리드 시퀀싱 시뮬레이터 비교 분석

탁해성*, 이상민*, 박기정**, 이도훈*, 조환규*
*부산대학교 컴퓨터공학과
**국가생명연구자원정보센터
e-mail:tok33@pusan.ac.kr, cse83319@pusan.ac.kr,
kjpark63@gmail.com, dohoon@pusan.ac.kr, hgcho@pusan.ac.kr

Analysis of Read Sequencing Simulator

Haesung Tak*, Sang-min Lee*, Kiejung Park**, Dohoon Lee*, Hwan-gue Cho*
*Dept of Computer Science & Engineering
**Korean Bioinformation Center

요 약

차세대 유전자 서열 시퀀싱 기법이 등장함에 따라 참조 유전자 서열로부터 리드를 생성하는 시퀀서의 기술이 다양화 되었다. 이전 시퀀싱 방식에 비해 비용 및 시간 측면에서 효율성이 증대 되었으나, 매핑도구의 검증을 위해서 다양한 생물학적 특이성을 반영하거나 비용이 소요되지 않는 방법을 연구하는 과정에서 리드 시퀀싱 시뮬레이터가 개발되었다. 본 논문에서는 현재 사용되고 있는 리드 시퀀싱 시뮬레이터에서 반영된 시퀀싱 기법을 분석하고 시뮬레이터의 기능적 특성을 분석하고자 한다. 이는 시뮬레이터 개발에 필요한 기능 설계 및 생물학적 특성을 반영하는데 활용하고자 한다.

1. 서론

차세대 유전자 서열 시퀀싱 기법이 등장함에 따라 기존에 참조 유전자 서열로부터 단편 서열(이하 리드)를 생성해 내는 시퀀서의 원천 기술이 다양화되었다. 원천 기술이 다양해짐에 따라 이전에 리드를 생성하기 위해 소요되었던 비용과 시간 측면에서의 효율성이 크게 증대되었다. 차세대 유전자 서열 시퀀싱에서 대표적으로 사용되는 시퀀서로는 Roche사의 454 시퀀싱을 활용한 GS FLX+, Illumina사에서 개발한 HiSeq, SOLiD 등이 있다 [1]. 참조 유전자 서열 샘플을 활용하여 시퀀서를 동작하면 이에 대하여 리드가 생성된다. 생성된 리드 정보를 활용하여 참조 유전자 서열에 매핑하여 참조 유전자 서열의 생물학적 특징을 알아보고자 하는 시도가 이루어지고 있다. 매핑도구는 참조 유전자 서열 및 리드가 입력파일로 받아, 오류/변이와 같은 생물학적 특정을 고려하여 단편서열을 참조 유전자 서열에 매핑한다. 현재 많은 수의 매핑도구가 개발되어 있으나, 매핑도구에 대한 검증을 시행하기 위해서 다수의 실험을 필요로 한다. 차세대 유전자 서열 시퀀싱 기술 이전의 리드 시퀀서에 비해 리드 생성에 있어서 가장 중요한 요수인 정확도와 비용 효율성이 증가되었으나, 독립적인 변인에 대한 변화 가능 여부와 다수의 실험을 진행함에 있어서 문제점들이 발생한다.

시퀀서의 한계를 극복하기 위해, 참조 유전자 서열에서 임의의 리드를 생성해 주는 시뮬레이터가 개발되었다. 시뮬레이터는 실제 시퀀서에서 포함하는 기계적인 오류 및 인위적인 발생 변수들을 고려하여 리드를 생성하며, 이를 통해 매핑도구의 정확성 및 효율성을 얻고자 한다. 본 논문에서는 현재 차세대 유전자 서열 시퀀싱 기법의 특징을 분석하고, 이를 바탕으로 기존에 개발된 시뮬레이터들의 리드 생성 방식 및 특성을 분석하고자 한다. 이러한 분석 결과는 추후에 개발할 시퀀싱 시뮬레이터의 개발 목표 설정에 활용하고자 한다.

