Kor. J. Appl. Microbiol. Biotechnol.
Vol. 24, No. 1, 132-136 (1996)

제놈 분석용 계산 Web 서버의 구성

박기정 · 이병욱 · 박용하*
한국과학기술연구원 생명공학연구소 유전자은행

Construction of a Computation Web Server for Genome Analysis. Kie-Jung Park, Byung-Wook Lee and Yong-Ha Park*. *Korean Collection for Type Cultures, Korea Research Institute of Bioscience and Biotechnology, KIST, P.O. Box 115 Yusung, Taejeon 305-600, Korea* — A computation server is needed to provide analysis programs to Korean biologists, especially genome researchers, on GINet. For each analysis program, we implmented an input form with HTML and a CGI program for interface between an input form and an analysis program with C language on GINet computation Web server. We made two construction methods of CGI programs for analysis programs, and implemented all CGI programs based on the methods followd by modifying each CGI program for specific processing of each analysis program. On the server ten programs are available now, which include most frequently used ones and those developed by our team, and most programs which will be ported or developed by our team will be available on the Web server.

인체유전자 프로젝트(human genome project)의 시작을 전후하여 미국, 유럽, 일본 등에서는 생물학에서 컴퓨터를 활용하는 이른 바, 생물정보처리(bioinformatics)가 본격적으로 한 분야로서 역할을 하고 있다. 대학이나 연구소 등에 설립된 bioinformatics 센터에서는 데이타베이스나 분석 프로그램들을 운영하면서 고용량의 분석계산도 전산망을 통해 지원하고 있다. 한편, 전산망 상의 컴퓨터 사용 환경에 대한 개선과 표준화 작업이 꾸준히 진행되면서 1990년대 초부터 WWW(World Wide Web)라는 개념의 도구가(1) 폭발적인 호응 하에 표준으로 자리잡아 가고 있다. 이는 전산망에 연결된 사용자 컴퓨터에서 전산망 상의 어떤 곳의 정보 자원도 쉽게 획득할 수 있도록 해준다. Web을 통해 제공되는 문서는 대부분 HTML(HyperText Markup Language)이라는 언어로(2, 3) 작성되며 사용자는 이러한 문서를 이해하여 format 시켜 주는 통역기(interpreter) 역할을 포함하는 Web browser 프로그램을 사용한다(Netscape이나 Mosaic 등). 단순한 문서의 제공 뿐 아니라 서버가 가지고 있는 실행 프로그램을 Web 사용자에게 제공할 때는 그 프로그램에 대한 CGI(Common Gateway Interface) 프로그램을(3, 4) 작성해야 한다. 일반적으로 기존의 실행 프로그램은 Web 환경과는 무관하므로, Web에서 사용자가 입력한 내용을 실행 프로그램이 요구하는 입력 형식에 맞추어 변환해서 이 실행 프로그램을 실행하고 그 결과를 Web 서버를 통해 사용자에게 보여 주기 위한 인터페이스가 필요한데, 이러한 역할을 CGI 프로그램이 하게 된다. 최근 bioinformatics 센터에서는 이러한 인터페이스를 통해 외부 사용자에게 생물정보분석 프로그램을 사용할 수 있도록 하고 있다.

국내에서도 1994년 제놈프로젝트가 시작되어 생명공학연구소에서 bioinformatics 연구를 수행하면서 제놈정보전산망 GINet(Genome Information Network of Korea)을 구성하였고 1차년도를 통해 여러 분석 프로그램과 데이타베이스를 활용할 수 있는 시스템을 구축하였다. 이 시스템의 효과적 활용을 위한 방안으로 Web을 통해 이러한 시스템을 서비스하기 위해 GINet 분석계산용 Web 서버를 설치하였다. 서버에 설치된 분석 프로그램을 Web에서 지원하기 위해 Web 사용자가 입력한 데이터를 이들 프로그램과 연결하고 실행결과를 사용자에게 제공하기 위한 CGI 프로그램들을 개발하고 설치하였다.

