
HMMs para Caracterização de Ilhas CpG

Heitor Baldo

Algoritmos em Bioinformática (IBI5037)

Prof. Alan M. Durham

2º Semestre, 2019

1 Modelos Ocultos de Markov para caracterização de ilhas CpG

Para a caracterização de ilhas CpG, construímos um HMM com oito estados, sendo A_+, C_+, G_+, T_+ os estados correspondentes às ilhas CpG e A_-, C_-, G_-, T_- os estados que não correspondem às ilhas CpG (NonCpG), totalmente conectado, como mostrado na figura 1 abaixo.

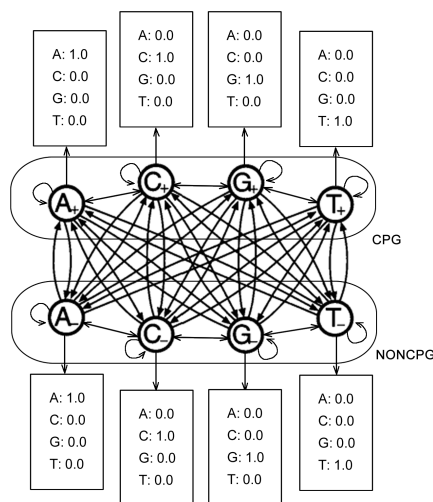


Figura 1. Arquitetura do HMM para ilhas CpG (imagem modificada de [1], p. 52). As caixas retangulares apresentam a probabilidade de emissão de cada nucleotídeo em cada estado.

2 Metodologia

Para implementarmos computacionalmente o HMM para ilhas CpG, iremos estimar seus parâmetros (probabilidades iniciais, de transição e de emissão) utilizando o framework ToPS [2], produzindo dois modelos, um tendo os parâmetros estimados pelo algoritmo de Baum-Welch, e o outro pelo algoritmo de Máxima Verossimilhança.

2.1 Estimação dos Parâmetros: Treinando o modelo com o Algoritmo de Baum-Welch

Como modelo inicial, utilizamos o modelo apresentado na figura 1, com as probabilidades iniciais, de transições e de emissões apresentadas na tabela 1 abaixo.

O modelo foi treinado separadamente: o modelo apenas com os estados (+) foi treinado com um conjunto de aproximadamente 15.000 nucleotídeos de ilhas CpG conhecidas (extraídos do arquivo CpG.fasta - cromossomo 1), e o modelo apenas com os estados (-) foi treinado com um conjunto de aproximadamente

	Transição								Emissão				Inicial
	A+	C+	G+	T+	A-	C-	G-	T-	A	C	G	T	
A+	0.22	0.25	0.35	0.18	—	—	—	—	1.00	0.00	0.00	0.00	0.20
C+	0.20	0.25	0.35	0.20	—	—	—	—	0.00	1.00	0.00	0.00	0.30
G+	0.22	0.28	0.28	0.22	—	—	—	—	0.00	0.00	1.00	0.00	0.30
T+	0.15	0.25	0.35	0.25	—	—	—	—	0.00	0.00	0.00	1.00	0.20
A-	—	—	—	—	0.23	0.27	0.25	0.25	1.00	0.00	0.00	0.00	0.25
C-	—	—	—	—	0.30	0.20	0.25	0.25	0.00	1.00	0.00	0.00	0.25
G-	—	—	—	—	0.25	0.25	0.25	0.25	0.00	0.00	1.00	0.00	0.25
T-	—	—	—	—	0.25	0.25	0.30	0.20	0.00	0.00	0.00	1.00	0.25

Tabela 1. Tabela de probabilidades do HMM apresentado na figura 1.

15.000 nucleotídeos de trechos normais de DNA¹, utilizando as mesmas probabilidades iniciais e de emissão da tabela 1 acima. O valor do parâmetro *maxiter* foi setado com o valor 300.

