대한임상미생물학회지 : 제3권 제1호 2000
Korean J Clin Microbiol 2000 ; 3 : 5-15

인터넷을 이용한 DNA 염기서열 분석

이 창 규

고려대학교 의과대학 임상병리학교실

DNA Sequence Analysis on Internet

Chang Kyu Lee, M.D.

Department of Clinical Pathology, Korea University College of Medicine, Seoul, Korea

1 서론

인터넷의 발달은 사회 생활 전반에 걸쳐 가히 혁명적인 변화를 가져오고 있다. 이제는 강원도 산골에서 자연과 더불어 살며 세계 각국과 무역을 하는 것이 가능하게 되었다. 분자생물학 분야에서도 지난 10여년간에 걸쳐 엄청난 양의 정보들이 쏟아져 나옴으로 이제는 더 이상 문헌을 찾고 교과서를 뒤지는 전통적인 방식으로는 새로운 첨단의 지식들과 발견들을 추적해 간다는 것이 불가능하게 되었다. 다행히도 수 많은 정보들이 전자 매체(electronic media)에 저장됨으로 적절한 software 와 인터넷 통신망만 있으면 이들을 효과적으로 이용할 수 있게 되었다. 특히 우리나라와 같이 아직 전반적인 연구의 infra가 잘 구축되어 있지 않은 환경에서 이러한 인터넷의 장점을 잘 활용하여 구미의 대학이나 연구기관들이 보유하고 있는 정보나 자원들을 적절히 이용할 수만 있다면 우리가 겪게 되는 많은 어려움들을 극복할 수 있다고 생각된다. 본문에서는 PC를 통한 DNA와 관련된 여러 정보들을 어떻게 얻고 활용할 수 있는가를 개략적으로 기술해 보고자 한다.

2. DNA database

DNA database는 처음에는 표준화된 형식으로 자료들을 장기간 집적하여 보관할 수 있다는 장점에서 시작되었으나 컴퓨터의 보편화와 조회 및 분석 softwae들의 개발로 이제는 아주 광범위하게 이용할 수 있는 중요한 자료가 되었다. 각 실험실에서 온라인으로 관련 염기서열들을 조회해 봄으로써 같은 작업의 중복을 피할 수 있을 뿐만 아니라 여러 종들의 유사 염기서열들을 비교해 봄으로 실험자가 얻은 염기서열에 대한 분석이 용이하게 되었다. 또한 문헌에 나와 있지 않은 수많은 실험 자료들을 이용할 수 있음으로 전 세계적인 실험의 장을 제공할 수 있다.

현재 전 세계적으로 운영되고 있는 대표적인 핵산 염기서열의 databank와 운영기관, Web상의 주소는 다음과 같다[1].

1. DNA database of Japan(DDBJ)
   DDBJ, National Institute of Genetics
   http://www.ddbj.nig.ac.jp
2. EMBL(European Molecular Biology Laboratory) database
   European Bioinformatics Institute
   http://www.ebi.ac.uk
3. GenBank
   National Center for Biotechnology Information
   http://www.ncbi.nlm.nih.gov//
4. Genome Sequence Database
   National Center for Genome Resources
   http://www.ncgr.org/gsdb

이들은 처음에는 모두 독자적으로 출발하여 운영되어 오다가 현재는 서로 협력하여 염기서열에 대한 자료를 모으고 매일 단위로 서로 정보를 교환하여 같은 염기서열들과 그 관련 정보들을 제공하고 있다. 이들 databank중 1980년에 설립된 EMBL은 현재는 영국의 Hinxton에 위치한 European Bioinformatics Institute가 그 뒤를 이어 EBI란 이름으로 주로 유럽지역에 분자생물학 분야의 정보를 제공하고 있다. 위에서 언급한 것처럼 이들은 상호 협력하에 같은 정보를 제공하고 있으나 기존의 자료를 추가로 교정하였거나, 덧붙인 경우는 이러한 사항들이 빠질 수 있으므로 매우 중요한 유전자 서열은 4개의 database를 모두 조사하는 것이 안전하다[1]. 이들 중 대표적인 DB인 GenBank는 NIH

원본접수 : 2000년 1월 25일
교신저자 : 이 창 규
(471-701) 경기도 구리시 교문동 249-1
한양대학교 구리병원 임상병리과
TEL : 0346) 560-2572 FAX : 0346) 560-2585
E-mail : jokang@email.hanyang.ac.kr

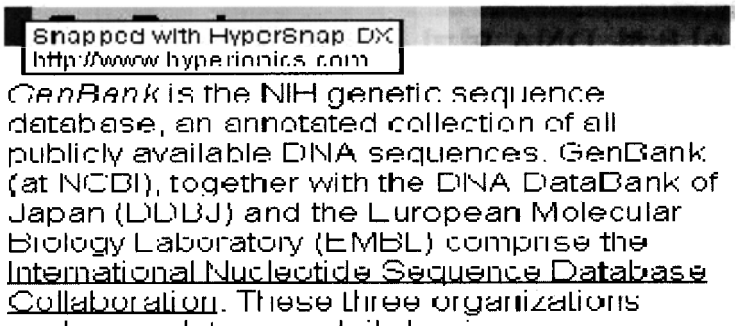

그림 2-1. NCBI에 있는 GenBank자료.

