Journal of Korea Multimedia Society Vol. 7, No. 3, March 2004(pp. 408-417)

웹 기반 통합 유전체 분석 시스템의 설계 및 구현

최범순[†], 이경희[††], 권해룡[†††], 조완섭[††††], 이충세[†††††], 김영창[††††††]

요 약

유전체 분석 과정은 여러 단계를 걸쳐 다양한 소프트웨어 분석 도구가 사용되는 복잡한 작업을 수반한다. 유전체 분석과 관련된 기존의 소프트웨어 도구들은 대부분 리눅스나 유닉스 기반 프로그램이므로 생물학자가 이들을 설치하고 사용하는데 어려움과 불편함이 많은 실정이다. 또한, 분석의 각 단계별로 생산되는 파일은 수작업을 통한 변환을 거쳐야만 다음 단계의 입력으로 사용될 수 있다. 근래에 웹을 기반으로한 도구들이 개발되고 있으나 한번에 하나의 서열을 처리하는 방식이므로 대량의 실험 데이터를 분석하는 경우에는 반복 작업으로 인한 시간과 노력이 요구되는 단점을 갖고 있다. 본 논문에서는 유전체 분석에 필요한 여러 도구들을 웹 환경에서 하나의 그래픽 사용자 인터페이스로 통합하여 생물학자들이 보다 쉽게 서열과 기능을 분석할 수 있도록 한 *WGAT(Web-based Genome Analysis Tool)*를 제안한다. WGAT는 리눅스 서버에 유전체 데이터 분석 프로그램을 구동하고, 클라이언트 웹(web)에서 데이터 파일과 분석에 필요한 선택사항들을 입력함으로써 한번에 여러 단계의 분석 작업을 수작업 없이 자동으로 처리할 수 있다. WGAT 시스템의 생산성을 분석하기 위하여 기존의 방식과 WGAT를 이용한 방식의 서열분석 처리 시간을 비교한 결과 서열 단편의 개수가 1000개인 경우 기존의 방식보다 20배 이상 분석 능력이 향상됨을 확인할 수 있었다.

The Design and Implementation of Web-Based Integrated Genome Analysis Tools

Beom-Soon Choi[†], Kyoung-Hee Lee[††], Hae-Ryong Kwon[†††], Wan-Seop Cho[††††], Chung-Sei Rhee[†††††], Young-Chang Kim[††††††]

ABSTRACT

Genome analysis process requires several steps of various software analysis tools. We propose WGAT(Web-based Genome Analysis Tool), which combines several tools for gene analysis and provides a graphic user interface for users. Software tools related to gene analysis are based on Linux or Unix oriented program, which is difficult to install and use for biologists. Furthermore, files generated from gene analysis frequently require manual transformation for next step input file. Web-based tools which are recently developed process only one sequence at a time. So it needs many repetitive processes to analyze large size data file. WGAT is developed to support Web-based genome analysis for easy use as well as fast service for users. Whole genome data analysis can be done by running WGAT on Linux server and giving sequence data files with various options. Therefore many steps of the analysis can be done automatically by the system. Simulation shows that WGAT method gives 20 times faster analysis when sequence segment is one thousand.

Key words: genome(유전체), gene information(유전정보), sequence analysis(서열 분석), integrated analysis system(통합 분석 시스템)

※ 교신저자(Corresponding Author) : 김영창, 주소 : 충북 청주시 흥덕구 개신동 산 48번지(361-763), 전화 : 043)261-2302, FAX : 043)268-2538
E-mail : youngkim@chungbuk.ac.kr
접수일 : 2003년 6월 3일, 완료일 : 2003년 8월 2일
[†] 준회원, 충북대학교 미생물학과 대학원
(E-mail : bschoi319@korea.com)
[††] 충북대학교 컴퓨터과학과 대학원
(E-mail : khlee@chungbuk.ac.kr)
[†††] 충북대학교 미생물학과 대학원
(E-mail : mrjesus@just.chungbuk.ac.kr)
[††††] 충북대학교 경영정보학과 조교수
(E-mail : wscho@chungbuk.ac.kr)
[†††††] 충북대학교 컴퓨터과학과 교수
(E-mail : csrhee@chungbuk.ac.kr)
[††††††] 충북대학교 생명과학부 교수, 바이오 연구소
※ 본 연구는 한국과학재단 목적기초연구(R01-2001-00097)와 학술진흥재단 기초과학연구(DS0031)지원으로 수행되었음.

1. 서 론

인간의 유전체 서열과 기능을 알아내기 위한 인간 유전체 프로젝트(Human Genome Project, HGP)는 국제협력사업으로 1990년에 시작되어 1999년에 완료되었다. 현재 바이러스, 식물, 동물등 지구상에 존재하는 생물체의 유전체 사업이 전 세계적으로 진행되고 있으며, 1995년 *Haemophilus influenzae*의 전체 유전체 서열 결정[1]을 시작으로 각종 생물체에 대한 유전체 사업이 가속화되고 있다. 2003년 5월 현재 239종의 생물체 유전체 사업이 완료되었으며, 577종이 진행 중에 있다.

유전체 사업의 궁극적 목표는 생물체의 유전체 서열을 밝혀내고 유전자를 찾아내어 유전자의 기능을 알아내는 것이다. INSD(International Nucleotide Sequence Databases)에는 이미 밝혀진 유전체 서열 데이터와 그 중 기능이 알려진 유전자 및 그의 기능에 관한 정보가 공개되어 있다. 생물학자들은 자신이 밝혀낸 서열에 대하여 이미 알려진 서열 중 유사한 것이 있는지, 어떤 종의 서열과 유사하며, 밝혀진 유전자 기능이 있는지 등을 확인하고자 한다. 생물학자들은 이러한 정보를 바탕으로 다음 단계의 연구를 진행하거나 새로운 연구를 시작할 수 있기 때문이다.

