# Lecture 12

CSE 527: Computational Biology
November 13, 2001
Notes: Scott Votaw

## The Motif-Finding Problem

Given: $n$ strings $x_1, x_2, \ldots x_n$, each with length $m$
each string has an instance of a motif of length $l << m$

Example:

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | $\Rightarrow$ | A | C | T | G | C | T | A | ***T*** | ***A*** | ***A*** | ***T*** | ***C*** | ***T*** | G | T | T | A | G | C |
| $x_2$ | $\Rightarrow$ | T | A | G | C | ***T*** | ***A*** | ***T*** | ***A*** | ***A*** | ***T*** | G | G | C | T | T | A | T | G | A |
| $x_3$ | $\Rightarrow$ | G | A | G | A | A | T | A | T | G | G | C | C | C | ***T*** | ***A*** | ***T*** | ***A*** | ***A*** | ***T*** |
| $x_n$ | $\Rightarrow$ | T | ***T*** | ***A*** | ***C*** | ***A*** | ***A*** | ***T*** | T | G | T | A | G | G | G | T | A | A | A | G |

The simplest motif model would simply designate an exact sequence of characters to match (e.g. "TATAAT"). A measure of the match at any location $k$ in the string would simply be the number of mismatches.

A slightly more advanced model of a motif of length $l$=6 (so-called "TATAAT" box) can be represented by a probability table. This is better than the previous model since the motif does not have to be an exact sequence. (This is a $0^{th}$ order Markov model since all positions are independent)

Hypothetical probability model for a TATAAT box

| | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **A** | 5 | 85 | 2 | 80 | 82 | 1 |
| **C** | 3 | 5 | 3 | 10 | 8 | 3 |
| **G** | 2 | 6 | 3 | 12 | 5 | 2 |
| **T** | 90 | 4 | 92 | 8 | 5 | 94 |

In order to give an estimate of the likelihood of a model, there must be a comparison model. The "background" model can assume equal distribution, or represent the actual distribution of each character (e.g. 42% GC, 58% AT).

Background model representing overall base pair frequencies

| | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **A** | 29 | 29 | 29 | 29 | 29 | 29 |
| **C** | 21 | 21 | 21 | 21 | 21 | 21 |
| **G** | 21 | 21 | 21 | 21 | 21 | 21 |
| **T** | 29 | 29 | 29 | 29 | 29 | 29 |

Each model ($M_{test}$ and $M_{background}$) has a probability of generating *every* possible string. For the $0^{th}$ order Markov models given above, the probability of a particular sequence occurring is simply the product of the probabilities for each character. This is true since the individual characters are independent in this type of model.

$p_{test}$(TATAAT) = .90 * .85 * .92 * .80 * .82 * .94
$p_{background}$(TATAAT) = .29 * .29 * .29 * .29 * .29 * .29

The natural way to compare the two models is to generate a ratio of the probabilities. This ratio is called the natural score.

$$\text{natural score} = \frac{\text{p(test model)}}{\text{p(background model)}}$$

Multiplying small numbers can be inaccurate computationally, so alternatively we can sum the logs.

$$\log \frac{\prod p_i}{\prod q_i} = \sum_{i=1}^{l} \log \frac{p_i}{q_i}$$

Defining $\theta = \log \frac{p_i}{q_i}$, finding a good motif is equivalent to finding a good $\theta$. We wish to maximize the value of $\theta$ so that the relative probability of our test model is greatest with respect to the background model.

