우 수 상

1. 작품명

Fragment Assembly를 위한 Mater 패키지 제작

2. 소프트웨어명

Mater
미생물체의 전체 염기서열을 random shotgun approach를 이용하여 밝혀낼 때 몇 가지 전산학적인 문제들을 수반한다. 이를 해결하기 위해 테크닉이 필요하며 그 중에서도 서열간의 forward, reverse의 mating 정보를 이용하는 것이 매우 중요하다. Mater 프로그램은 이러한 mating 작업을 한 눈에 볼 수 있게 하는 소프트웨어 패키지이며 소프트웨어명은 mating에서 착안했다.

3. 제작자

제 작 자 : 김명선 (성신여자대학교 전산학과 4학년)
지도교수 : 박현석 (세종대학교 컴퓨터공학과)
주 소 : http://my.dreamwiz.com/keystore
e-mail : keystore@orgio.net

4. 요약설명

지금까지 인간이나 다른 생물체의 전체 유전체 염기서열을 밝혀내는 작업은 크게 세가지 방법으로 진행되었다. Clone-by-clone approach, sequence tagged connector approach, random shotgun approach[1]가 그것인데 마지막의 random shotgun approach는 fragment assembly problem을 비롯한 여러 가지 전산학적인 문제들을 수반한다.

Ⅰ. 서 론

Whole-Genome random shotgun 방법은 어떤 한 생물체의 유전자 염기 서열을 밝혀내는 genome-project에서 오늘날 매우 많이 사용되는 방법이다. 이는 기존의 다른 방법들보다 훨씬 빠른 시간 안에 매우 큰 유전체의 염기서열을 분석할 수 있는 방법이다. 그러나 여기에는 전산학적, 통계학적인 방

법들이 그 바탕에 깔려있다. 전산학적인 뒷받침 중 대표적인 것이 fragment assembly problem인데 이는 NP-Hard problem중 하나인 SCS-problem (Shortest Common Super-string problem)이다[2]. 이를 위해선 또 다른 전산학적 방법을 사용하는데 오차를 허용하는 pattern matching이 그것이다. 이러한 전산학적 문제에 대한 해결책은 생물정보학(Bioinformatics)이라는 새로운 학문 분야를 탄생시켜 활발히 연구중이다.

II . The random shotgun approach

Random shotgun 방식은 Frederick Sanger를 중심으로 고안된 길이가 긴 유전체의 염기서열을 밝혀내기 위한 한 방법이다. 이는 긴 유전체 염기서열을 임의의 위치에서 여러 개의 짧은 조각들로 나눈 후 그 조각들 사이에 중복되는 부분들을 이용하여 다시 원래의 긴 서열을 만들어내는 방법이다.
이런 방법으로 얻은 염기 서열 조각들은 다중 염기 서열 정렬 과정을 통해 몇 개의 긴 염기서열(contig)로 합쳐진다. 이 과정에서 두 염기 서열 단편간의 유사도(homology)는 PAM이나 BLOSUM같은 Scoring-matrix를 통해 계산된다.

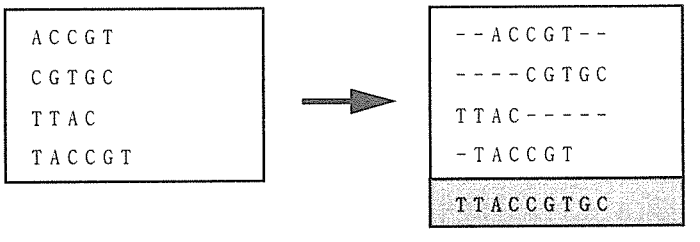



그림 1. 염기 서열 단편 정렬의 예 [3]
왼쪽이 각 염기 서열의 단편, 오른쪽 위가 다중 정렬(multiple alignment) 된 모습, 오른쪽 아래가 다중 정렬을 통해 얻은 하나의 긴 염기서열(contig)이다.

III . Random shotgun 방식에서 해결해야 할 어려운 점들

i ) 반복적인 염기서열

Random shotgun 방식은 유전체의 긴 염기서열을 매우 많은 수의 짧은 조각들로 나눈 후 그것들을 다시 하나의 긴 염기서열로 합친다. 그러나 생물체의 염기서열 중에는 일정 길이의 염기서열이 반복적으로 나오는 곳이 있다(repeated regions). 이런 경우엔 다음과 같이 염기 서열 조각들이 비정상적으로 결합하는 경우가 발생할 수 있다.