유전자 확인 및 기능 분석 작업은 관련된 데이터 양과 비교할 대상이 너무 많아서 수작업으로 하기엔 많은 시간과 노력이 필요하다. 주어진 유전체 서열에 대하여 기능 분석까지 완료하려면 (그림 1)에서 보는 바와 같이 5단계를 거쳐야 하며, 각 단계마다 다양한 프로그램과 데이터가 사용되고, 각 프로그램에서 입력으로 받아들이는 파일 형식이 상이하므로 수작업을 통한 데이터 변환이 필요하다.

생명 정보학(bioinformatics)은 컴퓨터 정보 기술을 이용하여 생물 정보 처리를 자동화하여 효율성을

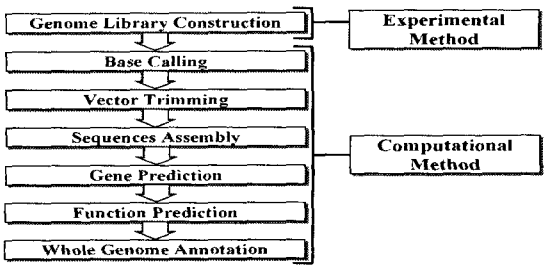


그림 1. 유전체 분석의 단계

증대시키는 것을 목적으로 하며, 포스트 지놈시대에 대규모의 생물 정보를 신속하게 처리하는데 필수적인 도구로 인식되고 있다. 실제로 생물 정보학의 여러 세부 분야 중에서 서열을 분석하여 유전자 기능을 예측하는 문제를 다루는 여러 가지 소프트웨어 도구와 DB가 개발되어 유전체 연구에서 핵심적인 역할을 수행해내고 있다.

기존의 유전자 서열 분석관련 도구들은 대부분 리눅스나 유닉스 기반의 프로그램이기 때문에 관련 프로그램을 이용하려면 프로그램을 서버에 설치해야 하고 명령어 사용법과 각종 옵션을 숙지해야 원하는 결과물을 얻어낼 수 있다. 특히, 유전자 분석 업무는 (그림 1)에서 보는 바와 같이 여러 단계를 거쳐 이루어지며, 각 단계에서 다양한 프로그램과 다양한 파일 형식 및 옵션을 사용해야 하기 때문에 컴퓨터 비전문가인 생물학자들의 불편함이 가중되고 있다. 최근 들어 웹으로 제공되는 분석 도구도 개발되고 있으나 대량의 서열을 분석하는 기능은 제공되지 않기 때문에 원하는 결과물을 얻기 위해서는 오랜 시간이 걸린다[2]. 따라서 윈도우 환경에서 유전자 기능 예측의 전체 과정을 편리하게 수행할 수 있는 통합 도구의 개발이 필요한 실정이다.

본 논문에서는 유전자 분석의 각 단계에서 필요한 분석 도구들을 웹 기반 GUI 환경에서 통합한 시스템인 WGAT (Web-based Genome Analysis Tool)를 제안한다. WGAT 시스템은 웹 기반으로 제공되기 때문에 윈도우 기반의 컴퓨터 환경에 익숙한 생물학자들이 보다 쉽게 사용할 수 있다. 즉, (그림 1)에서 나타난 5단계를 수작업이나 파일 변환 없이 웹 기반의 GUI 환경에서 편리하게 실행하여 생물학자들의 유전자 분석 효율성을 증진시킬 수 있다.

WGAT 시스템을 사용하여 유전자 분석을 실제로 수행한 결과 1000개의 유전자 단편을 분석할 때 기존의 통합되지 않은 도구들을 사용하는 방식의 경우보다 20배 이상 분석 능력이 향상됨을 확인할 수 있었다. 이러한 결과는 유전자 분석의 여러 단계들을 수작업 없이 자동으로 처리할 수 있도록 했다는 점과 대량의 서열을 한꺼번에 읽어서 처리한다는 점에서 기존의 방식보다 개선된 점을 보여준다. 물론, 사용자 편의성 측면에서도 웹 환경에서 그래픽 사용자 인터페이스를 제공함으로서 기능이 개선되었다.

논문의 구성은 다음과 같다. 제 2장에서 유전체

연구의 일반적인 방법과 관련 프로그램 및 기존의 통합 시스템에 대해서 설명한다. 제 3장에서는 유전체 연구에 필요한 프로그램들을 웹 환경의 GUI에서 하나로 통합하여 구축하는 방법에 대해서 설명한다. 제 4장에서는 본 연구를 통하여 구축한 시스템의 성능을 살펴보고, 제 5장에서 결론 및 향후 과제를 제시한다.

2. 관련연구

여기서는 유전체 분석 과정을 각 단계별로 살펴보고, 유전자 기능 예측에 사용되는 다양한 소프트웨어 도구들을 소개한다. 그리고 기존의 도구들을 특정한 목적으로 통합한 통합 시스템에 관해서 살펴본다.

2.1 일반적인 유전체 분석 과정

유전체 분석 작업은 다음과 같은 염기서열의 해독 과정과 유전자 발견 및 기능 예측 과정을 거쳐 이루어진다.