Para o treinamento e validação de ambos os modelos ((+) e (-)), foi utilizada a estratégia de validação cruzada k-fold, com $k = 6$, e com cada um dos 6 subconjuntos contendo aproximadamente 2.500 nucleotídeos, sendo que 5 subconjunto (totalizando 12.500 nucleotídeos) foram usados para treinamento, e um subconjunto utilizado para teste. Os erros nos treinamentos foram muitos pequenos e, por conseguinte, puderam ser desconsiderados.

O modelo final foi composto pelos dois modelos (+) e (-) treinados separadamente e as probabilidades de transição entre os dois modelos foram adicionadas posteriormente, utilizando como probabilidade total $p = 0.1$ para transições do estado (+) para (-) e vice-versa, então fizemos $\frac{p}{4} = 0.025$, e os valores das probabilidades de transição entre os estados de (+), tanto quanto dos estados (-), tendo valor $p' = 0.9$, e os valores dos parâmetros encontrados nos treinamentos separados foram ajustados, como mostra o modelo final apresentado na figura 2.

model_name="HiddenMarkovModel"	"A-" "C-": 0.25;	"G+" "T-": 0.025;	"C" "G+": 0;
state_names =	"C-" "C-": 0.28;	"T+" "T-": 0.025;	"G" "G+": 1;
("A+", "C+", "G+", "T+",	"G-" "C-": 0.09;	"C-" "A+": 0.025;	"T" "G+": 0;
"A-", "C-", "G-", "T-")	"T-" "C-": 0.28;	"G-" "A+": 0.025;	"A" "T+": 0;
observation_symbols =	"A-" "G-": 0.20;	"T-" "A+": 0.025;	"C" "T+": 0;
("A", "C", "G", "T")	"C-" "G-": 0.27;	"A-" "C+": 0.025;	"G" "T+": 0;
transitions =	"G-" "G-": 0.28;	"C-" "C+": 0.025;	"T" "T+": 1;
("A+" "A+": 0.22;	"T-" "G-": 0.15;	"G-" "C+": 0.025;	"A" "A-": 1;
"C+" "A+": 0.28;	"A-" "T-": 0.13;	"T-" "C+": 0.025;	"C" "A-": 0;
"G+" "A+": 0.32;	"C-" "T-": 0.27;	"A-" "G+": 0.025;	"G" "A-": 0;
"T+" "A+": 0.08;	"G-" "T-": 0.31;	"C-" "G+": 0.025;	"T" "A-": 0;
"A+" "C+": 0.17;	"T-" "T-": 0.19;	"G-" "G+": 0.025;	"A" "C-": 0;
"C+" "C+": 0.38;	"A+" "A-": 0.025;	"T-" "G+": 0.025;	"C" "C-": 1;
"G+" "C+": 0.20;	"C+" "A-": 0.025;	"A-" "T+": 0.025;	"G" "C-": 0;
"T+" "C+": 0.15;	"G+" "A-": 0.025;	"C-" "T+": 0.025;	"T" "C-": 0;
"A+" "G+": 0.15;	"T+" "A-": 0.025;	"G-" "T+": 0.025;	"A" "G-": 0;
"C+" "G+": 0.28;	"A+" "C-": 0.025;	"T-" "T+": 0.025;	"C" "G-": 0;
"G+" "G+": 0.37;	"C+" "C-": 0.025;	emission_probabilities =	"G" "G-": 1;
"T+" "G+": 0.10;	"G+" "C-": 0.025;	("A" "A+": 1;	"T" "G-": 0;
"A+" "T+": 0.11;	"T+" "C-": 0.025;	"C" "A+": 0;	"A" "T-": 0;
"C+" "T+": 0.34;	"A+" "G-": 0.025;	"G" "A+": 0;	"C" "T-": 0;
"G+" "T+": 0.28;	"C+" "G-": 0.025;	"T" "A+": 0;	"G" "T-": 0;
"T+" "T+": 0.17;	"G+" "G-": 0.025;	"A" "C+": 0;	"T" "T-": 1)
"A-" "A-": 0.23;	"T+" "G-": 0.025;	"C" "C+": 1;	initial_probabilities =
"C-" "A-": 0.24;	"A+" "T-": 0.025;	"G" "C+": 0;	("A+": 0.125; "C+": 0.125;
"G-" "A-": 0.28;	"C+" "T-": 0.025;	"T" "C+": 0;	"G+": 0.125; "T+": 0.125;
"T-" "A-": 0.15;	"A-" "G+": 0;	"A" "G+": 0;	"A-": 0.125; "C-": 0.125;
			"G-": 0.125; "T-": 0.125)

Figura 2. Parêmtros estimados com o Algoritmo de Baum-Welch.

