

Localización de motivos en biosecuencias mediante inferencia gramatical

Piedachu Peris

Departamento de Sistemas Informáticos y Computación.

Universidad Politécnica de Valencia.

pperis@dsic.upv.es

Inferencia Gramatical (GI)

- Consiste en aprender o inferir un lenguaje a partir de un conjunto de palabras.

$$S = \{aab, aaaab, aaaaaab\}$$

- Diferente algoritmo de GI \rightarrow diferente lenguaje:

$$L_a = \{a^n b : n \geq 1\}$$

$$L_b = \{a^n b : n \geq 2\}$$

$$L_c = \{(aa)^n b : n \geq 1\}$$

- Mayor alfabeto \rightarrow más difícil de aprender un lenguaje.

GI con biosecuencias

- Alfabeto: aminoácidos (o nucleótidos)

$$\Sigma = \{A, C, G, T\}$$

$$\Delta = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

- Palabras: proteínas o cadenas de ADN (cadenas de aminoácidos o de nucleótidos)

$$w = ATCTGATTCGGG, ATGGCGGATT, ATAGGCCCGGTA$$

$$W = \{MDAIKKM, GDAVKK, MDAAIKKM\}$$

- Mediante GI obtenemos un lenguaje que nos permite predecir el comportamiento de otras biosecuencias (palabras) que pertenezcan al mismo lenguaje.

Aplicaciones

- Detección de motivos coiled-coil.
- Detección de segmentos transmembrana.
- Detección de ADN codificante.
- Detección de dominios estructurales.

Método general

1. Proceso de la base de datos

$$W = \{MDAIKKM, GDAVKK, MDAAIKKM\}$$

2. Reducción del alfabeto: Dayhoff

MDAIKKM GDAVKK MDAAIKKM
ecbedde bcbedd ecbbbedde

3. Anotación (etiquetado). Al etiquetar decidimos qué información destacar:

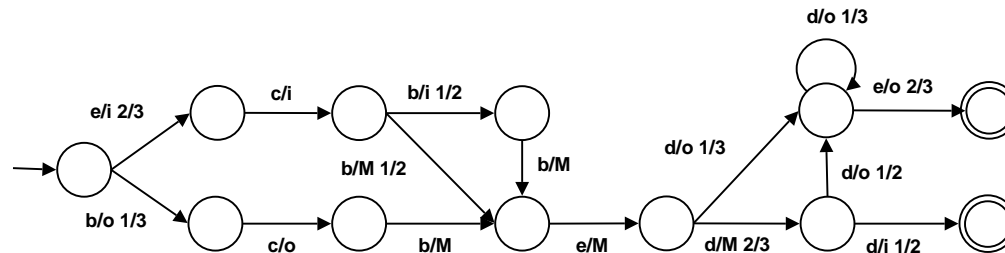
ecbedde bcbedd ecbbbedde
iiMM Moo ooMMMi iiiMMooo

Amino acid	Dayhoff
C	a
G, S, T, A, P	b
D, E, N, q	c
R, H, K,	d
L, V, M, I	e
Y, F, W	f
B, Z	g

Método general (II)

4. Proceso de GI: Inferencia de un transductor probabilístico:
input: proteína + etiquetado (cada símbolo con su etiqueta):

[ei] [ci] [bM] [eM] [dM] [do] [eo]
[bo] [co] [bM] [eM] [dM] [di]
[ei] [ci] [bi] [bM] [eM] [do] [do] [eo]



output: anotación de las palabras (proteínas): iiMMMOo ooMMMi
iiiMMOoo

5. Fase de test: devuelve la transducción más probable a partir de la secuencia de entrada, utilizando Viterbi.

input: MDAIKKKHL → ecbeddde

output: iiiMMOooo

Servidor online

- <http://www.dsic.upv.es/users/tlcc/bio/bio.html>