Samblaster 병렬화를 위한 DNA 리드 분할 방법

DNA Reads Partitioning Method for Parallelizing Samblaster

이 현 병, 송 석 일
한국교통대학교

Hyeonbyeong Lee, Seokil Song
Korea National University of Transportation

요약

NGS 분석과정에서는 정확도를 높이기 위해서 중합 효소 연쇄반응 (PCR)을 통해 DNA 리드들을 증폭한다. PCR 증폭은 DNA 리드의 중복을 발생시키며 중복으로 인해 NGS 과정의 정확성이 저하되는 문제가 발생한다. 본 논문에서는 DNA 리드 중복을 제거하는 도구중 하나인 Samblaster 의 병렬화를 위한 DNA 리드 분할 방법을 제안한다. DNA 리드를 분할하여 Samblaster를 병렬 실행 하도록 하여 수행시간을 단축 할 수 있도록 한다.

I. 서론

NGS 데이터를 생성하는 과정 중 중합 효소 연쇄 반응 (PCR, Polymerase Chain Reaction) 단계에서 DNA 리드가 중복 생성될 수 있다[1]. 중복 리드는 이후 분석 과정에서 오류나 성능저하의 원인이 될 수 있다. 이 중복 리드를 삭제하거나, 표시하는 과정이 반드시 필요하며, 이를 중복 제거라고 한다. 다음과 같이 다양한 중복 제거 기법이 중복 제거 정확도와 시간을 줄이기 위해 제안되었다.

Padre[2]는 접두부(prefix)가 동일한 리드들을 같은 그룹에 포함시키고 같은 그룹에 있는 리드들을 비트열로 변환하여 비트연산을 통해 나머지 부분의 유사도를 계산하여 중복 리드를 찾아낸다. 이 과정에서 MPI(Message Passing Interface)를 이용해서 각 그룹에 대한 병렬 처리를 수행한다. 병렬 형태로 처리가 가능하다. 다수 프로세스가 각각 중복 제거 결과 파일을 출력하고 이를 병합하여 하나의 중복 제거 결과 파일을 만든다.

Samblaster[3]는 유닉스의 파이프(Pipe)를 이용해서 이전 단계의 출력을 바로 입력으로 받아올 수 있어 이전단계 출력을 저장하고 다시 읽기 위한 디스크 IO를 줄일 수 있다. 중복제거를 위해서는 SAM 파일 형식의 CIGAR 필드를 사용하여 특정 리드의 정렬 위치를 계산하고 계산한 정렬위치를 이용해서 중복제거를 수행한다. Picard[4]는 널리 사용되는 DNA 분석 도구 중 하나이나 분산 병렬 실행을 지원하지 않고, Samblaster 와 같이 이전 단계의 입력을 받아서 점진적으로 중복제거를 하는 형태도 아니기 때문에 중복제거에 소요되는 시간이 매우 길다.

본 논문에서는 Samblaster 의 장점인 이전단계의 입력을 직접 입력으로 받는 특성을 가지며 동시에 다수 프로세스가 중복제거를 병렬로 처리할 수 있도록 Samblaster의 입력을 분할하는 방법을 제안한다.

Ⅱ. 제안하는 DNA 리드 분할 방법

본 논문에서는 Samblaster를 병렬로 실행하여 중복 제거 시간을 단축할 수 있도록 DNA 리드에 대한 파티셔닝 방법을 제안한다. 즉, 중복 제거 이전단계의 출력이 유닉스 파이프를 통해서 전달될 때 각 SAM 레코드를 주어진 수의 그룹으로 분할한다. 각 분할된 그룹에 대해 Samblaster를 병렬로 동시에 수행해서 중복제거에 소요되는 시간을 줄인다. 제안하는 분할 방법은 각기 다른 Samblaster 프로세스에 중복되는 SAM 레코드가 분산되지 않도록 설계한다. 이를 위해서는 Samblaster의 중복 제거 방법을 파악하여 중복가능성이 없는 SAM 레코드들만을 서로 다른 그룹에 분할해야 한다.

그림 1은 Samblaster의 페어드 엔드 (Paired-end) DNA 리드들에 대한 중복 제거 방법을 보여준다. 그림 1에서처럼 Qname 이 같은 페어드 엔드 리드의 첫 번째와 두 번째 리드의 위치 (FIRST_POS, SECOND_POS) 에 대해서 해시키 (Hash Key)를 생성한다. 해시키를 생성하는 과정은 다음과 같다. 먼저 각 리드의 POS 필드 값에 CIGAR 스트링을 적용하여 POS 값을 조정한 후 SAM 파일의 헤더를 참조하여 리드의 RNAME의 염색체 (Chromosome)의 누적 길이와 더한다.

