StatsM254 Statistical Methods in Computational Biology Lecture 8 - 04/24/2014

Lecture 8

Lecturer: Jingyi Jessica Li *Scribe: Ruyi Huang*

1 Introduction

In this lecture, we discussed about the Expectation-maximization algorithm (E-M algorithm) and the application of it. E-M algorithm is an iterative method for finding maximum likelihood or maximum of a posterior likelihood in statistical models, where the model depends on imcomplete dataset with unobserved value.[1],[2]. It is derived from Bayes Theorem, which defines that the probability of a hypothesis Y conditional on a given body of data X is the ratio of the unconditional probability of the conjunction of the hypothesis with the data to the unconditional probability of the data alone.

$$P(Y|X) = P(Y,X)/P(X) = P(X|Y)P(Y)/\int P(Y|x)P(x)dx$$

$P(Y|X)$ is the conditional probability of Y on X.

Based on the Bayes Theorem, R.A. Fisher brought up the numerical procedure of maximum likelihood extimation (MLE) early in 1912 and the method in 1922. [3]The making of maxium likelihood was one of the most important developments because Fisher further defined the notion of "likelihood" as a quantity for appraising hypothetical quantities on the basis of given data. The "maximum likelihood" gives estimates satisfying the criteria of "sufficiency" and "efficiency"[6]. The EM algorithm is a further developed version of MLE and the first implementation of the EM algorithm is used to estimate gene frequencies in a random mating population [4] in which the new method sacrificed some information to simplify the calculations and speed up the computation. The new method is shown to be equivalent to maximum likelihood and fully efficient in the statistical sense. Dempster and his co-workers generalized the method and sketched a convergence analysis for a wider class of problems.

With the EM algorithm, latent variables are usually used in addition to unknown parameters and known data observations, which means this model assumes there are missing values among the data or the model can be formulated more simply by assuming the existence of additional unobserved data points. And these characters make EM algorithm a powerful approach to translate the maximum likelihood for incomplete dataset and that would be perfect for studying and analyzing genetic problem in large population scale since it is really hard to track down all the individuals in certain huge population. The paper we discussed during the lecture[5] is focusing on applying EM algorithm in human blood type linkage and segregation analysis and the application of EM in allels linkage suggests that the EM algorithm could also be applied for searching and detecting other highly linked sequence such as gene or protein motif.

2 EM outline

2.1 Notion and Definition

By applying the EM algorithm, people want to solve the cases where there are two sets of data: a set of observed data X and a set of unobserved data Y, and want to estimate a vector of interested parameter θ. The maximum likelihood inference would be:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta; X)$$

ℓ: the log likelihood

Alternatively, the Bayesian inference is

$$\hat{\theta}_b = \arg\max_{\theta} p(\theta|X) \propto p(X|\theta)\pi(\theta)$$

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

Expectation step (E step):

Calculate the expetation of the likelihod function under the conditional distribution of observed data.

$$Q(\theta|\theta^{(n)}) = E[\log P(X, Y|\theta)|X, \theta^{(n)}]$$

* Filling in the missing data

Maximization step (M step):

Find the parameter that maximizes this quantity:

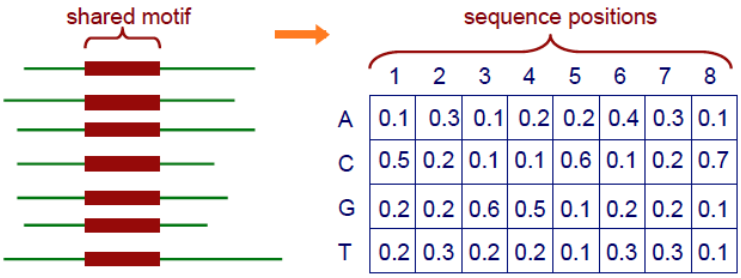
$$\theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)})$$

2.2 Motif Models

A sequence motif is a sequence pattern of biological significance, for example, the DNA sequences corresponding to protein binding sites and protein sequences corresponding to common functions or conserved pieces of structure are both sequence motifs.[8] [9]The motif model is important because it can help us improve our understanding of the regions of sequences that are "functional". In DNA sequences, the motifs can help us understand how the genes are regulated.[7]

Below are the basic concepts of motif study.

Positive specific weight matrix (PWM) Given a set of aligned sequences, it is straightforward to construct a profile matrix characterizing a motif of interest.



$$\begin{matrix} A \\ T \\ C \\ G \end{matrix} = \begin{bmatrix} \theta_{1A} & ... & \theta_{nA} \\ \theta_{1T} & ... & \theta_{nT} \\ \theta_{1C} & ... & \theta_{nC} \\ \theta_{1G} & ... & \theta_{nG} \end{bmatrix}$$

Figure1. The Positive specific weight matrix (PWM) for DNA sequence motif. $\sum \theta_{ij} = 1$ *nucletide position,* $i = 1.....W$, $j \in [A, T, G, C]$

