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.

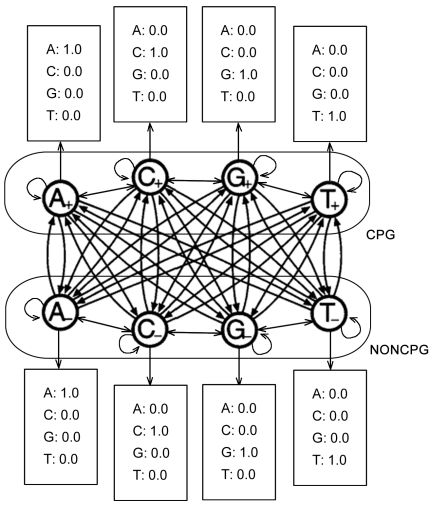

Figura 1. Arquitetura do HMM para ilhas CpG (imagem modificada de [1], p. 52). As caixas retangulares apresentam a probabilidade de emissão de cada nucleotídeo em cada estado.

2 Metodologia

Para implementarmos computacionalmente o HMM para ilhas CpG, iremos estimar seus parêmtros (probabilidades iniciais, de transição e de emissão) utilizando o framework ToPS [2], produzindo dois modelos, um tendo os parâmetros estimados pelo algortimo 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 DNA[1], 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 subcojunto (totalizando 12.500 nucleotídeos) foram usandos 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"
state_names =
   ("A+", "C+", "G+", "T+",
    "A-", "C-", "G-", "T-")
observation_symbols =
   ("A", "C", "G", "T")
transitions =
   ("A+" | "A+": 0.22;
 "C+" | "A+": 0.28;
 "G+" | "A+": 0.32;
 "T+" | "A+": 0.08;
 "A+" | "C+": 0.17;
 "C+" | "C+": 0.38;
 "G+" | "C+": 0.20;
 "T+" | "C+": 0.15;
 "A+" | "G+": 0.15;
 "C+" | "G+": 0.28;
 "G+" | "G+": 0.37;
 "T+" | "G+": 0.10;
 "A+" | "T+": 0.11;
 "C+" | "T+": 0.34;
 "G+" | "T+": 0.28;
 "T+" | "T+": 0.17;
 "A-" | "A-": 0.23;
 "C-" | "A-": 0.24;
 "G-" | "A-": 0.28;
 "T-" | "A-": 0.15;
"A-" | "C-": 0.25;
"C-" | "C-": 0.28;
"G-" | "C-": 0.09;
"T-" | "C-": 0.28;
"A-" | "G-": 0.20;
"C-" | "G-": 0.27;
"G-" | "G-": 0.28;
"T-" | "G-": 0.15;
"A-" | "T-": 0.13;
"C-" | "T-": 0.27;
"G-" | "T-": 0.31;
"T-" | "T-": 0.19;
"A+" | "A-": 0.025;
"C+" | "A-": 0.025;
"G+" | "A-": 0.025;
"T+" | "A-": 0.025;
"A+" | "C-": 0.025;
"C+" | "C-": 0.025;
"G+" | "C-": 0.025;
"T+" | "C-": 0.025;
"A+" | "G-": 0.025;
"C+" | "G-": 0.025;
"G+" | "G-": 0.025;
"T+" | "G-": 0.025;
"A+" | "T-": 0.025;
"C+" | "T-": 0.025;
"G+" | "T-": 0.025;
"T+" | "T-": 0.025;
"A-" | "C+": 0.025:
"C-" | "C+": 0.025;
"G-" | "C+": 0.025;
"T-" | "C+": 0.025;
"A-" | "G+": 0.025;
"C-" | "G+": 0.025;
"G-" | "G+": 0.025;
"T-" | "G+": 0.025;
"A-" | "T+": 0.025;
"C-" | "T+": 0.025;
"G-" | "T+": 0.025;
"T-" | "T+": 0.025)
emission_probabilities =
   ("A" | "A+": 1;
    "C" | "A+": 0;
    "G" | "A+": 0;
    "T" | "A+": 0;
    "A" | "C+": 0;
    "C" | "C+": 1;
    "G" | "C+": 0;
    "T" | "C+": 0;
    "A" | "G+": 0;
    "C" | "G+": 0;
    "G" | "G+": 1;
    "T" | "G+": 0;
    "A" | "T+": 0;
    "C" | "T+": 0;
    "G" | "T+": 0;
    "T" | "T+": 1;
    "A" | "A-": 1;
    "C" | "A-": 0;
    "G" | "A-": 0;
    "T" | "A-": 0;
    "A" | "C-": 0;
    "C" | "C-": 1;
    "G" | "C-": 0;
    "T" | "C-": 0;
    "A" | "G-": 0;
    "C" | "G-": 0;
    "G" | "G-": 1;
    "T" | "G-": 0;
    "A" | "T-": 0;
    "C" | "T-": 0;
    "G" | "T-": 0;
    "T" | "T-": 1)
initial_probabilities =
 ("A+": 0.125; "C+": 0.125;
  "G+": 0.125; "T+": 0.125;
  "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:

