MÔ HÌNH HÓA TRÌNH TỰ SINH HỌC

MODELING OF BIOLOGICAL SEQUENCES

TRƯƠNG THẾ QUANG(*)

THÔNG TIN

Ngày nhận bài: 19-8-2024
Ngày biên tập xong: 27-8-2024
Ngày duyệt đăng: 31-9-2024
Mã số: TCKH47-07-2024
ISSN: 2525 – 2429

TÓM TẮT

Lý thuyết mô hình Markov được ứng dụng phổ biến và có hiệu quả trong việc mô hình hóa các trình tự sinh học như dự đoán cấu trúc thứ cấp protein, giải trình tự sinh học, phát hiện gene. Trong bài viết này trình bày các định nghĩa về mô hình Markov, chuỗi Markov, mô hình Markov mở rộng, mô hình Markov ẩn, cũng như tính toán xác suất của các trạng thái ban đầu, kết thúc và xác suất chuyển đổi trạng thái trong chuỗi Markov. Ứng dụng lý thuyết mô hình Markov để mô hình hóa vùng CG trên DNA.

Từ khóa: *chuỗi Markov; mô hình Markov; mô hình Markov ẩn; trình tự sinh học; vùng CG.*
Key words: *Markov chain; Markov model; hidden Markov model; biological sequence; CG region.*

ABSTRACT: *Markov model theory is widely and effectively applied in modeling biological sequences such as protein secondary structure prediction, biological sequencing, gene discovery. This article presents the definitions of Markov models, Markov chains, extended Markov models, hidden Markov models, as well as the calculation of the probability of initial and final states and state transition probabilities in Markov chains. Application of Markov model theory to modeling CG regions on DNA.*

1. MỞ ĐẦU

Mô hình hóa sinh học bằng lý thuyết mô hình Markov (Markov model theory - MMT) được ứng dụng trở lại từ những năm 1950. Vào cuối năm 1970 Andrew Viterbi, Leonard Baum và các đồng nghiệp đã dựa vào MMT phát triển các thuật toán giải mã và ước lượng tham số [4, tr.1]. Anders Krogh, I. Saira Mian và David Haussler áp dụng mô hình Markov ẩn dự đoán một số gene trong DNA của vi khuẩn *E. coli* [2, tr.4768-4778]. Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, David Haussler đã ứng dụng mô hình Markov ẩn để mô hình hóa protein, công trình nghiên cứu này được trình bày tóm tắt trong bài báo "*Mô hình Markov ẩn trong tính toán sinh học: Ứng dụng để mô hình hóa protein*" [3, tr.1501-1531].

MMT được ứng dụng phổ biến và có hiệu quả trong tin sinh học như dự đoán cấu trúc thứ cấp protein, sắp hàng nhiều trình tự, phát hiện gene. MMT là mô hình thống kê cực kỳ linh hoạt có thể được sử dụng để mô hình hóa bất kỳ tập hợp các dữ liệu phần tử một chiều rời rạc. MMT được ứng dụng để mô hình hóa các trình tự protein trong nhiều trường hợp khác nhau, tùy thuộc vào chức năng của protein được biểu diễn bởi các trạng thái Markov. Đối với dự đoán cấu trúc protein, các trạng thái đã được chọn để đại

(*) TS. Trường Đại học Văn Lang, *quangtruongthe@gmail.com*

diện cho các vị trí trình tự tương đồng, các loại cấu trúc thứ cấp hoặc màng địa phương.

MMT gồm một số trạng thái, mỗi trạng thái phát sinh phần tử theo xác suất phát sinh, các trạng thái được kết nối bởi xác suất chuyển trạng thái. Bắt đầu từ một trạng thái ban đầu nào đó, một chuỗi các trạng thái được tạo ra bằng cách di chuyển từ trạng thái này đến trạng thái kế tiếp theo xác suất chuyển trạng thái cho đến trạng thái cuối cùng để tạo ra một trình tự quan sát [1, tr.221-222].

2. NỘI DUNG

Mô hình hóa trình tự sinh học.

2.1. Vùng CG

Trong trình tự DNA (Deoxyribonucleic Acid) nếu xác suất xuất hiện nucleotide là như nhau, thì xác suất xuất hiện của mỗi nucleotide a $\in$ {A, G, C, T} là p(a) $\approx$ 1/4. Như vậy, xác suất xuất hiện của một dinucleotide là $\approx$ 1/16. Thực tế, xác suất xuất hiện của dinucleotide trong trình tự DNA khác nhau, đặc biệt dinucleotide CG thường có xác suất xuất hiện nhỏ hơn 1/16.

