The Korean Journal of Microbiology, Vol. 43, No. 4, December 2007, p. 317-320

단 보

미생물 유전체 프로젝트 수행을 위한 Base-Calling 오류 감지 프로그램 및 알고리즘 개발

이대상[1] · 박기정[2]*

[1]한국폴리텍 바이오대학 바이오생명정보과

[2](주)스몰소프트 정보기술연구소

미생물 유전체 프로젝트를 수행하는 과정에서 발생하는 base-calling 오류를 포함하는 것으로 의심되는 유전자나 염기서열의 리스트를 보여 주는 프로그램을 개발하였다. 이 프로그램의 모듈들은 base-calling 오류로 의심되는 염기들의 후보군을 유전체 프로젝트를 수행하는 주요 단계에서 감지할 수 있도록 하였다. 이들 프로그램들은 초기 단계에서는 Phrap 파일에 존재하는 contig assembly 정보를 이용하여 base-calling 오류를 감지하는 모듈, 중간 단계에서는 상동성 검색 결과물로부터 frame shift 돌연변이의 진위 유무를 분석할 수 있는 모듈, 마지막 단계에서는, 이미 발표된 미생물 유전체와 같은 종으로부터 유래된 균주에 대한 유전체 프로젝트를 수행할 경우, 비교 유전체 분석 기법을 활용하여 base-calling 오류 가능성이 높은 서열의 후보군을 추출하여 해당 서열의 크로마토그램파일을 유전체 연구자가 볼 수 있는 모듈로 구성되어 있다.

Key words □ **assembly, base-calling error, contig, genome, homology, phrap**

미생물 유전체(genome) 프로젝트를 수행하는데 있어(1), Applied Biosystems사의 ABI 3730 sequencer와 같은 대용량 염기서열 분석 기기에서 산출되는 4가지 염기에 대한 농도 곡선(trace)으로 이뤄져 있는 크로마토그램(chromatograms) 파일들로부터 DNA 염기서열을 결정하는 과정을 base-calling이라 한다.

Base-calling 오류는, chromatogram 파일로부터 각각의 염기에 대한 trace를 이에 상응하는 염기를 할당하는 과정에서 발생하는 오류를 일컫는 말이다. 이상적인 크로마토그램의 곡선은 이들의 trace가 겹쳐져 있지 않고 peak가 일정하게 분포되어 있는 것이지만, 실제는 sequencing 반응, 젤 전기영동, trace 분석 과정에서의 오류, hairpin 구조나 GC-rich 지역과 같이 서열이 가지고 있는 내재적 특성 때문에 peak가 겹치거나, 한 쪽으로 치우치거나, 각각의 peak 사이의 간격이 일정치 못한 경우가 발생한다(4). 자동화된 프로그램으로 이러한 제반 실험을 파악해서 정확하게 base-calling하는데 많은 제약과 제한이 다르므로, base-calling 프로그램들은 경우에 따라 많은 오류를 발생하게 된다.

Base-calling의 품질은 특정 염기서열의 정확도와 이러한 서열들이 조립(assembly) 과정을 거쳐 만들어지는 contig 및 유전체 서열 정보 전체에 영향을 미치게 된다. 비록 DNA 서열의 정확도가 99.99%를 넘는다고 하더라도 오류를 포함하고 있는 염기의 개수는 유전체의 크기에 정비례하므로 무시할 수 없는 숫자에 이른다. 미생물 유전체의 길이가 5 Mb라고 가정할 때, 최소 500개의 base-calling 오류가 존재한다고 추정 할 수 있다.

*To whom correspondence should be addressed.
Tel: 82-42-864-2524, Fax: 82-42-385-9240
E-mail: kjpark@smallsoft.co.kr