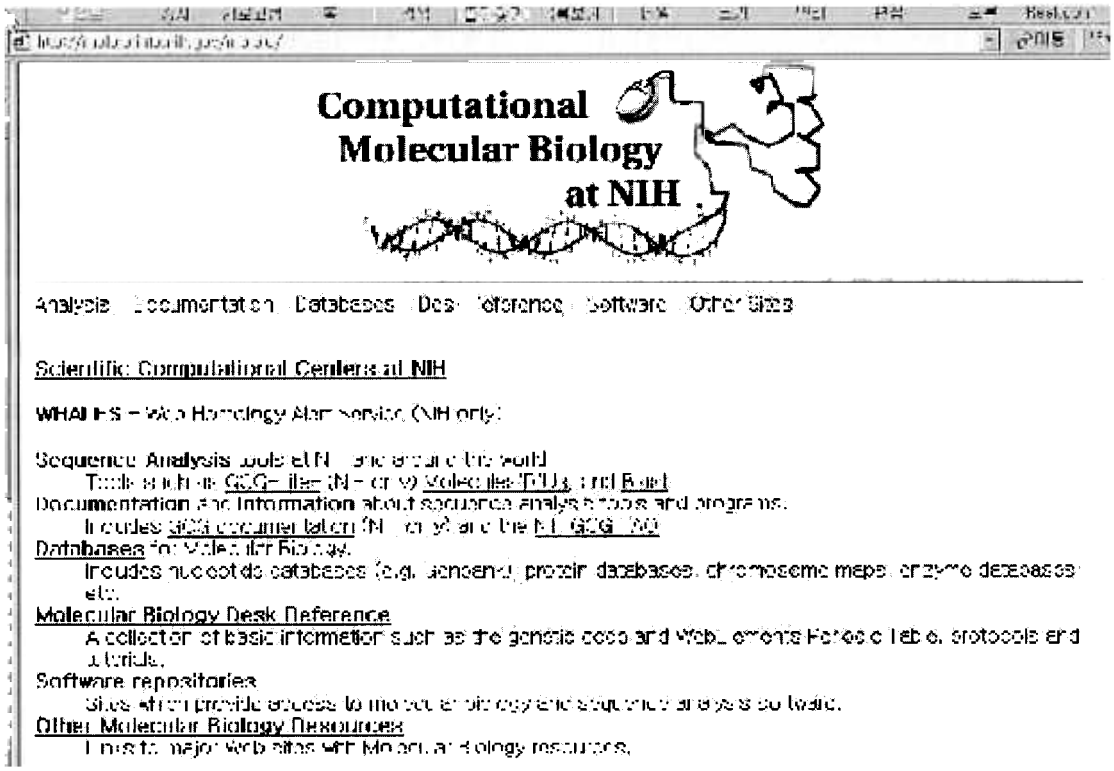

그림 2-2. NIH의 Biomolecular databases 들을 모아 놓은 site.

의 일부인 NCBI (National Center for Biotechnology Information)에서 관리하고 있으며 1999년 12월 현재 47,000 종에 걸쳐 30억개 염기서열을 갖고 있다.

이들 외에도 많은 분자생물학과 관련된 정보들이 있는데 이들에 대한 목차와 같은 database가 있는데 대표적인 것이 LiMB database이다. 이는 database들의 database로서 분자생물학 및 관련 자료들을 관심있는 사람들에게 제공하고자 만들어진 DB로 Web이전의 gopher형태로 운영되던 것이다. 현재는 WWW의 사용이 보편화 됨에따라 잘 이용되고 있지 않고 있으며 이에 상응하는 Website로서 http://molbio.info.nih.gov/molbio/ 가 있다.

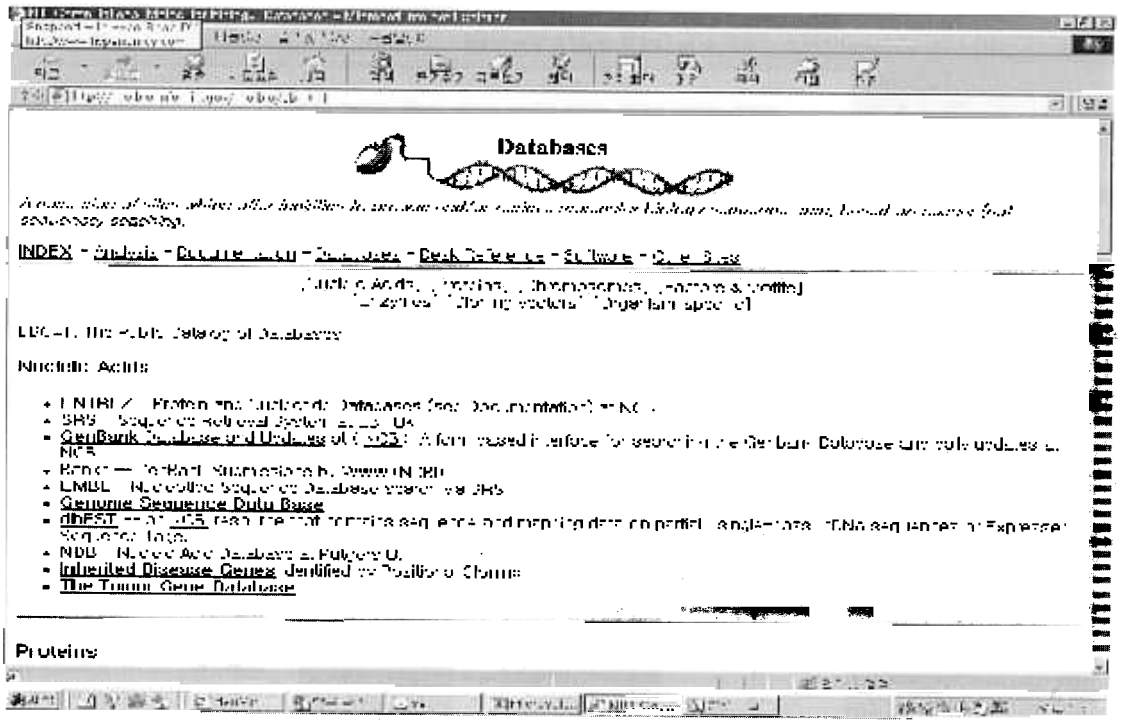

그림 2-3. NIH 의 Computational Molecular biology의 Database 내용

3. 핵산 서열분석에 흔히 사용되는 Softwares

1) Entrez

NCBI는 핵산의 염기 및 단백질의 아미노산 서열, 거대분자의 구조, 전체 genome, MEDILINE의 문헌 정보 등을 제공하고 있는데 Entrez는 이들을 유기적으로 검색하고 조회 할 수 있는 통합 시스템이다[2,3].