Hoạt động của dinucleotide CG trong gene, nucleotide C trong cặp CG thường bị biến đổi qua quá trình methyl hóa, nghĩa là một nguyên tử H được thay thế bởi một nhóm methyl CH_3 và sau đó methyl-C có xu hướng biến đổi thành T.

Tuy nhiên, ở đầu của gene quá trình methyl hóa bị ức chế tại các đoạn gene ngắn có chiều dài 100 ÷ 5000 bp, các đoạn này được gọi là những vùng CG. Vùng CG có đặc trưng là chứa nhiều cặp CG hơn những vùng khác (hình 1). Vì vậy, việc tìm kiếm các vùng CG trong hệ gene là một vấn đề quan trọng.

```
...ApCpCpApTpGpApTpGpCpApGpGpApCpTpTpCpCpApTpCpGpTpTpCpGpCp...
...| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ...
...TpGpGpTpApCpTpApCpGpTpCpCpTpGpApApGpGpTpApGpCpApApGpCpGp...
```

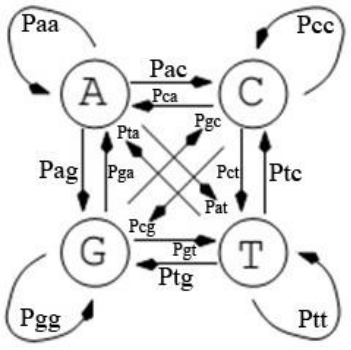
Hình 1. Vùng CG trên gene

Vùng CG đóng vai trò quan trọng trong các gene sinh vật có chứa 5-methyl-cytosine. Vùng CG trong promoter của các gene có chức năng ức chế khi sao chép nhiễm sắc thể X ở nữ giới và ức chế hoạt động của hệ gene ký sinh trùng.

Theo định nghĩa kinh điển của Gardiner Garden & Frommer, vùng CG là trình tự DNA có chiều dài khoảng 200 bp với thành phần C + G chiếm 50 % và tỷ lệ số lượng CG quan sát thực tế lớn hơn 0,60. Theo một nghiên cứu của D. Takai và P. A. Jones, ở nhiễm sắc thể 21 và 22 của người có chứa khoảng 1100 vùng CG trong khoảng 750 gene.

2.2. Mô hình Markov và chuỗi Markov

Cặp nucleotide đóng vai trò quan trọng trong trình tự DNA, do đó cần xây dựng mô hình xác suất của dinucleotide, trong đó xác suất của nucleotide đứng trước phụ thuộc vào xác suất của nucleotide đứng sau. Mô hình Markov và chuỗi Markov được ứng dụng để xây dựng mô hình xác suất cho vùng CG (hình 2).



Hình 2. Chuỗi Markov với các xác suất thay thế nucleotide

Cho tập trạng thái S = {s_1, s_2, ..., s_k}, tại thời điểm t quá trình Markov ở trạng thái $s_i \in$ S, bước qua thời điểm t + 1 quá trình chuyển sang trạng thái $s_j \in$ S với xác suất chuyển trạng thái p_{ij} là xác suất có điều kiện xảy ra trạng thái s_j khi trạng thái s_i đã xảy ra được tính theo (1).

$$p_{ij} = p(x_{t+1} = s_j \,|\, x_t = s_i);\ 1 \le i, j \le k;\ p_{ij} \ge 0 \quad (1)$$

Ma trận gồm các số hạng là xác suất chuyển trạng thái được gọi là ma trận xác suất chuyển đổi (2).

$$P = [p_{ij}] = \begin{array}{c|cccc} & s_1 & s_2 & \cdots & s_k \\ \hline s_1 & p_{11} & p_{12} & \cdots & p_{1k} \\ s_2 & p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_k & p_{k1} & p_{k2} & \cdots & p_{kk} \end{array} \quad (2)$$

Trong đó, p_{ij} là xác suất chuyển trạng thái $s_i \rightarrow s_j$. Như vậy tổng tất cả xác suất chuyển trạng thái bằng 1, thể hiện ở (3).