재료 및 방법

개발 및 설치 도구

분석 프로그램과 데이타베이스를 설치하고 계산용 Web 서버 운용를 위한 CGI 프로그램들을 개발·설치하기 위해 UNIX(IRIX 5.3) 워크스테이션인 'Indigo2'(CPU R4400, memory 64MB)를 사용하였으며, CGI 프로그램은 C 언어로(5) 작성하였다.

분석 프로그램의 선정

단백질과 DNA의 상동성 분석을 위해 가장 많이 사용하는 프로그램인 blast(6-9) 및 fasta(10-12) 패키지와 motif 데이타베이스인 prosite 검색용 프로그램 prosearch, TFD(transcription factor database)의 signal sequence를 검색하는 sigscan, 그리고 제한효소검색 프로그램 resenz를 비롯하여 자체 개발한 프로그램들을 Web 서버용 분석 프로그램으로 선정하였다(Table 1).

*Corresponding author.
Key words: WWW, Web, CGI, genome project, analysis program database, GINet

Table 1. Analysis programs on GINet computation Web server

프로그램 이름	프로그램 용도	구동 방식
blast	GB/PIR local homology 검색	명령행
fasta	GB/PIR homology 검색	명령행
align	1:1 global alignment	명령행
prosearch	prosite의 motif 검색	명령행
sigscan	TFD의 signal sequence 검색	대화식
revcomp	reverse/complementary strand 변환	명령행
dna2prot	6가지 reading frame으로 translation	명령행
gbidx	accession/locus로 GB 검색	명령행
piridx	accession/entry로 PIR 검색	명령행
resenz	제한 효소 자리 검색(REBASE)	대화식

분석계산용 Web 서버의 구성

분석계산 Web 서버는 서버 프로그램인 httpd와 사용자가 필요로 하는 분석 프로그램들, 그리고 분석을 위한 조건이나 데이타를 입력하기 위한 form을 사용자에게 제공하는 HTML 프로그램, 그리고 이 form에서 사용자 입력 데이타를 읽어 분석 프로그램을 실행시키고 그 결과를 사용자에게 보여주는 CGI 프로그램들로 구성하였다.

입력용 HTML 프로그램

각 분석 프로그램에 대해 간략히 소개하고 처리할 데이터와 옵션(각 분석 프로그램 실행 조건으로 사용자가 지정해 주는) 값을 사용자가 입력할 수 있도록 구성된 양식을 HTML로 작성하였다. 이 문서들은 서버가 아닌 전산망 상의 임의의 곳에 둘 수 있는데, GINet에서는 현재 계산 서버에 두었다.

CGI 프로그램

CGI 프로그램은 Web 서버에서 별도의 실행 프로그램을 통한 데이타처리가 필요한 경우에, 사용자로부터 데이타를 받아 그 실행 프로그램을 실행시켜 출력을 그 사용자에게 제공하고자 할 때 사용하는 프로그램이다 (Fig. 1). 현재 이 서버에서 제공하는 분석 프로그램은 실행 명령어에서 모든 옵션과 입출력 화일 이름을 주는 경우와 실행 중에 대화식(interactive)으로 옵션과 입출력 화일 이름을 주는 경우의 2가지로 나눌 수 있다(Table 1). 각 경우, 처리할 옵션을 받아들일 form은 HTML로 작성되어 있고, 이들 form에서 입력된 옵션과 데이타는 특정한 형식을 가지고 CGI 프로그램으로 넘어가게 된다. 이 형식에서는, 각 값을 변수 명과 그 변수에 대해 사용자가 입력한 값의 쌍으로 나타내며, 변수 명과 변수 값의 사이는 특수 기호에 의해 표시된다(4).

