다염기변이 및 메타유전체 염기서열 생성도구에 관한 연구

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

A Study on a tool to generate polymorphic genome and metagenome sequences

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

요 약

유전체학 (genomics)의 가장 기초적인 기반이 되는 것은 염기서열을 정확하게 결정해 내는 것이다. 많은 진핵생물들 (eukaryotes)은 두개의 상동염색체를 가지며 두개의 염색체의 염기서열에는 차이가 존재한다. 현재의 유전체 염기서열 결정방법으로는 염기변이가 많이 존재할 경우 유전체의 염기서열을 결정하기 어렵다. 특정한 장소에 서식하는 무수히 많은 미생물들의 유전체의 염기서열을 동시에 결정하는 문제도 미생물학에서 중요성을 인정받는 문제이지만, 미생물들간의 염기변이의 정도는 단일개체의 경우보다 복잡하며 염기서열을 효과적으로 결정하기 힘들다. 따라서 염기변이가 많은 생물들과 미생물들 집합의 염기서열을 결정할 수 있는 방법론의 개발이 시급한 실정이다. 본 논문에서는 조립된 다염기변이 유전체및 메타유전체의 염기서열의 정확성을 평가하기 위한 유전체 서열과 시뮬레이션에 기반한 read 들을 생성하는 도구를 개발하는 것을 목표로 한다.

1. 서론

현재 많은 생물들의 유전체 염기서열들이 밝혀져 왔고, 단일 유전체 염기서열을 밝히는데 소요되는 비용은 계속 하락하고 있는 추세이다. 이렇게 생물체의 염기서열을 밝혀내는 것이 용이하게 됨에 따라 더욱 많은 생물들의 염기서열들이 규명되어 나갈 것이다. 하지만 현재의 염기서열을 규명하는 기술 (genome sequencing technology)에도 한계점이 존재한다. 가장 근본적인 한계점은 한번에 높은 수준의 신뢰도를 가지고 sequencing 할 수 있는 base 의 수가 500~800 bp 에 불과하다는 점이다. 따라서 서열이 결정된 짧은 길이의 read 들을 가지고 computational approach 를 사용하여 서열들을 조립해 나가야 한다. 이러한 computational approach 측면에서 대표적인 한계점은 유전체에 염기변이가 많이 존재할 경우 유전체의 서열을 조립해 내기가 힘들다는 것이다. 진핵생물들 (eukaryotes)은 많은 경우에 한쌍의 상동 염색체를 갖는다. 두개의 상동염색체의 동일한 위치에 서로 다른 염기가 존재할 수 있다. 이러한 염기변이들을 단일염기변이 (Single Nucleotide Polymorphism)이라고 부른다. 이런 염기변이가 많이 존재할 경우 (염기변이율 > 0.4%)에는 유전체의 서열을 조립하기 힘들다고 보고되어 있다. 이미 Candida albicans, Ciona intestinalis, Ciona savignyi, Strongylocentrotus purpuratus 들의 유전체 프로젝트에서 이미 서열조립의 어려움이 나타났다 ([1], [2], [3], [4]). 가장 문제가 되는 것은 조립된 결과인 scaffold 들의 길이가 염기변이가 낮은 생물들의 scaffold 의 길이보다 너무 짧다는 것이다. 예를 들어 human genome project 의 N50 scaffold 길이는 2.5 Mb 였지만, Ciona intestinalis 의 N50 scaffold 길이는 190 kb 에 불과하였다. 이러한 유전체 서열의 연속성의 저하는 유전체학의 발전에 심각한 장애요인이 되고 있는 실정이다.

본 연구는 학술진흥재단의 BK21(2 차) 사업과 과학기술부 과학재단 특정연구개발사업(2007-03965)의 지원을 받아 수행되었습니다.

2. 메타유전체학 (Metagenomics)

미생물 연구에서 가장 큰 문제점 중에 하나는 실험실에서 배양하여 연구할 수 있는 미생물들의 수가 실제 자연계에서 존재하는 미생물들의 수보다 훨씬 적다는 것이다. 따라서 자연계에 존재하는 미생물들을 실제 서식환경 상태에서 현대 genomics 기법을 사용하여 연구하는 학문을 메타유전체학 (metagenomics)이라고 한다. Genome sequencing 기법의 측면에서는 특정 장소에서 서식하고 있는 수많은 미생물샘플을 채취하여 그 샘플에 서식하고 있는 무수히 많은 유전체의 서열을 동시에 조립하는 시도들이 이루어 지고 있다. 하지만 염기변이가 많은 유전체의 경우와 마찬가지로 염기변

이로 인해서 유전체 서열의 조립에 어려움을 겪고 있다. 하지만, 염기변이로 인한 유전체 서열 조립의 복잡성은 메타유전체의 경우가 단일 생물에 존재하는 염기변이로 인한 복잡성보다 심각하다.

3. 유전체 서열조립 (Genome sequence assembly)

기존의 유전체 서열조립 방법론은 인간과 침펜지와 같이 염기변이가 많지 않은 (0.1% ~ 0.2%) 생물에 적합하게 개발되었다. 최근에 Ciona savignyi 의 유전체를 조립하기 위해 개발된 방법론이 있지만, 이 방법론은 통상적인 것보다 2 배 가까운 sequence coverage 를 요구하며 염기변이가 극단적으로 많은 유전체의 염기서열을 조립하는데 사용될 수 있다. 다양한 염기변이율에 일반적으로 사용될 수 있고 (0.4% ~ 6%), sequence coverage 를 7 ~ 8 정도 요구하는 유전체 서열조립 방법의 개발이 필요한 실정이다.