염기서열 해독과정은 다음과 같다. 먼저, 생물체로부터 염색체 DNA를 추출하고, 추출한 염색체 DNA를 제한효소 또는 초음파 분쇄기를 이용하여 작은 단편으로 제작한 후, 이를 벡터에 삽입하여 유전체 서열 목록을 제작한다. 유전체 목록을 50~100kb 정도의 cosmid, BAC 라이브러리로 제작하고, 이를 다시 1~2kb 정도의 작은 단편의 서열 목록을 제작한다. 다음으로, 제작된 유전체 서열 목록의 양쪽 말단 부분은 자동 서열분석기를 이용하여 해독한다. 전체 유전체 크기의 7~10배 정도를 해독하면 중복된 부분이 생기며 이 중복된 부위를 찾아 재배열함으로서 긴 컨티그(contigs)를 제작할 수 있다. 이 과정에서 두 컨티그 사이에 갭(gap)이 생기게 되는데 갭 클로져(gap closure) 과정을 통해 모든 갭을 채우고 나면 거대한 하나의 전체 염기 서열이 되고 비로소 전체 염기서열 해독이 끝나게 되는 것이다[3].

다음으로, 유전자의 발견 및 기능 예측 과정은 다음과 같다. 염기해독이 끝나면 유전체 서열에서 유전자를 동정(identification)하고 기능을 예측하는 작업인 주해(annotation) 작업을 수행한다. Glimmer나 ORF finder 프로그램을 이용하여 유전자를 예측하게 되며, 예측된 유전자로부터 상동성 검색(homology search)을 이용하여 그 기능을 예측하게 된다[4,5]. 상동성 검색의 도구로서 가장 대표적인 BLAST를 이용하여 기존의 유사한 유전자중 기능이 알려진 것을 찾아 동정된 ORFs의 기능을 결정한다[6-8].

2.2 유전자 기능 예측 도구

유전자의 기능을 예측하기 위해서는 DNA 서열로부터 다중 컨티그(multi-contig) 파일을 생성하는 제 1단계(그림 2)와 다중 컨티그 파일로부터 기능을 예측하는 제 2단계(그림 3)로 구분한다.

제 1단계에서 컨티그 제작은 (그림 2)에서 보는 바와 같이 Phred-Phrap 프로그램을 이용하여 이루어진다. (그림 2)의 각 단계는 다음과 같다.

① Phred는 염기서열을 읽어들이는 작업(base calling)을 수행하는 프로그램으로써 자동 서열분석기가 만들어낸 이미지 파일을 읽어서 FASTA 형태의 파일로 변환하여준다. 즉, 그래프 이미지 내의 크로마토그램(chromatogram)의 곡선을 읽어서 적당한 염기를 지정함과 동시에 그 염기에 대한 정확성을 점수로 나타내어 파일로 제공한다. Phred가 만들어내는 정확도를 QV (Quality Value)라고 정의하며, 다음과 같이 계산한다.

$$QV = -10 \log P_e \qquad \text{(식 1)}$$

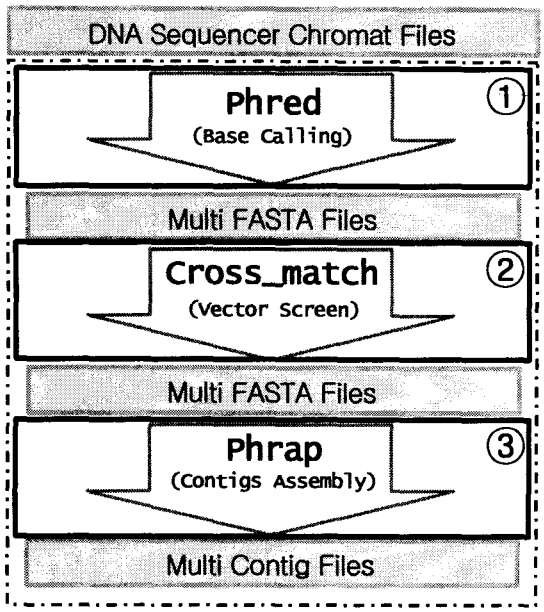


그림 2. 컨티그 형성 과정

P_e는 Probability Error의 약자이며 염기서열을 읽어 들일 때 에러가 날 확률을 말한다.

일반적으로 QV가 40 정도면 99.99%의 시퀀싱 정확도를 갖는다.

② Cross_match 프로그램은 이미 서열이 밝혀진 벡터와의 정렬을 통해 일치되는 서열을 찾고, 시퀀싱 정확도가 낮은 서열을 제거하여 순수한 서열을 생성한다.

③ 이렇게 얻어진 순수한 서열로부터 Phrap 프로그램을 이용하여 중복되는 부분을 찾아 서로 이어줌으로써 긴 컨티그를 만든다.

제 2단계인 기능 예측에서는 유전자의 동정을 위한 방법으로 (그림 3)에서와 같이 두 가지의 프로그램이 사용된다. 먼저, Glimmer는 미생물 유전체 사업의 유전자 동정에 대표적으로 사용되는 정확성이 높은 프로그램이다. Glimmer의 단점은 긴 서열에서만 유전자를 동정할 수 있다는 점이다. 다음으로 ORF finder는 상대적으로 짧은 서열을 6개 프레임(frame)으로 나누어서 유전자를 동정함으로서 짧은 서열에 대해서도 가능한 모든 유전자를 찾아낼 수 있으나 정확성은 떨어진다. 따라서 WGAT에서는 이 두 프로그램을 모두 사용할 수 있게 지원함으로써 상호 보완적으로 작업할 수 있게 하였다.

동정된 유전자는 주해과정을 거치게 되는데 가장 기본적인 방법은 이미 밝혀진 유전 정보와 비교를 통하여 미지의 유전자에 대해 기능을 유추하는 것이다. 대표적인 프로그램으로 NCBI에서 개발한 BLAST가 있다. BLAST에서는 동정된 유전자와 이미 밝혀진 유전 정보와의 비교를 통하여 기능을 예측한다.

