Facultad de Ingeniería, UNER

Modelización de sistemas biológicos por computadora

Guía de Trabajos Prácticos Nº 9

Modelos Ocultos de Markov: Etiquetado de ADN

1. Introducción

Los genomas de eucariontes son muy grandes y mucha de su estructura corresponde a regiones que no codifican para ningún producto funcional, estas son las llamadas secuencias no-codificantes. Algunas de estas secuencias no codificantes son secuencias espaciadoras entre los genes y otras (intrones) los interrumpen.

Uno de los problemas básicos en genómica es la predicción y anotación de genes. La cual constituye un primer paso, necesario para la utilización y comprensión de los genomas. Una forma particular de este problema es la clasificación de regiones codificantes y no codificantes.

En el caso de los genes humanos suele tratarse de detectar las señales de los sitios de splice para localizar las regiones codificantes. En los años 80 se desarrollaron programas usando sólo señales de splice y codones. Más tarde, en los 90, se añadieron técnicas estadísticas, lingüísticas y de aprendizaje automático. Entre estas últimas se encuentran los Modelos Ocultos de Markov (MOM), los cuales pueden ser considerados hoy en día como una de las técnicas clásicas en el etiquetado de secuencias.

Para el presente trabajo práctico se propone la solución de un problema de juguete en el cual cada porción de la secuencia puede tomar los estados codificante o no-codificante. En términos de un MOM se toma como salida del modelo cada una de las bases de la secuencia como se puede observar en la Figura 1. Este modelo constituye una versión sobre-simplificada de los modelos utilizados realmente en los problemas de etiquetado de ADN.

2. Actividades

1. Implemente el método de Viterbi para resolver el problema de identificación de la secuencia de estados mas probable, dados la secuencia de salida y los parámetros de un modelo (ver Algritmo 1).

2. Implemente el método iterativo de identificación de parámetros basado en el algoritmo de Viterbi de decodificación.

3. Utilizando las cuatro secuencias de salida de entrenamiento dadas en el archivo asociado a este documento obtenga los parámetros óptimos del modelo. Encuentre también las secuencias de estados mas probables correspondientes a dichas salidas.

4. Compare los parámetros estimados obtenidos con los parámetros del sistema real.

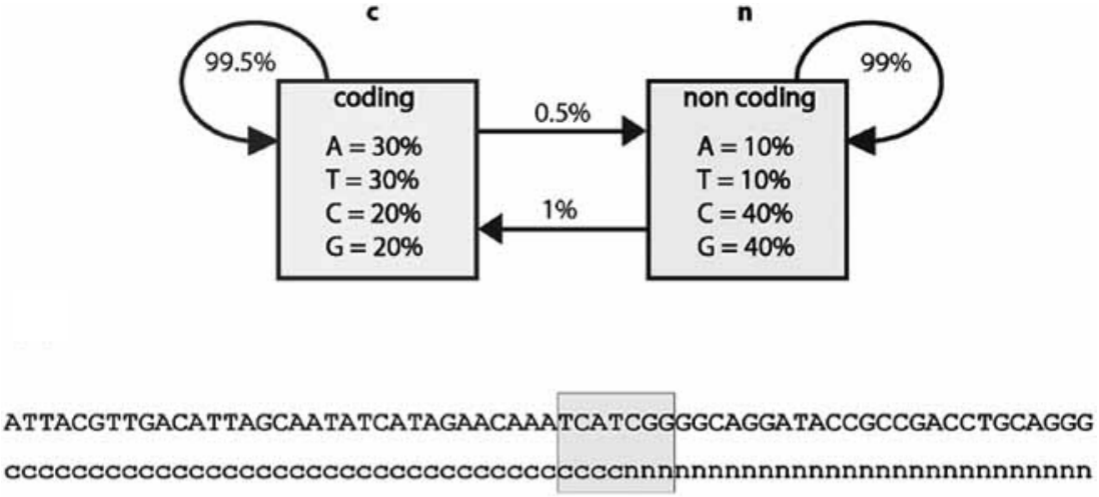

Figura 1: Esquema del MOM utilizado para resolver el problema de juguete propuesto.

3. Apéndice

Sean $\mathcal{E}$ el conjunto de estados posibles, $\mathcal{X}$ el conjunto de símbolos de salida posibles, x_i la observación en el instante i, L es el largo de la secuencia, $b_e(x)$ la probabilidad de emisión para la salida $x \in \mathcal{X}$ en el estado $e \in \mathcal{E}$, a_{ij} la probabilidad de transición del estado i al estado j, el algoritmo de Viterbi para encontrar la secuencia de estados mas probable $\mathbf{s}^* = [\mathbf{s_1^*}, \ldots, \mathbf{s_L^*}]$, y su correspondiente probabilidad p^* se describe a continuación:

Algoritmo 1 Algoritmo de Viterbi para Decodificación

1: $\gamma_1(e) \leftarrow \mathrm{b}_e(x_1) \cdot \pi_e$, $\forall e \in \mathcal{E}$ ▷ inicializa γ
2: $\Psi_1(e) \leftarrow 0, \forall e \in \mathcal{E}$ ▷ inicializa Ψ
3: **for** $1 < k \leq L$ **do** ▷ bucle principal
4: $\quad \gamma_k(e) \leftarrow \mathrm{b}_e(x_k) \cdot \mathrm{máx}_v(\gamma_{k-1}(v) \cdot a_{ve})$, $\forall e \in \mathcal{E}$
5: $\quad \Psi_k(e) \leftarrow \arg\max_v(\gamma_{k-1}(v) \cdot a_{ve})$, $\forall e \in \mathcal{E}$
6: $p^* \leftarrow \mathrm{máx}_v\, \gamma_L(v)$ ▷ obtengo la probabilidad de ocurrencia
7: $s_L^* \leftarrow \arg\max_v \gamma_L(v)$ ▷ obtengo el estado final de la secuencia mas probable
8: **for** $L > k \leq 1$ **do** ▷ "backtracking"
9: $\quad s_k^* \leftarrow \Psi_{k+1}(s_{k+1}^*)$

Referencias

[1] De Fonzo V., Aluffi-Pentini F. and Parisi V. *Hidden Markov Models in Bioinformatics.* Current Bioinformatics, 2007, 2, 49-61.