nature 平成23年3月10日 毎週木曜日発行 第471巻 第7337号
昭和58年6月8日 第三種郵便物認可

発行所：ネイチャー・ジャパン株式会社
東京都新宿区市谷田町2-37 千代田ビル

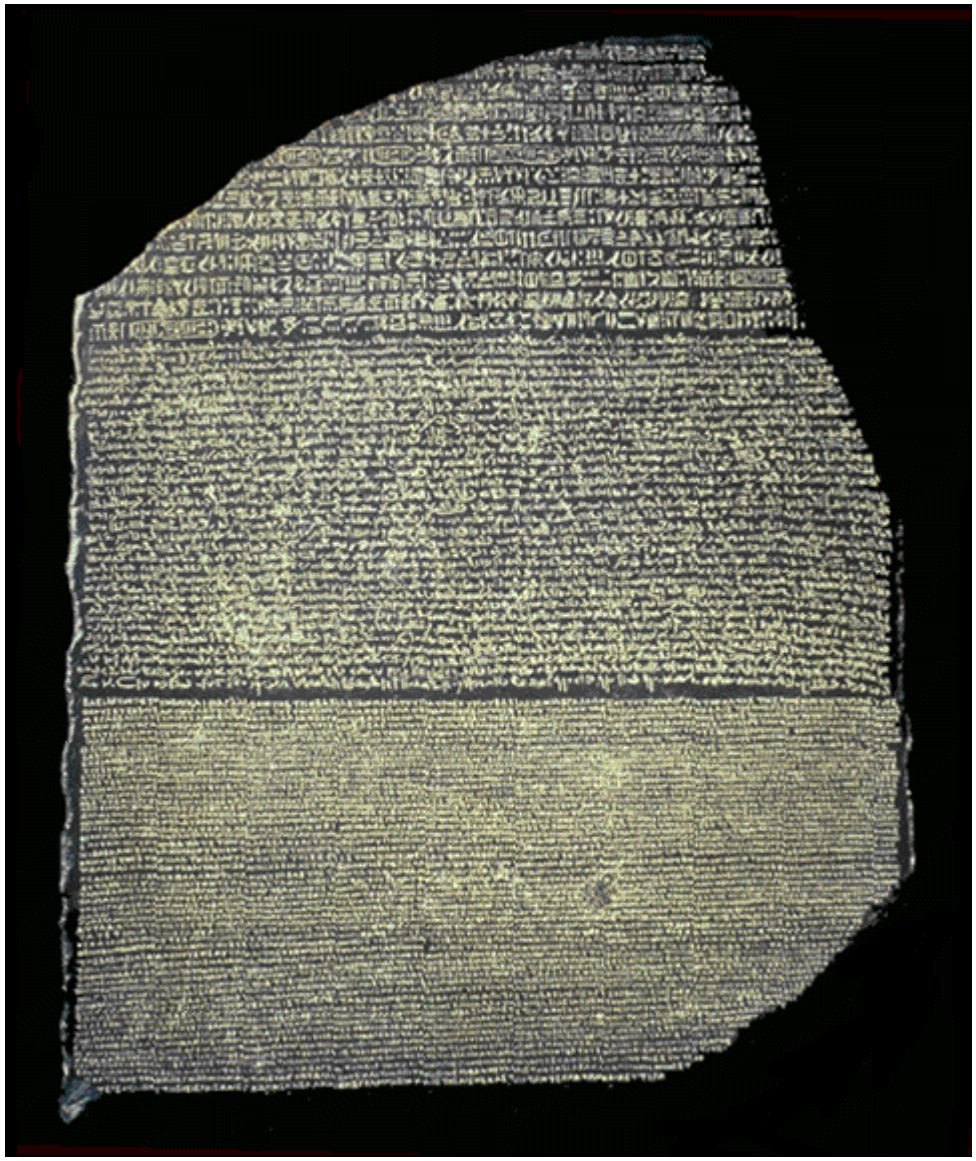
編集発行人：David Swinbanks
発売所：日本出版貿易株式会社

Printed in Japan 定価 9,560円 本体 9,105円

雑誌 29442-3/10

4910294420315
09105

The Human Genome is the "Rosetta Stone of life and death, health, and disease"