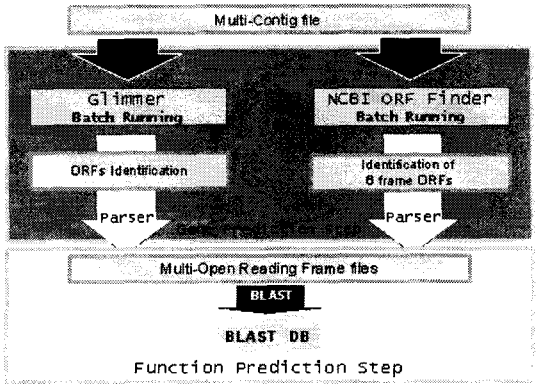


그림 3. 컨티그 서열로부터의 주해 방법

2.3 유전자 기능 예측을 위한 기존의 통합시스템

비교 유전체학적 분석 기법의 비약적 발전으로 현재 유전체 사업에서 필요한 주해과정, 컨티그 제작과정, 매핑(mapping)에 관한 단계별 프로그램은 인터넷상에 공개되어 있다. 또한, 각 유전체 사업의 목적에 따라 여러 단계의 프로그램들을 통합한 시스템들이 개발되고 있다. 유전체 연구를 위한 통합시스템으로는 Sanger Center에서 인간 유전체사업의 분석을 위하여 제작한 웹 기반의 Ensembl 프로그램(그림 4)이 대표적이다[9]. Ensembl 프로그램은 클론정보와 서열정보를 일시에 분석하여 데이터베이스화하는 장점이 있으나, 인간 유전체에 관련한 프로그램이므로 미생물 유전체의 목적에는 적합하지 않는 부분이 있다. Ensembl과 비슷한 공개 프로그램으로 Emboss[10]가 있다. 다양한 기능을 통합하여 패키지로 제공하지만 유전체 분석에는 적합하지 않다. 이러한 상황에서 볼 때 하나의 통합 시스템이 모든 유전체 사업에 적합할 수는 없으며, 각각의 유전체 사업의 고유 목적과 방법에 적합한 통합 시스템의 개발이 필요하다.

미국 JGI(Joint Genome Institute)에서는 미국 에너지성의 지원을 받아 다양한 생물체에 대한 유전체 사업을 진행 중에 있으며 유전체 분석에 대한 일부의 데이터를 외부로 공개하였다. 그러나 분석에 사용된 도구는 공개하고 있지 않다.

국내에서는 본 연구실을 포함한 몇몇 기관에서 미생물 유전체 사업을 진행하고 있으나 유전체 분석을

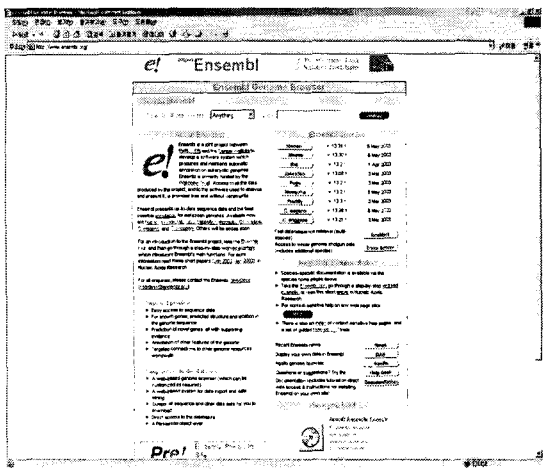


그림 4. Ensembl 홈페이지

위한 통합 시스템의 개발은 선진국에 비해 아직 미비한 상태이다. 이러한 국내의 상황에서 볼 때 본 연구에서 개발한 WGAT 시스템은 산업적 이용가치가 큰 미생물 유전체 분석 작업과 관련 분야의 활성화에 크게 기여 할 것으로 본다.

3. WGAT 시스템의 설계 및 구현

본 장에서는 WGAT 시스템의 설계, 전체 시스템의 구현과 각 모듈을 차례로 설명하고, 사용자 입장에서 WGAT 시스템의 동작 과정과 주요 특징을 설명한다.

3.1 시스템의 설계

본 연구에서는 사용자 편의성과 분석 효율성을 높이기 위해 유전자 분석 과정의 여러 단계에서 사용되는 기존의 소프트웨어 도구들을 통합하여 웹 환경으로 제공하고자 한다.

유전체 분석에 사용되는 대표적인 프로그램은 제 2.2절에 서술하였다. 위에서 서술한 프로그램들은 리눅스를 기반으로 하며, 전세계에서 폭넓게 진행중인 유전체 사업과 관련하여 가장 많이 사용되는 프로그램으로써 검증된 성능을 갖고 있다.

본 시스템에서는 유전체 분석의 각 단계에 필요한 프로그램들을 하나의 시스템에 설치하고, Perl을 이용하여 일괄처리할 수 있도록 한다. 각 프로그램의 구동에 필요한 옵션값과 입력 데이터는 PHP를 이용하여 웹을 통해 받게되며, 결과값 또한 웹을 통해 사용자에게 제공된다.

WGAT 시스템의 전체 구조는 (그림 5)와 같다. 그림에서 사용자는 자신의 PC를 이용하여 원격지 서버에 관리되는 WGAT 시스템에 접속한 후, 각 단계를 순서대로 거쳐 주어진 서열에 대한 주해 작업을 마치게 된다.