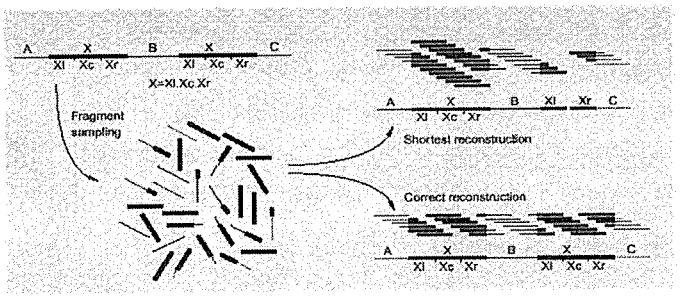



그림 2. repeated region[1]

단순히 가장 짧은 superstring을 생성하는 경우 잘못된 contig를 만들게 된다.

ii ) 불충분한 염기 서열 조각(Lack of coverage)

6-fold가 넘는 충분한 수의 염기 서열 단편들을 분석했더라도 여전히 서열 단편들이 만들어지지 않은 부분이 남아있게 된다(gap). 이러한 부분은 생물학적으로 cloning-vector로 단편들을 복제할 수 없는 부분(un-clonable sequence)이거나 몇 가지 기계적인 이유에 의해 해당 부분의 염기 서열 조각이 생성되지 않은 부분이다.

iii ) 서열 분석 에러

전자동 염기 서열 분석기를 통해 분석될 결과에도 에러가 있을 수 있다. 정교하게 조절된 실험 환경에서 처음 500bp정도 까지에서는 에러율이 약 1% 미만의 확률로 일어나지만 그 이후에는 급속도로 증가하여 650bp 정도에서부터는 에러율이 15%가 넘어가기도 한다. 각 염기 서열 조각들의 끝 부분끼리의 유사성을 통해 하나의 긴 컨티그를 만들어 가는 염기 서열 조각 재결합 문제(fragment assembly problem)에서 각 조각들의 끝부분에 에러가 많다면 잘못된 컨티그를 생성할 가능성이 매우 커지게 된다.

iv ) 방향성

DNA염기서열은 서로 상보적인 두 가닥의 서열이 나선형 구조를 이루고 있어서 분석한 염기 서열 조각의 실제 방향이 어떤 것인지 알아내기가 매우 어렵다. 이런 이유로 여러 개의 염기 서열 조각을 하나의 긴 컨티그(contig)로 재결합할 때 각각의 서열 조각들끼리의 유사성 검사를 최소 두 번씩 해야하는데 이는 문제의 시간 복잡도를 증가시키게 된다.

V) 키메라(chimera)

마지막으로 서로 멀리 떨어진 곳에 있는 두 개 이상의 짧은 서열 조각이 서로 결합하여 마치 하나의 서열 단편처럼 보이게 되는 경우도 생긴다.(chimeric fragment).

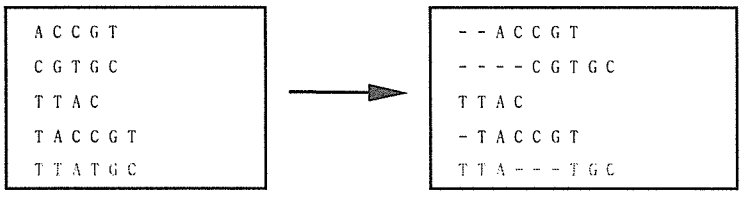



그림 3. 비정상적 염기 서열 단편(chimeric fragment)[4]

가장 아랫쪽에 있는 염기 서열 조각의 경우 서로 멀리 떨어진 두 염기단편의 일부분이 하나로 합쳐진 것임을 알 수 있다.

이러한 잘못된 서열 조각들에 대한 고려 없이 일반적인 fragment assembly problem에 대한 알고리즘을 적용한다면 정체 불명의 생물체가 만들어질 것이다.