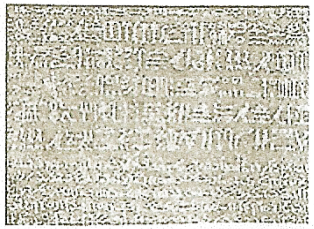
포스트지놈 시대의 생물정보학

http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

2001 년 인간 유전체 초안이 완성되었으며, 이 기념비적인 업적으로 30 억 개에 달하는 인간 게놈의 방대한 바이오디지털(bio-digital) 데이터가 공개되었다. 이로 인해 생물학 자체에는 크게 두 가지 변화가 일어났다. 첫째, 생물학이 정보과학이 된 것이다. 둘째, 생물학 자체가 생물학자, 화학자, 전산학자, 공학자, 수학자, 의학자 등이 협력하여 연구하는 새로운 협력의 장으로 바뀌었다는 점이다. 특히, 이러한 대량의 바이오디지털 데이터를 해석하는 문제로 인하여 바이오인포매틱스의 중대성이 더욱 강조되기도 하였다. 하지만, 아직 게놈프로젝트나 바이오인포매틱스란 용어가 생소한 독자도 있을 듯하니, 로제타스톤에 관련된 역사적인 사건과 인간게놈프로젝트를 비유하며 가벼운 마음으로 이 장을 시작할 수 있도록 하겠다.

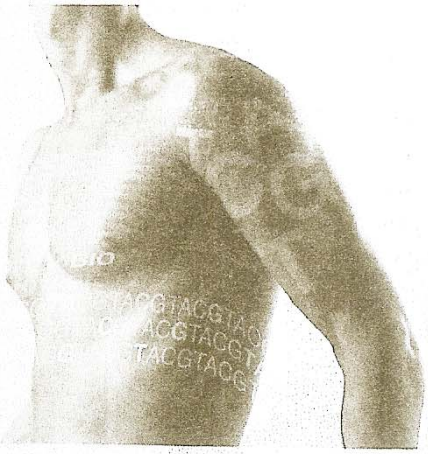
1.1

1799 년 나폴레옹이 이끄는 원정군이 이집트에 이르렀을 때, 이집트의 유적과 유물에 새겨져 있는 이상한 기호들을 수없이 발견했지만, 그 의미를 해독할 수는 없었다. 그러던 중 알렉산드리아에서 동쪽으로 60km 떨어진 '로제타' 라는 마을에서 요새를 쌓던 도트풀이라는 병사에 의해 '로제타스톤(Rosetta Stone)' 이 발견되게 되었다. 로제타스톤은 길이 1.25m, 너비 0.7m 인 검은 돌로, 석비를 본 순간 학자들은 돌에 새겨진 문자가 신의 메시지라고 생각하였으며, 4,000 년이나 쓰였던 상형문자를 풀 열쇠가 될 수 있을 것으로 기대하였다. 심지어는 그 당시 전쟁의 승패가 로제타스톤에 적혀진 신의 계시를 어느 나라가 먼저 해독하느냐에 달려 있다고까지 생각하였다. 언어, 수학, 역사, 고고학, 의학, 종교를 연구하는 유럽 각국의 학자들이 모두 암호해독에 경쟁적으로 매달리게 되었으나, 이집트 상형문자가 중국어와 같은 뜻글자라 가정하고 실마리를 찾았기에, 그 의미는 쉽게 풀리지 않았다. 그러던 중 프랑스의 천재 언어학자 샹폴리옹이 나타났으며, 마침내 이집트 상형문자를 푸는 기본 원리가 밝혀지게 되었다. 샹폴리옹은 신성문자의 몇몇은 형태를 나타냄과 동시에 음에 기반하고 있음을 가정하였고, 이러한 가설하에 알아낸 몇몇 문자를 기초로 전후관계를 따져 남은 문자를 추리하여 풀어나갔던 것이다. 비록 로제타스톤에 새겨진 내용이 처음에 생각했던 것처럼 중요한 신의 계시는 아니었지만, 로제타스톤의 비밀은 이와 같은 방법으로 23 년 만에 밝혀지게 되었던 것이다.



로제타스톤(Rosetta Stone)
(사진출처 : http://www.ancientegypt.co.uk/writing/rosetta.html)

2000 년 6 월 26 일 비영리 단체인 인간게놈사업 컨소시엄과 미국의 생명공학기업 셀레라지노믹스사가 인간게놈지도 초안을 발표, 신의 대륙에 착륙했다고 비유하며, 전 세계를 놀라게 하였다. 인간은 A(아데닌), T(티민), G(구아닌), C(시토신)란 네 가지 문자로 쓰여졌으며, DNA 서열에 의하여 결정되는 생명의 기본 프로그래밍 코드, 즉 '게놈(genome)' 정보를 모두 밝혀내기에 이른 것이다. 궁극적으로 게놈프로젝트가 가져다 준 것은 막대한 바이오디지털 정보이다. 마치 컴퓨터가 0과 1로 표현되는 정보를 저장하고 있듯이, 인간은 A, T, G, C라는 네 개의 문자로 표현된 30 억 개의 정보를 저장하고 있고 이 데이터를 컴퓨터로 옮겨 저장하거나, 그 의미를 분석하는 일이 가능해졌다.