WGAT 시스템은 각 단계의 분석에 필요한 도구들을 하나의 웹 프로그램으로 통합하여 자동화함으로써 신속한 일괄처리가 가능하도록 한다. 기존의 방식에서는 이러한 프로그램들이 서로 다른 플랫폼에서 서로 다른 데이터 양식을 사용함으로써 도구들을 사용하는데 많은 수작업이 반복적으로 요구되며, 그 결과 분석 시간이 증가하게 된다. 또한, WGAT에서는 웹 브라우저를 통하여 모든 서비스를 사용할 수 있도록 웹기반 GUI 환경을 제공하며, 처리 결과는 사용자의 PC로 다운로드 받을 수 있도록 한다.

3.2 전체 시스템 구현

WGAT 시스템은 클라이언트 서버 구조로서 <표 1>과 같은 환경에서 구현된다. 서버 운영체제는 리눅스 7.3이며, 웹 환경으로 개발되었기 때문에 클라이언트는 WWW가 가능한 모든 플랫폼에서 이용이 가능하다. (그림 5)에서 보는 바와 같이 각 단계별 프로그램에서 요구되는 입력 파일 형태가 다르기 때문에 각 단계에서 입력 데이터의 변환이 이루어진다. 그리고 한번에 하나의 서열만 처리하도록 제작된 프로그램의 경우 일괄처리가 가능하도록 확장한다. 시스템의 구현 언어로는 Perl 5.8.0, BioPerl 1.0.2가 사

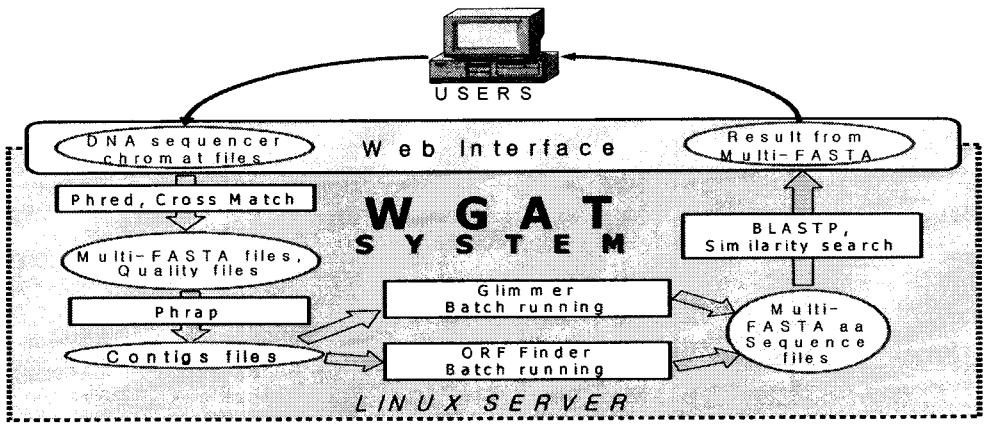


그림 5. WGAT 시스템 구조그림

표 1. 시스템 환경

		개발환경
OS	Server	Linux
	Client	Windows , Unix and Linux (Internet Explorer 5.0 & Netscape Navigator 4.0 이상)
Script Language		Perl 5.8.0, Bioperl 1.0.2
Interface		HTML, PHP, Javascript

용된다. 또한, 사용의 편리성을 제공하는 웹 인터페이스는 PHP와 Javascript를 이용하여 개발한다.

3.3 WGAT 시스템의 동작 과정

이 장에서는 WGAT 시스템을 사용하여 서열 분석을 수행할 때 WGAT 시스템의 동작 과정을 설명한다. 전체 과정은 다수의 유전체 서열 파일을 WGAT에 자동 입력하는 단계, 각종 옵션을 조정하는 단계, 대량의 서열에 대하여 유전자를 찾는 단계, 단백질 서열 분석 단계의 4단계로 이루어진다. 여기서는 각 단계를 상세히 설명한다.

1) 유전체 서열 파일의 자동 입력

유전체 분석 과정에서는 수많은 유전체 서열 파일이 1차 산물로 얻게 되는데 이 서열 파일을 효율적으로 관리 및 분석하기 위하여 서열 데이터를 데이터베이스에 저장할 필요가 있다. 유전체 분석을 위한 서열 파일의 개수는 수천 또는 수만 개에 이른다. 이들로부터 불필요한 부분을 제거(예를 들면, 벡터 서열과 잘못 읽혀진 서열)하거나, FASTA 형식으로 변환하는 일은 여러 단계를 거쳐야 한다. 수작업으로 처리한다면 각 단계마다 명령어와 옵션을 직접 입력해야 한다. 본 논문에서 제안하는 WGAT에서는 이러한 분석 과정에서 수작업을 최소화하여 몇 번의 조작만으로 여러 단계를 일괄처리 하게 된다.

본 시스템에서는 유전체 사업과 같은 많은 양의 유전자 파일을 처리하기 위해서 (그림 6)에서 보는 바와 같이 FTP를 이용하여 파일의 업로드를 수행한다. 이러한 일괄처리 방식은 각 단계에서 수작업이 수반되어야 했던 기존의 방식보다 서열 분석의 생산성을 높일 수 있게 된다.

2) 웹 기반의 Phred-Phrap을 이용한 옵션 설정

FTP를 통해 업로드된 서열 파일은 다음 단계에 사용되는 프로그램인 Phred-Phrap의 입력 값으로

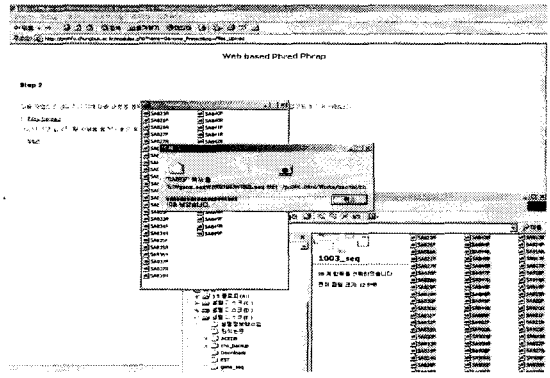


그림 6. FTP를 이용한 파일 업로드

자동으로 넘어가게 된다. Phred-Phrap은 리눅스 기반의 프로그램으로써 (그림 7)에 나타낸 것처럼 각종 옵션을 텍스트 모드에서 직접 입력해야 하므로 매우 불편하며, 각 옵션의 의미에 대해서도 숙지를 하고 있어야 하는 어려움이 있다.

