

# Gibbs Sampling and EM Algorithm: a Case Study on Statistical Methods in Motif Finding

Siyuan Ma

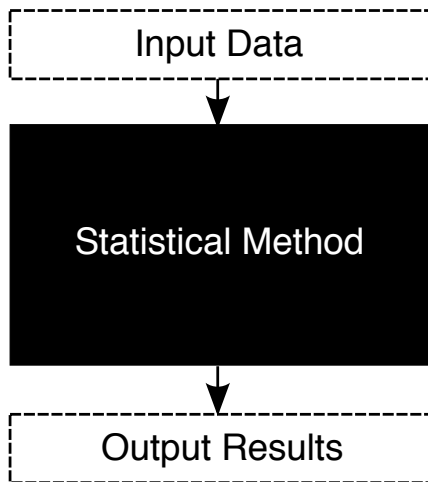
April 2, 2015

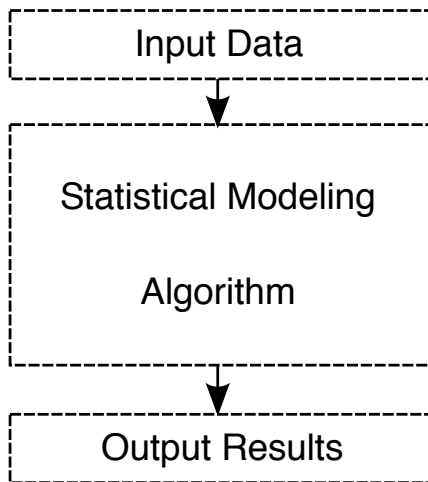
# Problem Statement

- The goal is to find common segments (motif) from a collection of DNA sequences.

```
cole1      taatgttttgctgtttttgtggcatcggcgagaaagagcgggtgggtgnaagactgtttttttagctgtttccaaaaaagggaagtcacagtcgttgcag
ecoarabop  gacaaaaacggtaacaaaagtgtctataatcacggcagaaaagtccacatgtgatatttgcacggcgccacaccttgcctatgcccatagcattttatccat
ecobgr1    acaaatcccantaaacttaattattggatctgttatataaactttataaattccaaaattacacaaagttataaactgtggcctggctcatcttttatcaat
ecocrp     cacaaagcgaagactatgctaaaaacgtcaggatgctacgttaatacattgatgactgcatgatgcaaaaggagtcacataccgtgcagtcacgttggatgc
ecocya     accgtgctacactgtatgtagccatctttcttaaggtaacacgcatgggttaattgatcagcttttagacatttttctgctgaaactaataaaacc
ecodecop  gtagaattattgaaaccagatcgcattacagtgatgcnaacttgtaagtagatttccctaaattgtaggtgatctcgaagtgctgtggagtagatgttagnata
ecogale   ggcataaaaaacggctaaatctttgtgtaagattccactaaattattccatgcaacatttgcacatcttgtatgctatgggttaattcataccataagcc
ecoilvbrp gctcggcgggttttttttatctgcaattcagtaaaaaagtgatcaaacctcaatttccctttgcaaaaaatttccatgtctccctgtaagctgt
ecolac    aacgcaat taatgtgagttagctactcaataggcaaccggctttacactttatgctccggctgctatgttgtgtggaattgtagcggatcaaaatttcaac
ecomale   acat ta ccccaattcigtacagagatcacacaaagcagcggggcgtagggcaagggaaggaaagggtgpcgctaaagaactagagtcogttta
ecomalk   ggggggggggggggggagacacggctctgtgaaclaaaccgggtcagtaaggaatttctgtagttgttgcraaaactcggggcatttatgtgagca
ecomalt   gatcagctgctttttaggtgagttgtaataaagatttggatttgacaagtgcaattcagacacataaaaaacgtcagcttgcattagaaagtttct
ecoompa  gctgcaaaaaagattaaacacacttatcaagacttttttcaatgcccagggagttcacactgtaagttcaactacgctgtagactttacacagcc
ecotnaa  tttttaaacttaaaattctacgtatattatcttataaaaaagcatttatattgctcccgaaagctgtgattcagttacatacttaaaactttcaga
ecouxu1  ccatagagtgaaattgttgtggtttaaaccattagaaattcgggattgacatgcttaccaaaaggtagaacttaacgctctcactccgattcagagc
pbr-p4    ctgcttaactatggcactcagacagattgactagaggtgcaccataggggtgtgnaatccgcacagatgctgaagggaanaatccgcatcagggctc
(tdc)     ggttttaactttaaactgtgtatattaaaggattttatgttaataacgatactctggaaagattgaaagttantttgtgaggtgcccacataccctgt
```

18 DNA Sequences. Article: Stormo and Hartzell, 1989



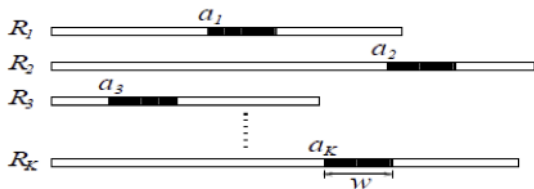


# Statistical Modeling

- This includes two component: modeling the **structure** of the data, and modeling the **mechanism** of the data generation process.

