CLASIFICACIÓN DE SECUENCIAS MEDIANTE K-MERS

Ignacio Bitrián 717901

Bioinformática

7 de Mayo de 2020

INDICE

INTRODUCCIÓN

¿Qué son los K-mers?

INTRODUCCIÓN

- Se entiende por k-Mers a todas las subsecuencias de tamaño k de una cadena de aminoácidos.
- En una secuencia de tamaño t, el número de k-mers es t+1-k.

Secuencia : ATGAACACTA

K	K-MERS
1	A, T, G, A, A, C, A, C, T, A
2	AT, TG, GA, AA, AC, CA, AC, CT, TA
3	ATG, TGA, GAA, AAC, ACA, CAC, ACT, CTA
4	ATGA, TGAA, GAAC, AACA, ACAC, CACT, ACTA
5	ATGAA, TGAAC, GAACA, AACAC, ACACT, CACTA
6	ATGAAC, TGAACA, GAACAC, AACACT, ACACTA
7	ATGAACA, TGAACAC, GAACACT, AACACTA
8	ATGAACAC, TGAACACT, GAACACTA
9	ATGAACACT, TGAACACTA
10	ATGAACACTA

UTILIZACIÓN

UTILIZACIÓN

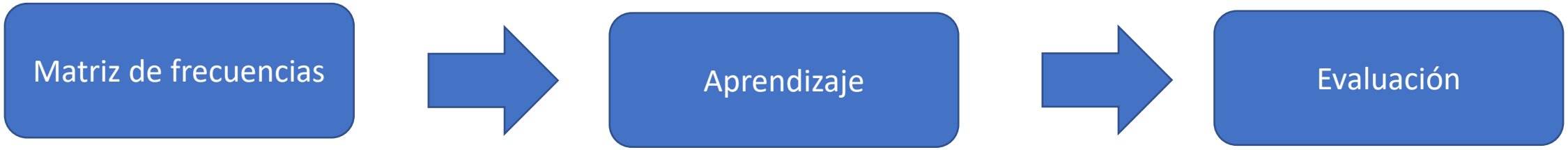
¿Cuál es la utilidad de los k-mers?

¿Qué tipo de problemas puede resolver?

METODOLOGÍA

¿Cómo se trabaja con los k-mers?

METODOLOGÍA



MATRIZ DE FRECUENCIAS

- El primer paso consistirá en rellenar la matriz de frecuencias con el número de apariciones de cada k-mer.

Secuencia : ATGAACACTA

K = 1			
A	T	G	C
5	2	1	2

MATRIZ DE FRECUENCIAS

- La matriz contendrá k dimensiones.

Secuencia : ATGAACACTA

K = 2				
	A	T	G	C
A	1	1	0	2
T	1	0	1	0
G	1	0	0	0
C	1	1	0	0

MATRIZ DE FRECUENCIAS

- Normalización de la matriz.
- Todos los componentes de la matriz deben sumar 1.

Secuencia : ATGAACACTA

K = 2				
	A	T	G	C
A	0.1	0.1	0	0.2
T	0.1	0	0.1	0
G	0.1	0	0	0
C	0.1	0.1	0	0

MATRIZ DE FRECUENCIAS

- El coste computacional de la generación de la matriz es lineal. El número de k-mers que hay que extraer en una secuencia de tamaño t es k+1-t.

- El espacio de la matriz crece exponencialmente. Sin embargo, se puede utilizar otra estructura. Ej. Estructura en árbol, etc.

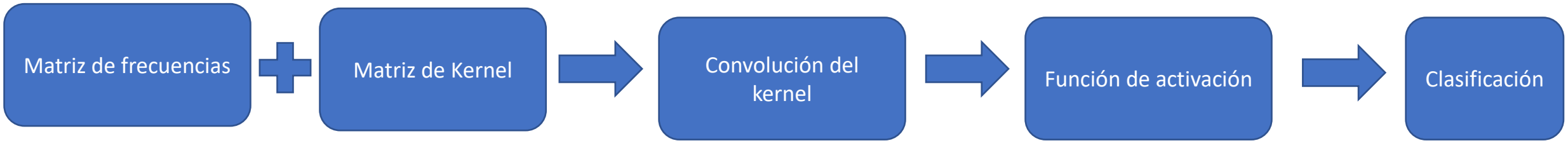
APRENDIZAJE

APRENDIZAJE

- Aplicación de un algoritmo de aprendizaje automático:
  - Regresión logística
  - Máquina de soporte vectorial (SVM)
  - Árboles de decisión
  - Bosques aleatorios
  - Redes neuronales
- Independientemente del tipo de algoritmo, se utiliza la matriz de frecuencias como datos de entrada para hacer aprender al sistema.

