Universidad de las Fuerzas Armadas “ESPE”

Departamento de ciencias de la computación

MODELOS DISC PARA ING ITIN

Ingeniería en tecnologías de la información

Algoritmos de análisis de secuencias

INTEGRANTES:

Gahona Patiño Jordan

Amagua Castañeda Christian

Altamirano Santacruz Mateo Marcelo

Tutor: Washington Eduardo Loza Herrera

NRC: 9901

SANGOLQUÍ, 04 DE JULIO DE 2022

Introducción:
El algoritmo de análisis de secuencias compara dos secuencias para identificar las similitudes y diferencias entre ellas. Uno de los métodos más utilizados para la alineación de secuencias es el algoritmo de alineación de Needleman-Wunsch.

El algoritmo de Needleman-Wunsch es un algoritmo de programación dinámica que busca la mejor alineación global entre dos secuencias. La alineación global busca alinear cada elemento de una secuencia con otro elemento de la otra secuencia, permitiendo la introducción de gaps (espacios) en las secuencias si es necesario.

Esta clase de algoritmos tiene muchas funcionalidades en todos los campos, las principales son:

- Comparación de secuencias biológicas: El algoritmo de secuencias es ampliamente utilizado en bioinformática para comparar secuencias de ADN, ARN y proteínas.
- Anotación de genes y proteínas: El algoritmo se utiliza para identificar regiones codificantes de genes en secuencias de ADN y para predecir la estructura y función de proteínas.
- Búsqueda de similitudes en bases de datos: El algoritmo se utiliza en bases de datos biológicas para buscar secuencias similares a una secuencia de consulta.
- Estudios de evolución molecular: El análisis de secuencias es fundamental para estudiar la evolución molecular.
- Diseño de fármacos y terapia génica: El análisis de secuencias es crucial en el diseño de fármacos y terapia génica. Permite identificar regiones específicas en el ADN o ARN.
- Reconocimiento de patrones en datos secuenciales: El análisis de secuencias se aplica también en campos como el procesamiento de lenguaje natural y la minería de dato

Tipos

En minería de datos, los algoritmos de análisis de secuencias se utilizan para descubrir patrones y relaciones en conjuntos de datos secuenciales, como secuencias de eventos, secuencias de transacciones o secuencias de palabras. Estos algoritmos se centran en encontrar asociaciones temporales y patrones de comportamiento en las secuencias.

Tenemos los siguientes tipos de algoritmos de análisis de secuencias utilizados en minería de datos:

- **Patrones frecuentes:** Estos algoritmos encuentran los conjuntos de elementos o eventos que aparecen con mayor frecuencia en las secuencias. Uno de los algoritmos más conocidos para este propósito es el algoritmo de extracción de patrones frecuentes (FP-growth). Puede utilizarse para descubrir patrones comunes en secuencias de transacciones, como la compra frecuente de ciertos productos.

- **Reglas de asociación:** Estos algoritmos buscan relaciones de co-ocurrencia entre diferentes elementos o eventos en las secuencias. El algoritmo Apriori es ampliamente utilizado para descubrir reglas de asociación en secuencias, lo que puede ser útil para encontrar patrones de comportamiento secuencial, como la secuencia de eventos de navegación en un sitio web.
- **Modelos ocultos de Markov (HMM):** Los HMM son modelos estadísticos utilizados para modelar secuencias y capturar las dependencias entre los elementos. Estos modelos se utilizan para el reconocimiento de patrones, el análisis de texto y la predicción de secuencias. El algoritmo de Viterbi se utiliza para determinar la secuencia más probable de estados ocultos en un HMM dado un conjunto de observaciones.
- **Análisis de series temporales:** Estos algoritmos se utilizan para analizar secuencias temporales y encontrar patrones, tendencias y anomalías. El algoritmo de suavizado exponencial, el modelo ARIMA (AutoRegressive Integrated Moving Average) y el modelo de Holt-Winters son ejemplos comunes de técnicas utilizadas para analizar series temporales secuenciales.
- **Minería de secuencias basada en grafos:** En esta aproximación, las secuencias se representan como grafos, donde los nodos representan los eventos y las aristas indican la secuencia temporal. Algoritmos como el algoritmo de caminos cerrados y el algoritmo de descubrimiento de secuencias cerradas (CloSpan) se utilizan para extraer patrones secuenciales significativos de estos grafos.

