팀 프로젝트

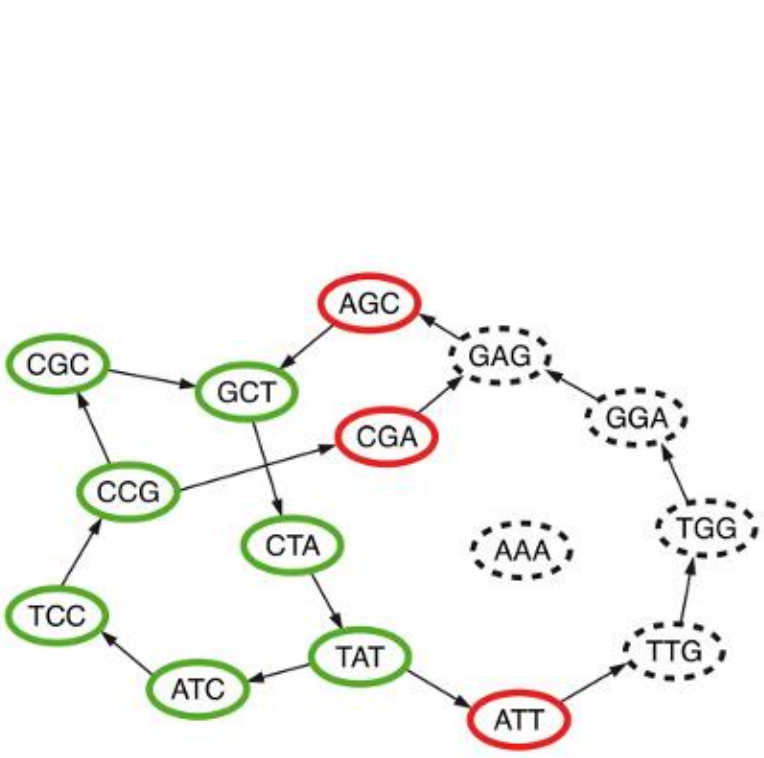
프로그램 시연 코드 및 결과 정리

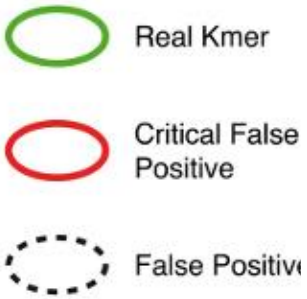
2018112053 황종익

Bloom Filter + De Bruijn Graph

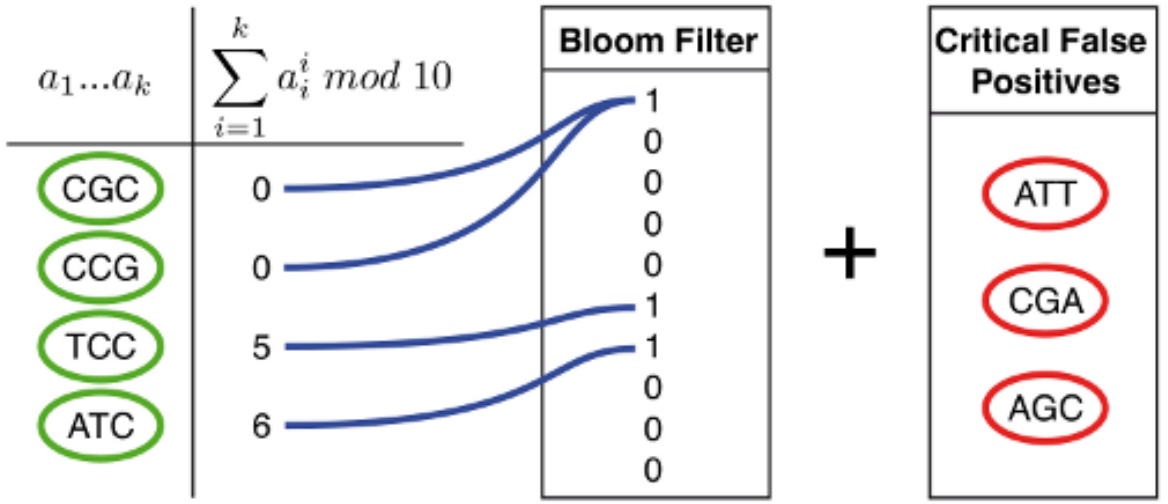
Read: AGATCGAGTG

3-mers: AGA GAT ATC TCG CGA GAG AGT GTG





해싱을 활용해
vector<bool> 배열에
비트마스킹을 적용!
이를 활용해 k-mer 존재여부 판단 가능



▪ 결과 및 성능

- 성능 평가 요소

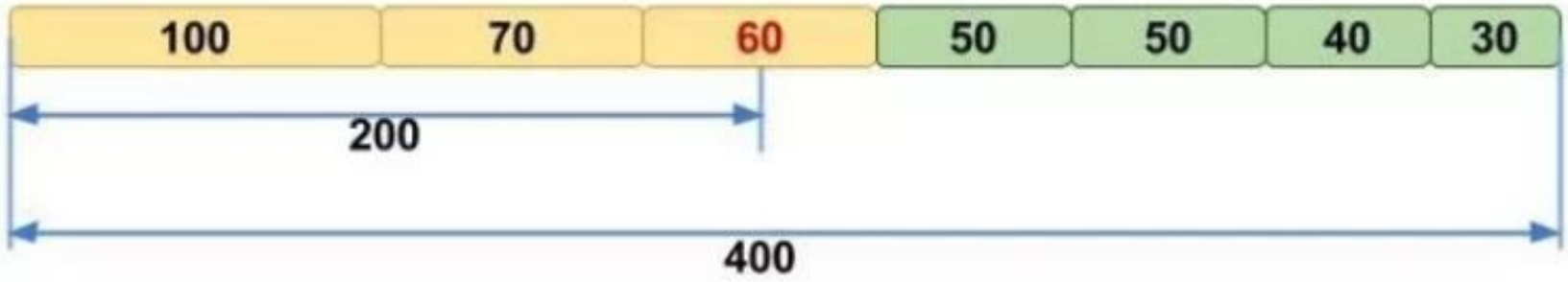
- contig에 대한 N50의 값
- myDNA와의 가장 긴 contig의 일치율
- 소모 시간

- N50이란 ?

de novo assembly의 품질을 정의할 때 사용되는 수치

- 예) N50 : 60

1a. Contigs, sorted according to their lengths.



1b. Calculation of N50 using sorted contigs.

```
kmers의 수: 154000
filterSize : 1669493 numHashFunctions : 7
블룸 필터 생성중...
거짓 긍정들을 찾아내고 있습니다...
그래프 순회 중...
initKmersSize : 1
Starting kmer: AAGTTGCGAGATGAGCGTGCATCG, 진행도 : 1/1
소모 시간 : 0.679seconds
De Bruijn 그래프 크기(바이트) : 1669493


      원본 myDNA의 길이      :20000
