathcal{C}_L|\Phi) = \sum_{i=1}^{L} \log P(C_i|\lambda)$$
$$+ \sum_{i=1}^{L} \sum_{j \in \mathcal{I}_i} \log \left\{ \sum_{Z_{ij}} P(S_{ij}^W|Z_{ij}, \Theta, \theta_0) P(Z_{ij}|C_i, \pi) \right\}.$$

$$\log P(\mathcal{C}_L|\mathcal{S}_L,\Phi)=\sum_{i=1}^{L} C_i \log P(C_i=1|S_i,\Phi) + \sum_{i=1}^{L}(1-C_i)\log P(C_i=0|S_i,\Phi).$$



- We can use gradient based optimization algorithms to find the local maximum. This discriminative approach has been widely used in the discriminative motif discovery to compensate for the model misspecification. Discriminative training for generative models gives good predictive performance in terms of classification accuracy in general. It usually suffers from overfitting to the training data when the size of training data is not sufficient, which is the main drawback of all the discriminative approaches for classification.

3. HYBRID MODELS



Hybrid model is made by the use of maximising log-likelihood which makes use of labelled seq and unlabeled seq in semi supervised learning. Hybrid models can be made using 2 approaches.

- Takes two likelihoods(Generative and Discriminative) and combines to log-likelihoods.
- Takes two probabilistic models (Generative and Discriminative)and then combines them using prior distribution.

3.1 Discuss the formula for making hybrid model

$$P(\Theta^G,\Theta^D)=P(\Theta^D|\Theta^G)P(\Theta^G) = \prod_{k=1}^{W}\{P(\Theta_k^D|\Theta_k^G)P(\Theta_k^G)\}.$$

; Θ^G and Θ^D for generative and discriminative models.

3.2 Method for making log-likelihoods maximise using gradient concept.

Also the concept of hash function used in sequence binding is mentioned in section 3.2.

3.3 Double stranded DNA

It says that in Generative model we have assumed one strand DNA but in hybrid model we can take upon double strand DNA.

4. Related work:

Section4 shows how the 6 diff probabilistic model compares to the hybrid model for defining its structure.

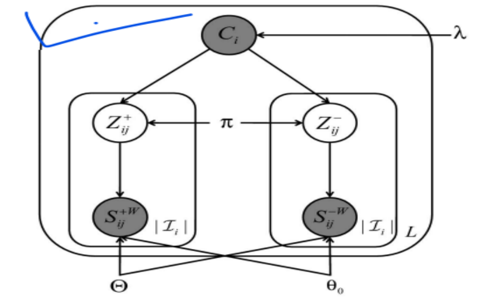
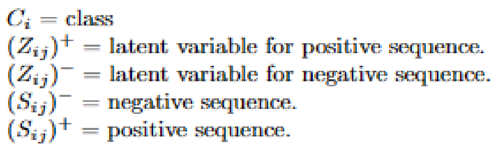


Fig. 3. Graphical representation of the generative model for a set of labeled DNA sequences.



C_i = class
$(Z_{ij})^{+}$ = latent variable for positive sequence.
$(Z_{ij})^{-}$ = latent variable for negative sequence.
$(S_{ij})^{-}$ = negative sequence.
$(S_{ij})^{+}$ = positive sequence.

1.OOPS:

Only subsequence is generated from the motif model and this model is designed by greedy algorithm and updated by Gibbs sampling (Heuristic).

2. ZOOPS:

Zero or One subsequence is generated from the motif model and it has also used gibbs sampling algorithm.

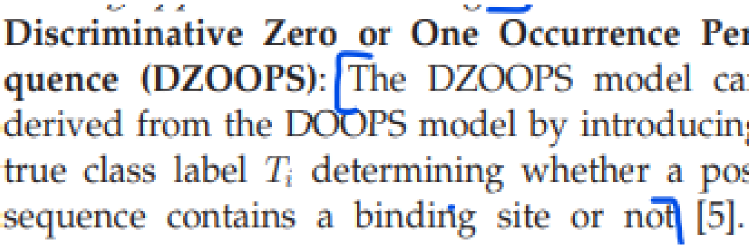
3.MOPS:

It's a further extension of ZOOPS model and it is also referred as Two component mixture model.