$$\sum_{j=1}^{k} p_{ij} = 1 \quad (3)$$

2.3. Định nghĩa mô hình Markov

Giả sử thời gian đồng nhất, mô hình Markov được định nghĩa là một hệ thống (S, P) bao gồm một tập hợp hữu hạn các trạng thái $S = \{s_1, s_2, ..., s_k\}$ và một ma trận xác suất chuyển đổi $P = [p_{sr}]$ với $\sum_{r \in S} p_{sr} = 1$, $\forall s \in S$. Xác suất chuyển trạng thái $s \rightarrow r$ được xác định theo (4).

$$p(x_{i+1} = r \,|\, x_i = s) = p_{sr} \quad (4)$$

Theo định nghĩa, mô hình Markov là một hệ thống gồm k trạng thái phân biệt $s_1, s_2, ..., s_k$. Tại thời điểm t bất kỳ, hệ thống có thể chuyển từ trạng thái s_i sang một trong k – 1 trạng thái còn lại hoặc chuyển trở lại chính trạng thái s_i. Như vậy, ở thời điểm t từ trạng thái s_i có k hướng chuyển trạng thái. Mỗi hướng chuyển trạng thái có một độ đo khả năng xảy ra gọi là xác suất chuyển trạng thái.

Gọi p_{sr} là xác suất chuyển trạng thái $s \rightarrow r$, nên hiểu p_{sr} là xác suất trạng thái r xảy ra với điều kiện trạng thái s đã xảy ra. Xác suất chuyển trạng thái p_{sr} không phụ thuộc vào thời gian t, độc lập với các trạng thái đã chuyển trước đó và duy nhất chỉ phụ thuộc vào trạng thái hiện tại. Quá trình mang tính ngẫu nhiên này được gọi là có thuộc tính Markov.

Kết xuất của hệ thống là một chuỗi các trạng thái tại các thời điểm t tương ứng. Ta biết được từng trạng thái ở thời điểm t nào, vì vậy mô hình này được gọi là mô hình Markov hiện OMM (Observed Markov Model) hay mô hình Markov.

2.4. Định nghĩa chuỗi Markov

Chuỗi Markov là một chuỗi các biến ngẫu nhiên $x_0x_1x_2 ... x_i ...$ Tập tất cả các giá trị của các biến này được gọi là không gian trạng thái S, giá trị x_i là trạng thái của quá trình (hệ thống) tại thời điểm *i*. Với mọi $i > 0$ và $s \in S$, xác suất chuyển trạng thái của hệ được xác định theo (5).

$$p(x_i = s \,|\, \cap_{j<i} x_j) = p(x_i = s \,|\, x_{i-1}) \quad (5)$$

Chuỗi Markov được gọi là đồng nhất, nếu xác suất chuyển trạng thái không phụ thuộc vào thời gian i. Tại thời điểm i hệ ở trạng thái x_i, sang thời điểm i + 1 hệ chuyển sang trạng thái x_{i+1} theo xác suất chuyển trạng thái nhất định, gọi là sự thay đổi tức thời.

Theo định nghĩa 2.4, chuỗi Markov là một quá trình ngẫu nhiên với thời gian rời rạc. Trong một quá trình như vậy, việc dự đoán tương lai không liên quan đến quá khứ, mà chỉ phụ thuộc vào thông tin ở hiện tại.

2.5. Xác suất trạng thái bắt đầu, kết thúc của chuỗi Markov

Gọi s_0 là trạng thái bắt đầu của mô hình Markov, xác suất trạng thái bắt đầu s_1 của chuỗi Markov được tính theo (6).

$$p(x_1 = s_1 \,|\, x_0 = s_0) = p(s_1) = p_{s_0 s_1} \quad (6)$$

Tương tự, gọi s_n là trạng thái kết thúc của mô hình Markov, trước nó là trạng thái s_{n-1}, xác suất trạng thái kết thúc s_n của chuỗi Markov được tính theo (7).

$$p(x_n = s_n \,|\, x_{n-1} = s_{n-1}) = p_{s_{n-1} s_n} \quad (7)$$

2.6. Xác suất của chuỗi Markov

Cho trình tự trạng thái $X = x_1x_2 ... x_L$. Xác suất của chuỗi Markov là xác suất của chuỗi khi qua từng bước trạng thái x_i, $i = 1 \div T$ thuộc trình tự trạng thái X và được tính theo (8) [2, tr.4770].