APRENDIZAJE

- Redes neuronales convolucionales:



APRENDIZAJE

- Para mejorar el resultado del sistema se debe aplicar validación cruzada.
- Cualquier algoritmo de clasificación que se desarrollase podría adaptarse para hacer aprender al sistema.

EVALUACIÓN

¿Cómo valoramos el sistema?

EVALUACIÓN

- Valorar el sistema utilizando secuencias que no han sido utilizadas para entrenar (datos de test).

EVALUACIÓN

- Precisión: Mide el porcentaje de verdaderos positivos
- Recall: Mide el porcentaje de positivos que se han devuelto
- F1score: Valor ponderado de la precisión y el recall
- Curva de ROC: Valoración del sistema con distintos umbrales de clasificación
- Curva de AUC: Medición agregada del rendimiento en todos los umbrales de clasificación posibles.

Posibles mejoras del algoritmo: Uso de la herramienta GKM

POSIBLES MEJORAS

Problemas:

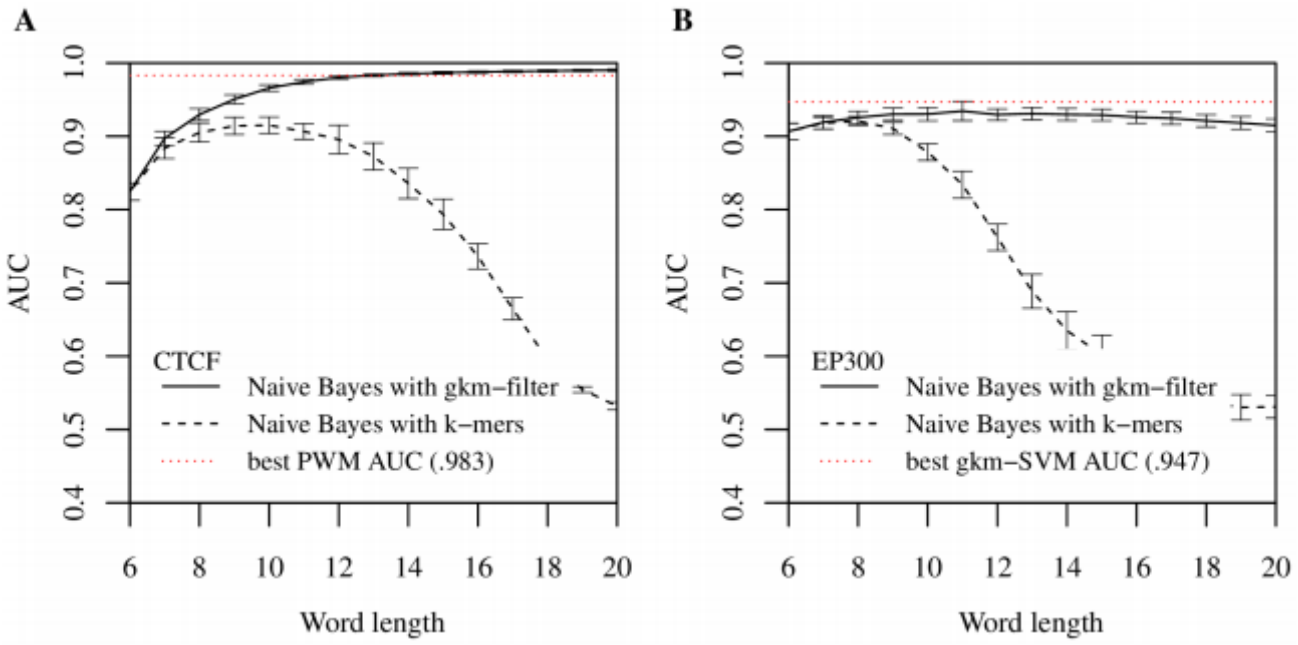
- Escoger tamaño de k.
- Frecuencia muy baja de aparición de k-mers cuando la k toma un valor alto.

Solución:

- No buscar coincidencias exactas.

HERRAMIENTA GKM

- Ambos tipos de clasificadores obtienen una eficacia similar en tamaños de k pequeños.
- A la hora de aumentar el tamaño de k la eficacia del sistema de k-mers con gaps aumenta mientras que la eficacia del sistema convencional decae.