IV. Mating 작업

위에서 말한 Whole-Genome random shotgun approach[5]에서 선결해야 할 몇 가지 문제에 대한 해결방안으로 각 염기 서열조각을 다른 것과 짝을 이루도록 하는 것이 있다.

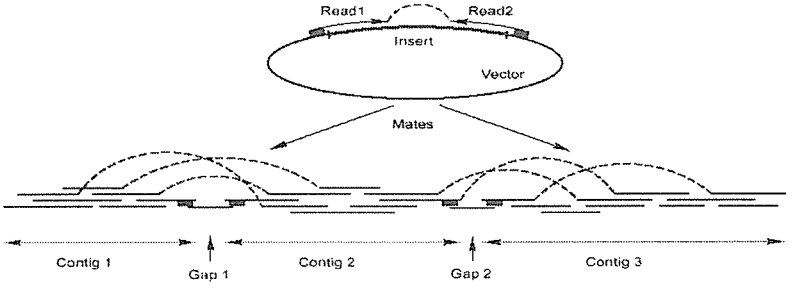



그림 4. mate[1]

cloning vector에 삽입된 유전자 조각(clone)의 염기 서열을 양쪽에서부터 읽어나가면 동일한 유전자 조각에 대한 염기 서열 분석 결과물이 두 개씩

짝을 이루게 된다.

(그림 4)에서와 같이 각 염기서열 조각들이 두 개씩 짝을 이루게 되면 그것을 통해 많은 정보를 추가적으로 알 수 있게 된다.
먼저 짝을 이루는 두 개의 염기 서열 조각 사이의 대략적인 거리를 알 수 있다. 각 유전자 조각들은 길이가 2k나 10k가 되도록 하였으므로, 짝을 이루는 두 염기서열 조각의 길이를 각각 *l1,l2*라 했을 때 그 둘 사이의 거리는 약 2000-(*l1+l2*) 혹은 10000-(*l1+l2*) 가 된다. 직관적으로 이러한 거리 정보는 일정 염기 서열이 반복적으로 나타나는 부분(repeated region)을 재구성할 때 이용될 수 있다. 즉, 만약 여러 개의 염기 서열 조각으로 하나의 긴 컨티그를 만들었을 때, 짝을 이루는 두 개의 염기 서열 조각 사이의 거리가 계산치보다 더 짧다면 이는 잘못 구성된 컨티그라 할 수 있다. 이런 경우 두 염기 서열 조각 사이의 거리를 계산치 만큼 늘여 놓으면 반복적으로 나타나는 염기 서열 부분이 원래의 그것과 같이 정상적으로 만들어질 수 있다. 컨티그들 사이의 순서 정보도 짝을 이루는 두 염기 서열 조각을 통해 얻을 수 있다. Random shotgun 방법으로 유전체에 대한 염기 서열 정보를 얻으려 할 때 부딪히는 어려운 점인 불충분한 염기 서열 데이터(Lack of coverage)로 인해 6-fold이상의 충분한 염기 서열 조각을 분석해도 일부분은 여전히 분석되지 않고 남아있게 된다. 이러한 갭(gap)부분은 별도의 방법을 통해서 메꿔야 하는데, 이 때 각 컨티그들의 순서 정보가 매우 중요하게 쓰인다. (그림 4)에서 볼 수 있듯이 컨티그들의 순서는 짝을 이루는 두 염기 서열 조각을 통해 얻어낼 수 있다.

참고 문헌

[1]. Gene Myers, "Whole-Genome DNA Sequencing"

[2]. D. Gusfield, Algorithms on strings, trees, and sequences - Computer Science and Computational Biology, 1997.

[3]. J. Meidanis / J. C. Setubal, "Introduction to Computational Molecular Biology", pp.107-108.

[4]. J. Meidanis / J. C. Setubal, "Introduction to Computational Molecular Biology", pp.108-109.

[5]. R.D. Fleischmann *et al.*, "Whole-Genome Random Sequencing and Assembly of *H.Influenzae*," Science, Vol. 269, No. 5,223, 1995, pp.496-512.

(1) 작품 설명

Mater 프로그램은 이러한 mating 작업을 한 눈에 볼 수 있게 하는 소프트웨어 패키지이다.