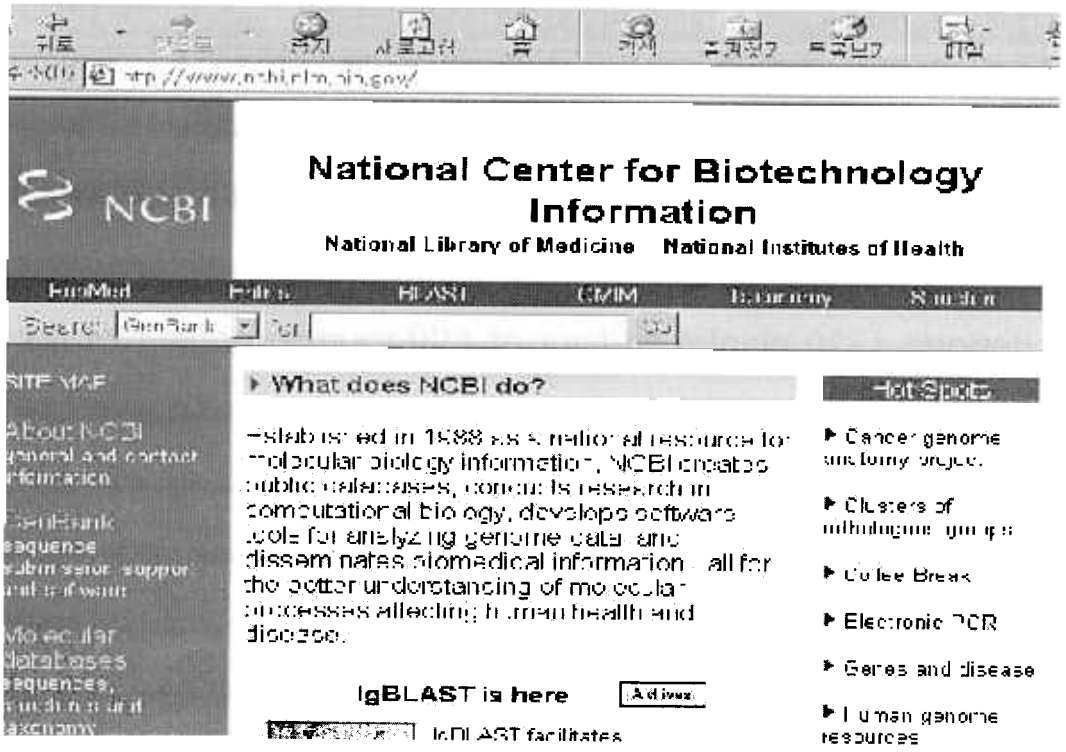

그림 3-1. NCBI home page.

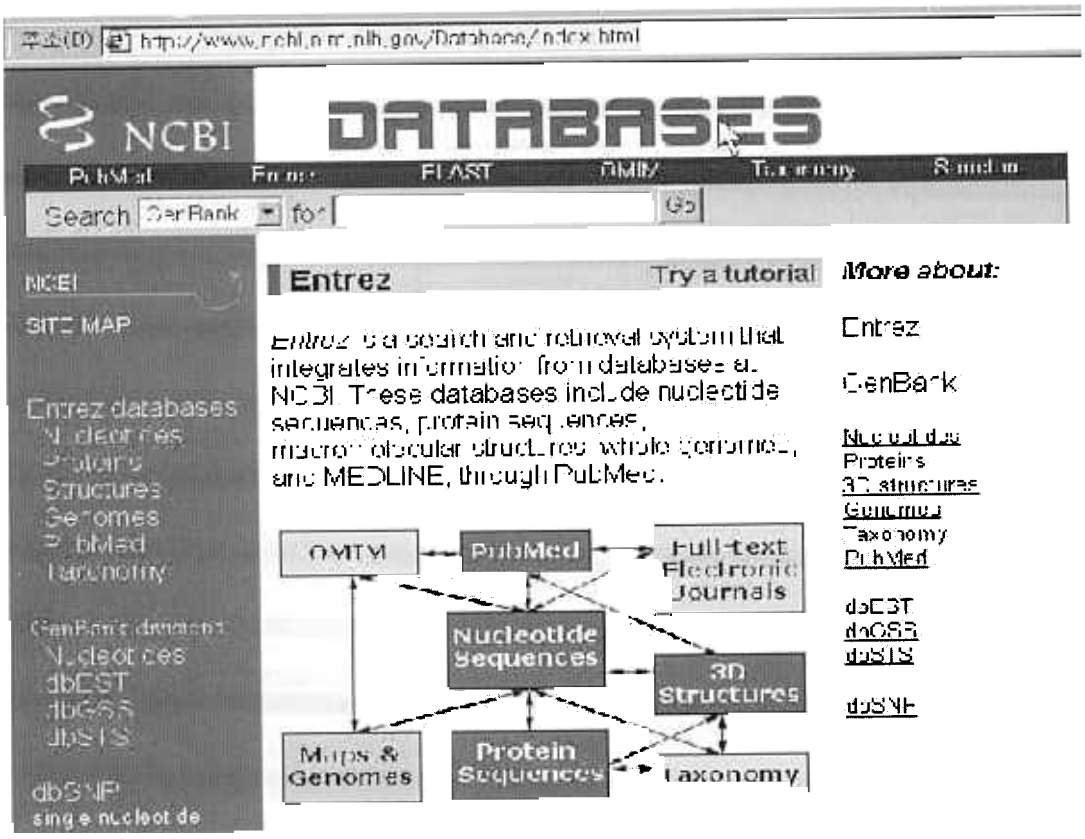

그림 3-2. ENTREZ의 초기 화면으로 여러 자료들이 유기적으로 연결되어 있음을 볼 수 있다.

