은닉마르코프모델을 사용한

인간 유전자 단백질 코딩 영역 예측

바이오인포매틱스 알고리즘 기말 프로젝트

2017-12-15

Team G

김승희

박윤주

신동희

정혜수

초록

생물의 DNA 에는 유전 정보가 담겨 있을 뿐만 아니라, 개체의 형질을 실제로 발현하는 데 필요한 다양한 기능을 담당하는 부분도 포함되어 있다. 복잡한 구조를 가지고 있는 진핵생물일수록 그 구성은 더욱 다양하다. 인간과 같이 많은 연구가 진행된 생물의 DNA 서열은 염기의 조성, 해당 서열이 유전자인지 아닌지 등의 정보가 이미 데이터베이스에 공개되어 있어 쉽게 이용이 가능하다. 그러나 일체의 사전정보없이 미지의 DNA 서열이 주어졌을 때, 염기 서열만으로 생물의 유전 정보가 암호화된 부분을 예측하기 위해서는 다른 방식의 접근이 필요하다.

우리는 이번 프로젝트를 통해 임의로 주어진 유전자에서 단백질을 암호화하고 있는 영역과 그렇지 않은 영역을 구분하기 위한 프로그램을 구축하고자 한다. 이미 개발된 유전자 예측 툴을 표방하여 HMM 알고리즘을 사용해, Python 환경에서 임의의 인간 유전자 서열에 대해 엑손과 인트론을 구분하는 모델을 가능한 근접하게 구현해볼 예정이다. 모델의 구현에 성공한다면, 나아가 기존에 나와있는 다른 툴과의 비교를 통해 우리가 구현한 모델의 성능 확인과 개선사항을 탐구해볼 것이다.

키워드: 엑손, 인트론, HMM, 유전자 예측

I. 서론

생물의 DNA 서열은 A, T, G, C 의 네 가지 염기로 이루어져 있으며 서열의 구성은 생물마다 서로 다르다. 기술의 발전으로 생물의 DNA 염기서열을 밝혀내는 데 성공한 연구자들은 본인들의 성과를 NCBI[1] 등으로 대중에 공개했다. 하지만, 생물의 DNA 서열을 낱낱이 식별했다 하더라도, 여전히 DNA 의 많은 부분이 미지의 영역으로 남아있다.

모든 DNA 서열이 유전정보를 담고 있는 것은 아니다. 각 DNA 서열은 유전정보의 암호화 외에도 다양한 기능을 수행할 수 있고, 각각의 역할에 따라 구분될 수 있다. 인간과 같은 복잡한 구조를 가진 진핵생물의 경우, DNA 서열은 보다 더 다양한 부분으로 이루어져 있다.

진핵생물의 DNA 서열은, 기능을 가지고 있어 RNA 로 전사될 수 있는 Intragenic DNA 와 그렇지 않은 Intergenic DNA 로 나누어진다[2]. 여기서 다시 Intragenic DNA 는 스플라이싱(Splicing)이라는 전사 조절을 통해, 마지막까지 살아남아 번역 과정을 거쳐 단백질로 합성되는 엑손(exon)과 그렇지 않은 인트론(intron)으로 구분될 수 있다[3].

생물 내에서 실질적인 역할을 수행하는 효소, 호르몬, 항체 등의 물질은 주로 단백질로 이루어져 있기 때문에, 생리를 이해하기 위한 DNA 서열 분석에서 엑손과 인트론의 구분은 매우 중요하다. 그러나 방대한 정보를 담고 있는 DNA 서열을 분석하여 임의의 DNA 서열에 대해 엑손과 인트론을 구분해내기란 결코 쉬운 일이 아니다.