이러한 base-calling 오류로 산출된 DNA 서열들은 frame shift, deletion, insertion 돌연변이로 오해될 가능성이 매우 높다. 특히, base-calling 오류로 오독된 염기가 open reading frame에서 stop codon을 발생시키게 된다면, 정확도가 높다고 알려진 유전자 예측 프로그램(3)을 사용하더라도 유전자 전체가 아닌 일부분을 예측 하는 상황이 발생하게 된다. 이러한 오류들을 검증하고 수정하는 추가적인 작업들이 유전체 프로젝트의 최종 단계인 annotation 과정이나 annotation 작업을 수행하기 이전에 필요하게 된다. 이러한 과정은 모두 수작업으로 이뤄지는 것이 일반적이므로 base-calling 오류의 가능성이 높은 염기를 알려주거나 유전체 연구자에게 해당 염기가 base-calling 오류인지 그렇지 않은지를 확인할 수 있는 리스트를 제시해 주는 프로그램의 필요성이 미생물 유전체 프로젝트를 수행하고 있는 연구자들 사이에서 요구되어 왔다. 이러한 필요성을 해결하기 위해, 유전체 프로젝트를 수행하는 각각의 단계에서 이러한 base-calling 오류들의 가능성을 지닌 것으로 의심되는 유전자들이나 염기서열들의 리스트를 보여줄 수 있는 프로그램들을 개발하게 되었다.

개발내역

프로그래밍 언어는 Visual C++, Java를 사용하였으며, 서열 상호간의 상동성 검색은 Linux (RedHat 9.0)를 기반으로 하여 BLAST (2)를 활용하여 수행하였다. Base-calling 오류로 추정되는 염기의 후보군을 찾아내기 위한 프로그램과 알고리즘을 다음과 같이 크게 세 가지 접근 방법을 사용하여 각각의 모듈로 개발하였다.

첫째, contig 정렬 과정에서 발생하는 base-calling 오류 감지 방법은 미생물 유전체 프로젝트의 시작단계에 있어, Phrap 파일(5) 안에 들어있는 base-calling 오류로 의심되는 염기를 찾기 위해 contig assembly 정보를 사용하였다. 만약, 특정 염기가 Phrap 파일 안에서의 품질 지수가 매우 낮고 contig 안의 보존염기(consensus base)와 같지 않을 경우 그 염기가 포함된 부분의 염기서열들의 크로마토그램을 볼 수 있도록 하였다(Fig. 1). 이러한 base-calling 오류들은 보존성이 높은 지역(highly consensus region)에서 발견되었으며, 그러한 contig의 크로마토그램을 조사하여 base-calling의 정확성을 조사할 수 있도록 하였다.

둘째, frame shift 돌연변이의 진위 분석을 위한 base-calling 오류 감지 방법은 유전체 프로젝트의 중간 단계에서 이용할 수 있도록 하였다. 각각의 frame shift 돌연변이가 실제로 존재하는 돌연변인지 아니면 base-calling 오류로 발생하는 오류인지 frame shift 돌연변이로 추정되는 리스트를 제공하여 유전체 연구자가 해당 염기가 포함되어 있는 크로마토그램 확인 과정을 통해 진

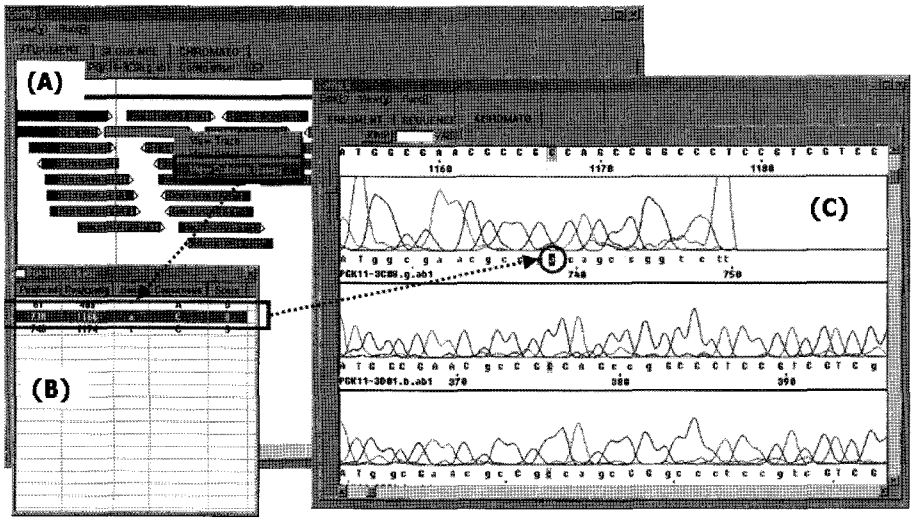

Fig. 1. Base-calling error detection program based on contig alignment information. The gray and dark gray colors in fragment alignment indicate high and low consensus regions, respectively (A). The dubious bases located at consensus alignment regions are searched for and the candidate list is shown (B). The chromatograms of suspicious bases are shown with the chromatograms of other sequences. Users can view chromatograms by clicking the dubious base list and determine its accuracy (C). The dotted arrows show process of finding dubious base from consensus alignment to chromatograms.