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

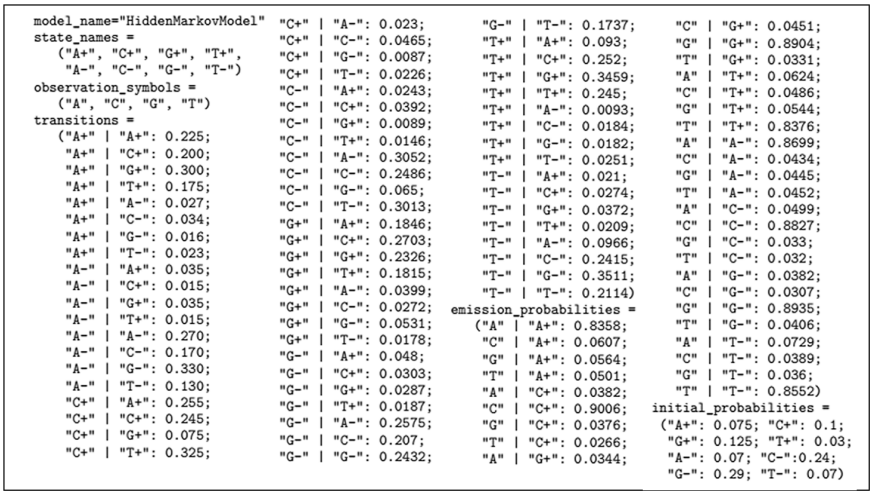

Figura 3. Modelo inicial para estimação dos parâmetros com o Algoritmo de Máxima Verossimilhança.

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 (−):

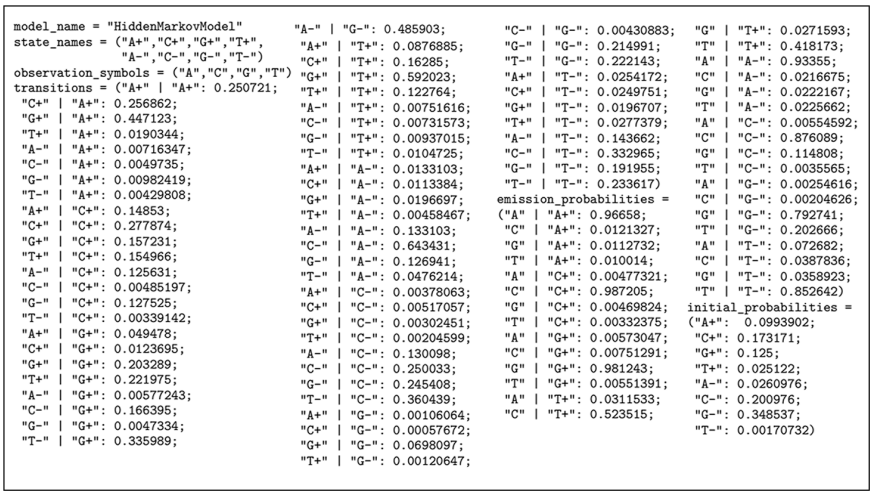

Figura 4. Parêmtros estimados com o Algoritmo de Máxima Verossimilhança.

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 algortimo 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.