임의의 DNA 서열이 주어졌을 때, 이 DNA 서열에서 유전정보를 담고 있는 부분은 어디인지, 최종적으로 단백질 합성이 되는 서열은 무엇인지 등의 유전자

예측을 수행할 수 있도록 돕는 도구들은 이미 개발되어 있다. NCBI BLAST[4]나 UCSC genome browser[5]와 같은 웹서비스는, 인간과 같이 DNA 서열이 비교적 잘 알려진 생물에 대해 유전자 예측 정보를 제공하고 있다. 그러나 충분한 연구가 진행되지 않은 생물, 혹은 아무런 정보 없이 주어진 DNA 서열에 대해서는 위에서 언급한 웹서비스의 도움을 받기가 어렵다. 미지의 서열의 분석을 위해서는 다른 접근법이 필요하다. GENSCAN[6,7]이나 Gene Finder[8,9] 등은 DNA 서열로부터 엑손 영역을 탐색할 수 있는 도구로, 은닉 마르코프 모델(Hidden Markov Model, 이하 HMM)[10]을 사용한 것이다.

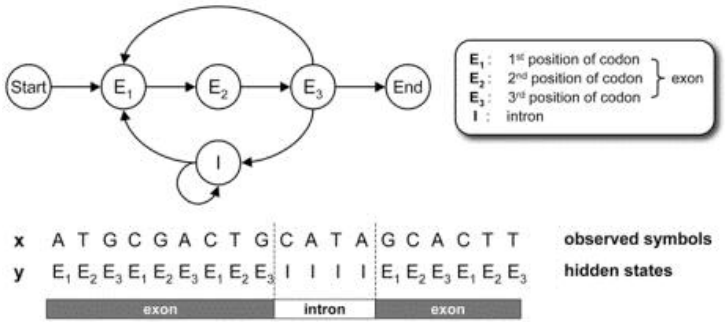


그림 1. 주어진 유전자 서열에서 exon 과 intron 을 예측하는 HMM 의 예시[11]

우리는 이번 프로젝트를 통해 임의로 주어진 유전자에서 단백질을 암호화하고 있는 영역과 그렇지 않은 영역을 구분하기 위한 프로그램을 구축하고자 한다. 본 프로젝트에서 대상이 될 생물은 인간으로, 앞서 언급한 유전자 예측 도구들의 바탕이 된 HMM 알고리즘을 사용하여, Python 환경에서 임의의 인간 유전자 서열에 대해 엑손과 인트론을 구분하는 모델을 가능한 근접하게 구현해볼 예정이다. 모델의 구현에 성공한다면, 기존에 나와있는 다른 툴과의 비교를 통해 우리가 구현한 모델의 성능의 확인과 개선사항을 탐구해볼 것이다.

II. 연구 방법

i. 은닉 마르코프 모델 (Hidden Markov Model, HMM)[10]

HMM 은 미지의 확률론적 과정을 모형화하는 이중의 확률론적 과정으로, 이는 관찰 가능한 기호를 발생시키는 다른 확률론적 과정을 통해 이루어진다. HMM 은 다음과 같이 2 개의 상태집합과 3 개의 확률 집합으로 구성되는 5 개의 요소를 갖는다.

- 은닉 상태 집합 (Hidden state set)
- 관찰 상태 집합 (Observable state set)
- 초기 확률 (Initial probability)
- 전이 확률 (Transition probability)
- 관찰 확률 (Emission probability)

이와 같이 HMM 은 관찰 가능한 상태와 은닉 상태 간의 확률적 관계를 이용하여 확장된 마르코프 프로세스라고 볼 수 있다. HMM 의 전이 관계는 다음과 같이 정의할 수 있다.

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j \mid q_{t-1} = i)$$

$$P(q_t = j \mid q_{t-1} = i) = P(q_{t+l} = j \mid q_{t+l-1} = i)$$

이를 실제로 사용하기 위해서는 다음과 같은 3 가지 문제를 해결하여야 한다.

- 확률 평가의 문제 (Probability Evaluation)[12]

  관측열 O={o_1, o_2, o_3, o_4, ...}과 모델 λ =(Π, A, B)에 대하여, 주어진 HMM 에서 관찰된 순서의 확률 P(O|λ)를 계산하는 문제로, 전향(forward)과 후향(backward) 알고리즘을 이용해 해결할 수 있다.

