UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAPEO DE SECUENCIAS DE ADN A ÁRBOLES FILOGENÉTICOS USANDO COMPRESIÓN LEMPEL-ZIV

PROPUESTA DE TEMA DE MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

Edgar Morales Gonzalez

MODALIDAD:
Memoria

PROFESOR GUÍA:
Gonzalo Navarro

SANTIAGO DE CHILE
2025

1. Introducción

El estudio de la diversidad genética y la identificación de organismos a partir de sus secuencias de ADN son pilares de la biología moderna. Con el avance de las tecnologías de secuenciación de nueva generación (NGS), el volumen y la variedad de los datos genómicos han crecido de forma sostenida, desplazando el foco desde la mera generación de datos hacia su análisis riguroso con herramientas computacionales. En este contexto, la *clasificación taxonómica* de lecturas —asignar una secuencia a su grupo biológico de origen— resulta esencial para la epidemiología, la metagenómica y el estudio de la biodiversidad. Conceptualmente, la tarea puede verse como ubicar una nueva secuencia dentro de un árbol filogenético en el que cada hoja representa un genoma conocido; desde la perspectiva algorítmica, equivale a identificar, para cada lectura, el subárbol más plausible que contenga su origen, bajo restricciones de precisión, tiempo de respuesta y uso de memoria.

Durante la última década, el enfoque predominante se ha basado en *k-mers* (subcadenas de longitud fija) extraídos de los genomas de referencia, ampliamente empleados en herramientas consolidadas como *Kraken* [9], *Centrifuge* [4] y *CLARK* [8]. Estos métodos habilitan consultas rápidas al indexar de manera exhaustiva subcadenas de los genomas, pero la elección de un único parámetro k introduce compromisos difíciles: valores pequeños incrementan falsos positivos por coincidencias espurias, mientras que valores grandes reducen la sensibilidad frente a lecturas cortas o con errores de secuenciación. Además, el crecimiento sostenido de los repositorios ha tensionado los costos de construcción y almacenamiento de índices.

En años recientes, la investigación ha explorado alternativas más adaptativas, entre ellas los *Maximal Exact Matches* (MEMs), que permiten coincidencias de longitud variable ajustadas al contenido de cada lectura. Este cambio de enfoque busca mejorar la calidad de la evidencia empleada para la asignación taxonómica y reducir la dependencia de parámetros fijos, apoyándose en representaciones comprimidas que posibilitan búsquedas eficientes. No obstante, persisten desafíos relevantes: representar colecciones genómicas de gran tamaño de manera compacta, realizar consultas informativas con bajo costo computacional y mantener un equilibrio claro entre exactitud biológica, tiempo de respuesta y memoria utilizada.

En esta memoria se desarrollará y evaluará un enfoque que combina evidencia de coincidencias exactas adaptativas con el soporte de índices comprimidos modernos; se intentará determinar su impacto en eficiencia espacio–tiempo y en precisión de clasificación mediante experimentación controlada; y se probará su desempeño frente a líneas base establecidas basadas en MEMs.

2. Situación Actual

Históricamente, el enfoque de *k-mers* (subcadenas de longitud fija k en una secuencia de ADN) ha dominado la clasificación taxonómica, y sustenta herramientas ampliamente utilizadas como *Kraken* [9] y derivados, así como *Centrifuge* [4] y *CLARK* [8]. Estos métodos indexan todas las subcadenas de longitud fija presentes en los genomas de referencia, lo que permite consultas rápidas; sin embargo, introducen limitaciones estructurales. En particular, la necesidad de escoger un único valor de k conlleva un compromiso conocido: valores pequeños incrementan la tasa de falsos positivos por coincidencias espurias, mientras que valores grandes reducen la sensibilidad frente a lecturas cortas o con errores de secuenciación. Con el crecimiento sostenido de las bases de datos genómicas, estas tensiones en precisión, memoria y tiempos de construcción del índice se han vuelto más críticas.