명령행(command line)에서 옵션을 주는 경우 blast와 fasta, prosearch 등을 이러한 방식으로 처리할 수 있으며 다음과 같은 규칙에 따른다.

i) form으로부터 CGI 프로그램으로 넘어온 옵션과 데이타를 분석하여 각 옵션 별로 할당된 값을 인식한다.

ii) 분석 프로그램이 표준 출력을 사용할 경우는 iv)의 명령행에서 이 표준 출력을 화일로 받고, 명령행에서 지정할 경우는 iv)의 명령행에서 이 화일을 지정한다.

iii) 명령행이 입력 화일을 요구할 때는 해당하는 입력 데이타를 입력 화일로 바꾸어 이 화일 이름을 명령행에서 사용한다.

iv) 각 분석 프로그램의 정해진 명령행 구성 규칙에 따라 인식된 각 옵션과 입출력 화일 이름으로 하나의 명령행을 완성한다.

v) 명령행을 실행한다. 이때 'system( )'을 사용한다.

vi) ii)의 출력 화일을 읽어 표준 출력으로 보낸다.

대화식으로 옵션을 주는 경우 resenz, sigscan 등이 이 경우에 해당하며 다음과 같은 규칙으로 처리한다.

i) form으로부터 CGI 프로그램으로 넘어온 옵션과

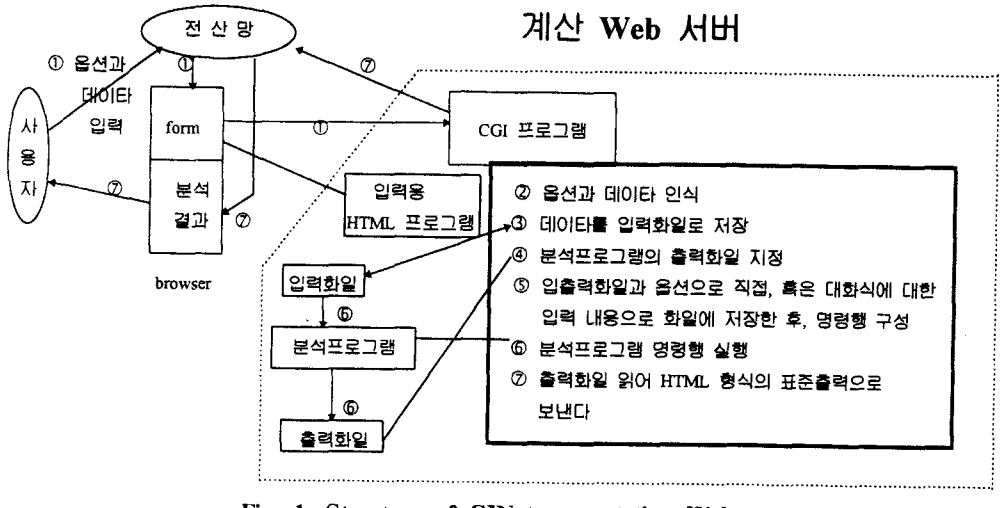


Fig. 1. Structure of GINet computation Web server.

데이타를 분석하여 각 옵션 별로 할당된 값을 인식한다.

ii) 분석 프로그램이 입력 데이타를 직접 요구하는 경우는 인식한 데이타를 값으로 주고, 입력 화일을 요구하는 경우는 입력 데이타를 화일로 작성하여 이 화일의 이름을 값으로 준다.

iii) 분석 프로그램이 표준 출력을 사용할 경우는 iv)의 명령행에서 이 표준 출력을 화일로 받도록하고, 출력 화일의 이름을 분석 프로그램이 요구하는 경우 화일의 이름을 iv)에서 분석 프로그램의 입력으로 한다.

iv) 대화식으로 입력하는 순서에 맞게, 인식된 각 옵션과 ii)와 iii)의 입출력 화일 이름이나 데이타를 하나의 화일에 입력하고 이 화일을 분석 프로그램의 표준입력으로 대화식이 이루어 지도록 하여 명령행을 구성한다.

v) 분석 프로그램 명령행을 실행한다. 이때 'system ( )'을 사용한다.

vi) iv)의 출력 화일을 읽어 표준 출력으로 보낸다.