(2) 개발배경

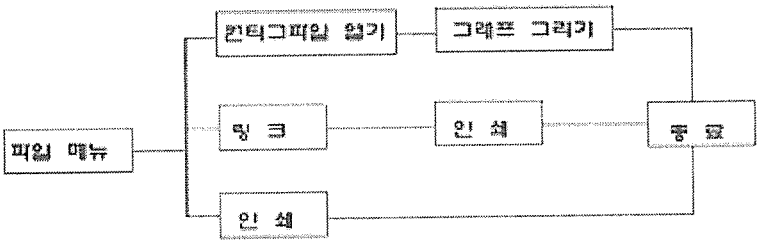
서울의대 유전자이식연구소를 모태로 창업한 바이오 벤처기업 '마크로젠(http://www.macrogen.com)'의 바이오 인포매틱스(Bioinformatics) 사업은 세종대 컴퓨터공학과 박현석 교수(본 프로그램의 지도교수)의 담당하에 이루어지고 있다. 지금까지 각 염기 서열조각을 다른 것과 짝을 이루도록 하는 메이팅(mating) 작업은 사람의 손을 통해 일일이 이루어져 많은 시간이 걸리고 있다. Mater는 작업시간을 단축하고 사람이 해야 할 작업을 줄이기 위해 개발되었다.

(3) 작품개요

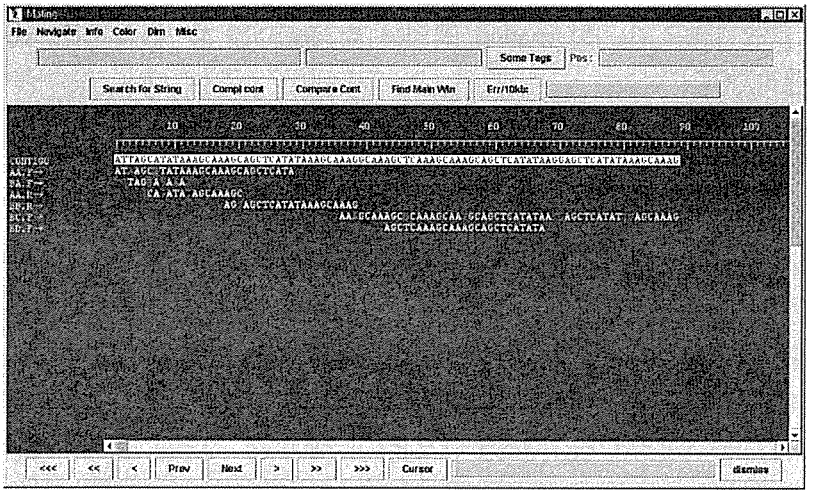
1) 특징

Random shotgun 방식은 Frederick Sanger를 중심으로 고안된 길이가 긴 유전체의 염기서열을 밝혀내기 위한 한 방법이다. 이는 긴 유전체 염기서열을 임의의 위치에서 여러 개의 짧은 조각들(read)로 나눈 후 그 조각들 사이에 중복되는 부분들을 이용하여 다시 원래의 긴 서열(contig)을 만들어내는 방법이다. 이런 방법으로 얻은 염기 서열 조각들은 다중 염기 서열 정렬 과정을 통해 몇 개의 긴 염기서열(contig)로 합쳐진다. 이 모든 일련의 과정들을 그래픽 화면으로 구성하였다.

2) 구성과 기능



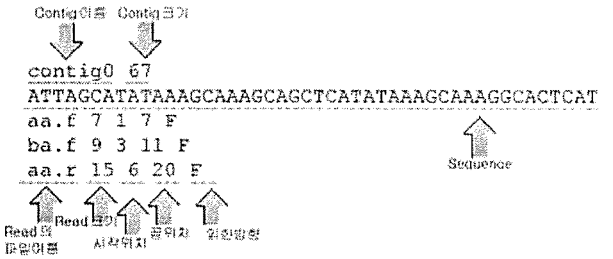

[그림 1] 프로그램의 구성




〔그림 2〕 시작 모듈

① 컨티그파일 열기 (Open Contig)