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1)$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$$

3. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

알고리즘 1. Forward algorithm

1. Initialization:

$$\beta_T(i) = 1$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

$$1 \le i \le N,$$

$$t = T-1, \cdots, 1$$

알고리즘 2. Backward algorithm

· 최적 상태열의 결정 문제 (Optimal State Sequence)[13]

관측열 O={o_1, o_2, o_3, o_4, ...}과 모델 λ =(Π, A, B)에 대하여 관측열을 가장 잘 설명하는 최적의 상태 순서 q=(q_1, q_2, ..., q_r)를 생성할 확률이 가장 높은 은닉 상태들 간의 순서를 찾는 문제로, 비터비(Viterbi) 알고리즘을 이용해 해결할 수 있다.

- Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N$$

$$\psi_1(i) = 0$$

- Recursion

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t)$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$$

$$2 \le t \le T, \quad 1 \le j \le N$$

- Termination

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

- Path (state sequence) backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1$$

알고리즘 3. Viterbi algorithm

· 매개변수 추정 문제 (Parameter Estimation)[14]

관측열 O={o_1, o_2, o_3, o_4, ...}에 대하여 P(O|λ)를 최대로 하는 모델 λ=(Π, A, B)의 매개변수를 결정하는 문제로 바움-웰치(Baum-Welch) 알고리즘을 이용하여 해결할 수 있다.

Step 1. Let initial model be λ_0.

Step 2. Compute new λ based on λ_0 and observation O.

Step 3. If $\log P(O|\lambda) - \log P(O|\lambda_0) < DELTA$ then stop.

Step 4. Else set $\lambda_0 \leftarrow \lambda$ and goto step 2.

알고리즘 4. Baum-Welch algorithm

ii. 사용 데이터

- GRCh37/hg19 human reference genome[5]
  2009 년 2 월에 발표된 reference genome 으로 13 명의 익명의 기증자로부터 받은 DNA 서열을 기초로 하여 만들어졌다. 해당 버전은 UCSC genome browser 와 Ensembl 에서 FASTA 파일 형태로 다운로드 받을 수 있다.
- Position of genomic data[5]
  UCSC table browser(http://genome.ucsc.edu/cgi-bin/hgTables)로부터 해당 reference genome 버전의 유전자 정보에 대한 BED (Browser Extensible Data) 파일을 다운 받았다. BED 는 주어진 genomic data 의 위치 정보를 UCSC genome browser 표준 포맷으로 제공하며, BED 파일이 가진 필드는 다음과 같다 (표 1).

필드 이름	설명
chrom	염색체의 이름
chromStart	유전자의 시작 위치
chromEnd	유전자의 끝 위치
name	유전자의 이름 (RefSeq identifier)
score	UCSC genome browser 에서 해당 트랙을 표시할 때 사용

strand	DNA strand 정의
thickStart	Start codon 의 위치
thickEnd	Stop codon 의 위치
itemRgb	UCSC genome browser 에서 해당 트랙을 표시할 때 사용
blockCount	해당 BED 라인의 블록(Exon)의 수
blockSizes	각 블록의 크기를 나타내며 쉼표로 구분 됨
blockStarts	각 블록의 시작 위치를 나타내며 chromStart 를 기준으로 계산

이 프로젝트에서는 단백질을 코딩하는 영역만 사용하기 위해, name 열의 RefSeq ID 가 NM 으로 시작하는 열에 대해 blockCount, blockSize, blockStart 를 파싱하여 유전자의 엑손과 인트론의 정확한 위치를 계산하여 모델에 사용하였다.

iii. 작업순서도

이 프로젝트의 모든 작업 과정은 아래의 순서대로 진행되었다.