En este contexto, una línea de investigación reciente ha propuesto reemplazar los *k-mers* por *Maximal Exact Matches* (MEMs) [1]. A diferencia de los *k-mers*, los MEMs se ajustan dinámicamente al contenido de cada lectura, extendiéndose hasta el máximo posible sin perder coincidencia con los genomas de referencia. De este modo se obtiene evidencia más robusta y menos dependiente de parámetros externos. El trabajo citado plantea un método de clasificación taxonómica basado en MEMs implementado sobre índices comprimidos derivados de la *Burrows-Wheeler Transform* (BWT); a partir de los MEMs encontrados, se asigna cada lectura a un nodo del árbol filogenético mediante el cálculo del *Lowest Common Ancestor* (LCA) de las ocurrencias detectadas.

De forma complementaria, en el área de estructuras de datos se conocen representaciones comprimidas y mecanismos de consulta pertinentes para este problema. Entre ellas destacan:

- **FM-index** [2]. Índice basado en la *Burrows-Wheeler Transform* (BWT) que permite conteo y localización directamente sobre el texto comprimido, con soporte eficiente para operaciones *rank/select* y una huella espacial menor que índices clásicos.
- **RLFM / r-index** [6]. Variante del FM-index que *factoriza la BWT en corridas*, almacenando símbolos y longitudes. Reduce significativamente el espacio cuando la BWT presenta alta repetición, manteniendo tiempos de búsqueda competitivos y operaciones rank/select eficientes.
- **LZ-index (LZ77)** [5]. Estructura que explota redundancia global mediante un parsing en frases; diferencia ocurrencias primarias y secundarias para localizar coincidencias sin expandir el texto original, capturando repeticiones largas a nivel de colección.
- **RMQ sucinto** [7]. Estructuras de *Range Minimum Query* que representan el árbol cartesiano del arreglo y responden mínimos en $O(1)$ con sobrecosto cercano a $2n$ bits, útiles para priorizar posiciones y evitar enumeraciones exhaustivas en arreglos de apoyo.
- **Wavelet Trees** [3]. Representación jerárquica de secuencias que soporta *access/rank/-select* y consultas de rango en $O(\log |\Sigma|)$; funciona como columna vertebral en diversos índices comprimidos y se extiende naturalmente a escenarios 1D/2D.

3. Objetivos

Objetivo General

Desarrollar y evaluar un método de clasificación taxonómica de secuencias de ADN basado en estructuras de datos comprimidas, que combine índices por corridas y compresión Lempel–Ziv con estructuras auxiliares de rango, con el fin de mejorar la eficiencia en espacio y tiempo respecto a las soluciones actuales basadas en *k-mers* y MEMs.

Objetivos Específicos

1. Adaptar el módulo de obtención de *Maximal Exact Matches*, para que opere sobre un índice BWT, preservando la exactitud de las coincidencias y compatibilizando su salida con el resto de estructuras del proceso.
2. Diseñar e implementar un procedimiento que, para cada MEM detectado, determine su intervalo en la BWT del texto y lo indexe en el eje Y del LZ-index, y que realice el mismo mapeo sobre el texto reverso para indexar los prefijos en el eje X, manteniendo la correspondencia entre los rangos de ambos ejes y las posiciones originales en el texto.
3. Diseñar e implementar una grilla bidimensional comprimida,basada en *Wavelet Tree* que permita responder a consultas rectangulares $[X_1..X_2] \times [Y_1..Y_2]$ con operaciones de conteo y listado (*range reporting*) en costo logaritmico y mantenga cotas de memoria comprimida.
4. Integrar estructuas de *RMQ* asociadas a niveles del *Wavelet Tree*, para idenificar eficientemente la posicion mínima en el texto dentro de cada subrango candidato, optimizando el acceso a las ocurrencias durante la búsqueda.
5. Replicar el flujo de búsqueda sobre el texto reverso para obtener las ocurrencias más a la derecha de cada *MEM*, asegurando simetría y completitud en la detección de coincidencias.
6. Evaluar experimentalmente el desempeño del método sobre colecciones genómicas repetitivas, comparando el tiempo de procesamiento, uso de memoria y precisión de clasificación con respecto a soluciones basadas en *k-mers* y *MEMs*.
7. Validar la implementación mediante un conjunto estructurado de pruebas unitarias e integrales, que cubran casos borde y escenarios representativos, garantizando la corrección funcional y la reproducibilidad de los resultados.