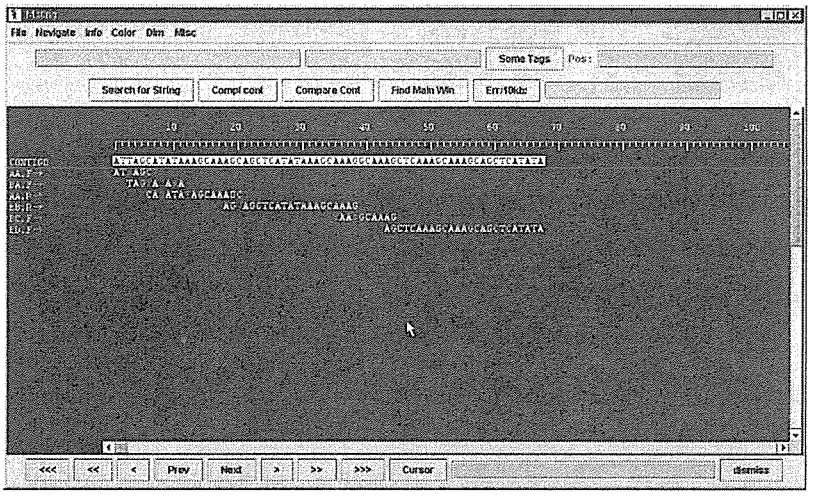
염기서열의 단편(read)들을 다중정렬(multiple alignment)하고 이를 통해 얻은 하나의 긴 염기서열(contig)을 화면에 출력한다. open contig의 입력은 확장자가 seq(sequence)인 파일을 사용하며 텍스트 파일(text file)로 일반 메모장 같은 프로그램을 사용하여 보거나 저장할 수 있다. open contig는 텍스트 파일(text file)로 이미 저장된 데이터를 읽어들여 화면에 출력한다. 컨티그 파일명은 컨티그 이름(Contig + 컨티그번호)으로 구성되며 입력 파일의 구성 다음과 같다.




〔그림 3〕 컨티그 파일 (contig0.seq)의 구성

리드파일은 파일명, 파일길이, 첫 번째 인자의 시작위치, 마지막 인자의 위치, 그리고 파일이 읽힌 방향 데이터를 가지고 있으며 실제 데이터는 현재 작업디렉토리 안의 Data디렉토리 안에 파일명과 동일한 파일이 있으며 그 파일 안에는 하나의 문자열로 구성된 염기서열의 단편(read)이 저장되어 있다. [그림 4]의 aa.f, ba.f, 그리고 aa.r은 각 파일을 표시하기 위해 쓰여졌을 뿐 파일엔 없는 부분이다. 파일 안에는 노란 부분의 리드(read) 데이터만이 있다. 리드 데이터는 염기를 구성하는 아데닌(A, Adenine), 티민(T, Thymine), 구아닌(G, Guanine) 그리고 시토신(C, Cytosine)의 각 첫 자로 이루어진 문자열이다.

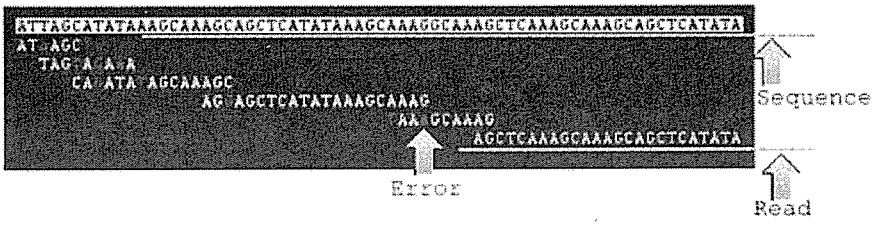

[그림 4] 리드파일 (*.f or *.r)의 구성



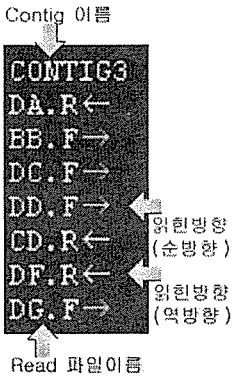

[그림 5] 컨티그 파일 출력 모듈

사용자가 입력파일을 선택하면 시퀀스(sequence)의 위치를 쉽게 알 수 있도록 숫자가 표현된 눈금자를 출력한 후 시퀀스를 화면에 출력하고 입력파일 안에 있는 리드 파일을 찾아 데이터를 읽어 들인 후 리드를 시퀀스에서의

위치에 맞게 정렬하여 출력한다. 이 때 시퀀스와 일치하지 않는 염기는 에러(error)의 의미로 붉은색으로 표현한다.