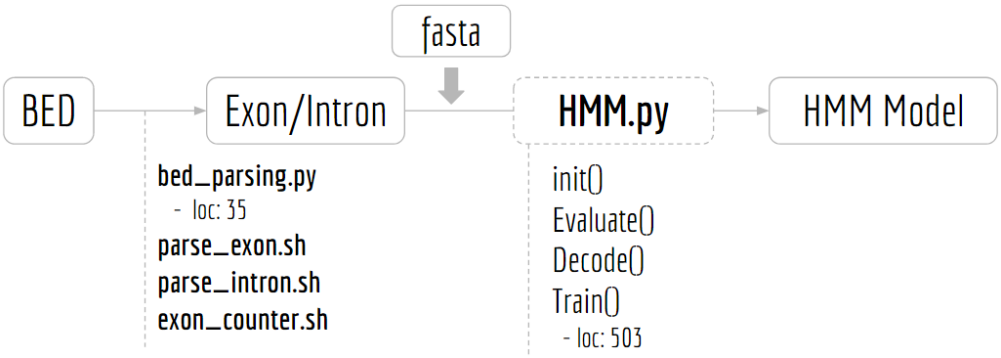


그림 2. Workflow

III. 결과

i. 엑손과 인트론의 GC contents 비교

이전 연구들에서 유전자 내의 엑손과 인트론은 염기의 비율이 서로 다른 것으로 알려져 있는데, 일반적으로 엑손에서의 GC content 는 50%를 상회하며, 인트론에서는 40% 아래로 떨어진다[16]. 우리는 실제 데이터에서도 이것이 성립하는지 확인하기 위해 엑손과 인트론의 평균 GC content 를 아래와 같이 나타냈다.

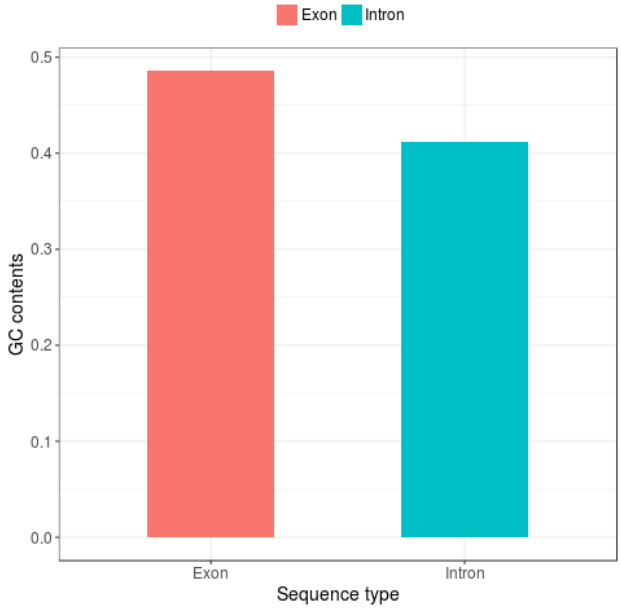
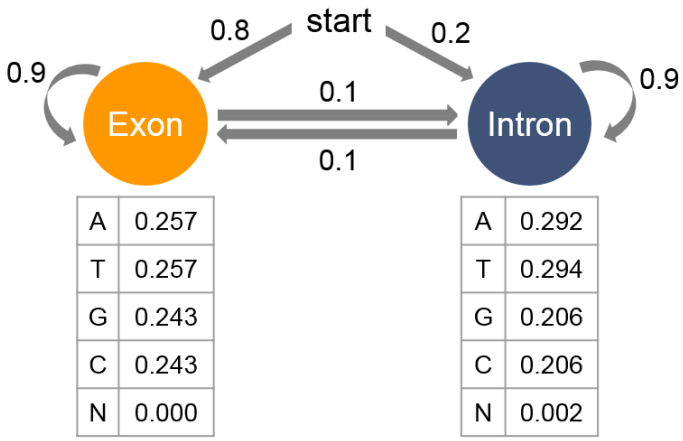


그림 3. 인간 유전체의 엑손과 인트론의 GC contents

ii. 전이 확률 및 관찰 확률 계산

단백질 코딩 영역을 예측하는 HMM 구현에 앞서, 우리는 다음의 3 가지 확률(초기 확률, 전이 확률, 관찰 확률)을 정의했다. 초기 확률 및 exon 과 intron 상태의 전이 확률은 기존 연구[15]로부터 가져왔는데, 초기 확률은 두 전이 상태에 대해 각각 0.5 이며, 전이 확률은 상태 유지할 시 0.9 이고

상태 전이 시 0.1 로 정했다. 반면 관찰 확률은 학습 데이터를 사용해서 구했다. 우선 각 exon 과 intron 상태에 대해서, 염기의 빈도수를 센 후 그 결과를 정규화시켜 전체 합이 1 이 되도록 계산하였다. 그 결과, 각 전이 상태에 대해서 정의된 관찰 확률들은 아래 그림과 같이 정의되었다.