2.2 Estimação dos Parâmetros: Treinando o modelo com o Algoritmo de Máxima Verossimilhança

Baseando-se no modelo precedente, escolhemos os parâmetros do modelo inicial do seguinte modo:

¹Extraídos de <https://www.ncbi.nlm.nih.gov/nuccore/CM000663.2?report=fasta>

model_name="HiddenMarkovModel"	"C+" "A-": 0.023;	"G-" "T-": 0.1737;	"C" "G+": 0.0451;
state_names =	"C+" "C-": 0.0465;	"T+" "A+": 0.093;	"G" "G+": 0.8904;
("A+", "C+", "G+", "T+",	"C+" "G-": 0.0087;	"T+" "C+": 0.252;	"T" "G+": 0.0331;
"A-", "C-", "G-", "T-")	"C+" "T-": 0.0226;	"T+" "G+": 0.3459;	"A" "T+": 0.0624;
observation_symbols =	"C-" "A+": 0.0243;	"T+" "T+": 0.245;	"C" "T+": 0.0486;
("A", "C", "G", "T")	"C-" "C+": 0.0392;	"T+" "A-": 0.0093;	"G" "T+": 0.0544;
transitions =	"C-" "G+": 0.0089;	"T+" "C-": 0.0184;	"T" "T+": 0.8376;
("A+" "A+": 0.225;	"C-" "T+": 0.0146;	"T+" "G-": 0.0182;	"A" "A-": 0.8699;
"A+" "C+": 0.200;	"C-" "A-": 0.3052;	"T+" "T-": 0.0251;	"C" "A-": 0.0434;
"A+" "G+": 0.300;	"C-" "C-": 0.2486;	"T-" "A+": 0.021;	"G" "A-": 0.0445;
"A+" "T+": 0.175;	"C-" "G-": 0.065;	"T-" "C+": 0.0274;	"T" "A-": 0.0452;
"A+" "A-": 0.027;	"C-" "T-": 0.3013;	"T-" "G+": 0.0372;	"A" "C-": 0.0499;
"A+" "C-": 0.034;	"G+" "A+": 0.1846;	"T-" "T+": 0.0209;	"C" "C-": 0.8827;
"A+" "G-": 0.016;	"G+" "C+": 0.2703;	"T-" "A-": 0.0966;	"G" "C-": 0.033;
"A+" "T-": 0.023;	"G+" "G+": 0.2326;	"T-" "C-": 0.2415;	"T" "C-": 0.032;
"A+" "A+": 0.035;	"G+" "T+": 0.1815;	"T-" "G-": 0.3511;	"A" "G-": 0.0382;
"A+" "C+": 0.015;	"G+" "A-": 0.0399;	"T-" "T-": 0.2114;	"C" "G-": 0.0307;
"A+" "G+": 0.035;	"G+" "C-": 0.0272;	emission_probabilities =	"G" "G-": 0.8935;
"A+" "T+": 0.015;	"G+" "G-": 0.0531;	("A" "A+": 0.8358;	"T" "G-": 0.0406;
"A+" "A-": 0.270;	"G+" "T-": 0.0178;	"C" "A+": 0.0607;	"A" "T-": 0.0729;
"A+" "C-": 0.170;	"G-" "A+": 0.048;	"G" "A+": 0.0564;	"C" "T-": 0.0389;
"A+" "G-": 0.330;	"G-" "C+": 0.0303;	"T" "A+": 0.0501;	"G" "T-": 0.036;
"A+" "T-": 0.130;	"G-" "G+": 0.0287;	"A" "C+": 0.0382;	"T" "T-": 0.8552;
"C+" "A+": 0.255;	"G-" "T+": 0.0187;	"C" "C+": 0.9006;	initial_probabilities =
"C+" "C+": 0.245;	"G-" "A-": 0.2575;	"G" "C+": 0.0376;	("A+": 0.075; "C+": 0.1;
"C+" "G+": 0.075;	"G-" "C-": 0.207;	"T" "C+": 0.0266;	"G+": 0.125; "T+": 0.03;
"C+" "T+": 0.325;	"G-" "G-": 0.2432;	"A" "G+": 0.0344;	"A-": 0.07; "C-": 0.24;
			"G-": 0.29; "T-": 0.07)