4.<u>DOOPS:</u>

Exactly one subsequence is generated by selecting latent position.

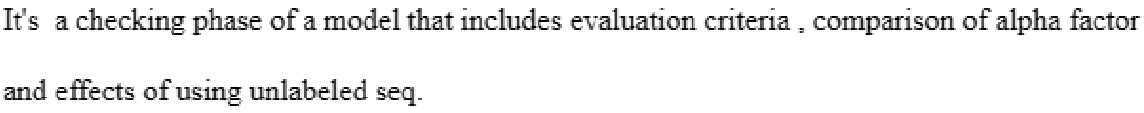
5.<u>DZOOPS:</u>



Discriminative Zero or One Occurrence Pe **quence (DZOOPS):** The DZOOPS model ca derived from the DOOPS model by introducin true class label T_i determining whether a pos sequence contains a binding site or not [5].

6 <u>DMOPS</u>: Similar to 5th model but it will check for multiple occurrences.

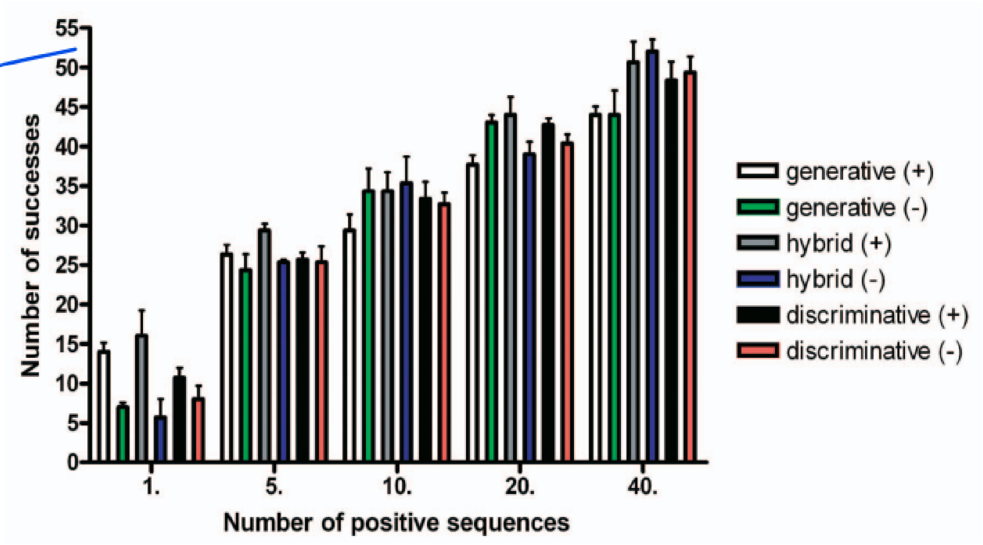
5. Results:

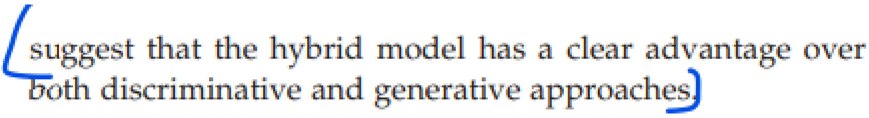


It's a checking phase of a model that includes evaluation criteria , comparison of alpha factor and effects of using unlabeled seq.

- <u>Comparison of alpha</u>:

  It tells that the alpha factor got affected if positive sequence is more compare to negative sequence and it need to be balanced to retain its adjusted value.





suggest that the hybrid model has a clear advantage over both discriminative and generative approaches.

- <u>Using unlabeled seq.</u>

  It reveals the importance of unlabeled seq on the value of alpha factor.

<u>6. Conclusions:</u>

Hybrid models are more effective and optimised for developing unique sequences in motif discovery.

REFERENCES:

(2013, October 6). pLogo: a probabilistic approach to visualizing sequence motifs. Retrieved February 5, 2023, from https://arep.med.harvard.edu/pdf/Oshea_nmeth_13.pdf

(2022, August 10). YouTube. Retrieved February 5, 2023, from https://doi.org/10.1093/bioinformatics/bti336

(2022, August 10). YouTube. Retrieved February 5, 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5557858&isnumber=5958719