1)과 2)의 vi)에서는 CGI의 출력을 위해 항상 다음 행을 가장 먼저 출력한다.

"Content-type:text/html"

" " (빈 행)

또한 HTML 상에서 각 line을 분리하기 위한 간단한 명령어들을 함께 출력한다.

결과 및 고찰

Fig. 2에 GINet의 분석계산용 Web 서버의 초기 화면이 나타나 있다. 이 초기 화면은 'http://grcsys1.geri.re.kr'의 위치에 의해 시작되는 GINet 전체 서버의 초기 화면에서 '생물학 관련 응용 프로그램들'을 선택하거나 직접 'http://grcsys1.geri.re.kr/software.html'의 위치를 지정함으로써 시작된다. Fig. 2의 'GINet 서버에서 실행되는 프로그램'은 분석 프로그램과 이를 실행하는 CGI 프로그램들이 모두 GINet 계산 서버에 있는 것이고, 화면을 아래로 이동하여 나타나는 '외국 Web 서버에서 실행되는 프로그램'은 각 프로그램을 서비스하는 외국의 각 서버에서 실행되는 것이다. 외국 서버를 이용하는 프로그램들은 Web을 통해 지원하는 형태의 분석 프로그램을 GINet 계산 서버에서 갖추지 못한 것들인데, 이 중에는 개발한 곳에서 몇 개의 서버에서만 그 프로그램을 사용하도록 제한한 것도 있다. 이들 프로그램들은 설치하거나 자체 개발하면 GINet 서버에서 직접 지원할 예정이다.

Fig. 3은 명령행 만으로 구동되는 프로그램에 대한 예로, blast의 입력 form과 분석후 서버로부터 전송된 결과를 보여준다. 이들 프로그램들은 현재 설정된 옵션에 대해서는 잘 작동하는 것으로 시험되었는데, 사용자가 증가할 경우 발생할 실행 시간의 문제 등이 발견되면, 프로그램의 일부 옵션은 Web에서는 사용되지 못하도록 입력 form에서 삭제하거나, 제한된 범위의 값만을 입력할 수 있도록 입력 form을 조정해야 할 것이다.

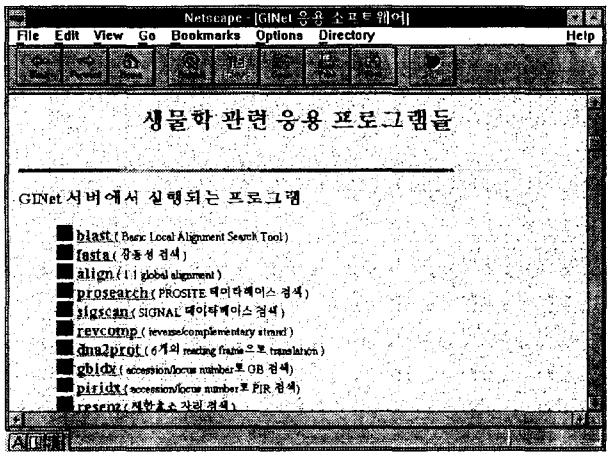
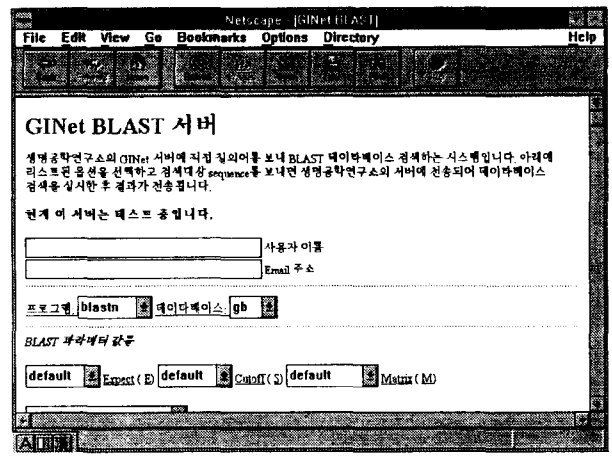
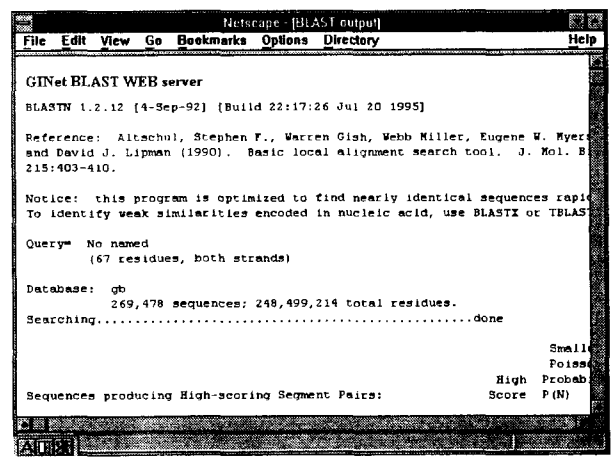