이러한 방법론을 개발하는 과정에서 조립된 서열의 정확성을 평가하기 위해서는 실제 유전체 서열과 read 들을 이용하기 힘들다. 조립된 유전체 서열과 원래의 유전체 서열을 비교하여 정확성을 평가해야하는데 실제 genome project 에서는 실제 유전체서열을 미리 알 수 없기 때문이다. 본 논문의 목표는 메타유전체를 포함하여 염기변이가 높은 유전체의 서열조립에 있어서 일반적인 해결책을 개발하는 것을 용이하게 할 수 있도록 simulation 을 통해 유전체 서열과 염기변이, 그리고 read 를 생성하는 도구를 개발하여 제공하는 데에 있다.

4. 확률모형 (Probabilistic model)

다양한 염기변이율에 일반적인 해결책을 제시하는 방법론을 개발하게끔 하는 것이 목표이기때문에 사용자로 하여금 염기변이율을 임의로 입력할 수 있게 하였다. 염기변이가 유전체상에서 실제 분포는 uniform distribution 이 아니라 coalescent theory 에 따라 compound Poisson 분포를 이룬다. 본 서열생성 도구는 compound Poisson 분포에 따라 염기변이를 유전체상에 생성해 낸다. 유전체상에 존재하는 base 들의 집합이 다음과 같고 $\mathsf{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, -\}$, k 가 다른 유전체의 index 이고, j 가 유전체상의 임의의 위치를 나타내는 index 라 할때, 각각의 위치에 임의의 base 가 존재할 확률은 다음과 같다.

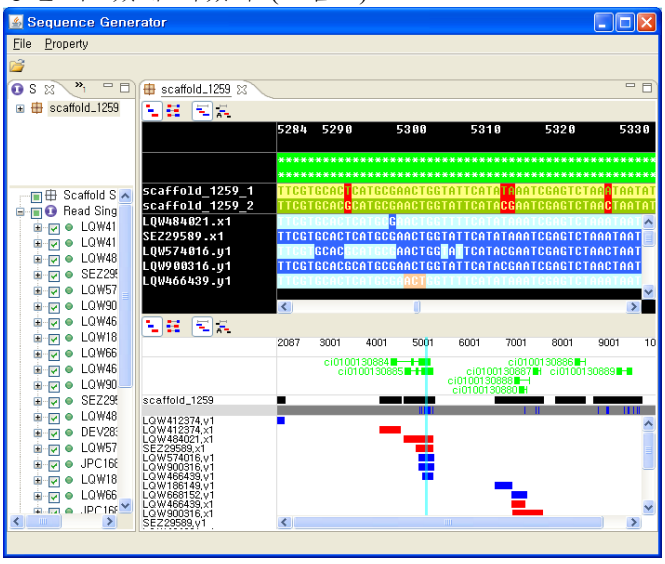
$$\Pr(S_{kj} = s), s \in \mathsf{A}.$$

또한 실제 genome project 의 sequencing error rate 를 반영하기 위해 실제 genome project 에서 사용된 read 들과 quality score 들을 입력으로 읽어드린다. I 와 j 를 genome assembly 상에서 임의의 위치를 나타내고 X 가 실제로 관찰된 base-call 이고 Y 가 실제 base 이고 q 가 base-call 에 해당하는 quality score 라고 할 때 sequence error rate 는 다음과 같은 확률을 따른다 [5].

$$\Pr(X_{ij} \neq y_{ij} \mid Y_{ij} = y_{ij}) = 10^{-\frac{q_{ij}}{10}}, y_{ij} \in \mathsf{A}.$$

각각의 read 가 특정한 위치에서 시작될 확률은 유전체서열상의 모든 위치에서 같은 확률을 갖는다.

메타유전체의 경우에는 유전체의 서열이 두개 이상 존재할 수 있기 때문에 사용자가 유전체의 서열의 갯수를 임의로 지정할 수 있게 하였다. 또한 미생물들 간의 recombination 이 존재하는 것을 반영하여 사용자가 interactive 한 방식으로 recombination 을 반영하게 만들었다. 서열을 생성하는 프로그램은 GUI 환경을 제공하며 JAVA 로 개발되어 어떠한 플랫폼에서도 작동할 수 있게 하였다 (그림 1).



(그림 1) 서열생성 도구

5. 결론

서열조립 방법론을 개발하는데 있어 정확성을 평가하여 지속적인 피드백을 받을 수 있는 것은 필수적이다. 이를 통해서 서열조립 방법론의 점진적인 개선을 이루어 나가는데 기여할 수 있다. 본 논문에서 공개된 서열생성 도구는 다염기변이 유전체서열과 메타유전체 서열을 조립하는 새로운 방법론의 정확성을 테스트하는 도구로서 이용될 수 있다. 서열생성 도구는 공개 소프트웨어로 http://embio.yonsei.ac.kr 을 통해서 download 받을 수 있다.

참고문헌

[1] Jones, T., et al. 2004. The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci.* **101:** 7329-7334.

[2] Dehal, P., et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298:** 2157-2167.

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

[4] Sea urchin genome sequencing consortium. 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*. 314: 941-952.

[5] Ewing, B. and Green, P. 1998. Basecalling of automated sequencer traces using Phred. II. error probabilities. *Genome Res.* **8:** 186-194.