WGAT에서는 리눅스 환경에 익숙하지 않은 사용자라도 쉽게 사용할 수 있도록 (그림 8)과 같이 윈도

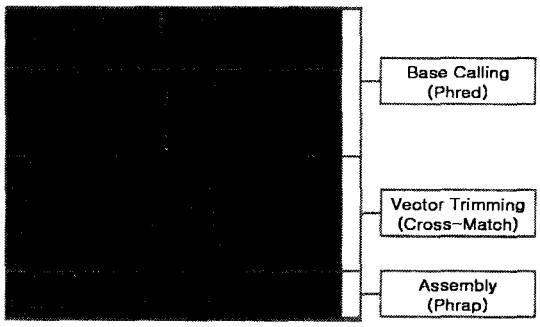


그림 7. 기존의 리눅스 환경에서의 Phred & Phrap 프로그램 실행화면

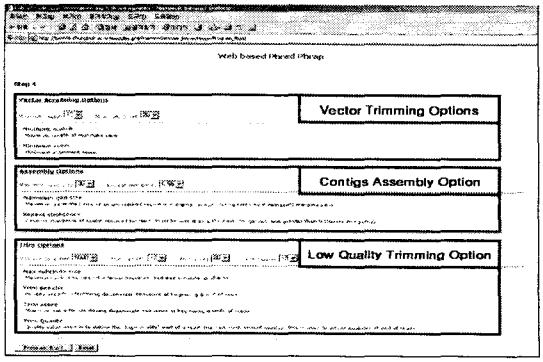


그림 8. WGAT에서 Phred & Phrap 실행 인터페이스

우 기반의 웹 브라우저에서 간단한 GQL (Graphic Query Language)을 사용하여 컨티그를 제작할 수 있으며, (그림 9)와 같이 결과값도 웹으로 제공함으로써 파일 형태로 쉽게 받을 수 있다. 특히 사용자가 다양한 옵션을 선택할 수 있도록 함으로써 보다 정확한 결과값(contig file)을 얻을 수 있는 환경을 구축하고, 최적화된 옵션 값을 디폴트로 저장하여 옵션 값의 변경 없이도 정확한 결과값을 얻을 수 있다.

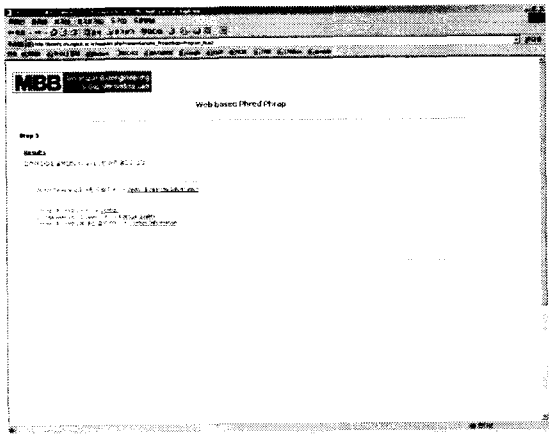


그림 9. WGAT에서 Phred & Phrap 실행 결과 화면

3) 유전자 예측

Phrap의 결과 컨티그로부터 유전자의 위치를 예측할 수 있다. 이 때 사용되는 Glimmer 프로그램은 리눅스 환경에서 동작하는 프로그램으로서 정확도는 높다고 알려져 있지만 (그림 7)과 같은 방식으로 조작해야 하므로 많은 불편이 있다. Glimmer는 내부적으로 4개의 모듈로 구성되며 이를 이용하여 유전자를 찾으려면 4개의 모듈을 각각 실행해야 한다. 하나의 서열로부터 유전자를 찾기 위해서는 매번 같은 작업을 반복해야 하므로 대량의 서열을 분석해야 할 때에는 많은 시간이 소요된다. WGAT 시스템에서는 (그림 10)과 같이 웹 브라우저를 통해 사용자의 요구를 입력받고, 정확한 결과 값을 얻기 위해 다양한 옵션을 직접 설정해 줄 수 있으며, 프로그램 수행결과를 웹브라우저로 출력함으로써 기존의 수작업을 수반한 다단계 분석 과정을 한번의 조작으로 단순화시키게 된다.

4) 유전자의 기능 예측

Glimmer에서 유전자로 예측된 서열은 기능이 밝

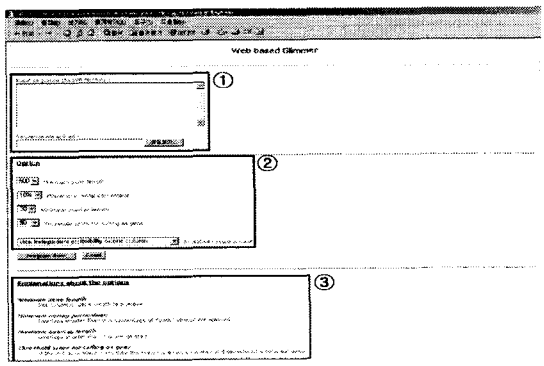


그림 10. WGAT에서 Glimmer 실행 인터페이스

① 직접 서열을 입력을 수도 있으며 파일로 업로드 받을 수 있다.
② 좀 더 정확한 유전자 예측을 위해 각 옵션을 쉽게 선택할 수 있다.
③ 초보자도 쉽게 사용할 수 있도록 위 옵션에 대한 설명이 되어 있다.