## Using E.M. to solve the Motif Finding Problem

Given: $n$ strings $x_1, x_2, \ldots x_n$, each with length $m$

Define: $y_{ik}$ = 1 if motif starts at position $k$ in string $i$
= 0 otherwise

$\theta = \log \frac{p_i}{q_i}$

We need:

1) p($x_i$ | $\theta$, $y_{ik}$ = 1)
2) a way to find $\theta'$ maximizing above given some data

**The E.M. (Expectation Maximization) algorithm**

**Given:** $\theta^t$ (an initial estimate of $\theta$)

**E Step: estimates** $E_{\theta^t}(y_{ik})$

(Calculate the expectation that the motif exists at location $k$ in each string $i$)

$$E_{\theta^t}(y_{ik}) = E(y_{ik} = 1 \mid x_i, \theta^t)$$
$$= 1 \cdot p(y_{ik} = 1 \mid x_i, \theta^t) + 0 \cdot p(\ldots$$

$$= \frac{p(x_i \mid y_{ik} = 1, \theta^t) \cdot p(y_{ik} = 1 \mid \theta^t)}{p(x_i \mid \theta^t)}$$

(by Bayes' rule)

Trick 1: The denominator is independent of $y_{ik}$, so it cancels out in the ratio
Trick 2: The $p(y_{ik} = 1 \mid \theta^t)$ is the prior belief, and can often be assumed a constant that can be factored out

**M Step: maximize $\theta$ given data**

$$Q(\theta \mid \theta^t) = E_{\theta^t}(\log(x, y \mid \theta)$$
$$= E\left(\log \prod_{i=1}^{n} p(x_{ij}, y_i \mid \theta)\right)$$

(Trick: only one $y_i$ is 1, the rest are zero)

$$= E\left(\log \prod_{i=1}^{n} \prod_{k=1} p(x_{ij} \mid \theta, y_{ik} = 1)^{y_{ik}}\right.$$
$$= E\left(\sum_{i=1}^{n} \sum_{k=1}^{m} y_{ik} \log(p(x_i \mid \theta, y_k = 1))\right)$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{m} E(y_{ik}) \log(p(x_i \mid \theta, y_k = 1))$$

(since $p(x_i \mid \theta, y_k = 1)$ is independent of expectation)

Goal: find $\theta$ maximizing Q function given data

**How do you start the E.M. algorithm? (i.e. how to determine initial $\theta^t$)**

Setting $\theta$s equal to the background is a bad idea since there will not be any progress on the first iteration.

Other possible starts:

- Start with random values and repeat many times
- Use prior knowledge
- MEME (Bailey and Elkan, San Diego Super Computer Center)
  1. Try all length $l$ substrings of $x_i$'s
  2. Do 2 iterations of E.M. for each seed
  3. Pick best few and do full iteration

## Model Selection

General problem: Given 2 models $M_1$ and $M_2$, which is better?

Given: $M_1(\theta_1)$, $M_2(\theta_2)$, and observed data D

What's $p(M_1 \mid D)$?

by Bayes rule: $p(M_1 \mid D) = \dfrac{p(D \mid M_1) \cdot p(M_1)}{p(D \mid M_1)p(M_1) \cdot p(D \mid M_2)p(M_2)}$

What's $p(M_1)$ and $p(M_2)$?

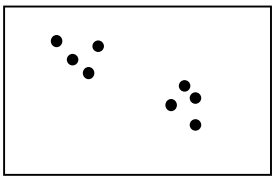
These are the a priori probabilities of each model being correct.

- could assume equal probability of both models (i.e., 50/50)
- could calculate some other a priori probability using knowledge

e.g., data is a distribution of 8 points

$M_1$ = 1 distribution with 1 $\mu$ and $\sigma^2$

$M_2$ = 2 populations with different $\mu$ and $\sigma^2$



We could use likelihood directly, but more complicated models fit better (e.g. $M_3$ = 8 populations with mean $\mu$ and $\sigma^2 = 0$). Intuition: penalize extra degrees of freedom. This can be done using a BIC score.

**BIC score (Bayesian Information Criterion)**

$$\text{BIC} = \text{likelihood} - \text{penalty for degrees of freedom} = 2 \log[\, p(x \mid \hat{\theta}) \,] \; - \; d \log n$$

(where $d$ = number of free parameters and $n$ = number of data points)