2. 차세대 유전자 서열 시퀀싱 기법 분석

유전자 시퀀싱은 염기 서열을 분석해 내는 과정을 말한다. 이러한 시퀀싱은 크게 각 리드 조각의 염기서열을 읽어내는 염기 결정 과정과 리드 정보를 바탕으로 전체 염기 서열을 밝혀내는 어셈블리 과정으로 나뉜다 [2]. 차세대 유전자 서열 시퀀싱 기술(이하 NGS)의 발전에 따라 이전에 사용되었던 시퀀싱 기법인 Sanger 시퀀싱 보다 효율적인 시퀀싱 기법이 등장하였다. 현재 가장 알려진 NGS 기법으로 454, Illumina, SOLiD가 있다. 이외에도 차세대 유전자 서열 시퀀싱 기법으로 SRMT 방식을 활용하여 단편서열을 생성하는 PacBio와 반도체를 활용하는 Ion Torrent 방식이 개발되었다. 본 논문에서는 분석 시뮬레이터에서 기반 시퀀서로 하는 NGS 기법 및 이전 시퀀싱 방식 중에서 신뢰도가 높은 Sanger 기법을 분석한다.

2.1 Sanger sequencing

NGS 이전의 시퀀싱 기법에는 Sanger가 개발한 효소 활용법과 Maxam과 Gilbert가 개발한 화학반응법이 있다. 이중에서 현재까지 사용되는 기법은 Sanger Sequencing 이다. Sanger 기법은 DNA 합성 중에 전구체들인 dNTP (dGTP, dATP, dTTP, dCTP)에서 2번탄소위치의 -OH기를 제거한 ddNTP를 첨가하면 반응이 멈추는 것을 바탕으로 한다. Sanger 시퀀서는 자동으로 시퀀싱이 될 수 있도록 ddNTP에 각기 다른 파장을 가지는 형광물질을 표지한 후 한꺼번에 반응하고, 이를 전기영동한 후 laser를 이용해서 염기 서열의 순서와 색을 구별한다. 이 기법은 시퀀서 내부적인 오차가 매우 작기 때문에 기계 자체에서 발생할 수 있는 오류/변이가 없다는 장점을 가진다. 하지만

리드 생성 비용이 여느 장비에 비교하여 너무 많이 소비되고 생성 시간이 오래 걸리므로 현재에는 비교 대상으로서의 기술로 활용된다.

2.2 454 - Pyrosequencing

Roche에 소속되어 있는 454의 시퀀싱 기술은 보통 GS FLX+ 시퀀서를 기준으로 말하며, 리드 길이가 600bp 또는 1,000bp로 긴 리드를 생성한다. 리드의 길이가 길 경우 어셈블리 과정이나 Resequencing 과정에서 매핑 난이도를 낮출 수 있다.

454는 하나의 bead에 하나의 DNA 사슬을 고정한 후에 emPCR로 증폭을 시킨다. 증폭이 끝나면 각 bead에는 수백만 개로 복제된 동일한 DNA 서열이 덮이게 되고, 각 bead는 PicoTiterPlate라는 장치에서 하나의 구멍에 하나씩 들어가서 시퀀싱에 개시된다. 454의 시퀀싱 기법은 Pyrosequencing chemistry라고 하며 Luciferase가 사용되어 염기의 빛을 측정하여 읽어낸다.

454는 참조 유전자 서열 없이 리드를 매핑하는 de novo 시퀀싱에 유용하다. 하지만 리드가 생성 양이 다른 기종에 비해서 상대적으로 적은 단점이 존재한다.

2.3 Illumina - Sequencing by synthesis

Solexa사에서는 SBS (Sequence by Synthesis)라는 새로운 기술을 개발하였으나, 2007년 Illumina에 합병되고 난 후 SBS 기반의 다양한 NGS 장비가 개발되었다. HiSeq 2000의 경우, 200bp의 리드를 11일 가량의 실행 시간으로 최대 540-600Gb까지 데이터 생산이 가능하다. 또한 저렴한 비용으로 많은 데이터를 생산하므로 대용량의 시퀀싱을 처리하는 연구에서 많이 사용된다.

