제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

효율적인 서열 분석을 위한 sequence finishing program의 구현

문상훈,정우철,김진
한림대학교 컴퓨터공학과

e-mail: shmoon@center.cie.hallym.ac.kr,
wcjung@center.cie.hallym.ac.kr,
jinkim@hallym.ac.kr

Implementation of Sequence Finishing Program for Efficient Sequence Analysis

Sang-Hoon Moon,
Woo-Cheol Jung, Jin Kim
Dept of Computer Science, Hallym University

요 약

Automated sequencer로부터 얻게된 서열은 PCR이나 sequencing의 영향 등으로 기존의 자료 또는 분자생물학자가 원하는 서열과는 조금씩 차이점을 나타내게 되고, 이를 보정하기 위해 수작업으로 처리하게 된다. 이는 아주 간단함 작업임에도 불구하고 30분에서 1시간, 많게는 몇 시간씩 걸리는 불편함을 감수하고 있다. 본 논문에서는 분자 생물학자들이 효율적으로 서열을 분석하게 하는 sequence finishing program의 구현에 관하여 논의하였다.

1. 서론

DNA의 염기서열을 알아내는 것은 분자생물학 연구에 있어서 가장 기본적인 정보를 제공해주는 중요한 실험이다. DNA의 sequencing은 1977년에 이르러 거의 동시에 개발된 두 가지 방법에 의해 가능하게 되었다. 그 두 가지 방법은 F. Sanger와 A. R. coulson이 개발한 chain termination 방법과 A. Maxam과 W. Gilbert의 chemical degradation 방법이다. 현재는 여러 가지 장점으로 chain termination 방법을 선호하게 되었다. Chain termination 방법이 일반적인 실험실에서 많이 사용되며, 현재는 여러 종류의 automatic sequencer가 시판되고 있고 훨씬 편하고 빠르고 정확하게 염기서열을 분석할 수 있다. 이러한 방법을 통해 얻게된 서열은 PCR이나 sequencing의 오차 등으로 인해 기존의 자료 또는 분자생물학자가 원하는 서열과는 조금씩 차이점을 나타내게 된다. 따라서 이들을 보정하기 위해 분자생물학자들을 기존 데이터베이스의 서열과 비교해서 얻은 서열들을 pairwise alignment해서 보정하는데, 여기에 noise등이 발생하게 되면 pairwise alignment등으로도 그 차이점으로 구분하기 어렵게 되어 일일이 눈으로 비교하게된다. 이 작업들에서

† 본 연구는 정보통신부 정보통신 선도 기반 기술 개발 사업의 지원에 의하여 이루어진 것임

분자생물학자들은 여러 개의 온라인 또는 오프라인 프로그램을 사용하여 수작업으로 연결해서 사용하게 되는데, 아주 간단함 작업임에도 불구하고 몇 시간씩 걸리는 불편함을 감수하고 있다.

이와 같은 finishing 작업들을 도와주는 프로그램으로 Staden의 gap4 program과 Sequencher를 포함해서 DNA Star Seqman과 ABI AutoAssemble, 그리고 Consed라는 finishing tool이 개발되어 있으나, 이들은 상업적이거나 Unix 환경으로 개발되어있다. 그러나, 대다수의 분자생물학자들이 Windows 환경의 GUI에 보다 익숙하므로, 이들을 위한 보다 쉬운 프로그램의 개발이 필요하다. 본 논문에서는 효율적인 서열분석을 위한 sequence finishing program의 구현에 대해 논의하였다.

2. Sequence Finishing Program

Sequence finishing program에서의 주요 기능들은 automated sequencing을 통해서 얻은 .SCF 파일(standard chromatogram format)과 서열정보가 저장된 .txt 파일을 불러와서 base calling된 trace를 chromatogram을 통하여 보여 줄 수 있도록 구현되었다. 또한 trace 파일과 텍스트 파일의 서열 정보를 비교하여 서로 다른 부분을 표시하도록 하고 이를 보다 쉽게 편집하기 위해서 chromatogram 또는 텍스트 파일의 일부분을 드래그하면 같은 부분을 표시하고 편집할 수 있게 하였고, 편집 후에 프린터로 출력할 수 있도록 하고 있다.