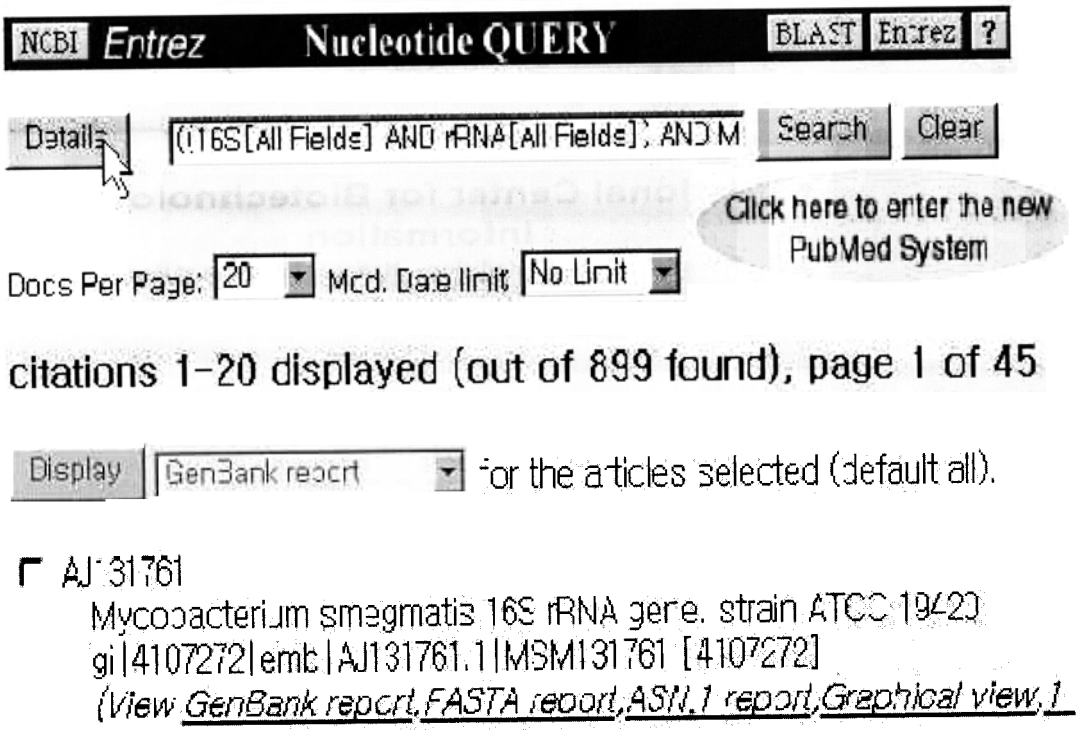

그림 3-3. ENTEREZ 에서 Mycobacterium의 16S rRNA의 DNA 염기서열의 검색

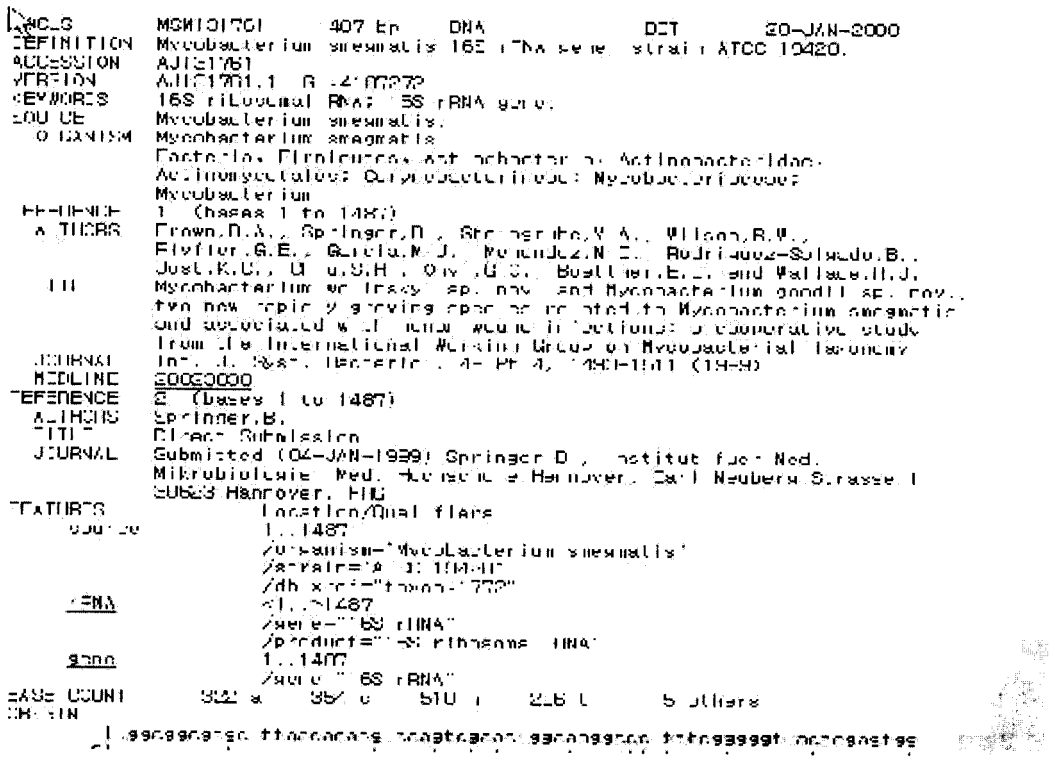

그림 3-4. ENTEREZ에서 Mycobacterium의 16S rRNA의 DNA 염기서열의 검색 결과를 Genbank 양식으로 표현한 것

이 ENTREZ를 이용하여 Mycobacterium의 16S rRNA의 염기서열을 검색하고자 하면 대화창에 이를 입력한다.

2) BLAST

BLAST (Basic Local Alignment Search Tool)로 NCBI에서 제공하는 고속 similarity search 프로그램으로 속도로 인한 예민도의 손상을 최소화하였고 핵산, 단백질, 핵