Illumina의 NGS 기법은 브릿지 증폭이라고 해서 슬라이드 위에 DNA 단편을 고정시킨 후에 최대 1,000 분자까지 증폭시켜 같은 서열의 DNA 단편 집단을 형성한다. 이 집단을 클러스터라고 표현하며, 이를 주형으로 네 종류의 형광 표식 염기를 사용한 염기 합성반응인 SBS를 수행한다. 다른 시퀀싱 기법처럼 용액 안에서 증폭시키는 것이 아니라 판 위에 고정시킨 후에 판 위에서 구부러지면서 증폭되어 서열집단을 형성하는 것이 특징이다. 형성된 클러스터는 집단 별로 시퀀싱이 이루어져 각 리드 정보를 형성한다. Illumina의 출력 형식인 FASTQ 파일은 현재 가장 많은 분석 소프트웨어에서 널리 대응하고 있다.

2.4 SOLiD - Sequencing by ligation

라이프 테크놀로지스의 SOLiD 시리즈에서는 emulsion PCR 과정 이후의 Ligation을 사용한 시퀀싱이 특징이다. 이 기법에서는 간격을 두면서 두 개씩 염기를 읽는데, Primer reset을 통해 독립적으로 다섯 번을 반복하기 때문에, 각 염기를 중복하여 읽어서 정확도를 높인다.

SOLiD 후속 기종인 5500 시리즈에서는 ECC (Exact Call Chemistry) 기법이 적용되어 여섯 번째 프라이머가 추가되었고 따라서 리드 상에서 상당수의 염기를 서로 다른 프라이머로 세 번씩 독립적으로 읽음으로써 정확도를 향상시켰다. 5500의 다른 특징은 여섯 개의 lane으로 구성된 FlowChip에서 각 lane 별로 서로 다른 시퀀싱을 동시에 수행하는 Pay-Per-Lane sequencing 방식도 가능하기 때문에 시간과 비용을 절약할 수 있다. SOLiD 시리즈에서는 csfasta 형식으로, 5500 시리즈에서는 기본적으로 35, 60, 75bp 길이로 XSQ (eXtensible SeQuence)라는 바이너리 파일 형식의 데이터가 생성된다.

표 1 시퀀싱 기법 비교 분석 결과

Sequencer	Read 길이	Read/Run	생성시간	Cost / 1M base
Sanger	400~900bp	-	20m~3h	$2400
454	700bp	5M~	2h	$1
Illumina	50~250bp	3B~	1~10d	$0.05~0.15
SOLiD	85, 100bp	1.2~1.4B	1~2w	$0.13
PacBio	2900bp	35~75T	30m~2h	$2

2.5 PacBio - SMRT(Single Molecule, Real-Time)

Pacifie Biosciences사의 SMRT(Single Molecule, Real-Time)는 제 3세대 시퀀서로 잘 알려져 있다. 이 기술은 기존 NGS 시퀀싱 과정에서 필요로 하는 PCR DNA 증폭 과정이 생략된 것이 특징이다. 즉, DNA를 1 분자 상태에서 시퀀싱을 함을 의미하며, 평균 리드 길이는 1000 bases 이상으로 다른 시퀀서에 비해 길고, base throughput은 현재까지 약 90Mb이다. 하지만 기존의 시퀀서와 같이 형광 원리를 사용기 때문에, 이미지 처리를 과정에서 필요로 하는 관련 부품과 시약 등이 여전히 사용되며 오류 발생 가능성을 해결하지 못했다.

2.6 시퀀서 특징 비교 분석

현재 개발된 시뮬레이터는 각 도구별로 특정 시퀀서의 알고리즘 및 오류 모델을 적용한다. 표 1은 개발된 시뮬레이터로 부터 파악된 5가지의 시퀀서 기계의 동작 특성을 나타낸 것이다. 각 시퀀서는 생성하는 리드를 생성할 때 각 리드의 길이와 생성 시간, 생성되는 리드의 양, 그리고 비용에 이르기까지 각기 다른 특성을 나타낸다. 단편서열을 생성하는 방법이 각기 다르기 때문에 소비되는 비용 및 생성시간이 차이가 발생하는데, 이를 통해 Sanger부터 PacBio에 이르기 까지 생성되는 리드 길이의 변화와 비용 효율성이 다른 것을 확인할 수 있다.