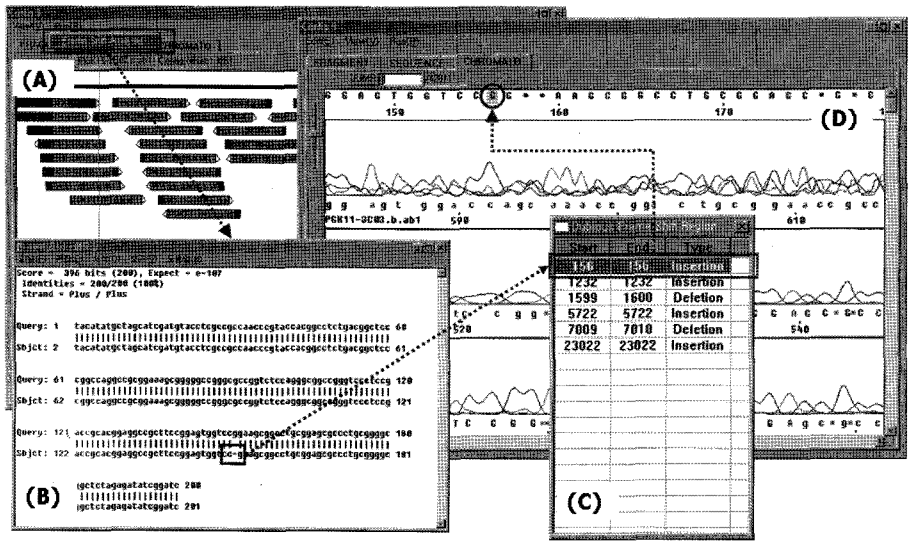

Fig. 2. The frame shift mutation analysis program. Frame shift mutation analysis starts from homology search (A). From BLAST search result, this program predicts the positions of frame shift mutations (B). The list of frame shift positions (base insertion or deletion) is shown (C). Finally, detailed chromatograms for each frame shift candidate base can be viewed by further clicking (D). The process of finding dubious base in the region of frame shift mutation are indicated by dotted arrows.

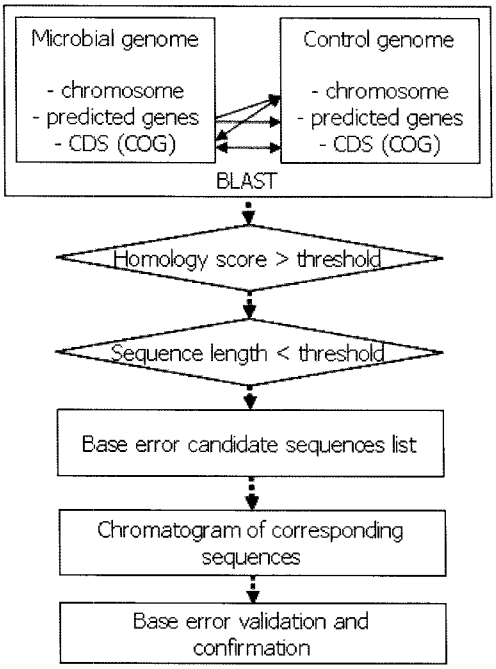

Fig. 3. The algorithm for base calling error detection in annotation step.

위파악을 하는데 도움을 줄 수 있도록 하였다. 이러한 frame shift 돌연변이의 진위유무를 파악할 수 있는 후보군 리스트를 확보하기 위해 contig 서열들 상호간에 그리고 비슷한 유전체 서열들 간에 BLAST를 이용하여 상동성 비교를 수행하였다. 상동성 검색 결과 특정 염기서열이 다른 유전자나 염기서열과 매우 비슷하고 frame shift 돌연변이로 추정된다면, 이들 frame shift 돌연변이의 진위를 확인할 수 있는 가장 효과적인 방법이 frame shift 위치로 추정되는 부분의 contig 서열들에 포함되어 있는 다른 염기서열들을 확인해 보는 방법 일 것이다(Fig. 2). 이러한 크로마토그램을 검증하는 과정을 통해 base-calling 오류의 진위 유무를 연구자들이 평가할 수 있도록 하였다.