Estructura

Se presenta una descripción de la estructura básica de cada uno de los algoritmos de análisis de secuencias:

- **Patrones frecuentes:**

Entrada: Conjunto de secuencias o transacciones.

Paso 1: Construcción de la estructura de datos (por ejemplo, un árbol FP-growth) que permite el descubrimiento eficiente de patrones frecuentes.

Paso 2: Exploración de la estructura de datos para encontrar los patrones frecuentes en las secuencias.

Salida: Conjunto de patrones frecuentes junto con su frecuencia de aparición.

- **Reglas de asociación:**

Entrada: Conjunto de secuencias o transacciones.

Paso 1: Cálculo de los conjuntos de elementos frecuentes en las secuencias (utilizando el algoritmo Apriori o técnicas similares).

Paso 2: Generación de reglas de asociación basadas en los conjuntos de elementos frecuentes.

Salida: Conjunto de reglas de asociación, que consisten en un antecedente y un consecuente, junto con medidas de soporte y confianza.

- **Modelos ocultos de Markov (HMM):**

Entrada: Conjunto de secuencias de observaciones y el número de estados ocultos del HMM.

Paso 1: Inicialización de los parámetros del modelo HMM, como las probabilidades de transición y las probabilidades de emisión.

Paso 2: Estimación de los parámetros del modelo utilizando algoritmos de aprendizaje, como el algoritmo de Baum-Welch.

Paso 3: Utilización del algoritmo de Viterbi para encontrar la secuencia de estados ocultos más probable dado el conjunto de observaciones.

Salida: Secuencia de estados ocultos más probable y los parámetros aprendidos del modelo HMM.

- **Análisis de series temporales:**

Entrada: Conjunto de secuencias temporales.

Paso 1: Visualización y análisis exploratorio de las series temporales para identificar patrones y tendencias.

Paso 2: Modelado de las series temporales utilizando técnicas como suavizado exponencial, ARIMA o Holt-Winters.

Paso 3: Evaluación del modelo y predicción de futuras secuencias temporales.

Salida: Modelos ajustados y predicciones de las series temporales.

- **Minería de secuencias basada en grafos:**

Entrada: Conjunto de secuencias o transacciones.

Paso 1: Construcción de un grafo donde los nodos representan los eventos y las aristas indican las secuencias temporales.

Paso 2: Extracción de patrones secuenciales significativos utilizando algoritmos como el algoritmo de caminos cerrados o el algoritmo CloSpan.

Salida: Conjunto de patrones secuenciales significativos y su frecuencia de aparición en el grafo.

Ejemplo

Un ejemplo de algoritmo de análisis de secuencias es el algoritmo BLAST (Basic Local Alignment Search Tool), que se utiliza ampliamente para buscar similitudes entre secuencias biológicas en bases de datos genómicas.

El algoritmo BLAST utiliza un enfoque heurístico para encontrar alineamientos locales entre una secuencia de consulta y una base de datos de referencia. Aquí se presenta un ejemplo simplificado de cómo funciona el algoritmo BLAST:

1. Preprocesamiento de la base de datos: Antes de realizar la búsqueda, la base de datos de secuencias se prepara y se indexa para acelerar el proceso de búsqueda. Esto implica dividir la base de datos en fragmentos más pequeños y generar tablas de índice que contienen información sobre las secuencias y sus ubicaciones.

2. Preparación de la secuencia de consulta: La secuencia de consulta se procesa para identificar regiones o patrones clave que se utilizarán para buscar similitudes en la base de datos. Estos patrones se conocen como "semillas" o "palabras" y generalmente son secuencias cortas y conservadas.