A	0.257
T	0.257
G	0.243
C	0.243
N	0.000

A	0.292
T	0.294
G	0.206
C	0.206
N	0.002

그림 7.모델의 전이 확률 및 관찰 확률

iii. 확률 평가

우리는 확률 평가와 최적 상태열에 대한 결과를 확인하기 위해 각 시퀀스의 길이가 300 인 인간 염색체 22 데이터를 활용하였으며 각 데이터의 특징은 표 2 와 같다.

표 2. 테스트 데이터를 이용한 확률 평가 결과의 예

Chromosome	Start	End	Sequence	Evaluation
chr22	25115000	25115300	tcagtgttgt...atatcactc	2.429456e-179
chr22	24115035	24115335	caagcccagc...agcttgcag	3.956869e-188
chr22	19863040	19863340	cacacttcag...cctgccaac	3.251544e-185
chr22	19881780	19882080	agcagggagt...tgagcctgg	8.957380e-179
chr22	26138119	26138419	ggggaaggga...tgattcctt	5.067118e-184
chr22	28247656	28247956	gtggcatttt...actttgcag	2.042505e-178
chr22	28247656	29280054	catgaggccg...agatcgtgc	1.086878e-188

iv. 최적 상태열 결정

우리 모델을 사용하여 관측 시퀀스에 대하여 다음과 같은 최적의 상태열을 제시했다.

표 4. 테스트 데이터를 이용한 최적 상태열 결정의 예

Observed sequence	agcagggagt...tgagcctgg with length 300
Evaluation	8.957380e-179
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII with the highest probability -441.154422196	

v. 매개 변수 추정

엑손과 인트론 각각의 영역을 구별하는 HMM 모델을 학습하기 위해 Baum-welch 알고리즘을 토대로 구현하였다. 학습에 사용된 training data set 는 염색체 21 의 Exon bed 파일 전체의 데이터 시퀀스를 정제하여서 생성했다. 각각의 시퀀스는 엑손과 인트론이 교차적으로 이어진다. 아래 표 5 와 6 에는 학습 이전의 transmission 확률과 emission 확률이 기술되었다. 모든 state 와 symbol 에 대하여 학습 이후 모델 확률의 변화가 관찰된다. 트레이닝 전의 초기 확률은 exon 과 intron 두개의 state 와 ‘A’,’T’,’G’,’C’,’N’ 5 개의 symbol 에 대해서 설정되었다. 학습 이후에는 ‘N’의 확률이 0 으로 나타났으며, 기존의 연구결과(Eddy, S. R. (2004). What is a hidden Markov model? [15]와 유사한 intron 구간의 GC 의 발생빈도가 낮고 AT 와 비교했을 때 A:T:G:C 는 4:4:1:1 의 상대적빈도를 보여주고 있다. 이를 통해 training 의 결과가 성공적인 것을 알 수 있다.

표 5 Transition 확률

STATE	학습 이전		학습 이후	
To / From	E	I	E	I
E	0.90	0.10	0.92	0.07
I	0.10	0.90	0.02	0.97

표 6. Emission 확률

	학습 이전					학습 이후				
Symbol / State	A	T	G	C	N	A	T	G	C	N
E	0.257	0.257	0.243	0.243	0.000	0.069	0.381	0.180	0.371	0.0
I	0.292	0.294	0.206	0.206	0.002	0.362	0.419	0.127	0.092	0.0