마지막으로 유전체 프로젝트의 마지막 단계에서의 base-calling 오류 감지 방법은 비교 유전체 검색 기법을 활용하였다. 이 방법의 경우 현재 진행 중인 미생물과 같은 속(Genus)이나 종(Species)에 속하는 미생물의 유전체의 annotation 정보가 발표된 것이 있을 경우 활용도가 높으며, 이미 발표된 유사성이 높은 미생물의 유전체 정보가 없을 경우에는 사용할 수 없는 단점을 가지고 있다. 미생물 유전체 프로젝트가 진행 중인 유전체의 정보를 유사도가 높은 미생물 유전체 정보(control genome)와 상호 비교를 수행하는 것이다. 세부적인 알고리즘은 Fig. 3에서 같이 유전체 프로젝트가 진행 중인 미생물로부터 예측되어 나온 모든 유전자들을 대조군 염색체 서열과 BLAST를 이용하여 상동성 비교를 하고, 또한 대조군 유전체의 예측된 유전자들과도 상동성 비교를 수행하는 것이다. 이와 동시에 coding sequence정보로부터 산출되어 나온 정보를 기반으로 ontology 분석(6)을 수행한 그룹들 상호간의 상동성 검색을 수행한다. 이러한 BLAST 검색 결과물들 가운데 특정 점수(threshold) 이상의 상동성 점수를 가진 후보군을 가려낸 후, 이들 서열의 길이가 특정 길이 이하의 점수를 가진 서열들을 모두 base-calling 오류를 가진 후보군으로 모으는 작업을 수행한다. 이러한 후보군들로부터 해당 서열의 크로마토그램 파일을 살펴 base-calling 오류의 진위를 살펴보고 확인할 수 있도록 하였다.

본 논문에서 개발된 프로그램들과 알고리즘이 가지는 장점은 유전체 프로젝트 과정에서 불가피 하게 발생하는 base-calling 오류에 대한 감지를 유전체 프로젝트 진행의 각 단계에서 유전체 연구자에게 오류 가능성을 가진 염기의 후보군의 리스트를 제시한 것이다. 또한 이들 프로그램들이 유전체 프로젝트가 완료된 후가 아니라 진행 중에 오류를 가진 염기를 수정하여 유전체 서열의 정확도와 유전체 annotation의 품질을 높이는데 기여할 것으로 사료된다. 개발된 프로그램은 독립적으로 사용될 수 있고 또한, 웹 기반 미생물 유전체 annotation 시스템(paper submitted, Lee and Park, 2007)에 통합 될 수 있도록 개발하였다. 향후 사용자 interface는 추가로 개발 할 예정이고 업그레이드 시킬 계획이다.

감사의 말

본 연구는 산업자원부의 중기거점기술개발 사업의 지원에 의해 수행되었습니다.

참고문헌

1. 이대상, 태홍석, 박기정. 2003. 유전정보분석시스템. 전자공학회지 30, 68-78.
2. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
3. Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636-4641.
4. Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. *Genome Research* 8, 186-194.
5. Green, P. Phrap Documentation: Algorithms, http://www.phrap.org
6. Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 29, 22-28.

(Received October 15, 2007/Accepted October 22, 2007)

ABSTRACT: A Base-Calling Error Detection Program for Use in Microbial Genome Projects
Daesang Lee[1] and Kiejung Park[2]* ([1]Department of Bioinformatics, Korea Bio Polytechnic, Nonsan 320-905, Korea, [2]Information Technology Institute, SmallSoft Co. Ltd., Daejeon 305-343, Korea)

In this paper, we have developed base-calling error detection program and algorithm which show the list of the genes or sequences that are suspected to contain base-calling errors. Those programs detect dubious bases in a few aspects in the process of microbial genome project. The first module detects base-calling error from the Phrap file by using contig assembly information. The second module analyzes frame shift mutation if it is originated from real mutation or artifact. Finally, in the case that there is control microbial genome annotation information, the third module extracts and shows the candidate base-calling error list by comparative genome analysis method.