In binding site we observe:

$$X_1, ..., X_W, \text{where } X_i \in [A, T, G, C]$$

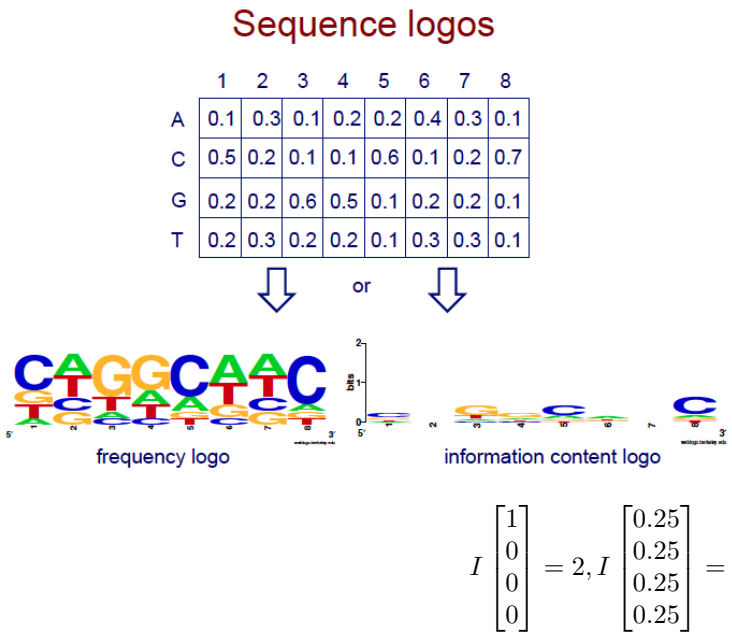
The basic assumption is: $X_i \perp X_j$, $i \neq j$

Likelihood:

$$L(\theta, X) = P(X_1, ..., X_n|\theta) = \sum_{i=1}^{W} \theta_{iX_i}$$

$$I(\theta) = bit \rightarrow \text{ defined as } 2 + \sum_{j\in(A,T,C,G)} \theta_{ij} \log_2 \theta_{ij} \in [0, 2]$$

Based on the likelihood calculation, we can get the PWM value equal to 2 when one nucleitide is dominent in the specific position and no dominace when the PWM value equal to 0



$$I\begin{bmatrix}1\\0\\0\\0\end{bmatrix} = 2, I\begin{bmatrix}0.25\\0.25\\0.25\\0.25\end{bmatrix} = 0$$

Figure2. The logo plot generated by the bit of the likelihood of certain nucleitide appers in specific position

Prodict motif from observed binding sites

Data: n nucleotides $X_1, ..., X_n$ Discrete $(\theta = (\theta_A, \theta_T, \theta_C, \theta_G))$, X_i are the observed binding site:
prior $\pi(\theta) \sim \text{Dirichlet}(\alpha_A, \alpha_T, \alpha_C, \alpha_G)$

"Conjugated prior" to multinominal

$$\pi(\theta) = \frac{\Gamma(\alpha_A + \cdots + \alpha_G)}{\Gamma(\alpha_A) \cdots \Gamma(\alpha_G)} (\theta_A^{(\alpha_A - 1)})(\theta_T^{(\alpha_T - 1)})(\theta_C^{(\alpha_C - 1)})(\theta_G^{(\alpha_G - 1)})$$

$\Gamma(\alpha) = \int_0^\infty \mu^{\alpha-1}\epsilon^{-\mu}d\mu$
$\Gamma(n) = (n-1)!$
$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

Posterior

$$p(\theta|X) \propto p(X|\theta)\pi(\theta) \propto \theta_A^{n_A+\alpha_A-1}\theta_T^{n_T+\alpha_T-1}\theta_C^{n_C+\alpha_C-1}\theta_G^{n_G+\alpha_G-1}$$

$^*n_i = \sum_{i=1}^{n} I(X_i = i)$
$\theta|X \sim Dirichlet(\tilde{\alpha})$, where $\tilde{\alpha} = (n_A + \alpha_A, n_T + \alpha_T, n_C + \alpha_C, n_G + \alpha_G)$.

Posterior mean:

$$\hat{\theta}_i = E[\theta_i|X] = \frac{n_i + \alpha_i}{n + \alpha_0}, \quad i \in \{A, T, C, G\},$$

where $\alpha_0 = \sum_{i \in \{A,T,C,G\}} \alpha_i$.

3 Possible Extensions

The possible extension usage of the EM and motif model would be trying to see the motif model in protein motif detection. That would be harder than DNA motif model because comparing with DNA which only have 4 nucleotides, there are more than 20 different kinds of peptides in protein to constitute the motif.

Based on the EM approach, we can also estimate the width of the motif to find multiple modif in a long sequence.

4 Conclusions

Based on the paper, the EM algorithm is useful when it is difficult to optimize the likelihood directly, the likelihood can be decomposed by the introduction of latent values and it is also easy to optimize the function (with respect to Θ)

The major disadvantages of EM are slow convergences near the max and the maximazation will be trapped at local max.

References

[1] Dempster, N. M. Laird, D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1. (1977), pp.1-38

[2] Sundberg, Rolf . "Maximum likelihood theory for incomplete data from an exponential family". *Scandinavian Journal of Statistics* 1 (2): 4958.(1974)

[3] R. A. Fisher and the Making of Maximum Likelihood 1912 1922. *Statistical Science* 1997, Vol. 12, No. 3, 162]176

[4] R. Ceppellini, "The estimation of gene frequencies in a random mating population", *annuals of Human Genetics*, 1955, 20:97-115

[5] Ott J. "Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis." *annuals of Human Genetics*, 1977, 40(4):443-54

[6] Little, Roderick J.A.; Rubin, Donald B. " Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics." *New York John Wiley Sons* 134-136.

[7] Martin-Lf, Per The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scand. J. Statist. 1* (1974), no. 1, 318.

[8] Einicke, G.A.; Falco, G.; Malos, J.T. (May 2010). "EM Algorithm State Matrix Estimation for Navigation". *IEEE Signal Processing Letters* 17 (5): 437440.

[9] Timothy L. Bailey. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers" *UCSD Technical Report CS94-351*