3. 구현환경

본 시스템에서는 sequencing을 통해서 얻은 서열정보를 수정이 용이하고 데이터의 가공이 용이하게 한다. 웹 상의 데이터베이스를 통해서 서열정보를 얻을 수 있고 또한 텍스트 파일로 저장된 레퍼런스도 불러올 수 있는 기능을 가지고 있다. 이 프로그램은 Visual C++를 사용하여 윈도우즈 운영체제에서 사용할 수 있도록 개발하였다.

4. 기능

이 시스템의 주요 구성은 automated sequencer를 통해서 얻은 .SCF 파일과 텍스트 파일들을 불러오는 부분, 이를 서로 비교하는 부분과 비교를 통해 얻은 차이점을 수정하는 편집 부분, 그리고 이들을 서로 연결해 주는 assembly 부분과 마지막으로 그 결과를 chromatogram으로 출력하는 출력부분으로 구성되어

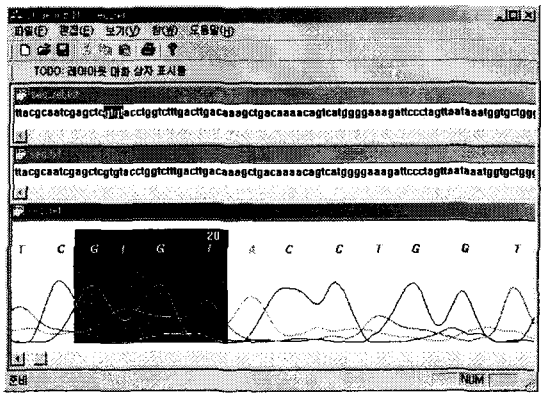

그림 1. Sequence finishing program

있다. 그림 1은 본 프로젝트에서 개발한 sequence finishing program을 보여준다.

4.1 텍스트 파일과 chromatogram 파일 불러오기

그림 1은 sequencing program의 전체를 보여준다. 실험을 통해서 얻은 텍스트 파일과 레퍼런스로 사용되는 텍스트 파일을 불러올 수 있고, 실험으로 얻은 chromatogram을 포함하고 있는 .SCF 파일을 불러 올 수 있다. 파일 불러오기는 *MDI form*을 사용하여 여러 개를 동시에 불러 올 수 있고 이를 사용하여 동시에 여러 파일을 볼 수 있고 편집도 가능하다.

4.2 Read Discrepancy의 발견

Automated sequencing을 통해 얻은 텍스트 파일과 기존의 레퍼런스 파일을 비교하여 서로 다른 부분을 붉은 선으로 표시하여 분자생물학자가 보다 알기 쉽게 표시해 주는 기능을 가지고 있다. 이를 바탕으로

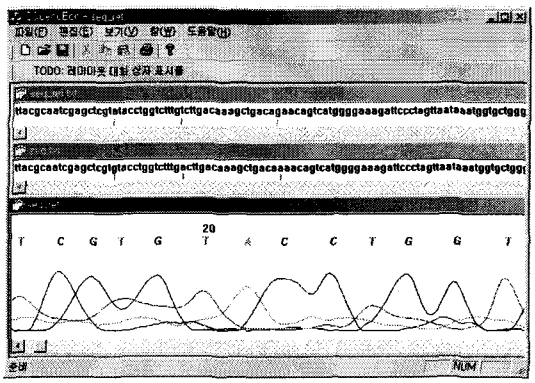

그림 2. Read discrepancy의 발견

하여 분자생물학자는 틀린 부분을 일일이 눈으로 확인하는 수작업 과정을 통하지 않고 보다 쉽게 발견함으로써 이를 발견하는데 소비되는 시간을 절약할 수 있다. 그림 2는 sequencing 오류로 인해 발생하는 read discrepancy를 발견하는 화면이다.