# Data Structure

- We have  $K$  observations  $R_1, R_2, \dots, R_K$  ( $K$  DNA sequences).
- These sequences have a common subsequence (motif) of length  $w$ .
- Use  $A = (a_1, a_2, \dots, a_K)$  to denote the starting locations of these subsequences.



Mathematical representation of the data. Article: Jun Liu, 2001

# Data Generation Mechanism

- Given  $A$  (where the motif starts on each sequence):
  - For basepairs outside of the motif, the appearance of A, C, T, G follows the same **Multinomial Distribution** with **parameters**  $(\theta_{01}, \theta_{02}, \theta_{03}, \theta_{04})$ .
  - For the  $j$ -th basepair inside the motif, the appearance of A, C, T, G follows a multinomial distribution with parameters  $(\theta_{j1}, \theta_{j2}, \theta_{j3}, \theta_{j4})$ .

# Data Generation Mechanism

- Given  $A$  (where the motif starts on each sequence):
  - For basepairs outside of the motif, the appearance of A, C, T, G follows the same **Multinomial Distribution** with **parameters**  $(\theta_{01}, \theta_{02}, \theta_{03}, \theta_{04})$ .
  - For the  $j$ -th basepair inside the motif, the appearance of A, C, T, G follows a multinomial distribution with parameters  $(\theta_{j1}, \theta_{j2}, \theta_{j3}, \theta_{j4})$ .
- What this is really saying is:
  - Sequences outside the motif are completely random.
  - Basepairs inside the motif appears in a certain pattern.



## Now pause and think...

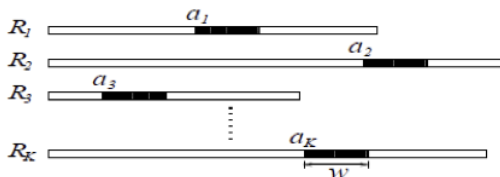
- What are we hoping to get out of this characterization of the problem?

## Now pause and think...

- What are we hoping to get out of this characterization of the problem?
  - What are the **parameters**?
  - What is the **data**?
  - How do we learn about the parameters using the data?

## Now pause and think...

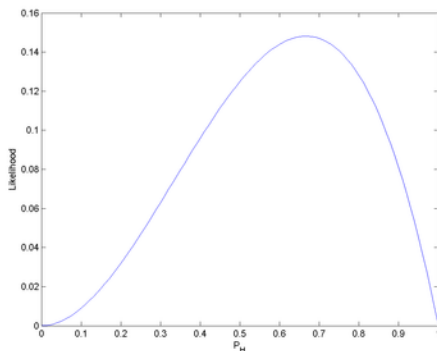
- What are we hoping to get out of this characterization of the problem?
  - What are the **parameters**?
  - What is the **data**?
  - How do we learn about the parameters using the data?



Mathematical representation of the data. Article: Jun Liu, 2001

## Likelihood: a criterion for choosing the best parameters

- Given the **data structure**, the **data generation mechanism** (usually with some **parameters**), we want to pick the set of parameters that best explain the data.
- One way to do this is through the likelihood criteria:



A likelihood function. Credit: wikipedia

# Likelihood: a criterion for choosing the best parameters

- In words, the likelihood for a (data, parameters) pair is how likely the data would appear, given that the parameters are what's generating the data.

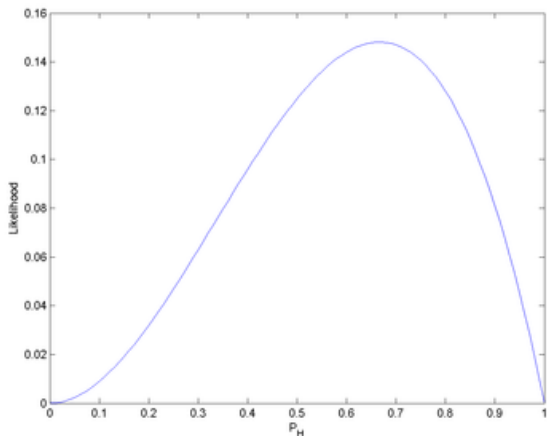
## Likelihood for the motif finding problem

$$P(R|A, \theta_{01}, \theta_{02}, \theta_{03}, \theta_{04}, \dots, \theta_{w1}, \theta_{w2}, \theta_{w3}, \theta_{w4}) = \prod_{j=1}^w \prod_{b=1}^4 \theta_{j,b}^{n_{j,b}} \times \prod_{j \notin \text{Motifs}} \prod_{b=1}^4 \theta_{0,b}^{n_{0,b}}$$

## Likelihood for the motif finding problem

- The goal is to find the combination of  $A$  and  $\theta$ s that maximizes the likelihood on the previous slide.
- This isn't an easy task!
- Gibbs sampling and the EM algorithm are essentially just different ways of tackling this problem.

# Comparing Gibbs sampling and the EM algorithm



A likelihood function. Credit: wikipedia