CIGAR 필드는 각 DNA 리드가 참조 DNA 시퀀스에 어떻게 정렬되어 있는지를 나타내는 것으로 총 8개의 연산 (M, I, D, N, S, H, P, =, X)과 숫자가 결합되어 나타난다[5]. 첫 번째와 두 번째 리드의 POS에 대해서 CIGRAR를 적용하여 보정된 POS 값을 하위 32bit와 상위 32bit로 결합해서 최종 해시키를 생성한다. 이 해시키를

기반으로 각 리드에 대한 해시를 수행하여 중복제거를 한다.

```
first.pos += rname.chromosome.length
second.pos += rname.chromosome.length

hash_key = lower32(first.pos) + upper32(second.pos)
```

▶▶ 그림 1. Samblaster 의 해시키 생성 방법

본 논문에서 제안하는 Samblaster 병렬 수행을 위한 DNA 리드 분할 방법은 그림 2와 같다. 먼저 이전 단계에서 출력되는 SAM 레코드들을 스트림 형태로 하나씩 입력받으면서 페어드 엔드를 구성하는 첫 번째와 두 번째 SAM 레코드를 입력받을 때 까지 기다린다. 페어드 엔드의 첫 번째와 두 번째 레코드를 입력 받으면 Samblaster 와 유사하게 첫 번째와 두 번째 SAM 레코드의 RNAME에 따른 누적 염색체 길이와 POS 필드 값을 더해서 POS 값을 조정한다.

다음으로 조정된 POS 값을 상수 사용자에 의해서 주어진 상수인 K (10의 배수) 로 나누어서 K 단위 이하의 값은 제거한다. 이것은 CIGAR 필드를 적용했을 때 POS의 조정 값이 K 이하 인 것을 가정하는 것이다. 즉, K를 1000으로 했다면 POS가 2000 ~ 2999 인 SAM 레코드들은 모두 2라는 POS 값을 갖게 되어 하나의 그룹에 속하게 된다. 동일한 예에서 CIGAR 에 의해 조정되는 값이 100 이하라면 POS가 2000 ~ 2999 인 SAM 레코드들은 CIGAR 에 의해 POS 가 조정되어도 대부분 같은 그룹에 포함되게 된다. 본 논문에서 제안하는 해시키 생성 알고리즘은 그림 2와 같다.

```
paired_end = first + second
paired_end.first.pos += rname.chromosome.length

hash_key = floor(paired_end.first.pos/K)
```

▶▶ 그림 2. 제안하는 DNA 시퀀스 분할 해시키 생성 방법

Ⅲ. 성능평가

본 논문에서 제안하는 DNA 리드 분할 방법을 검증하기 위해서 다음과 같이 실험을 수행하였다. SAM 파일을 다수의 SAM 파일들로 분할하고 분할 각 파일들에 Samblaster를 적용하여 중복제거를 수행한다. 각 파일들에 중복제거를 수행한 결과와 분할 전의 SAM 파일에 대해 Samblaster를 적용한 중복 제거 결과를 비교하여 중복 제거 정도가 얼마나 유사한지 측정한다. 또한, 분할 전의 SAM 파일에 대한 Samblaster 수행시간과 각 분할된 SAM 파일들에 대해서 병렬로 Samblaster를 실행한 시간을 측정하여 수행시간이 얼마나 향상될 수 있는지 비교한다.

실험에 사용한 DNA 시퀀스 파일은 ERR000954.fastq (516M)이다. 이를 참조 시퀀스에 정렬하여 SAM 파일을 생성한다. 생성한 파일을 제안하는 분할 방법을 이용해서 총 10개의 파일로 분할한 후 각 파일에 대해서 Samblaster를 적용한다. 이 과정에서 분할 전 SAM 파일에 대한 Samblaster 적용 결과와 분할된 파일들에 Samblaster를 병렬로 적용했을 때의 시간 및 중복 제거 결과를 측정하였다.

표 1. 실험 결과

입력 파일	중복제거율 (%)	실행시간 (seconds)
분할 전 SAM 파일	0.42	24.186
분할된 10개의 SAM 파일	0.418	5.2

표 1은 실험 결과를 보여주고 있다. 표에서 보는 것처럼 중복제거율은 0.002% 의 차이가 있었으며 변이 검출 결과는 모두 동일하였다. 또한, 실행시간은 약 5배 정도 빨라졌음을 볼 수 있었다.

Ⅳ. 결론

본 논문에서는 NGS 분석과정에서 발생하는 DNA 리드 중복 제거 실행 속도를 높이기 위한 방법을 제안한다. 특히 기존 중복 제거 도구중 하나인 Samblaster를 병렬화하기 위하여 DNA 리드 파일을 분할하는 방법을 제안하였다. 또한, 제안하는 DNA 리드 분할 방법을 실험을 통해 실행 시간과 중복제거의 정확도를 검증하였다. 중복제거율의 경우 분할 전 파일과 분할 후 파일사이에 0.002%의 차이가 있었다. 분할된 파일들에 대해서 Samblaster를 병렬로 실행했을 때는 실행시간을 약 5배가량 단축시킬 수 있었다.

■ 참 고 문 헌 ■

[1] González-Domínguez, J. and Bertil S., "ParDRe: faster parallel duplicated reads removal tool for sequencing studies." Bioinformatics 32.10 (2016): 1562-1564.
[2] Faust, Gregory G., and Ira M. Hall. "SAMBLASTER: fast duplicate marking and structural variant read extraction." Bioinformatics 30.17 (2014): 2503-2505.
[3] https://broadinstitute.github.io/picard/index.html
[4] https://www.illumina.com/science/technology/next-generation-sequencing.html
[5] https://samtools.github.io/hts-specs/SAMv1.pdf