4.3 Editing

Read discrepancy를 발견한 후에, 그 부분의 trace들을 직접 분자 생물학자가 눈으로 확인할 수 있게 하였다. 그림 3은 서열 데이터와 trace를 직접 눈으로 확인할 수 있게 구현한 화면이다.

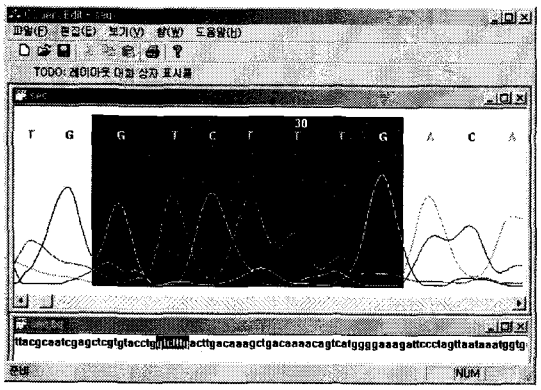

그림 3. 서열 데이터의 편집

아래의 서열과 대응되는 trace들을 검게 반전 시켜 표시함으로써, 원하는 서열과 데이터가 어떻게 틀린지 또는 automated sequencer가 만들어 낸 오류는 아닌지 직접 확인 할 수 있게 하였고, 이를 바탕으로 서열 데이터를 쉽게 편집할 수 있게 하였다.

4.4 Assembly

이렇게 수정된 약 500 ~ 700 bp 정도의 데이터들을 서로 연결할 수 있는 assemble 기능을 추가하였다. 각각의 contig들의 중첩되는 부분을 바탕으로 하여 분자 생물학자가 직접 눈으로 확인하여, 노이즈 또는 다른 부분들을 편집 기능을 사용하여 직접 제거함으로써 하나의 단일 서열로 assemble 하는 기능이다. 현재는 사람이 직접 눈으로 확인해서 수정하는 수작업을 사용하고 있다. 이는 정확하다는 장점이 있으나 프로그램이 추구하는 automation에는 부합하지 않으므로 기존에 개발된 assembly program을 참고하여 이를 자동화해 나갈 계획이다. 특히 TIGR에서 개발한 Assembler를 윈도우즈 환경에서 사용할 수 있도록 Integration할 계획이다. 그림 4는 본 프로그램을 사용하여 직접 assemble한 화면이다(그림 4에서는 보다 알기 쉽게 표현하기 위해 연결된 부분을 임의로 블록 설정하여 표현하였다).

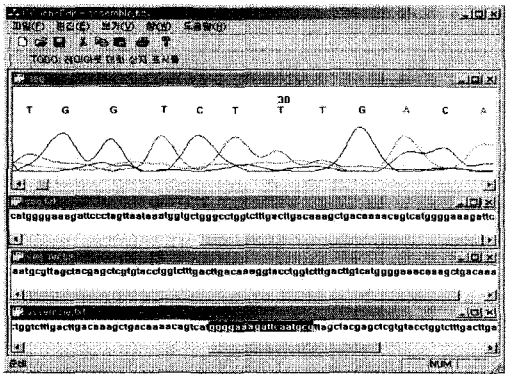

그림 4. Assemble 화면

4.5 결과 출력

컴퓨터 화면상에서 편집한 서열 정보를 가진 텍스트 파일과 trace 정보를 가지고 있는 chromatogram 파일을 출력하는 기능을 가지고 있다. 출력물에는 automated sequence에서 읽은 서열 정보와 *chromatogram* 정보와 읽어 들인 파일정보를 보여준다. 이를 출력하여 500bp가 넘는 분량의 서열 데이터를 쉽게 분석할 수도 있다.