Fig. 2. GINet computation Web server.

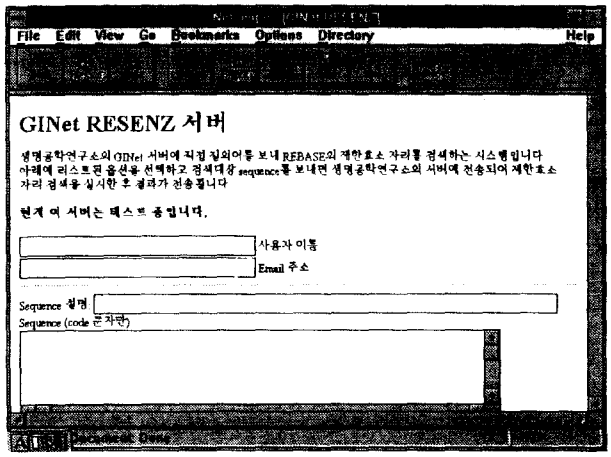


(a) Input form

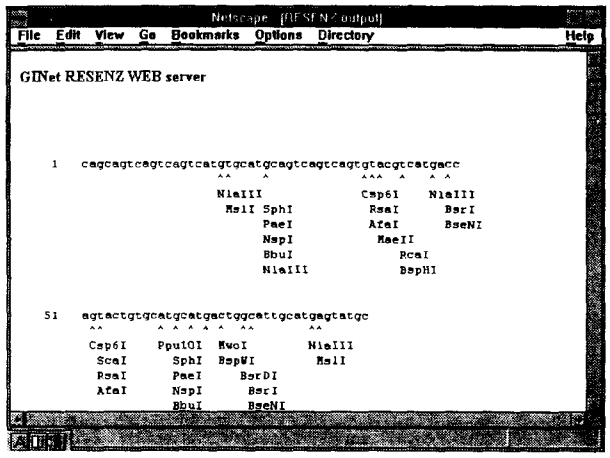


(b) Result

Fig. 3. GINet BLAST Web server.



(a) Input form



(b) Result

Fig. 4. GINet RESENZ Web server.

Fig. 4는 대화식으로 구동되는 프로그램에 대한 예로, resenz의 입력 form과 분석후 서버로부터 전송된 결과를 보여준다. 이들 프로그램은 대부분 대화식 도중 입력하는 값에 관계없이 대화의 형태가 동일하게 유지되는 것이나, sigscan은 signal 검색과 journal 검색의 각 경우 서로 다른 대화 형태를 취하게 된다. 이런 경우 각 경우에 대해 하나씩의 입력 form을 가지도록 구성하였고, CGI 프로그램도 각 경우에 대해 별도로 제작하였다. 대화식 구동 프로그램은 대부분 실행시간이 짧은 것이므로, 옵션의 제한보다는 분석 프로그램에서 지원하는 옵션의 종류를 확장하여 사용자가 더 다양한 옵션을 사용할 수 있는 방향으로 발전될 것이다.