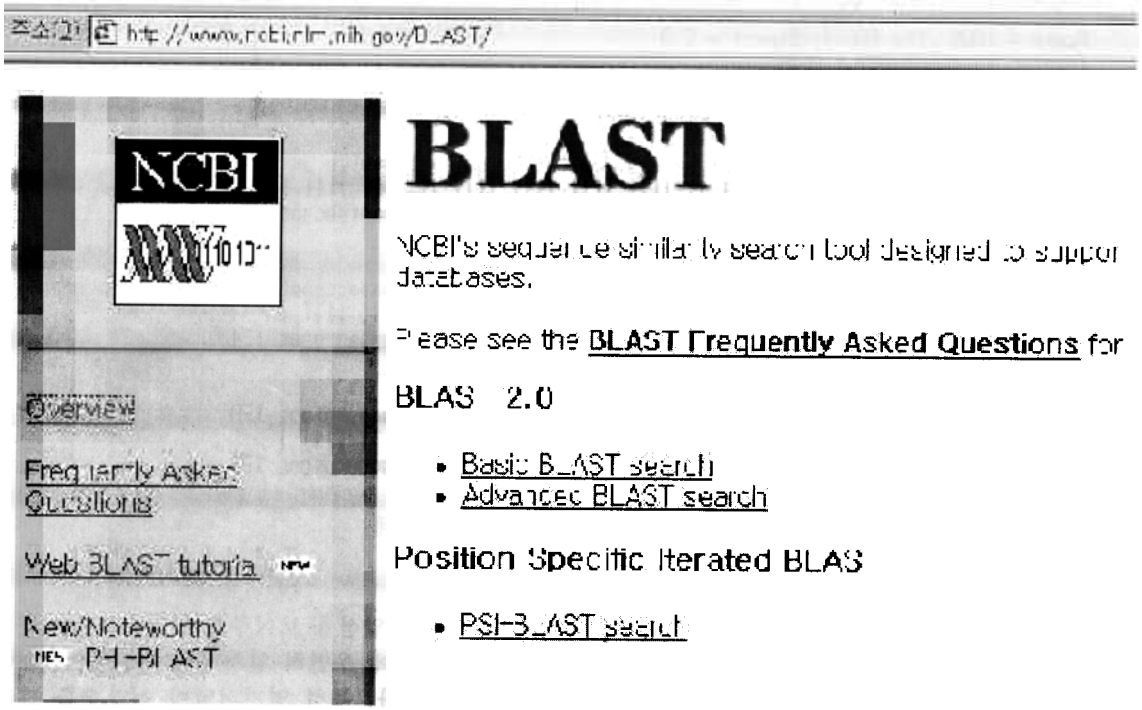

그림 3-5. BLAST의 초기화면

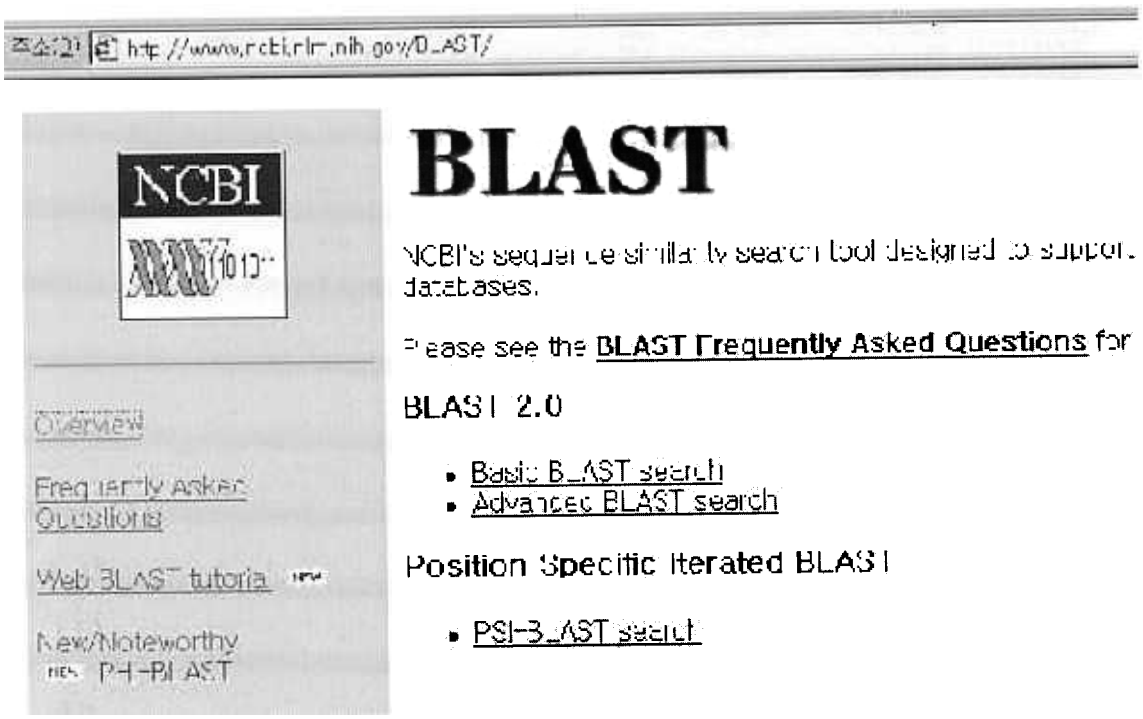

그림 3-6. BLAST에서 DNA염기서열 검색결과이며(상) 그 중 가장 상동성이 높은 Mycobacterium ratisbonense의 결과를 구체적으로 표시한 것(하).

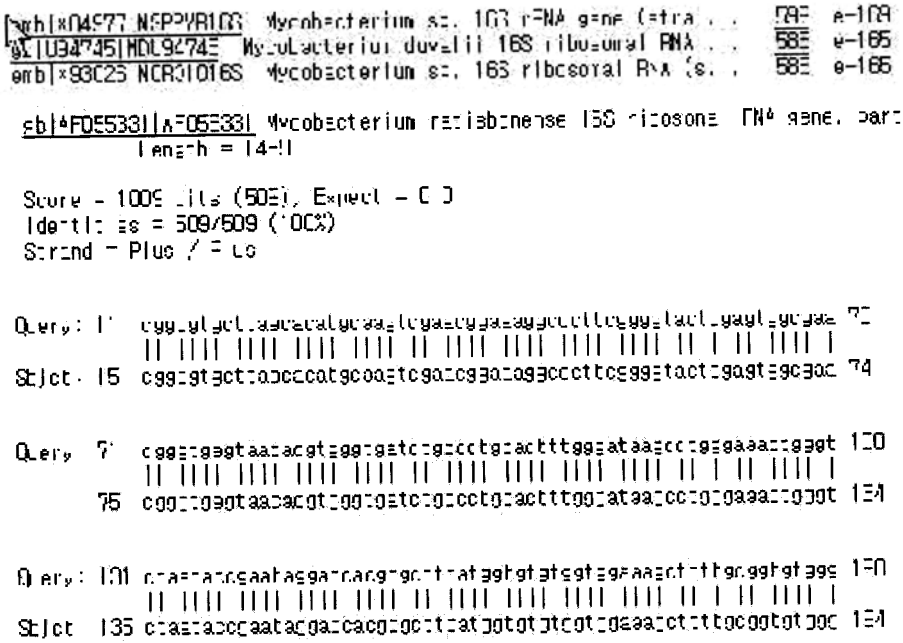

그림 3-7. 두 종류의 염기의 배열의 예

산의 단백질로의 translation (단백질과 비교한 핵산) 검색을 지원하고 있다.

이 예에서는 조회한 509개의 염기가 Mycobacterium ratisbonense의 16S rRNA의 509개와 동일한 배열을 보여 100%의 동일성을 보여주고 있다. 그런데 문제는 수많은 자료 중에 우연히 일치하여 자료들이 찾아 질 수 있다는 것이다. 특히 염기서열이 적을수록 그러한 가능성이 커지게 되는데 이를 나타내는 것이 E-value

(Expect) 값이다. 이는 검색할 때 생길 수 있는 random back-ground noise를 나타내며 이 값이 낮을수록 보다 stringent하다고 말할 수 있다. 만약에 위의 염기서열 중 일부인 20개를 조회하여 보면 찾은 염기서열의 일치율은 100%이지만 E값은 0.005가 되는데, 이는 200번 검색을 하였을 때 1번 정도는 이것이 우연히 나올 수 있다는 의미이다. 따라서 염기의 수가 길어질수록 E 값은 작아지고 보다 stringent하다고 말할 수 있다.

3) Multiple alignment program

우리가 실험이나, 문헌에서 얻은 몇 종류의 염기 서열들이 서로 공통된 구조나 기능을 갖고 있는지를 찾아 비교하고, wild strain에 비해 염기서열의 deletion이나 insertion, mutation등이 있는지를 알고자 할 때 필요한 것이 align이다. 여기에 사용되는 방법들로 global 또는 local alignment, pairwise 또는 multiple alignment 등이 있다[4].

Align방법의 알고리즘을 이해하기 위한 예로 두개의 염기서열을 그림 3-7의 'Path graph'를 살펴 보자. 우선 비교하고자 하는 9개, 8개의 염기를 가로 세로로 배열하고 그래프에서 좌 상단에서 우 하단을 향하여 갈 때, 여러 개의 경로가 가능할 수 있다. 굵은 실선은 그 경로중의 하나를 표시한 것이다.

이러한 경로들의 질을 평가하기 위해 짝을 이룬 각 염기마다 점수를 부여하게 되는데 가로와 세로가 동일한 염기이면 2점, 일치하지 않으면 0점, 짝이 없으면 ?1점을 부여하여 총 점수를 내게 된다. 위의 배열의 경우는

가로 T T - A C T T G C C
세로 A T G A C - - G A C
점수 0 2 -1 2 2 -1 -1 2 0 2

가 되어 총 점수는 7점이 된다. 가장 가능성이 높은 배열을 찾기 위해서는 가장 높은 점수를 찾는 경로를 찾게 된다.