3. 리드 시퀀싱 시뮬레이터

시퀀서의 특징을 활용하여 이를 임의로 생성하는 목적으로 생성된 시뮬레이터를 분석하였다. 실제 생물정보학에서 인용되거나 발표된 시뮬레이터를 선정하였다. 2007년부터 현재까지 발표된 논문 중에서 실행 가능한 11개의 시뮬레이터를 선정하였다. 선정된 시뮬레이터의 경우 공통적으로 참조 유전자 서열로부터 리드를 생성한다는 기능적 목표는 동일하지만, 리드의 생성 시 발생하는 오류 확률과 기계적 특성으로 인해 발생하는 오류 확률을 각각 다르게 반영한다. 선정된 시뮬레이터는 MetaSim [3], wgsim [4], Flowsim [5], Mason [6], ART [7], GemSIM [8], Grinder [9], pIRS [10], PBSIM [11], RSVSIM [12], wessim [13] 이다.

각 시뮬레이터의 특성을 비교하기 위해 실제 리드 생성에 대한 비교를 진행하는데 몇 가지 문제점이 존재한다. 시뮬레이터에서 리드를 생성함에 있어서 위치를 선정하는 방식과 실제 유전자의 발현 부분을 고려하는 지의 여부가 달라서 같은 참조 유전자 서열을 입력으로 넣는다고 하더

라도 같은 결과를 얻기 힘들다. 또한 같은 시뮬레이터에서 동일한 변수에 대해 다른 출력을 나타낸다. 이러한 특징으로 인해 각각의 특징 파악하는데 초점을 맞추어 연구를 진행하였다.

본 논문에서 선정한 시뮬레이터의 기능적 특성을 분석하여, 이를 통해 추후에 개발할 리드 시퀀싱 시뮬레이터의 기능적 특성을 반영하는데 사용하고자 한다. 리드를 생성하는 시뮬레이터의 경우 가장 중요하게 보는 요소로 리드 생성 알고리즘, 리드 생성 간에 적용될 생물학적 특성, 입출력 양식, 그리고 부수적으로 각 시뮬레이터에서 특징적으로 반영하는 부분을 중점적으로 분석하였다. 시뮬레이터에 따라 동일한 기능을 제공하는 시뮬레이터의 경우, 기존에 개발된 시뮬레이터의 한계를 극복하기 위한 지표로 활용하고 있음을 알 수 있다.

표 2에서는 시뮬레이터에서 기반으로 하는 시퀀서 정보를 확인할 수 있다. 표 2의 결과를 통해 시뮬레이터는 각기 개발 목적에 따라 특정한 시퀀서를 선정하거나 다수의 시퀀서를 선정한 것을 알 수 있다. 이전에 개발된 시뮬레이터들이 기본적으로 리드 생성 방식의 안정성을 가진다고 평가되는 Sanger 기법을 이용하여 리드를 생성하는 시뮬레이터가 있는 것을 확인할 수 있다. NGS 기법이 발전됨에 따라 각 시퀀서의 기술 발전이나 오류 모델을 적용한 시뮬레이터가 개발되기도 하였다. Wessim의 경우 GemSIM의 리드 생성 코어 기술을 반영하지만, 최근 개발된 시뮬레이터의 경우 더 이상 Sanger 기법에 대한 리드 생성을 하지 않는 것을 확인할 수 있다. wgsim과 flowsim, pIRS, PBSIM의 경우 기반으로 하는 시퀀서를 하나로 고정하여 개발되었다. PBSIM의 경우 차차세대 시퀀싱 기법으로 등장한 PacBio의 리드 생성 알고리즘과 오류모델을 적용한 것을 알 수 있는데, 이는 다른 시퀀서의 리드 생성 방식과 달리 길이가 긴 리드를 생성한다는 점에서 특징적으로 볼 수 있다. RSVSIM의 경우 기반 시퀀서를 고려하지 않고 통계 프로그래밍 언어인 R 및 SV 모델을 기반으로 리드를 생성하는 점에서 다른 시뮬레이터와 다름을 확인할 수 있다. 이러한 결과를 통해 기반으로 하는 시퀀서가 중복되더라도 각 시뮬레이터에서는 오류모델을 구현 방식의 차이가 있고, 같은 시뮬레이터에서 동일한 조건으로 리드를 생성하였을 때 동일한 결과를 얻을 수 없으므로 각 시뮬레이터의 차별성을 시퀀서 선택만으로 한정할 수 없다.

