이배체 유전체들의 서열비교를 위한 유전체 염기서열 생성도구 개발

김종현*, 박치현**, 박상현**
*School of Medicine, University of Pennsylvania
**연세대학교 컴퓨터과학과
e-mail : jongk@mail.med.upenn.edu

Development of a tool to generate diploid genome sequences for whole-genome alignments.

Jonghyun Kim*, Chihyun Park**, Sanghyun Park**
*School of Medicine, University of Pennsylvania
**Dept. of Computer Science, Yonsei University

요 약

현대 유전체학 기술의 진보는 생물학적으로 중요한 의미를 갖는 생물들의 유전체 서열의 규명 genome sequencing)에 힘입은 바 크다. 기존의 유전체 서열결정법은 주로 염기변이율이 낮은 생물들에 초점을 맞추어 왔다. 하지만 염기변이율이 높은 생물들의 유전체 염기서열을 결정할 필요가 높아짐에 따라 이를 위한 방법론에 대한 연구가 활발히 진행되고 있다. 염기변이율이 높은 생물들의 이배체 (diploid) 유전체 서열이 효과적으로 결정될 수 있을 경우 기존의 유전체 서열비교의 방법론에도 변화가 요청되고 있는 실정이다. 기존의 유전체 서열비교 (whole-genome alignment) 방법론은 반수체 (haploid) 유전체들의 서열비교을 위해 개발되었지만, 염기변이율이 높은 생물들의 유전체 서열비교에는 반수체 유전체들 비교에 특화된 도구들이 필요하다. 또한 현재 서열비교를 시각화하는 소프트웨어들도 반수체 유전체 비교를 위해 개발된 실정이다. 본 논문의 목표는 이배체 유전체 서열을 비교하는 방법론을 개발을 용이하기위해 이배체 유전체의 서열을 생성하는 도구를 개발하는 것이다. 개발된 도구는 실제 일어날 수 있는 염기변이와 genomic rearrangement 를 사용자의 입력을 받아 다수의 생물들의 유전체 서열을 생성해 낸다. 이를 통해 이배체 유전체 서열을 비교하는 도구의 개발을 용이하게 하는데 초점을 맞추고 있다.

1. 서론

현재 인간 [1] 과 다염기변이 생물들 ([2], [3]) 의 이배체 유전체의 복원문제에 관한 연구가 활발히 진행되고 있다. 이배체 유전체는 특히 염기변이가 높은 경우에는 유전체의 서열을 조립하기가 힘들다고 알려져 있지만 [3], 일단 서열을 조립하면 이배체 유전체 서열을 복원하기는 염기변이가 낮은 경우보다 용이하다. 앞으로 점차 많은 다염기변이 유전체의 염기서열을 결정할 필요가 커질 것이고, 또한 이런 다염기변이 유전체 서열조립으로부터 이배체 유전체의 서열을 유추해 내는 것도 유전체 프로젝트의 중요한 부분이 될 것이다. 이럴 경우에 기존의 유전체 프로젝트와 가장 큰 차이는 하나의 유전체 서열만 공개하는 것이 아니라 두가닥의 이배체 유전체의 서열이 유전체 프로젝의 산물로 공개될 것이다. 최근에 멍게 (Ciona intestinalis) 한 개체와 인간 한 명의 이배체 유전체 서열이 공개되었고 [1] 이를 통해 미래의 유전체 프로젝트의 한 모습을 엿볼 수 있다.

본 연구는 학술진흥재단의 BK21(2 차) 사업과 과학기술부 과학재단 특정연구개발사업(2007-03965)의 지원

2. 유전체 서열비교 (whole-genome alignment)

진화생물학적인 측면에서 중요한 생물의 이배체 유전체들의 서열을 유추해 낼 경우 이배체 유전체 서열을 다른 이배체 유전체의 서열에 정렬시키는 방법 (whole-genome alignment)을 통해 두개 이상의 다른 생물들 사이에 기능적으로 중요한 역할을 하는, 변하지 않고 있는 부분 (evolutionarily conserved region)을 발견할 수 있다.