Caso práctico: Un modelo de clasificación de lncRNA y mRNA basado en k-mers

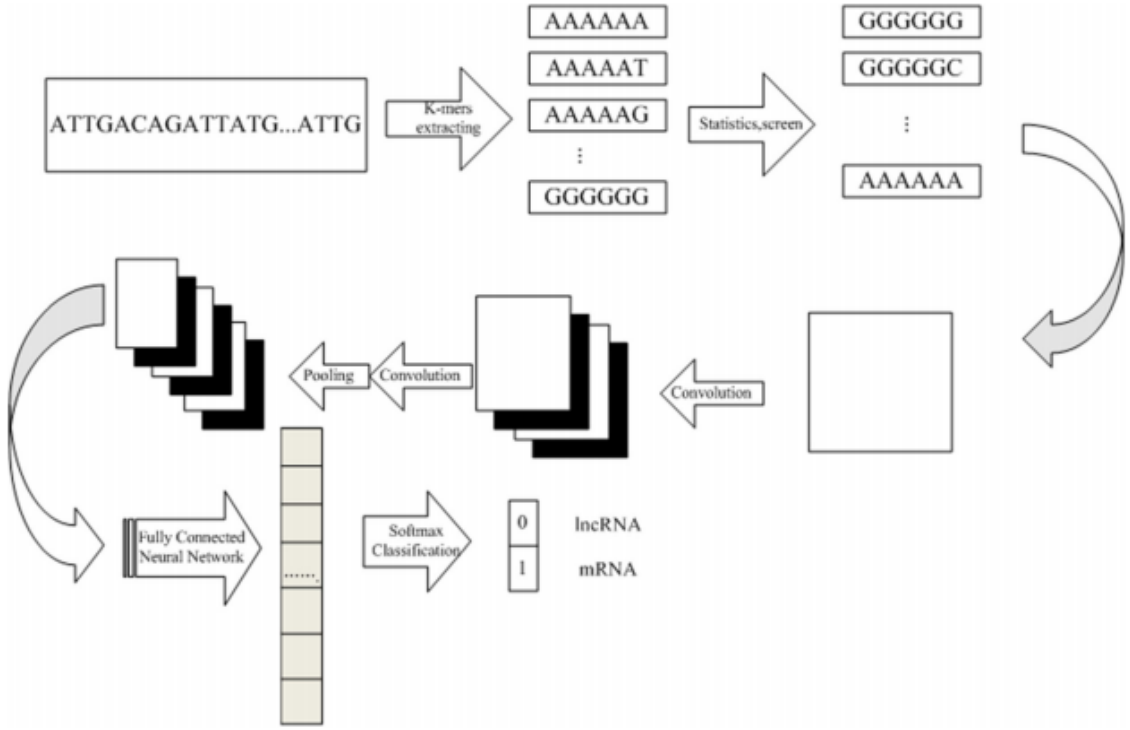
CASO PRÁCTICO

- Diferencias entre lncRNA y mRNA
  - lncRNA: de procesos, y también juegan un importante papel en la regulación de la expresión de los genes en tres niveles como es la epigenética, transcripcional y post-transcripcional.
  - mRNA: ne la importante función de determinar cómo se unirán los aminoácidos de una proteína y actúa de una manera similar a una plantilla.

CASO PRÁCTICO

Metodología

- Metodología:



CASO PRÁCTICO

Múltiples k

- Utilizando una única k la mayor exactitud del sistema obtenida es 0.76 para k = 4.
- Para mejorar la exactitud del sistema se ha probado entrenar el sistema combinando varios valores de k.

*k*value	number of k-mers	matrix form	model accuracy	precision rate	recall rate	F_1 score	calculating time (s/epoch)
1 + 2 + 3	84	17 × 20	0.9872	1.00	1.00	1.00	6
2 + 3 + 4	336	12 × 28	0.9738	1.00	1.00	1.00	57
2 + 3 + 5	1104	24 × 46	0.9798	1.00	1.00	1.00	217

- Los resultados obtenidos son muy superiores.

CASO PRÁCTICO

Comparación

- Comparación de modelos:

model	model accuracy	precision rate(P)	recall rate(R)	F_1score
CNN	0.9872	0.9993	0.9955	0.9974
RF	0.8820	0.8949	0.8867	0.8925
LR	0.7020	0.7247	0.7183	0.7218
DT	0.8030	0.7873	0.7852	0.7869
SVM	0.7020	0.7245	0.7158	0.7179

- Las redes neuronales convolucionales (CNN) obtienen los mejores resultados.

CONCLUSIONES

¿Qué nos aporta este método?

CONCLUSIONES

- Diseño e implementación sencilla: El esquema del algoritmo es sencillo de entender y permite la realización de variaciones para mejorar los resultados.
- Resultados: La evaluación de los resultados obtenidos por este método suele ser muy positivo llegando a ser en ocasiones casi perfecto como en el ejemplo mostrado.
- Coste computacional asequible: La generación de la matriz de frecuencias se puede realizar en tiempo lineal.

BIBLIOGRAFÍA

- Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, *10*(7).
- Wen, J., Liu, Y., Shi, Y., Huang, H., Deng, B., & Xiao, X. (2019). A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. *BMC bioinformatics*, *20*(1), 469.