게놈: 인간의 손에 넘어온 신의 언어

1.2

미지의 암호를 풀어낸다는 개념에서, 인간게놈지도 초안은 21 세기의 로제타스톤에 비유될 수 있다. 그렇지만, 앞절에서의 로제타스톤과 인간게놈프로젝트의 비유는 복잡도 면에서 서로 너무 차이가 난다는 점에서 다소 무리였다. 다만, 인간게놈프로젝트를 잘 이해하지 못하는 독자에게 그 의미를 단순화시켜, 정보학의 측면에서 쉽게 이해할 수 있도록 도입한 것이다. 로제타스톤의 글자처럼, 인간게놈프로젝트의 바이오디지털 데이터, 즉 게놈도 단순한 데이터일 뿐이라는 점을 강조하려 한 것이다. 다만, 인간게놈프로젝트를 잘 이해하지 못하는 독자에게는, 프로젝트의 의미를 단순화시킴으로써, 쉽게 이해할 수 있는 효과가 있었을 것이다. 로제타스톤의 문자처럼, 인간게놈프로젝트의 바이오디지털 데이터도 단순한 데이터일 뿐이라는 점을 강조하려 한 것이다.

http://www.hugo-international.org/abt_history.php

복잡도 면에서는 좀 무리한 비유였을지 모르는 반면, 로제타스톤 사건이 현재의 인간게놈 프로젝트와 또 한 가지 중요한 유사점이 있기는 하다. 문제를 풀어내려는 데 다양한 학제간의 연구가 시도되고 있다는 것이다. 마치 200년 전 다양한 분야의 학자들이 로제타스톤의 비밀을 풀기 위해 모여들었던 것처럼, 게놈이라는 신의 언어에 담긴 비밀을 해독하기 위해 의학, 생물학, 물리학, 전산학, 화학, 수학자들이 몰려들고 있다. 그리고 게놈정보의 방대함과 복잡함으로 인해, 이 암호를 해독하기 위한 가장 중요한 도구로서, 컴퓨터의 힘을 응용할 수 있는 바이오인포매틱스의 중요성이 강조되고 있다.

그렇다면 바이오인포매틱스(bio-informatics, 혹은 바이오정보학)란 무엇인가? 그 전에 '정보학(Informatics)'에 대해 생각해 보자. 정보학이란 정보기술(information technology)과 다소 비슷하게 사용되는 용어로, 컴퓨터를 이용한 정보조직 시스템 및 정보의 발생, 전달, 수집, 축적, 처리 등과 관련된 이론과 실제 운용에 관한 학문이다. 따라서 정보학은 그것이 적용되는 분야에 따라 간호정보학, 경영정보학, 문헌정보학, 언론정보학, 의료정보학 등 많은 다른 학문과 결합할 수 있다. 마찬가지로, 생명현상에서 발생하는 많은 정보를 다루는 정보학(Informatics)이 바로 바이오인포매틱스인 것이다. 간단하게 말하자면, 생물학과 전산학, 수학, 의학, 공학의 '컨버전스'가 바로 바이오인포매틱스이다. 바이오인포매틱스 분야는 분자생물학이 몇십 년 전 생물학 전 분야에 새로운 기술과 철학을 도입하여 각광받은 것처럼 여러 가지 기술혁명을 일으키며 수년 안에 융합 학문 분야의 가장 중요한 핵심 분야로 부각될 것이 자명하다.

한편, 바이오인포매틱스는 크게는 게놈프로젝트에 관련된 계산분자생물학(Computational Molecular Biology)을 말하기도 하지만 최근 이 분야의 연구규모가 방대해짐에 따라 점차 계산분자생물학과는 분리되어 독립된 학문으로 떨어져 나오고 있으며, 대량의 서열 데이터를 분석하는 게놈정보 학문으로 구체화되는 경향도 있는 듯하다. 어쨌든, 2000년 발표된 인간게놈프로젝트의 완성을 기점으로 바이오인포매틱스의 중요성이 더욱 강조된 것만은 분명하다. 다음 절에서는 현재의 게놈 혁명이 일어나기까지의 역사적인 사건에 대해서 정리해 보기로 하자.

