

FlinkBWA

Uso de tecnologías Big Data para el alineamiento de secuencias genéticas

Grado en Ingeniería Informática
Universidad de Santiago de Compostela

Autora: Silvia Rodríguez Alcaraz

Tutor: Juan C. Pichel Campos
Cotutor: José M. Abuín Mosquera

19 de julio de 2019

Tabla de contenidos

1 Introducción

- Alineadores de secuencias genéticas
- Burrows-Wheeler Aligner
- Apache Flink
- Objetivos

2 Gestión del proyecto

- Metodología
- Gestión del tiempo

3 Diseño

- Diagrama de clases
- Diagramas de secuencia

4 Implementación

- Requisitos y medios utilizados
- Funcionamiento
- Caso especial

5 Pruebas

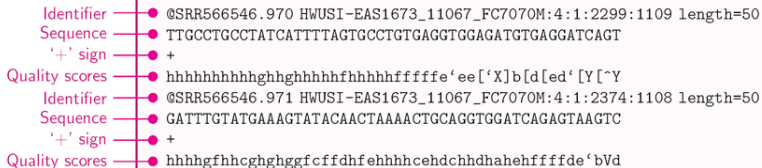
- Especificación de las pruebas
- Resultados

6 Conclusiones

7 Trabajo futuro

Burrows-Wheeler Aligner (BWA)

Burrows-Wheeler Aligner (BWA)