시뮬레이터가 반영하는 리드 생성 방식 이외에도 참조 유전자 서열에서 리드를 생성하기 위해 고려하는 부수적인 데이터 특징들이 존재하는데 이는 표 3에서 확인할 수 있다. 표 3의 분석 결과를 통해 대다수의 시뮬레이터가 참조 유전자 서열의 이중 나선 구조를 고려하는 것을 볼 수 있다. 표 3에서는 각 시뮬레이터에서 입력 받는 참조 유전자 서열 데이터는 FASTA로 되어있는 것을 확인할 수 있다. 이는 현재 유전자 서열 정보를 제공해주는 NCBI(National Center for Biotechnology Information) [13]와 1000 Genome Project [14]에서 FASTA 형식으로 되어 있기 때문이다. 또한 출력 파일의 양식을 보면 리드 생성 간에 유전자의 Quality Score 반영 여부에 따라 생

표 2 시뮬레이터 기반 시퀀서 비교 분석

Simulator	Sanger	454	Illumina	SOLid	PacBio	Etc.
MetaSim	O	O	O	-	-	
wgsim	-	-	O	-	-	
flowsim	-	O	-	-	-	
Mason	O	O	O	-	-	
ART	-	O	O	O	-	
GemSIM	O	O	O	O	-	
Grinder	O	O	O	-	-	
pIRS	-	-	O	-	-	
PBSIM	-	-	-	-	O	
RSVSIM	-	-	-	-	-	SV
wessim	O	O	O	O	-	

표 3 시뮬레이터 생물학적 특성 및 입출력 형식 분석

Simulator	Paired-end	Quality	Input	Output
MetaSim	O	-	FASTA	FASTA
wgsim	O	O	FASTA	FASTQ
flowsim	-	O	FASTA	SFF
Mason	O	O	FASTA	FASTQ,SAM
ART	O	O	FASTA	FASTQ
GemSIM	O	O	FASTA	FASTQ,SAM
Grinder	O	O	FASTA	FASTQ
pIRS	O	O	FASTA	FASTQ
PBSIM	-	O	FASTA	FASTQ
RSVSIM	O	-	R,FASTA	FASTA,CSV
wessim	O	O	FASTA	FASTQ

표 4 시뮬레이터 개발 특징 및 주요 특성 분석
(Lang. = 개발 언어, Open. = 오픈 소스)

Simulator	Lang.	Interface	Open.	Feature
MetaSim	Java	CLI GUI	-	Genome Evolution Model
wgsim	C	CLI	O	SAMtools
flowsim	Haskell	CLI	O	SFF Output
Mason	C++	CLI	O	Haplotype,SAM, SeqAn [14]
ART	C++	CLI	O	Quality Profiles
GemSIM	Python	CLI	O	Haplotype,SAM
Grinder	Perl	CLI GUI API	O	Amplicon, Metagenomics
pIRS	C++ Perl	CLI	O	GC-content, Profile based
PBSIM	C	CLI	O	CCS, CLR
RSVSIM	R	CLI	O	SV Mode
wessim	Python	CLI	O	GC-content

성되는 출력 파일이 다름을 확인할 수 있다. Quality Score를 반영하는 시뮬레이터라도 기본적으로 리드를 생성할 때 FASTA 형식으로 결과를 저장하는 기능을 보장한다.

표 4는 각 시뮬레이터의 개발언어 및 실행 환경 그리고 기타 특징에 대한 분석결과를 나타낸다. 현재 개발되고 있는 시뮬레이터들은 임의의 리드를 생성하고자 하는 사용자들에게 편의를 제공하기 위해 주로 오픈소스의 형태로 개발되었다. 대다수의 매핑도구가 리눅스 환경에서 동작하므로 시뮬레이터도 기본적으로 리눅스에서 동작가능하며, MetaSim과 Grinder의 경우 GUI를 활용할 수 있게 개발되어 윈도우나 기타 운영체제에서 동작이 용이하다. Mason의 경우 매핑도구의 개발에 많이 사용되는 SeqAn 라이브러리 [15]를 활용하는데 이는 추후에 개발될 리드 시퀀싱 시뮬레이터에 활용하고자 한다.