실제 상황에서는 개발된 software를 이용하게 되는데 이때 일정한 입력 양식이 필요하다. 흔히 FASTA 양식을 사용하며, 첫 줄은 이름이나 관련 정보를 기술하고 다음 줄부터 염기서열을 입력한다. 그리고 첫 줄의 시작은 부등호 '〉'로 시작하고, 염기서열부분은 각 줄이 80문자 이내이어야 하며 숫자는 사용하지 않는다. 또한 사용할 수 있는 문자는 A, C, G, T이고 G 또는 A인 경우는 R로, T 또는 C인 경우는 Y로, G 또는 T인 경우는 K로, A 또는 C인 경우는 M으로, G 또는 C인 경우는 S로, A 또는 T인 경우는 W로, AGCT가 정해지지 않은 경우는 N으로 표기한다. Gap인 경우는 '? 로 표기하고 아미노산의 경우는 정해지지 않은 경우 X로 표기한다. 또한 소문자도 사용이 가능하다.

Multiple align을 위한 프로그램으로 무료로 쓸 수 있는 것 중의 하나가 CLUSTAL W이다. 이는 프로그램은 다운 받아 쓸 수도 있고 외국의 대형 연구소의 서비스

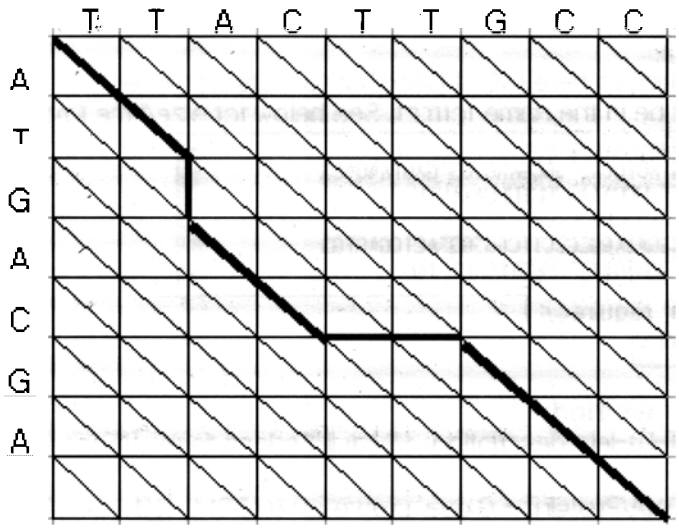

Fig. An alignment path graph

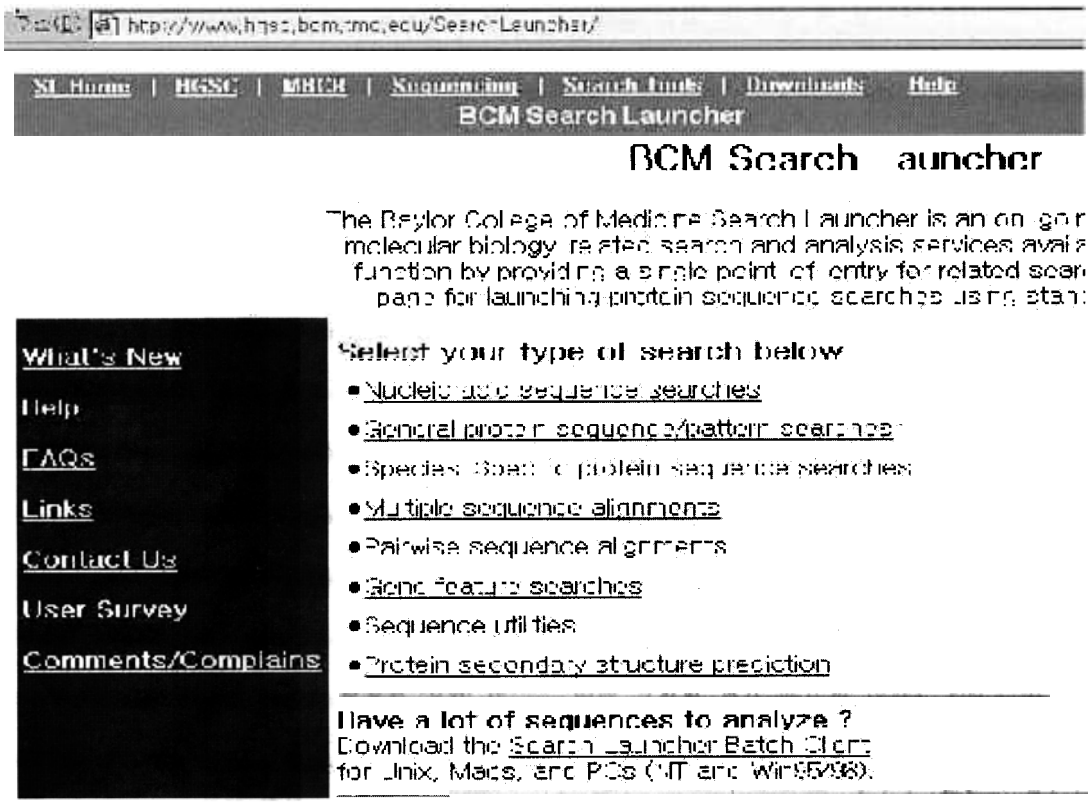

그림 3-8. 베일러 대학의 BCM 서비스 홈페이지

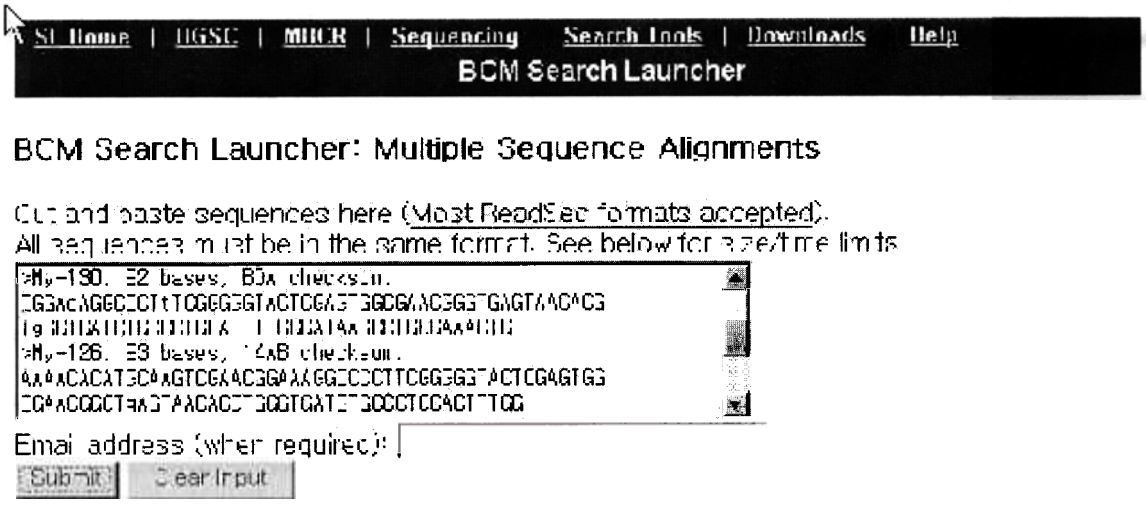

그림 3-9. 3종류 염기서열의 align을 위해 입력한 예

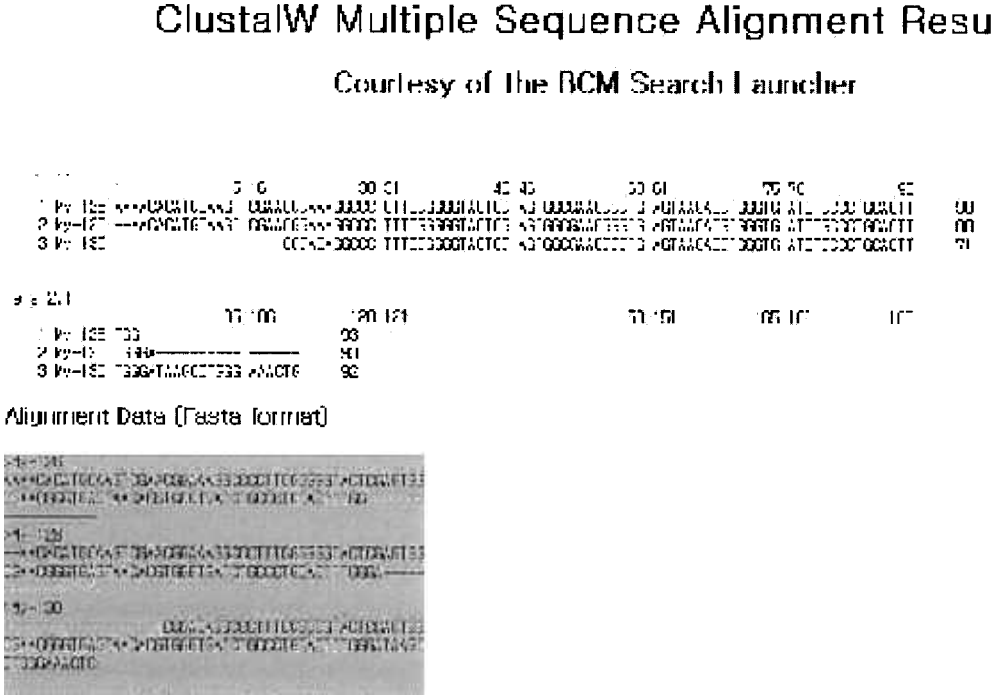