CGI 프로그램의 설계와 구현은 이처럼 '명령행 구동'과 '대화식 구동'의 2가지 경우로 나누어 지며, 각 경우는 위의 '재료 및 방법'에서 언급한 공통적인 절차에 따라 이루어 진다. 각 분석 프로그램도 개별적으로 특이적인 명령행 형태나 입력 형태를 가지고 있지만, 위의 공통적인 구성 절차와 개별적인 특이 절차는 정형화 시킬 수 있을 것으로 보인다. 이러한 정형화가 정리되면 임의의 분석 프로그램에 대해 HTML로 구성한 입력 form에서 사용하는 옵션과 데이타 종류, 최종적으로 구성할 명령어의 형태와 대화식의 순서 및 구성 내용 등에 대한 정보를 입력으로 하여, 그 분석 프로그램에 대한 C 언어 CGI 프로그램을 출력으로 생성하는 'CGI 생성기'를 만들 수 있을 것이다. 'CGI 생성기'가 제작되면 추가 또는 변경되는 분석 프로그램과 입력 form에 대한 CGI 프로그램의 제작을 더욱 효율적으로 할 수 있을 것이고, 여러 site에서 간단한 분석 프로그램 들을 Web을 통해 제공하는 시스템을 구성하는 데 활용할 수 있을 것이다.

요 약

국내 제놈 프로젝트와 생물학 연구자들을 지원하기 위한 GINet 서버에서 분석 프로그램들을 Web을 통해 사용할 수 있도록 GINet 분석계산용 Web 서버를 구성하였다. 각 분석 프로그램의 구조를 분석하여 필요한 데이타와 옵션의 종류 및 사용 형태를 정리한 후, 사용자가 이들을 입력할 수 있는 입력 form을 HTML로 작성하였다. 입력된 옵션과 데이터는 Web 서버에 전송되는데, 이들을 분석하여 해당 실행 프로그램을 실행한 후 분석 결과를 Web 사용자에게 전송해 주도록 CGI 프로그램들을 작성하였다. 특히 분석 프로그램의 형태를 2가지로 나누어 각 경우 CGI 프로그램을 구현하는 공통적인 절차를 구성하였고, 여기에 각 분석 프로그램마다 개별적으로 다른 부분은 개별적으로 보완하는 방식으로 각 CGI 프로그램을 구현하였다. 현재 이 서버에서 직접 지원하는 분석 프로그램은 사용 빈도가 가장 높은 프로그램들과 자체 개발한 프로그램들을 포함한 것이고, 추후 설치되거나 자체 개발되는 프로그램들의 대부분도 이 계산 Web 서버에서 제공될 수 있도록 시스템이 구성될 것이다.

참고문헌

1. 김용운 外 21명. 1995. 가자! Web의 세계로. *WWW Korean Forum.*
2. sunil@uel.co.uk. *An internet document.* HTML, the complete-ish guide.
3. *An internet document* "HTML+Discussion Document", November 8, 1993.
4. 김현석. 1995, March. 고수준 HTML 문서를 만들자. 마이크로소프트웨어 **137**: 494-500.
5. 황희융 편역. 1994. February. C 언어 기초+α. 교학사.
6. Altschul, S.F., W. Gish, W. Miller, E. Myers and D.J. Lipman. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**: 403-410.
7. Karlin, S. and A.F. Altschul. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA.* **90**: 5873-5877.

8. Karlin, S., A. Dembo and T. Kawabata. 1990. Statistical composition of High-Scoring segments from Molecular Sequences. *The Annals of statistics.* **18**: 571-581.
9. Karlin, S. and S.F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA.* **87**: 2264-2268.
10. Lipman, D.J. and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* **227**: 1435-1441.
11. Wilbur, W.J. and D.J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA.* **80**: 726-730.
12. Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* **85**: 2444-2448.

(Received 21 September 1995)