4. 결론 및 추후 연구 방향

참조 유전자 서열로부터 리드를 생성하는 시뮬레이터가 개발됨에 따라 매핑 도구에서도 다양한 실험을 위해 시뮬레이터들을 활용하게 되었는데, 각각의 시뮬레이터가 각기 다른 목적으로 개발되어 다수의 사용자가 사용하는 시뮬레이터는 많지 않다. 본 논문에서는 현재 개발된 시뮬레이터를 분석하고 이에 대한 특징을 분석하였다.

분석한 시뮬레이터들은 각각이 기존 시퀀서를 활용하여 시간 및 비용 효율적인 기능에 치우치거나 시퀀서의 특징을 반영하는데 그침을 알 수 있다. 하지만 다양한 실험을 위해 리드 매핑 결과에서 적용된 오류/변이율을 유사하게 적용하거나 유사 발현 구조를 나타내는 리드를 생성해주는 시뮬레이터가 존재하지 않는다. 이러한 한계점을 바탕으로 추후에 개발할 리드 시퀀싱 시뮬레이터에서 기본적으로 많이 활용되는 시퀀서에 대한 리드 생성뿐만 아니라 SAM 파일을 활용하여 리드를 생성하는 기능까지 구현하고자 한다.

감사의 글

본 연구는 KRIBB 기관주요사업의 연구비 지원에 의해 수행되었습니다.

참고문헌

[1] L. Liu, Y. Li, S. Li, and at el., “Comparison of next-generation sequencing systems,” Journal of Biomedicine and Biotechnology, vol. 2012, 2012.

[2] 원정임, 홍상균, 공진화, 허선, 윤지희, "NGS 데이터를 이용한 대용량 게놈의 디노버 어셈블리," 한국정보과학회 2012한국컴퓨터종합학술대회, 제39권, 제1호(C), pp. 25-27, 2012.

[3] D. C. Richter, F. Ott, A. F. Auch, and at el., “Metasim - a sequencing simulator for genomics and metagenomics,” PLoS ONE, vol. 3, no. 10, pp. e3373, 2008.

[4] H. Li, B. Handsaker, A. Wysoker, and at el., “The sequence alignment/map format and samtools,” Bioinformatics, vol. 25, no. 16, pp. 2078-2079, 2009.

[5] H. Manuel, “Mason - a read simulator for second generation sequencing data,” Technical Report FU Berlin, 2010.

[6] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “Art: a next-generation sequencing read simulator,” Bioinformatics, vol. 28, no. 4, pp. 593-594, 2012.

[7] K. McElroy, F. Luciani, and T. Thomas, “Gemsim: general, error-model based simulator of next-generation sequencing data,” BMC Genomics, vol. 13, no. 1, pp. 1-9, 2012.

[8] F. E. Angly, D. Willner, F. Rohwer, and at el., “Grinder: a versatile amplicon and shotgun sequence simulator,” Nucleic acids research, vol. 40, no. 12, pp. 1-8, 2012.

[9] X. Hu, J. Yuan, Y. Shi, and at el., “pirs: profile-based illumina pair-end reads simulator,” Bioinformatics, vol. 28, no. 11, pp. 1533-1535, 2012.

[10] Y. Ono, K. Asai, and M. Hamada, “Pbsim: Pacbio reads simulator-toward accurate genome assembly,” Bioinformatics, vol. 29, no. 1, pp. 119-121, 2013.

[11] C. Bartenhagen and M. Dugas, “Rsvsim: an r/bioconductor package for the simulation of structural variations,” Bioinformatics, vol. 29, no. 13, pp. 1679-1681, 2013.

[12] S. Kim, K. Jeong, and V. Bafna, “Wessim: a whole-exome sequencing simulator based on in silico exome capture,” Bioinformatics, vol. 29, no. 8, pp. 1076-1077, 2013.

[13] “National Center for Biotechnology Information,” http://www.ncbi.nlm.nih.gov/

[14] 1000 Genomes, “1000 Genomes - A Deep Catalog of Human Genetic Variation,” http://www.1000genomes.org/

[15] FU Berlin, “Seqan c++ libarary,” http://www.seqan.de/.