그림 3-10. 3 종류의 염기서열을 ClustalW 1.8 을 이용하여 allign한 결과

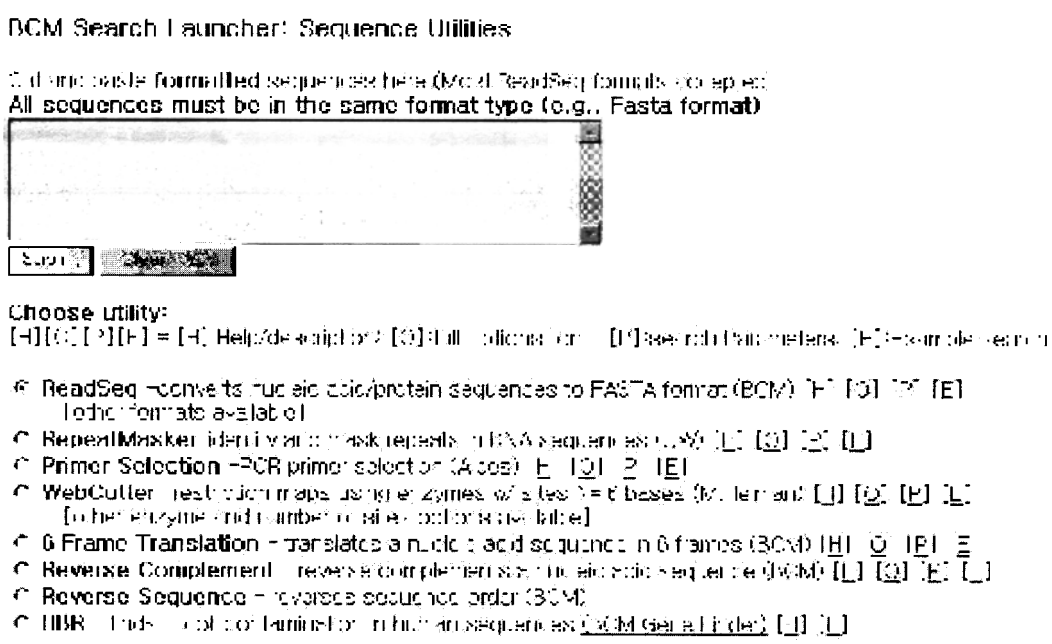

그림 3-11. 베일 대학 BCM의 무료 Software들

를 이용할 수도 있다.

DOS, Windows 95/NT사용자는 ftp:/ftp.ebi.ac.uk/pub/software/dos/clustalw에 접속하여 프로그램을 받을 수 있다. 이것이 번거로운면 외국 연구소에 인터넷을 통해 접속하여 실시간으로 서비스를 제공받을 수 있다.

다음의 예는 미국의 미국의 Bayler 대학 분자생물학 분야에 접속하여 BCM Search Launcher 도구 중 multiple aligment 프로그램을 이용한 것이다. 이 때 입력 양식이 프로그램에서 원하는 양식과 정확히 맞아야 결과를 얻을 수 있는데, 프로그램 상단부에 원하는 양식으로 전환해 주는 프로그램까지 친절하게 구비되어 있다.

4) 그 외의 프로그램들

베일러대학의 BCM에서는 염기서열을 제한 효소로 자르거나 primer를 선정할 수 있는 프로그램, 염기서열을 아미노산서열로 전환하는 프로그램들을 제공하고 있다. 또한 들어가면 각종 laboratory protocol들을 제공하고 있어 누구나 쉽게 참고 할 수 있도록 되어 있다.

이 밖에도 NCBI에서는 E-PCR(Electronic PCR)의 도구를 제공하고 있으며, 염기서열의 분석에 요긴한 Open Reading Frame (OFR) Finder, Taxonomy 등의 무료 서비를 제공하고 있다.