Figura 3. Modelo inicial para estimação dos parâmetros com o Algoritmo de Máxima Verossimilhança.

A partir deste modelo inicial (figura 3), geramos um conjunto rotulado com 66.000 nucleotídeos (44 sequências com 1.500 nucleotídeos cada, utilizando o comando *simulate* do ToPS) e aplicamos o método k-fold com $k = 10$, com 6.600 nucleotídeos cada subconjunto, sendo 9 utilizados para treinamento e um para teste. No conjunto de treino, aplicamos o algoritmo de Máxima Verossimilhança para estimação dos parâmetros, utilizando a estratégia de validação cruzada, e obtemos o modelo final com um erro médio de 0.04 para a probabilidade de transição entre si dos estados de (+) tanto quando dos estados de (-), e 0.005 para a probabilidade de transição entre os estados (+) e (-):

model_name = "HiddenMarkovModel"	"A-" "G-": 0.485903;	"C-" "G-": 0.00430883;	"G" "T+": 0.0271593;
state_names = ("A+", "C+", "G+", "T+",	"A+" "T+": 0.0876885;	"G-" "G-": 0.214991;	"T" "T+": 0.418173;
"A-", "C-", "G-", "T-")	"C+" "T+": 0.16285;	"T-" "G-": 0.222143;	"A" "A-": 0.93355;
observation_symbols = ("A", "C", "G", "T")	"G+" "T+": 0.592023;	"A+" "T-": 0.0254172;	"C" "A-": 0.0216675;
transitions = ("A+" "A+": 0.250721;	"T+" "T+": 0.122764;	"C+" "T-": 0.0249751;	"G" "A-": 0.0222167;
"C+" "A+": 0.256862;	"A-" "T+": 0.00751616;	"G+" "T-": 0.0196707;	"T" "A-": 0.0225662;
"G+" "A+": 0.447123;	"C-" "T+": 0.00731573;	"T+" "T-": 0.0277379;	"A" "C-": 0.00554592;
"T+" "A+": 0.0190344;	"G-" "T+": 0.00937015;	"A-" "T-": 0.143662;	"C" "C-": 0.876089;
"A-" "A+": 0.00716347;	"T-" "T+": 0.0104725;	"C-" "T-": 0.332965;	"G" "C-": 0.114808;
"C-" "A+": 0.0049735;	"A+" "A-": 0.0133103;	"G-" "T-": 0.191955;	"T" "C-": 0.0035565;
"G-" "A+": 0.00982419;	"C+" "A-": 0.0113384;	"T-" "T-": 0.233617;	"A" "G-": 0.00254616;
"T-" "A+": 0.00429808;	"G+" "A-": 0.0196697;	emission_probabilities =	"C" "G-": 0.00204626;
"A+" "C+": 0.14853;	"T+" "A-": 0.00458467;	("A" "A+": 0.96658;	"G" "G-": 0.792741;
"C+" "C+": 0.277874;	"A-" "A-": 0.133103;	"C" "A+": 0.0121327;	"T" "G-": 0.202666;
"G+" "C+": 0.157231;	"C-" "A-": 0.643431;	"G" "A+": 0.0112732;	"A" "T-": 0.072682;
"T+" "C+": 0.154966;	"G-" "A-": 0.126941;	"T" "A+": 0.010014;	"C" "T-": 0.0387836;
"A-" "C+": 0.125631;	"T-" "A-": 0.0476214;	"A" "C+": 0.00477321;	"G" "T-": 0.0358923;
"C-" "C+": 0.00485197;	"A+" "C-": 0.00378063;	"C" "C+": 0.987205;	"T" "T-": 0.852642;
"G-" "C+": 0.127525;	"C+" "C-": 0.00517057;	"G" "C+": 0.00469824;	initial_probabilities =
"T-" "C+": 0.00339142;	"G+" "C-": 0.00302451;	"T" "C+": 0.0032375;	("A+": 0.0993902;
"A+" "G+": 0.049478;	"T+" "C-": 0.00204599;	"A" "G+": 0.00573047;	"C+": 0.173171;
"C+" "G+": 0.0123695;	"A-" "C-": 0.130098;	"C" "G+": 0.00751291;	"G+": 0.125;
"G+" "G+": 0.203289;	"C-" "C-": 0.250033;	"G" "G+": 0.981243;	"T+": 0.025122;
"T+" "G+": 0.221975;	"G-" "C-": 0.245408;	"T" "G+": 0.00551391;	"A-": 0.0260976;
"A-" "G+": 0.00577243;	"T-" "C-": 0.360439;	"A" "T+": 0.0311533;	"C-": 0.200976;
"C-" "G+": 0.166395;	"A+" "G-": 0.00106064;	"C" "T+": 0.523515;	"G-": 0.348537;
"G-" "G+": 0.0047334;	"C+" "G-": 0.00057672;		"T-": 0.00170732)
"T-" "G+": 0.335989;	"G+" "G-": 0.0698097;		
	"T+" "G-": 0.00120647;		

