Identificação de regiões codificantes e não-codificantes: *Hidden Markov Models*

Bioinformática

António Sousa - up201208681

Bruno Cabral - up201202369

Ricardo Santos - up201203540

12 de Maio de 2018

1 Introdução

Tipicamente uma sequência de ADN é composta por um ou mais genes - região(ões) codificante(s) - intercalados por regiões não-codificantes. Por isso, antes de anotar a função de cada um dos genes que compõem uma sequência de ADN, que resulta da sequenciação de um genoma (ADN de uma espécie) ou de um metagenoma (ADN ambiental de várias espécies), é necessário identificar onde começa e acaba cada região codificante, *i.e.*, gene ou exão, assim como não-codificante, *i.e.*, intrão. Para esta finalidade existem duas abordagens possíveis: (1) métodos baseados na pesquisa de similaridade; e (2) métodos *ab initio* (Zhu et al. (2010)). Em (1) os genes/exões são identificados através de pesquisas de similaridade, e.g., usando BLAST (Altschul et al. (1990)), com outros genes/exões presentes em base de dados genéticas. Em (2) os genes/exões são identificados de novo com base em características distintivas que caracterizam as regiões codificantes e as separam das não-codificantes: regiões ricas em AT (regiões codificantes); regiões ricas em GC (regiões não-codificantes); codões de iniciação (ATG - o mais frequente) e terminação (TAA, TAG, TGA); uso/viés de codões (Zhu et al. (2010)).

Programas que procuram identificar regiões codificantes através de métodos *ab initio*, tais como o FragGeneScan (Rho et al. (2010)) e o GLIMMER (Salzberg et al. (1998)), implementam *hidden Markov models* (HMM) e *interpolated Markov models*, respectivamente, que procuram prever as regiões codificantes com base nos dados usados para treinar o modelo, viés de codões (entre outros parâmetros) e composição de sub-sequências de tamanho K que compõem um conjunto de genes. No primeiro caso, o HMM encontra o melhor caminho de *hidden states* para uma sequência de codões através do algoritmo de programação dinâmica Viterbi (Rho et al. (2010)).

Na prática, no caso de análise mais simples, um gene procariótico começa por um codão de iniciação, prolongado por vários codões e interrompido (a região codificante) por um codão de terminação (em média com um tamanho $\approx$ 950 bp) (Rho et al. (2010)). Ambos, os programas têm em consideração as seis possíveis grelhas de leitura nas quais

um gene pode começar e terminar. Enquanto os programas que se baseiam em (1) apenas conseguem identificar genes previamente identificados nas base de dados, programas em (2), nomeadamente aqueles que usam HMMs, são capazes de identificar novos genes.

2 Objetivos do trabalho

Este trabalho teve como objetivo a implementação de um *hidden Markov model* para prever regiões codificantes, *i.e.*, genes, presentes em sequências de ADN. Com este propósito o HMM foi construído com base nas transições de nucleotídeos (n=16) de um modelo de treino (constituído apenas por genes), para encontrar o melhor caminho de *hidden states* capaz de gerar a sequência de nucleotídeos observada através dos algoritmos de Viterbi e Forward.

3 Implementação

O nosso trabalho encontra-se dividido em duas partes. Uma parte dedicada à implementação do algoritmo de Viterbi e de Forward e outra relativa à estimação de parâmetros. Na primeira parte implementou-se o algoritmo de Viterbi e de Forward recebendo cada um destes uma HMM.

A HMM contém a seguinte informação:

- Número de observações;
- Número de estados;
- Sequência pretendida;
- Matriz de transição;
- Matriz de emissão;

- Array com as transições iniciais.

Depois da HMM estar preenchido, o algoritmo de Viterbi vai nos retornar o caminho mais provável para gerar a sequência pretendida. Relativamente ao algoritmo de Forward, este vai nos indicar a probabilidade da sequência pretendida.

Na segunda parte implementou-se o algoritmo de Baum-Welch de modo a estimar os parâmetros de uma determinada HMM, passando também algumas sequências de modo a treinar este mesmo algoritmo. É ainda necessário indicar o valor *threshold* bem como o número máximo de iterações. Servindo estes valores para terminar o cálculo das novas matrizes de transição e emissão, retornando as novas matrizes.

De modo a testar os algoritmos, encontram-se alguns ficheiros de texto com o *input* necessário para avaliar estes mesmos:

```
(input_viterbi.txt, input_forward.txt, input_bw.txt)
```

Forma de executar:

```
javac ‘algoritmo_desejado’.java && java  ‘algoritmo_desejado’
< input_’algoritmo_desejado’.txt

(ex.: javac viterbi.java && viterbi < input_viterbi.txt)
```

4 Conclusão

Durante este trabalho foi implementado um *hidden Markov model* com base nas transições de nucleotídeos (n=16) capaz de prever regiões codificantes e não-codificantes presentes numa dada sequência de ADN. A implementação bem sucedida do HMM foi feita através do algoritmo de programação dinâmica Viterbi e do algoritmo Forward para encontrar mais eficientemente o melhor caminho de *hidden states*.

Referências

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman

1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Rho, M., H. Tang, and Y. Ye

2010. Fraggenescan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191–e191.

Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White

1998. Microbial gene identification using interpolated markov models. *Nucleic Acids Research*, 26(2):544–548.

Zhu, W., A. Lomsadze, and M. Borodovsky

2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, 38(12):e132–e132.