5. 문헌 정보 검색

도서관도 정보화 시대를 맞아 digital library로의 기능을 갖고 시간적 공간적 여러 제약들을 극복하고 독자들에게 자료들을 제공하고 있다. 대부분의 큰 기관의 도서관들은 전자저널 site license를 갖고 있어서 이용자는 ID와 password 또는 할당된 IP address를 사용하여 접속하면 대부분의 저널들을 손쉽게 찾을 수 있다. 대표적인 전자 저널 제공업체로는 Ovid, High Wire Press, Wiley, Web of Science, ProQuest Medical Library, IDEAL-Academic Press 등이 있다.

1) Medline 검색

메디라인은 미국 국립도서관(NLM)의 문헌들의 데이터베이스로서 의학, 간호, 치의학, 수의학, 건강관리, preclinical science의 분야들에 대해 미국을 비롯한 70여 개국에서 발행되는 4000여종 이상의 생물의학 저널에 대한 초록 및 서지정보를 갖고 있다. 미국 국립의학 도서관(NLM)에서는 1997년 6월 26일부터 MEDLINE을 무료로 개방하고 있어 NLM의 Web site(/entrez/)로 접속하면 무료 검색엔진인 Pubmed를 이용하여 전문을 볼려면 사용자 등록을 해야 하며 구독료를 지불해야 한다.

2) 저널의 무료 full-text 검색

임상 미생물 및 감염과 관련하여 American Society for Microbiology (ASM)에서 발행하는 대부분의 저널들은 HighWire Press(http://highwire.stanford.edu/lists/freeart.dtl)에서 제공하고 있다. 이곳은 미국 스탠포드 대학의 연구팀에 의해 수행된 생명과학 관련 전자저널프로젝트로 비영리 학회 및 대학의 학회 및 협회의 저널들을 전자 저널 형태로 제공하고 있다. Journal of Clinical Microbiology, Journal of Bacteriology, Journal of Virology, Antimirobial Agents and Chemotheraphy, Clinical Microbiology Review, Infection and Immunity 등의 저널은 발행된지 11개월 후면 누구나 접속하여 전문을 볼 수 있다.

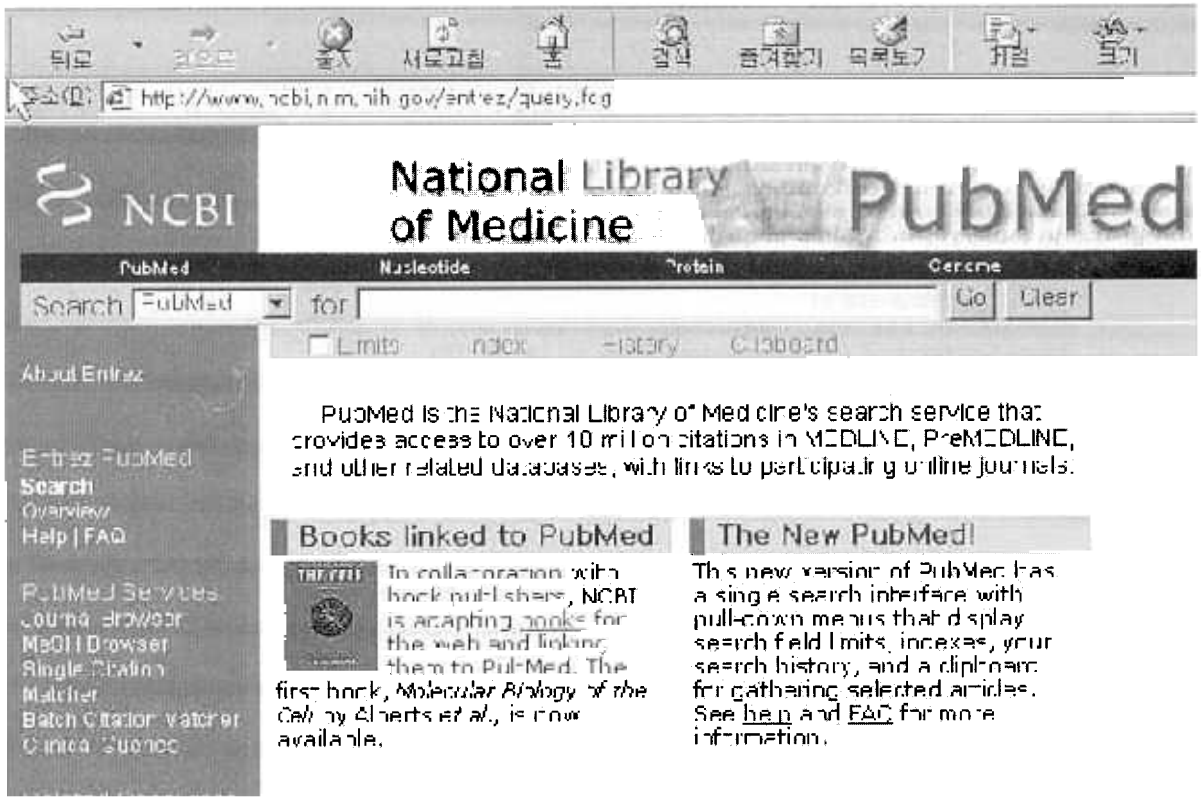

그림 5-1. 대표적인 메디라인 검색 프로그램인 PubMed.

6. 맺음말

흔히들 새 천년을 지식 정보화 사회라고 부른다. 우리는 그 가운데 정보의 홍수 속에서 살고 있다. 홍수가 나 떠내려 가도 내가 조금만 수영을 할 수 있으면 내가 원하는 방향으로 조금씩 나아 갈 수 있는 것처럼, 분자생물학 분야도 우리가 조금만 더 관심과 노력을 기울이면 훨씬 적은 부담으로 우리가 원하는 것들을 얻어 갈 수 있으리라 생각된다.

참 고 문 헌

1. Burks C. *Molecular biology databases. In : Bishop MJ and Rawlings CJ, ed. DNA and protein sequence analysis. 1' st ed. Oxford: Oxford University Press, 1997: 1-30.*
2. Gilbert D. *Free software in molecular biology for Macintosh and MS Windows computers. In : Misener S and Krawets SA, ed. Bioinformatics. 1' st ed Humana Press, 2000: 154-5.*
3. Ostell JM. *The NCBI software tools. In : Bishop MJ and Rawlings CJ, ed. DNA and protein sequence analysis. 1' st ed. Oxford: Oxford University Press, 1997: 31-43.*
4. Altschul SF. *Sequnce comparison and alignment. In : Bishop MJ and Rawlings CJ, ed. DNA and protein sequence analysis. 1' st ed. Oxford: Oxford University Press, 1997: 137-67.*