복구된 가장 긴 contig의 길이 :20000


다음에 나오는 아래의 정확도는 원본 길이의 1/4배정도가 복구되었을 때만 의미가 있는 값입니다.
그보다 짧다면 더 아래의 N50값을 확인해주세요.
DNA 일치율 : 100%
N50 : 20000
(유의미한 경우에 속한다.) DNA 일치율 : 100%
```

N=20000, M=2000, L=1000일 때

N=25000, M=2000, L=1000일 때

```
kmers의 수: 154000
filterSize : 1669493 numHashFunctions : 7
블룸 필터 생성중...
거짓 긍정들을 찾아내고 있습니다...
그래프 순회 중...
initKmersSize : 4
Starting kmer: GTGCCACGGTGTAGACGTGTGTCG, 진행도 : 1/4
Starting kmer: TCGTGACAGCACAGCTGCAGCACT, 진행도 : 2/4
Starting kmer: GAGACACTGTATGCTCGCGAGAGA, 진행도 : 3/4
Starting kmer: GCTGTCACGCACTGATCACGACGC, 진행도 : 4/4
소모 시간 : 0.724seconds
De Bruijn 그래프 크기(바이트) : 1669493


      원본 myDNA의 길이      :25000
복구된 가장 긴 contig의 길이 :16614


다음에 나오는 아래의 정확도는 원본 길이의 1/4배정도가 복구되었을 때만 의미가 있는 값입니다.
그보다 짧다면 더 아래의 N50값을 확인해주세요.
DNA 일치율 : 66.456%
N50 : 16614
(유의미한 경우에 속한다.) DNA 일치율 : 66.456%
```

test_data_set 결과 : 9번째(마지막) contig가 가장 김. 145의 길이.