Evaluación

La evaluación se realizará de forma experimental sobre colecciones genómicas de referencia y conjuntos de lecturas reales y simuladas. El desempeño se analizará en tres dimensiones: (i) **tiempo de consulta por lectura**, estimado a partir de corridas repetidas; (ii) **uso de memoria** del índice y de sus estructuras asociadas, incluyendo el *peak* residente durante las consultas; y (iii) **comparación de soluciones** frente a métodos de referencia basados en *k-mers* y *MEMs*, considerando exclusivamente **recursos computacionales y tiempos de ejecución**, bajo un protocolo común (mismos datasets, parámetros y entorno). Este procedimiento busca establecer con claridad si la propuesta mejora el compromiso **espacio–tiempo** respecto de las alternativas, **sin** evaluar métricas de calidad de clasificación.

4. Solución Propuesta

La solución propuesta busca combinar las ventajas de los MEMs con estructuras comprimidas basadas en Lempel–Ziv y consultas de rango sobre grillas. En términos generales, el procedimiento será el siguiente:

A partir de una lectura, se obtendrán sus *Maximal Exact Matches* (MEMs) utilizando $ropeBWT2$. Cada MEM será particionado en todos los puntos posibles de prefijo y sufijo. Para cada sufijo, se realizará una búsqueda en la BWT del texto, lo que produce un rango lexicográfico. Este rango será intersectado con un *bitmap* que identifica qué posiciones corresponden a inicios de frases en el LZ-index, mapeándolo al eje Y de una grilla. Simétricamente, para cada prefijo reverso del MEM, se realizará una búsqueda en la BWT del texto reverso, obteniendo un rango lexicográfico de sufijos. Dicho rango se intersecta con un *bitmap* que marca qué posiciones corresponden a finales de frases, mapeándolo al eje X de la grilla.

Con estos rangos en X e Y, el problema se traduce en una consulta sobre una grilla bidimensional. Esta grilla estará implementada mediante un *Wavelet Tree*, que permite responder consultas de rango 2D. A diferencia del LZ-index tradicional, que enumera todas las ocurrencias en la grilla, aquí se integrarán estructuras de *Range Minimum Query* (RMQ) de $2n$ bits en cada nivel del Wavelet Tree. De este modo, cada consulta sobre un rango bidimensional se resolverá en $O(\log n)$ rangos candidatos; en cada uno se ejecutará un RMQ que devolverá la posición mínima del texto en dicho rango. Entre todos los candidatos se seleccionará el mínimo global, correspondiente a la ocurrencia más a la izquierda del MEM en ese corte.

El mismo procedimiento se repetirá sobre una grilla construida para el texto reverso, permitiendo obtener también la ocurrencia más a la derecha. De esta forma, resuelve de manera eficiente las posiciones extremales de cada MEM, integrando BWT, LZ-index, Wavelet Trees y RMQs en una arquitectura híbrida. El objetivo es reducir el espacio ocupado respecto al LZ-index clásico, al evitar la enumeración completa de ocurrencias, y mejorar el tiempo de respuesta para la clasificación taxonómica sobre colecciones repetitivas de secuencias genómicas.

5. Plan de Trabajo (Preliminar)

El trabajo se desarrollará en fases progresivas a lo largo de dos periodos académicos: una etapa inicial de preparación durante las ultimas semanas del semestre de primavera y el desarrollo completo durante el semestre de otoño siguiente.