1.3 [illegible]

많은 사람들이 유전자에 대하여 연구해 오던 중 1953년 미국의 제임스 D. 왓슨(James D. Watson)과 영국의 프랜시스 크릭(Francis Crick)에 의해 DNA의 3차 구조를 밝혀 분자생물학의 새로운 장을 열었다. 이 때부터 DNA에 대한 연구는 본격화되기 시작하였다. 인간의 모든 유전정보를 담고 있는 게놈을 분석하려는 시도는 1988년 미국 에너지성(DOE: Department of Energy)과 미국 국립보건원(NIH: National Institutes of Health)에서 논의가 시작되었지만, 인체게놈을 완벽하게 분석하자고 미국 에너[illegible]이 제안했을 때는 그 중요성에 대해 많은 논쟁이 있었다.

이 프로젝트는 인간이라는 생물학적 정보를 디지털화하는 것으로 비유할 수 있었다. 컴퓨터 프로그램이 0과 1로 이루어진 것처럼 인간의 게놈은 A, C, G, T라는 기호로 표기되는 네 종류의 염기가 특정한 순서로 배열되어 이루어진 신이 만든 프로그램 코드이며 이러한 서열을 완전 구명하는 것을 목표로 한 방대한 계획이었던 것이다. 하지만 15년에 걸쳐 연구비 30억 달러를 투자하기 때문에, 달 착륙에 비유되는 초대형 사업구상이었다. 1989년 미국 국립인체게놈연구소(NHGRI: National Human Genome Research Institute)가 노벨 수상자인 왓슨을 초대 소장으로 발족하였다. 그 후 1993년부터 프랜시스 콜린스 박사가 인간게놈프로젝트를 주도해 나갔다. 1990년 다국적 인간게놈사업 컨소시엄(HUGO: The Human Genome Organisation)을 구성해 미국 외에 18개국 350여 개의 연구소가 참여한 가운데 2005년까지 인간유전자 지도를 완성하겠다는 취지로 인간게놈프로젝트가 시작되었다.

[그림 1.3] 크레이그 벤터 (J.Craig Venter) (사진출처 : http://www.forbes.com)

한편, 1998년 5월 분자생물학계의 권위자인 크레이그 벤터(J.Craig Venter) 박사가 셀레라지노믹스(Celera Genomics)라는 벤처기업을 설립하였다. 미국 정부 게놈프로젝트의 일원이던 벤터 박사는 자신의 새로운 분석방식이 거부되자 독자적으로 셀레라지노믹스사를 설립하게 된다. 그리고 그는 미국의 재정지원을 전혀 받지 않고 2000년까지 독자적으로 인간게놈프로젝트를 완료하겠다고 선언하였다. 이로 인해 인간게놈프로젝트는 민간 대 공공사업의 경쟁으로 변질되었다.

하지만, 2000년 3월 미국 클린턴 대통령과 영국 블레어 총리는 인간게놈 분석결과를 인터넷에 공개하겠다고 선언하였다. 결국 2000년 6월 26일 미국 백악관 기자회견에서 인간 유전정보인 게놈의 분석을 둘러싸고 경쟁을 벌여온 인간게놈프로젝트(HGP: Human Ge-

nome Proj… 진영과 셀레라지노믹스사는, 서로 악수하며, 유전자 염기서열 초안을 완료했다고 공동 발표하게 된다.

[그림 1.4] 인간유전체 서열이 발표된 네이처(*Nature*, 934-942, Feb 2001)와 사이언스(*Science*, 1304-1351, Feb 2001)

이로 인해 세계의 언론은 섣부른 낙관과 함께 인간게놈프로젝트의 쾌거를 보도하였다. 하지만, 발표된 인간게놈지도는 아직 띄어쓰기가 되어 있지 않은 인쇄본만을 가지고 있는 것에 비유될 수 있을 뿐이다. 그 의미를 판독하는 데는 몇십 년, 아니 몇백 년이 걸릴지도 모른다는 점에서 이것은 또 다른 새로운 시작에 불과하다. 마치 로제타스톤을 해독할 당시처럼 수많은 학자들이 또다시 수많은 실험과 연구결과를 발표할 것이고, 새로운 데이터에 대한 수많은 통합 분석결과들을 바이오인포매틱스 전문가들이 발표할 것이고, 이러한 과정에서 진리를 찾기까지 수많은 시행착오를 겪을 것이다.