repair_dna_1000.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
0번째 contig, 길이: 70
CTCGCCACCAAGGTGATATCCGTACAAAGCTTAGTAACGTAGCTGTGCTGTTCAGAATACAGACGTAACA
1번째 contig, 길이: 53
ACAGACGTAACATTTCAGGTCATACTGATGTTTATGCAAACACTTGCTTGTGC
2번째 contig, 길이: 50
ACAGTACGCAGGAACCAACGGGAGTCGTCGCTACAGTCGACTAAGCTTGG
3번째 contig, 길이: 113
GAAGCATGCTCCAAATTCCCGTCGTTGAGAGCTCGCCTCATGCCAGGTGAACTCTATTTGGCCACTGAAG
CATATTCTGTCCATAGCAAGCGAACGGATTACCTATTATTGCG
4번째 contig, 길이: 50
GCAATTCGACATAACGGTATGTATGAACTTTGGGTGTCGCACAGTACGCA
5번째 contig, 길이: 50
CTTGTATCGTTAACGTTCCCCCGTGGTTATACCACCCACTTTAGCTACGT
6번째 contig, 길이: 100
TCTATATATTGGAAAATTGACAATACACTCTGCTGCGTATAAGAAAGGCTGGGAGCTGAGCCTACCGTCC
GAGAGCACCAAGGTGCTGCCTCCGCCTCTC
7번째 contig, 길이: 72
TGCTCTCGTTGCATGGTCATTGCTCGCATTTGGCCAAGTGCGCCTGTTGAGGCTGCTCCGAGACTAACT
CGG
8번째 contig, 길이: 50
CCCGTTATCCCTCCTCATCTAAACGGGTGTGGAAGTTGTCGCTGAGAAGC
9번째 contig, 길이: 145
TGGTTCCATCTTATACGCATGGATACTTTGGCAATTAATAAAAGCGAGAGCGTGGATGCAGAGGCATTGG
TCTGATTACACCAGCAGTACGCCTGTTTTGAGGGGCAGCTCCTGATGACGCTGCCGCCCAGTCCACCGAC
GCGTA
```

Microsoft Visual Studio 디버그 콘솔

```
kmers의 수: 1080
filterSize : 10653 numHashFunctions : 6
블룸 필터 생성중...
거짓 긍정들을 찾아내고 있습니다...
그래프 순회 중...
initKmersSize : 10
Starting kmer: CTCGCCACCAAGGTG, 진행도 : 1/10
Starting kmer: ACAGTACGCAGGAAC, 진행도 : 2/10
Starting kmer: GCAATTCGACATAAC, 진행도 : 3/10
Starting kmer: CCCGTTATCCCTCCT, 진행도 : 4/10
Starting kmer: CTTGTATCGTTAACG, 진행도 : 5/10
Starting kmer: GAAGCATGCTCCAAA, 진행도 : 6/10
Starting kmer: TGCTCTCGTTGCATG, 진행도 : 7/10
Starting kmer: ACAGACGTAACATTT, 진행도 : 8/10
Starting kmer: TCTATATATTGGAAA, 진행도 : 9/10
Starting kmer: TGGTTCCATCTTATA, 진행도 : 10/10
소모 시간 : 0.018seconds
De Bruijn 그래프 크기(바이트) : 10671


      원본 myDNA의 길이      :1000
복구된 가장 긴 contig의 길이 :145

String Matching 중...
String Matching 완료!

다음에 나오는 아래의 정확도는 원본 길이의 1/4배정도가 복
구되었을 때만 의미가 있는 값입니다.
그보다 짧다면 더 아래의 N50값을 확인해주세요.
DNA 일치율 : 14.5%
N50 : 72
(유의미한 경우에 속한다.) N50의  값  : 72
```