〔그림 6〕 시퀀스(sequence)와 리드(read)



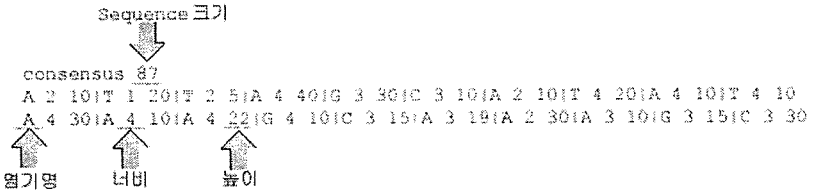

화면의 좌측엔 열린 컨티그(contig)의 이름과 각 리드(read)들의 파일명과 파일이 어떤 방식으로 읽혀졌는지에 대한 정보가 출력되어지는데 → 경우에는 파일의 시작에서부터 순차적으로 읽은 경우이고 ← 경우에는 파일의 끝에서부터 순차적(거꾸로)으로 읽은 것을 표현한 것이다.

〔그림 7〕 리드파일 이름과 파일이 읽혀진 방향

② 그래프 그리기

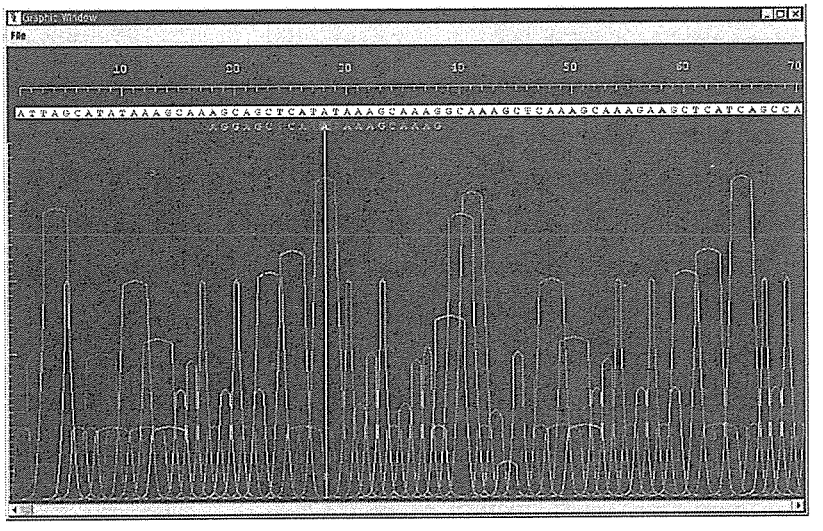
Open Contig 메뉴가 선택되어진 상태에서 화면에 나타난 리드(각 문자열) 중 하나를 선택한 후 특정 염기(한 문자)에서 오른쪽 마우스를 클릭했을 때 보여지는 화면이다.

그래프 입력파일은 고정 문자열인 consensus, 시퀀스 크기, 시퀀스를 구성하고 있는 각각의 염기에 대한 염기명과 그래프를 그리기 위한 너비, 그리고 높이에 대한 데이터를 가지고 있다. 하나의 염기에서의 구분자는 공백문자이고 염기와 염기의 구분자는 |이다. 그래프 파일은 컨티그 파일마다 가지며 컨티그 파일 안의 시퀀스에 대한 그래프 정보를 표현하고 그 위치는 현재작업디렉토리 아래 Graph 디렉토리이다. 그래프 파일명은 컨티그 파일명과 동일하고 확장자는 grp(graph)를 사용한다. grp파일의 구성은 다음과 같다.




〔그림 8〕 그래프 파일 (*.grp)의 구성

시퀀스를 구성하는 각각의 염기에 대해 정규분포곡선을 화면에 출력하는데 Open Contig 메뉴에서 열린 기본 화면에 그래프가 추가되어 출력된다. 선택되어진 염기임을 표시하기 위해 붉은 점으로 대칭축을 흰 라인으로 표시하였다.