혀진 유전 정보와 비교를 통해 그 기능을 밝혀야 한다. 이 작업은 BLAST를 사용하여 이루어진다. NCBI에 접속하여 서열 비교연산을 수행하게 되면 결과를 얻을 때까지 소요되는 시간이 네트워크의 트래픽에 따라 차이가 있으나 일반적으로 오래 걸리며, 한번에 하나의 서열만 비교할 수 있다는 단점이 있다. NCBI 데이터베이스를 WGAT 시스템 안에 복사하여 설치한 후, 이용하면 여러 서열을 한번의 처리에 의해 비교, 분석할 수 있기 때문에 매우 효과적이다. 유전체 사업과 같이 대량의 데이터를 처리해야 하는 경우에는 많은 시간과 노력이 요구되므로 필요한 서열 DB와 BLAST를 에 설치하고, (그림 11)과 (그림 12)에서 보는 바와 같이 사용자는 웹 기반의 WGAT 시스템을 이용하므로 보다 쉽고 빠르게 대량의 서열을 처리할 수 있다.

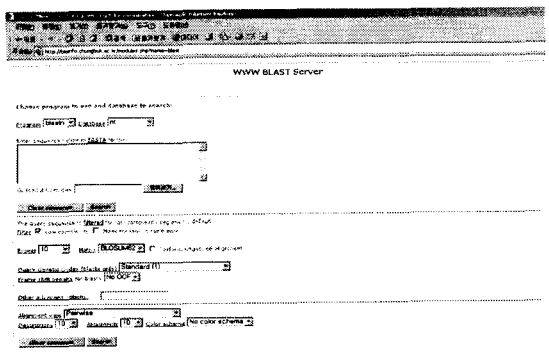


그림 11. 로컬 BLAST 옵션 인터페이스

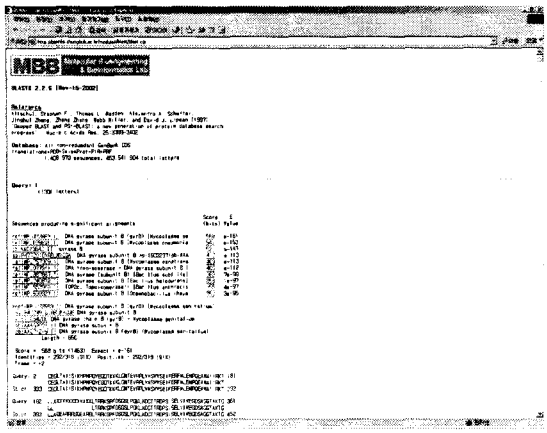


그림 12. WGAT를 이용한 다중서열에 대한 기능 예측의 결과 화면

4. WGAT 시스템의 성능 분석

이 장에서는 개발된 WGAT를 이용하여 서열 분석을 할 때 생산성이 얼마나 향상되는가를 기존의 방식과 비교하였다.

(그림 13)은 서열의 양에 따라서 기존 방식과 WGAT 방식의 유전체 서열 데이터 분석 소요시간을 보여주는 도표이다. 비교는 수작업을 동반하는 기존 방식과 본 연구에서 개발한 WGAT 시스템을 사용하는 방식을 비교하였다. 기존의 방식에서는 주어진 양의 서열 데이터를 여러 도구를 이용하여 각 단계를 거치면서 분석하는데 걸리는 시간의 합으로 측정하였다. 그래프에서 보는 바와 같이 WGAT를 사용한 분석 시간이 기존의 방식에 비하여 소요시간이 매우 적음을 알 수 있다. WGAT 시스템을 이용하여 분석할 경우에는 서열 입력과 옵션 선택을 한번만 해주면 분석이 자동으로 끝나지만, 수작업일 경우에는 각 단계에서 파일의 양식을 변환해서 다음 단계로 입력시키는 지루한 과정을 반복 수행해야 한다. 이 비교는 단지 유전자를 찾는 단계에서만 수행하였으며, 기능 예측까지 포함한다면 그 차이는 더욱 클 것으로 예상된다.

결과적으로 기존의 방법으로 대량의 서열을 분석하기 위해서는 많은 시간과 노동력이 필요하기 때문에 유전체 연구를 위한 WGAT 시스템의 개발은 서열 데이터 분석을 보다 효율적으로 수행하는데 기여할 것이다.

5. 결 론

본 논문에서는 유전체 사업의 수행에 필수적인 소프트웨어 도구들을 하나의 웹 기반 GUI 시스템으로 통합하여 생물학자들의 편리성과 분석의 생산성을 높인 웹 기반 유전자 분석 시스템 WGAT를 제안하였다.