The diagram illustrates the FASTQ format structure. It consists of two identical blocks, each representing a single sequencing read. Each block contains four lines: an identifier line, a sequence line, a '+' sign line, and a quality scores line. The identifier line includes a sample ID, a platform, a run ID, and alignment coordinates. The sequence line contains the raw sequencing data (A, C, G, T). The '+' sign line is a simple separator. The quality scores line contains a series of characters representing the confidence of each base call.

```
Identifier ● @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence ● TTGCCTGCCTATCATTTTAGTGCGCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign ● +
Quality scores ● hhhhhhhhhghghghhhhhfhhhhhhfffffe'ee['X]b[d[ed['Y[~Y
Identifier ● @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence ● GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign ● +
Quality scores ● hhhhghfhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

Figura: Ejemplo del formato FASTQ

Apache Flink

Características

- Plataforma *Big Data* creada por la *Apache Software Foundation*.
- Aparición en el mercado en 2015.
- Permite el procesamiento de datos por lotes y en flujos (*streaming*).

Apache Flink

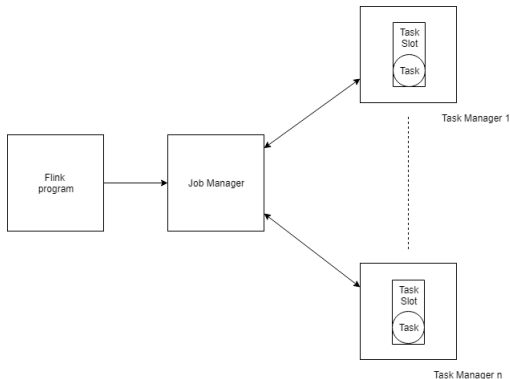


Figura: Arquitectura de *Flink*

Objetivos

- 1 Estudio del arte: tecnologías *Big Data*.
- 2 Formación en *Apache Flink* y en aplicaciones de alineamiento genético.
- 3 Diseño modular del nuevo alineador BWA paralelo.
- 4 Implementación de *FlinkBWA*.
- 5 Análisis del rendimiento de la aplicación en un *cluster*.

Metodología

Factores tenidos en cuenta para escoger la metodología

- Inexperiencia de la desarrolladora con las tecnologías.
- Aplicación orientada a un ámbito muy específico.
- Carácter innovador: la plataforma *Flink* sólo lleva 4 años en el mercado.
 - Poca documentación.
 - Escasa comunidad de usuarios.

Metodología

Alta incertidumbre \Rightarrow Metodología ágil: **Scrum**

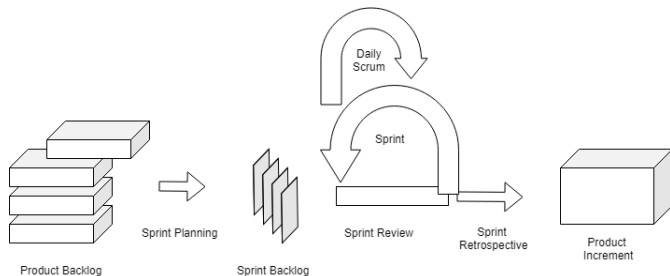


Figura: Ciclo de un *sprint*

Gestión del tiempo

Sprints

- *Sprint 1*: planificación del proyecto.
- *Sprint 2*: diseño de la aplicación y configuración del entorno de trabajo.
- *Sprint 3*: implementación de la aplicación.
- *Sprint 4*: ejecución de la aplicación en el *cluster*, corrección de errores y pruebas de rendimiento.

Gestión del tiempo

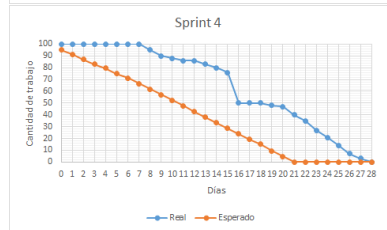
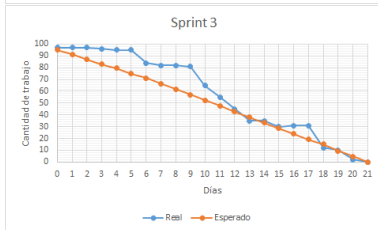
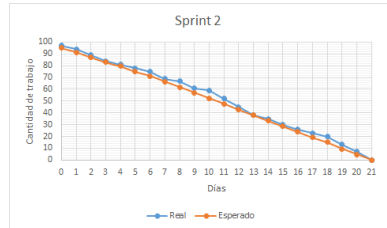
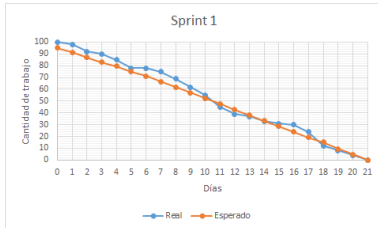