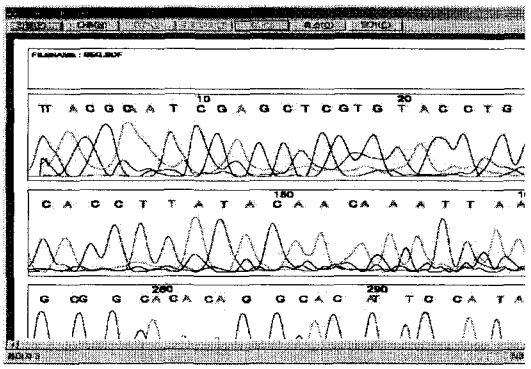

그림 5. chromatogram의 화면출력

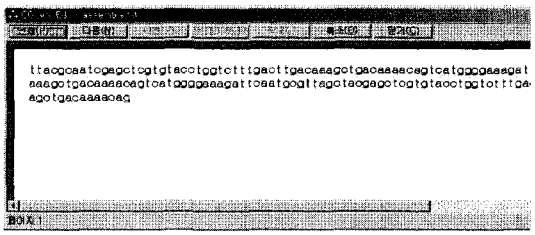

그림 6. 출력된 텍스트 파일 서열 데이터

출력물에서 trace 데이터는 여러 단으로 나누어 서열의 처음부터 마지막까지의 정보를 보여준다.

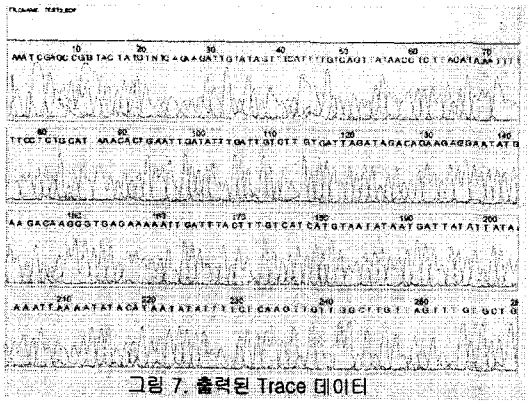

그림 7. 출력된 Trace 데이터

그림 5와 6은 화면에 출력된 서열 데이터 정보를 보여주며 그림 7은 직접 프린트된 출력물을 보여준다.

5. 결론

Sequence finishing이라는 작업은 서열 분석을 위해서는 꼭 필요한 작업이다. 모든 분자 생물학자들이 서열 분석을 위해서 이러한 과정을 거침에도 불구하고, 거의 모든 실험실이 수작업을 통해서 이를 분석하고 있는 실정이다. 외국에는 consed와 같은 여러 sequence finishing program이 개발되고 사용되고 있다. 그러나, 이들은 학교나 연구소를 제외한 기업에서는 유료로 제공되어 비용 문제를 야기하고 있고 이러한 프로그램들 또한 Unix 환경에서 구현되어 있어 컴퓨터 환경에 친숙하지 못한 일부 분자 생물학자들에게는 유명 무실한 프로그램이 되고 있다.

그러므로, 분자 생물학자들에게 보다 친숙한 프로그램을 무상으로 제공하고 서열 분석을 보다 더 효과적으로 분석할 수 있는 기회를 제공한다는데 의미가 있다.

그러나, 여러 개의 read들을 하나로 assemble하는 부분은 아직도 사람의 간섭을 필요로 하고 있고 입력 파일을 .SCF 파일만으로 한정하고 있으므로 이러한 점들을 보완해서 .AB1, .ZTR과 같은 여러 포맷들도 입력으로 받아들일 수 있게 하는 추가기능이 필요하다. 또한 앞으로는 primer 설계나 읽어들인 서열을 단백질 서열로 변환하는 기능, 그리고 vector 이미지를 그리는 기능들도 추가할 예정이다.

참고문헌

[1] http://www.dnai.co.kr

[2] 김용성, "Visual C++ 6 완벽 가이드", 영진출판사

[3] David Gordon, *et al*, 1998. Consed: A Graphical Tool for Sequence Finishing, Genome research: 195-202