1.4 바이오디지털 정보들

앞절에서는 게놈의 역사에 대해서 대강 살펴보았는데, 이번에는 대표적인 바이오디지털 정보들은 어떤 것이 있을지 설명해 보겠다. 이를 위해서는 먼저 게놈이란 무엇인지에 대한 올바른 이해가 다시 한 번 필요하다.

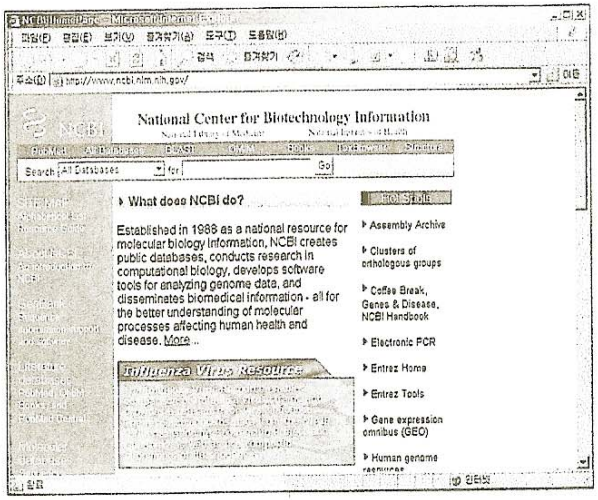
생명체는 **세포**로 이루어져 있다. 세포는, 단세포 생물에게는 그 자체가 하나의 생물 개체가 되고, 다세포 생물에게는 한 개체를 이루는 여러 조직(tissues)들을 구성하는 기본 단위가 된다. 세포는 세포막, 유전물질, 핵, 리보솜, 미토콘드리아, 엽록체, 세포질, 그리고 기타 세포 구성물질들로 이루어진다. 그 중에서도 핵[1] 안에는 데옥시리보핵산(deoxyribonucleic acid), 즉 DNA가 들어 있는데, 이것이 유전자의 물질적 실체이다. 그림 1.5에서 볼 수 있듯이 핵 속에 1번부터 22번, 그리고 성을 결정하는 염색체 X, Y 이 모두 23쌍의 염색체를 갖게 되는데 이 23개의 염색체 세트를 게놈(genome)이라 한다. 한 개의 염색체에는 수천 개의 유전자가 들어 있으며, 게놈은 다시 약 30억 쌍의 염기로 구성된다. 인간의 유전정보는 무게가 1천억 분의 1g에 불과한 DNA 가닥에 담겨 있는 것이다.

결론적으로 **게놈**이란 한 생명체가 가지고 있는 전체 DNA를 말하는 것이며, '게놈'이란 용어는 'gene'과 'chromosome'의 합성어인 'genome'을 독일식으로 발음한 것으로 우리나라에서는 게놈이라는 용어로 통일하여 사용하고 있다.

[그림 1.5] **핵 속의 염색체, 그리고 염색체를 구성하는 DNA 이중나선**
(사진출처 : http://www.alzheimers.org/unraveling/images/large/DNA-HIGH.jpg)

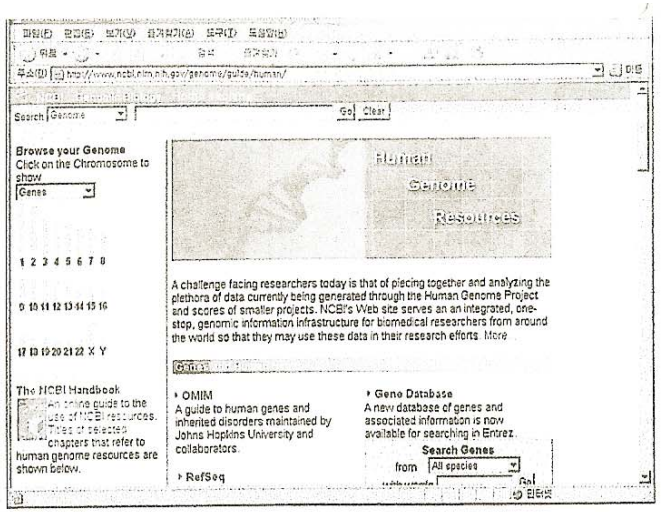
[1] 세포는 크게 진핵세포와 원핵세포로 나눌 수 있는데, 일반적으로 인간을 포함한 대부분의 다세포 생물은 진핵세포로 이루어져 있고, 미생물 등의 단세포 생물은 원핵세포로 이루어져 있다. 진핵세포와 원핵세포의 대표적인 차이점은 핵의 유무이다. 핵이 있는 진핵세포는 핵막으로 둘러싸인 핵 안에 유전물질을 갖는 반면에, 핵이 없는 원핵세포는 핵이 따로 존재하지 않고 세포질에 유전물질이 그대로 존재한다.