$$p(X) = p(x_1) \prod_{i=2}^{L} p_{x_{i-1} x_i} \quad (8)$$

Chứng minh phương trình (8) như sau:

$$\forall i = \overline{(2, L)}: \; p(x_i, x_{i-1}) = p(x_i \,|\, x_{i-1}) p(x_{i-1}) \quad (9)$$

$$\Rightarrow p(X) = p(x_L, x_{L-1}, ..., x_1)$$

$$p(X) = p(x_L \,|\, x_{L-1}, ..., x_1) p(x_{L-1} \,|\, x_{L-2}, ..., x_1) ... p(x_1)$$

$$p(X) = p(x_L \,|\, x_{L-1}) p(x_{L-1} \,|\, x_{L-2}) ... p(x_2 \,|\, x_1) p(x_1)$$

$$p(X) = p(x_1) \prod_{i=2}^{L} p_{x_{i-1} x_i}$$

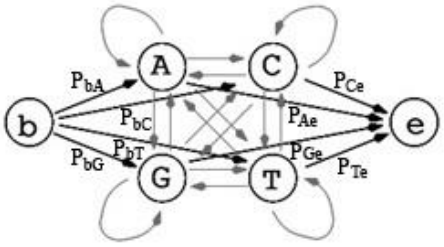
2.7. Mô hình Markov mở rộng

Mô hình Markov mở rộng là mô hình Markov có tập trạng thái bao gồm cả trạng thái

bắt đầu và kết thúc được minh họa ở Hình 3 với trạng thái bắt đầu được ký hiệu là b (begin) và trạng thái kết thúc được ký hiệu là e (end).

Ví dụ, ma trận xác suất chuyển đổi các nucleotide trong mô hình Markov mở rộng với 6 trạng thái A, C, G, T, b, e được thể hiện ở (10).

$$P = \begin{array}{c|cccccc} & A & C & G & T & b & e \\ A & 0,1795 & 0,2735 & 0,4255 & 0,1195 & 0 & 0,002 \\ C & 0,1705 & 0,3665 & 0,2735 & 0,1875 & 0 & 0,002 \\ G & 0,1605 & 0,3385 & 0,3745 & 0,1245 & 0 & 0,002 \\ T & 0,0785 & 0,3545 & 0,3835 & 0,1815 & 0 & 0,002 \\ b & 0,2495 & 0,2495 & 0,2465 & 0,2495 & 0 & 0,002 \\ e & 0 & 0 & 0 & 0 & 0 & 1,000 \end{array} \quad (10)$$



Hình 3. Mô hình Markov mở rộng

2.8. Xác suất trình tự DNA thuộc vùng CG

Ứng dụng phương trình (6.8) để tính xác suất trình tự DNA thuộc (hoặc không thuộc) vùng CG sau quá trình chuyển trạng thái. Ở đây các phần tử nucleotide trên trình tự DNA có thể rơi vào một trong hai trạng thái là "thuộc vùng CG" hoặc "không thuộc vùng CG".

Muốn vậy cần phải xác định ma trận chuyển đổi tương ứng. Ma trận chuyển đổi thường được tính toán từ tập các trình tự DNA. Ma trận chuyển đổi P^+ đối với các nucleotide thuộc vùng CG sau quá trình chuyển trạng thái từ s → r, được xác định theo phương trình (11).

$$P_{sr}^{+} = \frac{c_{sr}^{+}}{\sum_{t'} c_{sr'}^{+}} \quad (11)$$

Trong đó, c^+_{sr} là số vị trí nucleotide trên trình tự DNA thuộc vùng CG xuất hiện ở trạng thái r, sau quá trình chuyển trạng thái từ s → r.