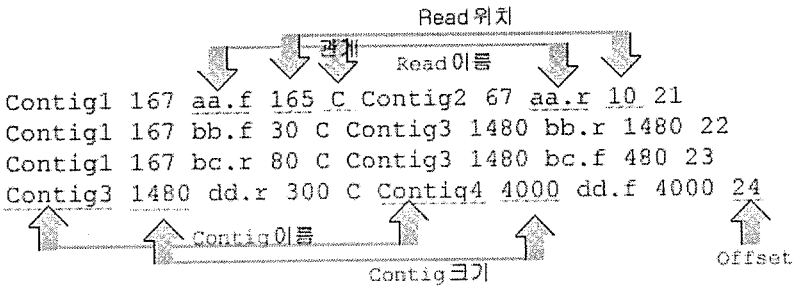



〔그림 9〕 그래프 출력 모듈

③ 링크 (Link)

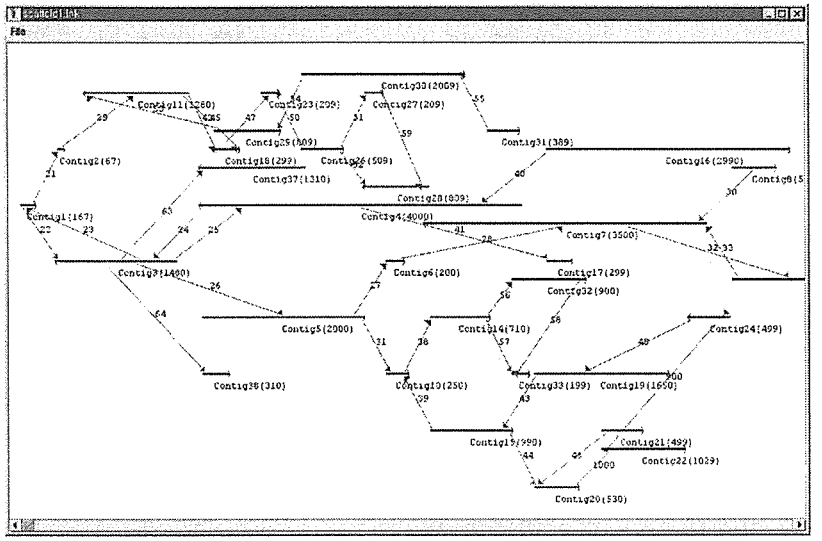
한 생물체의 전체 유전자 서열을 Random Shotgun 방식으로 밝히는데 있어서 생물학적인 이유에서나 실험적인 이유에서 밝혀지지 않는 부분들이 존재할 수 있다 이러한 부분에 의해서 Random Shotgun방식으로 접근했을 때

하나의 전체 서열로 나오는 것이 아니라 여러 개의 긴 염기서열(Contig)로 결과가 나오게 된다. 이 때 Assembler에 의해 나오게 되는 여러 가지 정보를 이용하여 이들 컨티그에 존재하는 같은 클론(Clone)에서 나온 fragment(mate)정보를 활용하게 된다.
긴 유전체 염기서열을 임의의 위치에서 여러 개의 짧은 조각들로 나눈 후 그 조각들 사이에 중복되는 부분들을 이용하여 다시 원래의 긴 서열을 만들어내는데 링크는 각 염기서열조각을 다른 것과 짝을 이루도록 하는 것이다. 링크 파일명은 scaffold + 파일번호로 구성되어지며 확장자는 lnk(link)를 사용한다. lnk파일의 구성은 다음과 같다.