기존의 유전체 프로젝트의 목표는 이배체 유전체의 서열을 유추해내는 것이 아니라 반수체 유전체 서열들만을 유추해 내었기 때문에 현재 개발된 whole-genome aligner 들은 반수체 유전체 서열들간의 비교만이 가능하다 ([4], [5]). 현재 개발된 comparative browser 들도 반수체 유전체 서열간의 비교를 시각화해주는 것에 초점을 맞추고 있다 ([6], [7]). 따라서 이배체 유전체들간의 서열비교에 최적화된 whole-genome aligner 와 이런 alignment 들을 시각적으로 나타내 주는 genome comparative browser 의 개발이 시급한 실정이다.

3. 이배체 유전체 서열의 생성

본격적인 algorithm 개발과 시각적인 도구를 개발 하기에 앞서 algorithm 과 시각화 도구의 성능을 테스트 할 도구의 개발이 선행되어야 할 것이다. 본 논문의 목표는 이배체 유전체들간의 서열비교를 위한 algorithm 의 개발을 용이하게 할 이배체 유전체 서열을 생성하는 도구를 개발하는 것이다. 이 도구를 사용하여 사용자가 원하는 대로 genomic arrangement 를 simulation 할 수 있다. 사용자와 원할한 상호작용을 위해 command line 을 기반으로 하여 원하는 염기변이의 형태와 genomic rearrangement 를 만들 수 있다 (그림 1). FASTA 포맷의 DNA sequence 를 읽어들여서 먼저 전체적인 heterozygosity rate 를 결정한 다음에 사용자의 지정에 따라 multibase indel 들과 multi base substation 들을 지정한다. 이때 각각의 indel 과 substitution 의 길이도 사용자에 의해 지정될 수 있다. 또한 transposition 과 inversion 과 같은 genomic arrangement 들도 사용자가 지정한 길이에 따라 유전체 상의 임의의 위치에 발생하게 된다. 사용자의 지정에 따라 만들어진 이배체 유전체 서열은 FASTA 포맷으로 각각의 파일에 저정된다.

```
[root@vodka bio_scaffold]#
[root@vodka bio_scaffold]# ./compGen start

-------------------------------------------------------------------------------
first you have to execute input operation because cionaV1.fasta file have to be opend!
-------------------------------------------------------------------------------

------------ menu ---------------
1) modify sequence in [%] heterozygosity rate
     command> -p [percentage]
ex)  command> -p 1.2

2) inversion
     command> -i [number of inversions] [length of inversion]
ex)  command> -i 3 7

3) transposition : copy a random sequence and insert at random position
     command> -t [number of transpositions] [lengh of transposition]
ex)  command> -t 5 3

4) substitution
     command> -s [number of substitutions] [length of substitution]
ex)  command> -s 6 3

5) deletion
     command> -d [number of deletions] [length of deletion]
ex)  command> -d 8 3

6) save input file
     command> input filename
     file name:  [filename].fasta
ex)  commnad> input filename

7) save 3 output files : .fasta  format
     command> output filename
     file name:  [filename]_1.fasta    /    [filename]_2.fasta    /   [filename].fasta
ex)  command> output filename

8) exit the loop
     command> quit
------------ menu ---------------

command>
```

(그림 1) 이배체 염기서열 생성도구

4. 결론

본 논문을 통해서 공개된 소프트웨어는 이배체 유전체간의 whole-genome aligner 를 위한 알고리즘개발에 직접적으로 활용되어 알고리즘의 정확성을 검증하는데 이용될 것이다. 또한 서로 다른 이배체 유전체간의 whole-genome alignment 를 시각화시켜주는 genome comparative browser 의 개발에 직접적으로 활용되고 있다. 이배체 서열 생성도구는 제한없이 사용할 수 있으며, embio.yonsei.ac.kr 을 통해서 download 받을 수 있다.

참고문헌

[1] Levy, S., et al. 2007. The diploid genome sequence of an individual human. *PLoS Biology*. **5:** e254.

[2] Kim, J. H., Waterman, M. S., Li, L. M. 2007. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res*. **17:** 1101-1110.

[3] Vinson, J., et al. 2005. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res.* **15:** 1127-1135.

[4] Brudno, M., et al. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. **13:** 721-731.

[5] Schwartz, S., et al. 2003. Human-Mouse alignments with BLASTZ. *Genome Res*. **13:** 103-107.

[6] Engels, R., et al. 2006. Combo: a whole genome comparative browser. *Bioinformatics.* **22:** 1782-1783.

[7] Frazer, K. A., et al. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. **32:** W273-279.