결론

본 연구에서는 엑손 인트론 구간 식별하는 HMM 의 세 가지 문제 평가, 디코딩, 학습에 대한 알고리즘을 구현하였다. 구현한 HMM 모델로 인간의 22 번 염색체 DNA 데이터셋을 분석하였다. 주어진 모델을 통해 관측 DNA 시퀀스가 발생할 확률이 계산되었고, 동일한 시퀀스의 최적의 상태열을 제안하였다. 염기 서열의 길이가 짧을 때는 미세한 확률 값을 가진 evaluation 결과가 비교적 쉽게 도출되었으나, 염기 서열의 길이가 길어질수록 확률이 극도로 낮아져 0 으로 수렴하는 결과가 나왔다. 이는 평가 단계에서 확률의 계산이 곱셈에 의해 이루어지는 것이 원인으로, 염기 서열의 길이가 길어지면 그만큼 값이 거듭해서

곱해지기에, 점차로 아주 낮은 확률 값이 나오게 된 것이라 추정하고 있다. 유전자 예측 모델 중 하나인 GenScan[6]의 경우, 이 문제를 방지하기 위해 Explicit State Duration HMM 을 사용하여 모델에 length distribution 을 도입함에 따라 길이에 의해 발생하는 문제를 어느 정도 해소시키는 것으로 볼 때 추후 모델 설계에 있어서 이 부분을 고려해야 할 것으로 보인다. 미지의 DNA 서열에서 엑손과 인트론의 구분을 하기 위해서는 다음의 사항을 고려할 필요가 있다. ⅰ) 평가 단계에서 확률의 계산이 곱셈에 의해 이루어지는 것을 고려하여, 서열의 길이에 따라 극단적으로 낮아지는 확률 값을 방지하기 위해 length distribution 을 도입하여야 한다. ⅱ)DNA 는 이중나선으로 구성되어 있으므로, 상보적인 두 strand 를 고려한 위치정보를 사용하여야 한다.

디코딩의 경우 엑손과 인트론 사이에서의 전이가 빈번하게 발생하지 않았다. 각 시퀀스 별 발생확률과 최적의 상태열을 제시하였음에 의의를 둔다. 그리고 전체 데이터 셋을 트레이닝 데이터로 사용하여 학습시킨 후의 모델의 확률은 기존의 연구결과와 유사한 결과를 보였기 때문에 적절히 모델링 했다고 판단된다. 그렇지만 트레이닝 데이터셋을 변경하면서 학습시킬 때의 발생된 문제는 연속적인 데이터셋을 트레이닝을 할 경우에 중간에 한 symbol 에 대하여 확률이 0 이 되는 경우에는 계산식에서의 결함으로 인해 그 이후에 트레이닝에 오류가 나는 것을 확인하였다. 그래서 추후에 HMM 모델을 발전 시키기 위하여 확률 계산 부분을 보완할 필요가 있다.

참고 문헌

[1] https://www.ncbi.nlm.nih.gov/

[2] Tropp, Burton E. (2008). Molecular Biology: Genes to Proteins. Jones & Bartlett Learning

[3] Alberts, Bruce (2008). Molecular biology of the cell. New York: Garland Science

[4] Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, & Madden T.L. (2012) "Domain enhanced lookup time accelerated BLAST." Biol Direct. 2012 Apr 17;7:12

[5] UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.

[6] Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of molecular biology, 268(1), 78-94.

[7] http://genes.mit.edu/GENSCAN.html

[8] Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proceedings of the National Academy of Sciences, 94(2), 565-568

[9] http://rulai.cshl.org/tools/genefinder/

[10] Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics. 37 (6): 1554–1563. doi:10.1214/aoms/1177699147. Retrieved 28 November 2011.

[11] Byung-Jun Yoon, Hidden Markov Models and their Applications in Biological Sequence Analysis

[12] Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77 (2), p. 257–286, February 1989

[13] Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 16.2. Viterbi Decoding". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press

[14] Frazzoli, Emilio. "Intro to Hidden Markov Models: the Baum-Welch Algorithm". Aeronautics and Astronautics, Massachusetts Institute of Technology. Retrieved 2 October 2013.

[15] Nature Biotechnology 22, 1315–1316 (2004) doi:10.1038/nbt1004-1315

[16] Cell Rep. 2012 May 31;1(5):543-56. doi: 10.1016/j.celrep.2012.03.013. Epub 2012 May 3.