〔그림 10〕 링크 파일(scaffold1.lnk)의 구성

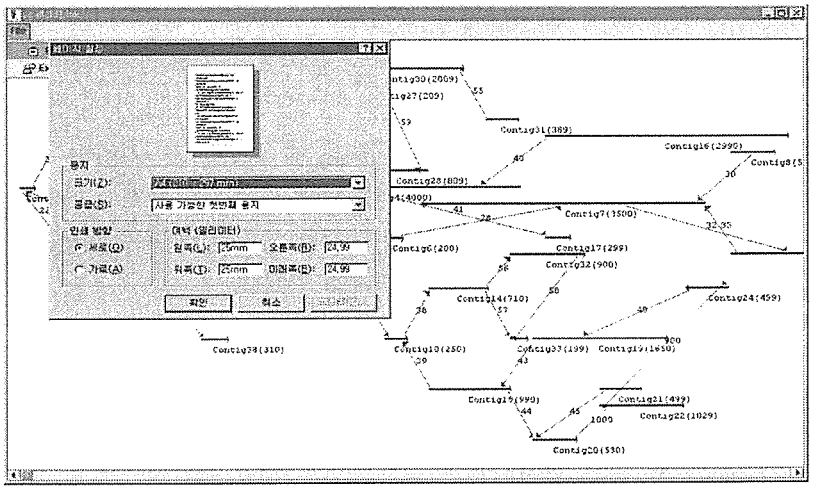
링크 입력파일은 메이트(mate, 서로 연관성이 있는 두 염기서열(contig))를 표현하기 위한 것으로 메이트는 두 개의 컨티그 파일과 그것들의 정보로 구성된다. 각각의 데이터를 분리하기 위한 분리자로 공백문자를 사용하며 한 라인에 하나의 메이트 정보를 표시한다. 파일의 구성은 컨티그 파일명, 컨티그 크기, 리드 이름, 리드 위치, 두 컨티그의 관계, 컨티그 파일명, 컨티그 크기, 리드 이름, 리드 위치, 그리고 offset이다.




〔그림 11〕 링크 모듈

④ 출력 (Print)

현재 화면을 프린터를 통해 출력한다. 사용자는 인쇄용지의 방향을 선택할 수 있으며 선택사항에 대한 입력이 끝나면 현재 화면이 프린터를 통해 출력된다.



〔그림 12〕 링크 출력 모듈

3) 효과

Fragment assembly 작업을 할 때 생물체의 유전체를 알아내는 것은 매우 중요하며 인간이 현미경과 시험관만으로 유전정보를 분석하는 것은 많은 어려움이 따른다. 이를 위해 컴퓨터를 이용해 유전정보를 분석하는 생물정보학(Bioinformatics)이 인간 게놈프로젝트 바람을 타고 생명공학 연구의 주류로 등장하고 있다.

게놈연구란 한 생물체가 지닌 모든 유전자 정보 지도를 그리는 작업을 말한다. 인간의 경우 세포핵에 이중 나선형으로 꼬여 있는 23쌍, 46개의 염색체에 모든 유전정보가 담겨 있다. 유전정보를 담고 있는 물질이 DNA(디옥시리보핵산)이며, DNA는 A(Adenine, 아데닌), G(Guanine, 구아닌), C(Cytosine, 시토신), T(Thymine, 티민) 등 네 가지 염기의 다양한 조합으로 이루어져 있다. 인간 게놈프로젝트의 목표는 30억 개에 이르는 인간의 염기배열 구조를 밝혀내는 것이다. 즉, 어느 염색체, 어떤 염기 서열에, 어떤 유전정보를 가진 염기가 존재하는지를 밝혀내는 작업이다.

게놈연구 결과물은 단순한 염기서열을 밝혀낸 것일 뿐, 그 자체로서는 기능이나 의미는 알 수가 없다. 이 때문에 생물정보학(Bioinformatics)이 급부상하고 있다. 생물정보학은 컴퓨터기록이 0과 1의 조합인 것과 마찬가지로 유전정보도 A, G, C, T의 네가지 염기로 이루어진 점을 이용, 컴퓨터 모의실험으로 염기서열의 역할을 추정하거나 염기들의 유형을 분석해 낸다. 포스트게놈 연구방향은 기능 유전체연구와 비교 유전체연구가 활발해질 전망이다. 기능 유전체연구는 특정 유전자의 기능을 연구, 유전자 질병의 원인을 규명해내는 과정이다. 포스트게놈으로 달라지는 삶은 우선 유전자 정보가 공개돼 유전자조작기술이 발전, 앞으로 맞춤형 의약품 산업시대가 도래할 것이다. 이와 함께 수 만개 유전자 정보를 고속으로 처리하는 기술이 실용화돼 기존 진단법보다 훨씬 뛰어난 유전자 진단법이 나올 전망이다. Mater 프로그램은 이와 같은 게놈 전체를 밝히는데 유용한 소프트웨어 툴로써 휴먼 프로젝트의 레벨을 한단계 상승시키는데 기여할 것이다.