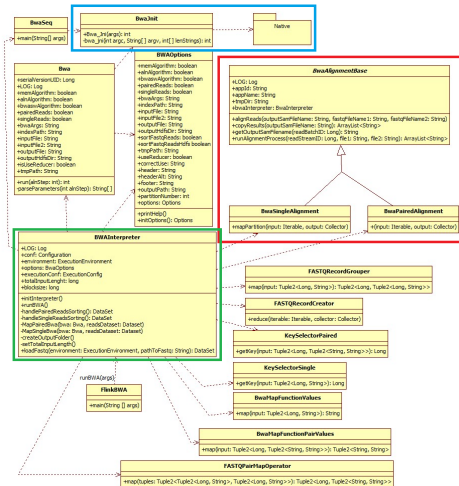


Diagrama de clases



Patrón *Facade* o Fachada.

Patrón *Template*
Method o Método
Plantilla.

“Controlador”.

Figura: Diagrama de clases

Diagramas de secuencia

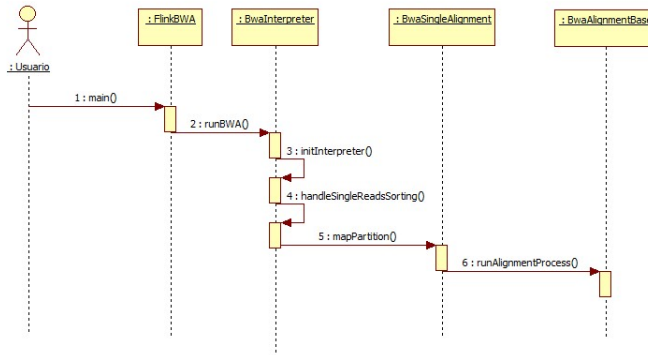


Figura: Diagrama de secuencia: *Single Reads*

Diagramas de secuencia

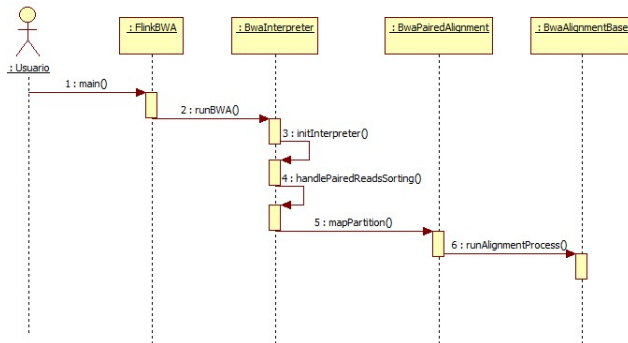


Figura: Diagrama de secuencia: *Paired Reads*

Requisitos y medios utilizados

Requisitos de Flink

- Entorno Unix
- Maven (> 3.1.1)
- Java 8.x

Herramientas de desarrollo

- IDE: IntelliJ IDEA Ultimate 2019.
- Control de versiones: Git

Entorno: *Cluster Big Data 1* del CiTIUS

- 16 Servidores Dell EMC
 - 2 x Intel Xeon E5-2630 v4 (2,2Ghz 10c)
 - 384 GB de RAM: 12 x 32GB RDIMM
 - 32 TB HDD
- *Apache Flink 1.7.2*
- Hadoop 2.7.3: sistema de ficheros HDFS.

Caso especial

Uso de 2 ficheros de entrada

- Necesidad de combinar ambos ficheros en un *Dataset*.
- Orden del contenido del *Dataset*.

2 soluciones

- SortHDFS: operación de preprocesado.
- Operaciones *Join* y *sortByKey* de *Flink*.

Especificación de las pruebas

Parámetros de *Flink*

- **Número de Task Managers:** tantos como particiones se indiquen sobre el fichero de entrada.
- **Memoria por Task Manager:** 30 GB.

Parámetros del programa

- **Algoritmo:** MEM.
- **Número de ficheros de entrada:** uno (*Single Reads*).
- **Particiones:** 4, 8, 16, 32 y 64.
- **Ordenamiento:** no.
- **Reducción:** no.

BWA secuencial

Tiempo de cómputo: 66.43 minutos.

FlinkBWA

Pruebas			
Nivel paralelismo	Tiempo (minutos)	Media (minutos)	Desviación Típica
4	16,54	17,2133	0,55530
	17,2		
	17,9		
8	10,31	10,2700	0,08641
	10,35		
	10,15		
16	7,23	7,2500	0,03559
	7,22		
	7,3		
32	5,47	5,4333	0,04497
	5,37		
	5,46		
64	5,06	4,7233	0,23977
	4,52		
	4,59		

Cuadro: Resultados *FlinkBWA*

Eficiencia temporal

- *FlinkBWA* mejora el tiempo del BWA secuencial original.

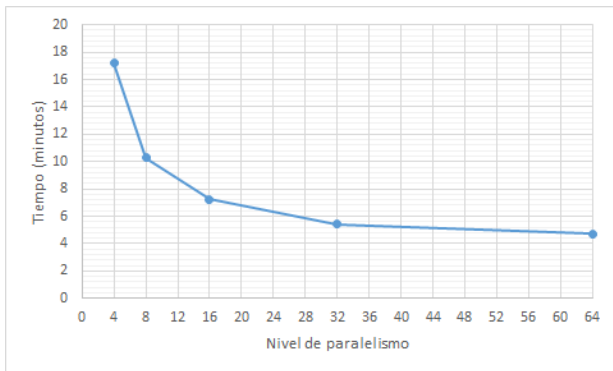


Figura: Tiempo vs. Nivel de paralelismo

Escalabilidad

- El programa es escalable.
- Caso ideal: ley de *Amdahl*

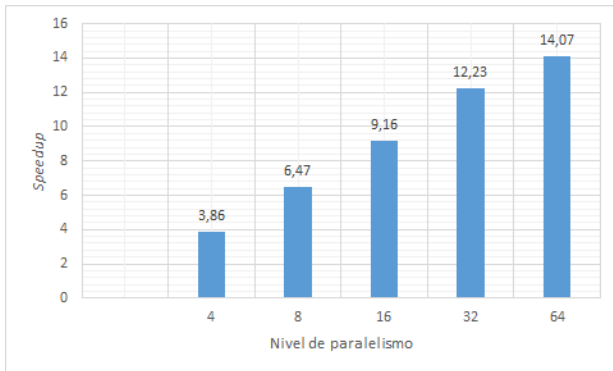


Figura: *Speedup*

Trabajo futuro

- Reducir el consumo de RAM.
- Implementar el ordenamiento del fichero SAM de salida.
- Tener en cuenta posible versión *streaming*.

FlinkBWA

Uso de tecnologías Big Data para el alineamiento de secuencias genéticas

Grado en Ingeniería Informática
Universidad de Santiago de Compostela

Autora: Silvia Rodríguez Alcaraz

Tutor: Juan C. Pichel Campos
Cotutor: José M. Abuín Mosquera

19 de julio de 2019