1. **Revisión y ganancia de contexto:** Profundizar en estructuras de datos comprimidas relevantes para la clasificación taxonómica, incluyendo BWT, LZ-index, Wavelet Trees y RMQ. Revisar literatura reciente sobre métodos basados en *MEMs* y definir las métricas y colecciones de evaluación.
2. **Implementación del módulo de MEMs:** Adaptar un módulo de obtención de *Maximal Exact Matches* para operar sobre un índice BWT y generar salidas compatibles con las estructuras de mapeo posteriores.
3. **Mapeo de rangos y construcción del índice:** Implementar el procedimiento de mapeo lexicográfico de cada MEM en los ejes X e Y del LZ-index y construir la grilla bidimensional comprimida basada en *Wavelet Tree*.
4. **Integración de RMQ y simetría:** Incorporar estructuras *Range Minimum Query* para optimizar la localización de ocurrencias mínimas y replicar el flujo de búsqueda sobre el texto reverso, garantizando simetría en la detección.
5. **Evaluación experimental:** Ejecutar pruebas sobre colecciones genómicas repetitivas, midiendo tiempo de consulta, uso de memoria y comparando los resultados con métodos de referencia basados en *k-mers* y *MEMs*.
6. **Validación y redacción final:** Realizar pruebas unitarias e integrales, analizar los resultados experimentales y redactar la memoria final con la discusión y conclusiones del trabajo.

Referencias

[1] Draesslerová, Dominika, Omar Ahmed, Travis Gagie, Jan Holub, Ben Langmead, Giovanni Manzini y Gonzalo Navarro: *Taxonomic Classification with Maximal Exact Matches in KATKA Kernels and Minimizer Digests*. En Liberti, Leo (editor): *22nd International Symposium on Experimental Algorithms (SEA 2024)*, volumen 301 de *Leibniz International Proceedings in Informatics (LIPIcs)*, páginas 10:1–10:13, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, ISBN 978-3-95977-325-6. `https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SEA.2024.10`.

[2] Ferragina, P. y G. Manzini: *Opportunistic data structures with applications*. En *Proceedings 41st Annual Symposium on Foundations of Computer Science*, páginas 390–398, 2000.

[3] Grossi, Roberto, Ankur Gupta y Jeffrey Scott Vitter: *High-order entropy-compressed text indexes*. En *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, página 841–850, USA, 2003. Society for Industrial and Applied Mathematics, ISBN 0898715385.

[4] Kim, Daehwan, Li Song, Florian P Breitwieser y Steven L Salzberg: *Centrifuge: rapid and sensitive classification of metagenomic sequences*. Genome research, 26(12):1721–1729, 2016.

[5] Kreft, Sebastian y Gonzalo Navarro: *On compressing and indexing repetitive sequences*. Theoretical Computer Science, 483:115–133, 2013, ISSN 0304-3975. `https://www.sciencedirect.com/science/article/pii/S0304397512001259`, Special Issue Combinatorial Pattern Matching 2011.

[6] Mäkinen, Veli y Gonzalo Navarro: *Succinct suffix arrays based on run-length encoding*. Nordic Journal of Computing, 12(1):44–66, 2005, ISSN 1236-6064.

[7] Navarro, Gonzalo: *Wavelet trees for all*. Journal of Discrete Algorithms, 25:2–20, 2014, ISSN 1570-8667. `https://www.sciencedirect.com/science/article/pii/S1570866713000610`, 23rd Annual Symposium on Combinatorial Pattern Matching.

[8] Ounit, Rachid, Steve Wanamaker, Timothy J Close y Stefano Lonardi: *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers*. BMC genomics, 16(1):236, 2015.

[9] Wood, Derrick E y Steven L Salzberg: *Kraken: ultrafast metagenomic sequence classification using exact alignments*. Genome biology, 15(3):R46, 2014.