3. Búsqueda de semillas: El algoritmo BLAST busca las semillas de la secuencia de consulta en la base de datos indexada. Utiliza una estrategia de búsqueda eficiente que permite identificar rápidamente las regiones de la base de datos que contienen semillas similares.

4. Expansión de alineamientos: Una vez que se encuentran las semillas, se realiza una expansión para extender los alineamientos locales en ambas direcciones de las semillas. Esto implica comparar las secuencias cercanas a las semillas y asignar puntajes a las coincidencias y a las diferencias.

5. Cálculo de puntajes y estadísticas: Finalmente, se calculan puntajes de similitud y se generan estadísticas para evaluar la significancia de los alineamientos encontrados. Esto se realiza comparando los puntajes observados con los puntajes esperados al azar.

El algoritmo BLAST se ha optimizado y mejorado en diversas variantes a lo largo de los años para aumentar su velocidad y precisión. Se utiliza ampliamente en estudios genómicos y de secuenciación para identificar secuencias similares, anotar genes y realizar análisis comparativos entre diferentes organismos.

Ejemplo en python:

```
from Bio.Blast import NCBIWWW

from Bio.Blast import NCBIXML

# Secuencia de consulta

sequence = "ATCGATCGATCG"

# Realizar la búsqueda BLAST en la base de datos "nt" de NCBI

result_handle = NCBIWWW.qblast("blastn", "nt", sequence)

# Analizar los resultados

blast_records = NCBIXML.parse(result_handle)

for record in blast_records:

    for alignment in record.alignments:

        print("Secuencia relacionada: ", alignment.title)

        for hsp in alignment.hsps:

            print("Score:", hsp.score)

            print("E-value:", hsp.expect)

            print("Secuencia alineada:", hsp.query)

            print("Secuencia alineada:", hsp.match)

            print("Secuencia alineada:", hsp.sbjct)
```

Conclusiones

Los algoritmos de análisis de secuencias son herramientas esenciales en el campo de la bioinformática y la genómica. Permiten realizar tareas fundamentales como el alineamiento

de secuencias, la búsqueda de similitudes y la predicción de estructuras secundarias. Los algoritmos mencionados en este informe son solo una muestra de la amplia gama de algoritmos disponibles, y su elección depende de la tarea específica y las características de las secuencias analizadas.

Recomendaciones

- Seleccione el algoritmo adecuado según sus datos y objetivos: Existen varios algoritmos de análisis de secuencias, y cada uno tiene sus fortalezas y limitaciones. Es importante comprender los requisitos y características específicas de sus datos, así como los objetivos del análisis, para elegir el algoritmo más adecuado.Considera el rendimiento computacional.
- Considere factores como el tipo de secuencias, la longitud de las secuencias, la estructura temporal y los patrones que espera descubrir. Investigue y evalúe diferentes algoritmos antes de decidir cuál utilizar, ya que esto puede tener un impacto significativo en los resultados y la eficiencia del análisis.

Referencias Bibliográficas

EMBnet Colombia - Centro de Bioinform�tica del Instituto de Biotecnolog�a. (s/f). Edu.co. Recuperado el 4 de julio de 2023, de http://bioinf.ibun.unal.edu.co/documentos/algoritmos/algor.php

Cruz, I. B., Martínez, S. S., Abed, A. R., Ábalo, R. G., & Lorenzo, M. M. G. (2007). Redes neuronales recurrentes para el análisis de secuencias. Revista Cubana de Ciencias Informáticas, 1(4), 48-57.

Minewiskan. (s/f). *Algoritmos de minería de datos (Analysis Services - Minería de datos)*. Microsoft.com. Recuperado el 4 de julio de 2023, de https://learn.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions

York, A. M. B. (2019, diciembre 20). *Algoritmo para el análisis de secuencias - Data scientist: Minería de datos esencial*.