유전체 분석 과정에서는 6개 이상의 프로그램들을 직접 구하고 설치하여 각 프로그램에서 요구하는 입력 형태로 데이터를 제공해야 하므로 컴퓨터에 대해 전문적인 지식이 부족한 생물학자 입장에서는 어려운 일이 아닐 수 없다. WGAT 시스템에서는 각 단계에서 필요한 도구들을 통합하고, 각 단계 사이의

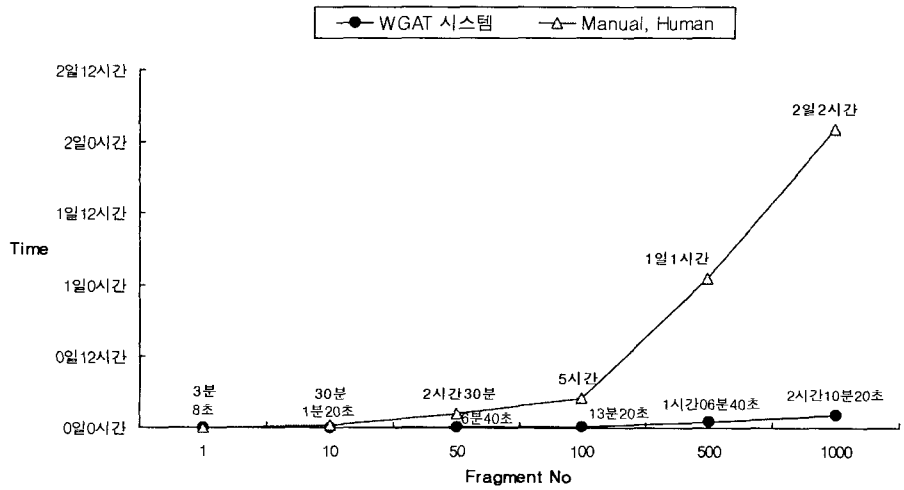


그림 13. DNA 컨티그 수에 따른 유전체 ORF 동정시간 비교

데이터 변환 및 이전을 자동화함으로써 생물학자들의 편리성과 생산성을 제고하였다. WGAT 시스템의 개발로 인해 산업적으로 이용 가치가 큰 미생물 유전체 분석 사업의 활성화에 크게 기여 할 것으로 사료된다.

향후 계획으로는 WGAT 시스템으로부터 얻은 데이터를 데이터베이스로 저장하여 관리함으로써 데이터 마이닝 등 다양한 분석을 수행할 수 있는 기반을 구축하며, 또한 기능 예측에 유용한 프로그램을 지속적으로 본 시스템에 추가해 나감으로써 보다 정확한 결과를 낼 수 있도록 확장해 나갈 것이다.

참 고 문 헌

[1] R.D. Fleischmann et al., Whole-genome random sequencing and assembly or *Haemophilus influenzae* Rd. science. Vol.269, pp. 496-512, 1995.

[2] L. Catherine, A Web interface generator for molecular biology program in Unix. *Bioinformatics*, Vol.17, No.1 pp. 73-82, 2001.

[3] S. Batzoglou et al., Arachne: A Whole-Genome Shotgun Assembler. *Genome Research.* Vol.12, No.1, pp. 177-189, 2002.

[4] A. L. Delcher et al., Improved microbial gene identification with GLIMMER. *Nucleic Acids Research,* Vol.27, No.23, pp. 4636-4641, 1999.

[5] S. Salzberg et al., Microbial gene identification using interpolated Markov models. *Nucleic Acids Research,* Vol.26, No.2, pp. 544-548, 1998.

[6] S. F. Altschul et al., Basic local alignment search tool. *J. Mol. Biol.* Vol.215, pp. 403-410, 1990.

[7] S. F. Altschul et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* Vol.25, pp 3389-3402, 1997.

[8] T. L Madden, R.L Tatusov, and J. Zhang, Applications of network BLAST server. *Meth. Enzymol.* Vol.266, pp. 131-141, 1996.

[9] T. Hubbard et al., The Ensembl genome database project. *Nucleic Acid Res.* Vol.30, pp. 38-41, 2002.

[10] P. Rice, L. Longden, and A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics.* Vol.16, No.6, pp. 276-277.

[11] B. Ewing, and P. Green, Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research.* Vol.8, pp. 186-194, 1998.

[12] Brent, Ewing et al., Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research.* Vol.8, pp. 175-185, 1998.

[13] P. Green http://bozeman.mbt.washing-ton.edu/phrap.docs/phrap.html, 1996.

최 범 순

2000년 충북대학교 생명과학부 미생물학 학사 졸업
2003년 충북대학교 대학원의 미생물학 및 생명공학 석사과정 졸업

관심분야 : 생명정보학, 유전체분석, 웹 어플리케이션

이 경 희

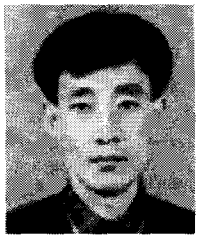
1999년 2월 충북대학교 전산학 석사
2001년 2월 충북대학교 전산학 박사과정 수료
2002년 9월 ~ 현재 서원대학교 강의전담 교수

관심분야 : XML, 객체데이터베이스, 질의처리

권 해 룡

2003년 충북대학교 생명과학부 미생물학 학사 졸업
2003년 ~ 현재 충북대학교 대학원 미생물학 및 생명공학 전공

관심분야 : 생명정보학, 유전체분석, 웹 어플리케이션



조 완 섭

1985년 경북대학교 이학사
1987년 KAIST 공학석사
1996년 한국과학기술원에서 전산학 공학박사 취득
현재 충북대학교 경영정보학과 조교수

관심분야 : 데이터베이스, 데이터 분석 및 마이닝, 생명정보학

이 충 세

1979년 Univ. of South Carolina 컴퓨터과학 석사
1990년 Univ. of South Carolina 컴퓨터과학 공학박사
현재 충북대학교 컴퓨터과학과 교수

관심분야 : 알고리즘, 병렬처리, 결함허용

김 영 창

1978년 서울대학교 이학사
1980년 서울대학교 미생물학 전공 이학석사
1986년 서울대학교 미생물 유전학 이학박사
현재 충북대학교 자연과학대학 생명과학부 교수

관심 분야 : 분자 유전학, 유전자 재조합, 생명정보학