Figura 4. Parêmtros estimados com o Algoritmo de Máxima Verossimilhança.

3 Resultados

Para estimar a capacidade de previsão dos nossos modelos finais (figura 2 e figura 4), aplicamo-os à um conjunto de 10 sequências, todas com aproximadamente 3.000 nucleotídeos, todas contendo ilhas CpG putativas.

Utilizamos o algoritmo de Viterbi para a decodificação (comando *viterbi_decoding* do ToPS). O modelo treinado com o algoritmo de Baum-Welch (figura 2) identificou todas as ilhas CpG em todas as 10 sequências, porém 12 novas ilhas CpG foram previstas (falsos positivos), e 10 trechos foram caracterizadas como não sendo ilhas CpG erroneamente (falsos negativos).

O modelo treinado com o algoritmo de Máxima Verossimilhança (figura 4) conseguiu identificar apenas cinco das ilhas CpG, e de forma não muito precisa. Além disso, 112 novas ilhas CpG foram previstas (falsos positivos), revelando uma grande imprecisão do modelo.

Todavia, para identificação de ilhas CpG em sequências pequenas (< 200 nucleotídeos), o modelo treinado com o algoritmo de Máxima Verossimilhança mostrou-se mais preciso do que o modelo treinando com o algoritmo de Baum-Welch, rotulando corretamente todas as 10 sequências curtas contra apenas uma do modelo estimado com Baum-Welch.

4 Conclusão

O modelo treinado com o algoritmo de Baum-Welch mostrou-se muito mais preciso do que o modelo treinado com o algoritmo de Máxima Verossimilhança para caracterização de ilhas CpG em sequências longas, porém o inverso mostrou-se verdadeiro para caracterização de ilhas CpG em sequências curtas (< 200 nucleotídeos).

Referências

- [1] R. DURBIN, S. R. EDDY, A. KROGH, G. MITCHISON, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
- [2] A. Y. KASHIWABARA, I. BONADIO, V. ONUCHIC, F. AMADO, R. M., A. M. DURHAM, *ToPS: A Framework to Manipulate Probabilistic Models of Sequence Data*. PLoS Comput. Biol. 2013 Oct; 9(10): e1003234.