Tương tự, ma trận chuyển đổi P^- đối với các nucleotide của trình tự DNA không thuộc vùng CG sau quá trình chuyển trạng thái từ s → r, được xác định theo phương trình (12) với c^-_{sr} là số vị trí nucleotide trên trình tự DNA không thuộc vùng CG xuất hiện ở trạng thái r, sau quá trình chuyển trạng thái từ s → r.

$$P_{sr}^{-} = \frac{c_{sr}^{-}}{\sum_{t'} c_{sr'}^{-}} \quad (12)$$

Ví dụ, ma trận chuyển đổi P^+ đối với các nucleotide của trình tự DNA thuộc vùng CG (13).

$$P^+ = \begin{array}{c|cccc} & A & C & G & T \\ A & 0,1795 & 0,2735 & 0,4255 & 0,1195 \\ C & 0,1705 & 0,3665 & 0,2735 & 0,1875 \\ G & 0,1605 & 0,3385 & 0,3745 & 0,1245 \\ T & 0,0785 & 0,3545 & 0,3835 & 0,1815 \end{array} \quad (13)$$

Và ma trận chuyển đổi P^- đối với các nucleotide của trình tự DNA không thuộc vùng CG (14).

$$P^- = \begin{array}{c|cccc} & A & C & G & T \\ A & 0,2995 & 0,2045 & 0,2845 & 0,2095 \\ C & 0,3215 & 0,2975 & 0,0775 & 0,0775 \\ G & 0,2475 & 0,2455 & 0,2975 & 0,2075 \\ T & 0,1765 & 0,2385 & 0,2915 & 0,2915 \end{array} \quad (14)$$

Cho trình tự DNA $X = x_1x_2 \ldots x_L$. Xác suất trình tự DNA thuộc vùng CG sau quá trình chuyển đổi trạng thái (mô hình$^+$) được tính theo (15).

$$p(X|+) = p(x_1)\prod_{i=2}^{L} p^{+}_{x_{i-1}x_i} \quad (15)$$

Xác suất trình tự DNA không thuộc vùng CG sau quá trình chuyển đổi trạng thái (mô hình$^-$) được tính theo (16).

$$p(X|-) = p(x_1)\prod_{i=2}^{L} p^{-}_{x_{i-1}x_i} \quad (16)$$

Tính log-odds ratio theo (17) [5, tr.4].

$$S(X) = \log_2 \frac{p(X|+)}{p(X|-)} = \sum_{i=2}^{L} \log_2 \frac{p^{+}_{x_{i-1}x_i}}{p^{-}_{x_{i-1}x_i}} = \sum_{i=2}^{L} \beta_{x_{i-1}x_i} \quad (17)$$

Với $\beta_{x_{i-1}x_i}$ là logarit cơ số 2 của tỷ lệ xác suất chuyển đổi tương ứng. Ví dụ, theo các ma trận chuyển đổi (13) và (14) thì $\beta_{AA} = \log_2(0,1795/0,2995)$. Logarit cơ số 2 thường được sử dụng, trong trường hợp này đơn vị tính là bit.

Nếu mô hình$^+$ và mô hình$^-$ khác biệt đáng kể và log-odds ratio > 0 thì xác xuất trình tự X ở mô hình$^+$ cao hơn mô hình$^-$.

Có thể sử dụng một giá trị ngưỡng c* và kiểm tra theo (18) để xác định xem trình tự X có là trình tự con của vùng CG hay không.

$$\phi^*(X) = \begin{cases} 1 \ nÕu \ S(X) > c^* \\ 0 \ nÕu \ S(X) \le c^* \end{cases} (18)$$

Ở đây, $\phi^*(X) = 1$ ứng với trình tự X thuộc vùng CG và (18) được gọi là trắc nghiệm Neyman-Pearson [1, tr.213-220].

2.9. Định nghĩa mô hình Markov ẩn

Mô hình Markov ẩn HMM (Hidden Markov Model) là một hệ thống M = (Σ, Q, P, E, U) gồm:

1. Bảng Σ, trình tự $X = x_1x_2 \dots x_T$ gồm các phần tử quan sát lấy từ bảng Σ, phần tử quan sát vào thời điểm t là b_t;

2. Tập trạng thái Q, đường dẫn $\Pi = \pi_1\pi_2 \dots \pi_L$ gồm các trạng thái ẩn thuộc Q, trạng thái ở thời điểm t là k_t;

3. Ma trận xác suất chuyển đổi $P = [p_{ij}]$ gồm các số hạng là xác suất chuyển trạng thái p_{ij}, với $1 \le i, j \le L$;

4. Hàm mật độ xác suất phát sinh $E = \{e_k(b)\}$;

5. Hàm mật độ xác suất khởi đầu của mỗi trạng thái $U = \{u_i\}$ [1, tr.224].

2.10. HMM đối với vùng CG

Cho bảng Σ = {A, C, G, T} gồm các phần tử là nucleotide; Tập trạng thái Q = {0, A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-}, trong đó 0 là trạng thái bắt đầu hoặc kết thúc; Ma trận xác suất phát sinh E gồm các số hạng là $e_k(b)$:

$e_k(b)$	0	A_+	C_+	G_+	T_+	A_-	C_-	G_-	T_-
A	0	0,40	0	0	0	0,05	0	0	0
C	0	0	0,10	0	0	0	0,40	0	0
G	0	0	0	0,10	0	0	0	0,50	0
T	0	0	0	0	0,40	0	0	0	0,05

Ma trận xác suất chuyển đổi P gồm các số hạng là p_{ij}:

p_{ij}	0	A_+	C_+	G_+	T_+	A_-	C_-	G_-	T_-
0	0	0,0725	0,1638	0,1788	0,0755	0,1322	0,1267	0,1226	0,1279
A_+	0,0010	0,1762	0,2683	0,4171	0,1175	0,0036	0,0055	0,0085	0,0024
C_+	0,0010	0,1672	0,3599	0,2680	0,1839	0,0034	0,0073	0,0055	0,0038
G_+	0,0010	0,1576	0,3319	0,3671	0,1224	0,0032	0,0068	0,0075	0,0025
T_+	0,0010	0,0773	0,3476	0,3759	0,1782	0,0016	0,0071	0,0077	0,0036
A_-	0,0010	0,0003	0,0002	0,0003	0,0002	0,2994	0,2046	0,2844	0,2096
C_-	0,0010	0,0003	0,0003	0,0001	0,0003	0,3214	0,2974	0,0778	0,3014
G_-	0,0010	0,0002	0,0002	0,0003	0,0002	0,2476	0,2455	0,2974	0,2076
T_-	0,0010	0,0002	0,0002	0,0003	0,0003	0,1766	0,2385	0,2914	0,2914

Nếu biết đường dẫn Π phát sinh ra trình tự X (gồm đoạn CGGC thuộc vùng CG và đoạn CAT không thuộc vùng CG), tính được các xác suất chuyển trạng thái p_{ij} và xác suất phát sinh $e_k(b)$ như sau [1, tr.224-225]:

Đường dẫn Π	C_+	G_+	G_+	C_+	C_-	A_-	T_-
p_{ij}	0,1638	0,2680	0,3671	0,3319	0,0073	0,3214	0,0296
Trình tự X	C	G	G	C	C	A	T
$e_k(b)$	0,10	0,10	0,10	0,10	0,40	0,05	0,05

3. KẾT LUẬN

Lý thuyết mô hình Markov được ứng dụng phổ biến và có hiệu quả trong việc mô hình hóa các trình tự sinh học như dự đoán cấu trúc thứ cấp protein, giải trình tự sinh học, phát hiện gene. Mô hình Markov ẩn là mô hình thống kê

cực kỳ linh hoạt được sử dụng để mô hình hóa các trình tự sinh học một cách hiệu quả. Mô hình Markov ẩn được ứng dụng để tìm kiếm trình tự tương đồng, dự đoán cấu trúc protein và phát hiện gene. Trong bài viết này trình bày các định nghĩa về mô hình Markov, chuỗi Markov, mô hình Markov mở rộng, mô hình Markov ẩn, cũng như tính toán xác suất của trạng thái ban đầu, kết thúc, xác suất của chuỗi Markov và xác suất chuyển đổi trạng thái trong chuỗi Markov. Ứng dụng lý thuyết mô hình Markov để mô hình hóa vùng CG trên DNA.

TÀI LIỆU THAM KHẢO

[1] Trương Thế Quang (2018), *Tin sinh học (Bioinformatics)*, Nxb Đại học Quốc gia Thành phố Hồ Chí Minh.

[2] Anders Krogh, I. Saira Mian and David Haussler (1994), *A hidden Markov model that finds genes in E. coli DNA*, Nucleic Acids Research, 22:22.

[3] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander and David Haussler (1994), *"Hidden Markov models in computational biology: Applications to protein modeling*", Journal of Molecular Biology, 235.

[4] Alexander Churbanov and Stephen Winters-Hilt (2008), *Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory*, BMC Bioinformatics, 9:224.

[5] Prashant K. Srivastava, Dhwani K. Desai, Soumyadeep Nandi and Andrew M. Lynn (2007), *HMM-ModE – Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences*, BMC Bioinformatics, 8:104.