바이오 전공자가 아니거나 최근의 바이오 데이터베이스 현황을 모르는 사람이라면 지금쯤, 그렇다면 이러한 방대한 양의 서열 데이터가 도대체 어디에 모아지고, 저장되어 있는지 궁금해질 것이다. 완성된 게놈프로젝트의 결과를 비롯하여 세계 각지에서 발생하는 생물학적 서열 데이터들의 대부분은 분석결과와 함께 전 세계의 연구자들이 공유할 수 있도록 공개된다. 대표적인 사이트 중 하나가 미국 국립보건원(NIH: National Institutes of Health) 산하의 국립생물공학정보센터(NCBI: National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/)이다.



[그림 1.13] NCBI의 사이트

예를 들면 앞절에서 언급하였던, 인간게놈프로젝트의 결과물, 즉 30억 개의 DNA 서열이나 분석자료 및 연구내용들도 바로 이 사이트를 통해서 여러분들이 직접 다운로드받을 수도 있을 것이다. NCBI의 http://www.ncbi.nlm.nih.gov/genome/guide/human/를 직접 방문하여 인간게놈 데이터를 확인해 보기 바란다.



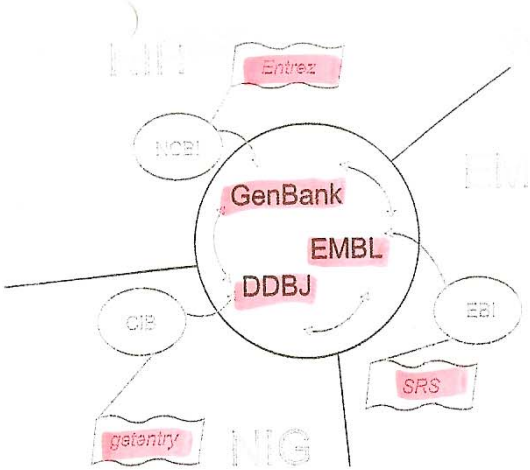
[그림 1.14] 인간게놈 데이터

우리를 더욱 흥미롭게 하는 것은, 사실상 이러한 데이터를 컴퓨터에 저장하고 또 여러 가지 알고리즘을 사용해서 분석하는 일이 가능하다는 점이다. 기계학습기법이나, 인공지능에서 쓰이는 다양한 알고리즘들이 세포 안에 있는 바이오디지털 데이터에도 그대로 적용될 수 있다는 점은 놀라운 사실이다. 이들 알고리즘에 의한 서열분석방법에 대해서는 1.6절에서 간단히 소개하기로 하자.

한편, NCBI 사이트에 등록되는 서열 데이터들은 사이트내의 GenBank라는 서열 데이터베이스에 저장된다. GenBank에 있는 모든 정보는 자신의 연구결과를 공개하고자 하는 학자나 연구자들이 직접 투고함으로써 구축되는데, 이 사이트는 공개된 모든 서열정보를 보관하고 있을 뿐 아니라 각종 생물학 관련 논문정보를 검색할 수 있는 기능도 제공하고 있다. 뿐만 아니라, 서열정보를 자체적으로 분석/가공한 결과를 통해 얻은 새로운 정보, 예를 들면 COGs 같은 정보도 제공하고, 이 외에도 연구자들이 직접 간단한 분석작업을 할 수 있도록 몇 가지 소프트웨어도 서비스하고 있다. GenBank 데이터 형식에 관해서는 7장에서 자세히 학습하게 될 것이다.

이와 비슷하게 서열정보를 보관, 관리하는 곳으로 유럽의 유럽분자생물실험실(EMBL: European Molecular Biology Laboratory, http://www.ebi.ac.uk/embl)과 일본의 일본DNA 데이터뱅크(DDBJ: DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp)도 있으며, 이들 세 사이트는 국제협약에 의해 서로 동일한 데이터를 유지하고 관리한다.

http://www.ncbi.nlm.nih.gov/



[그림 1.15] GenBank, EMBL, DDBJ 데이터베이스의 관계
(사진출처 : http://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Databases/collab.gif)

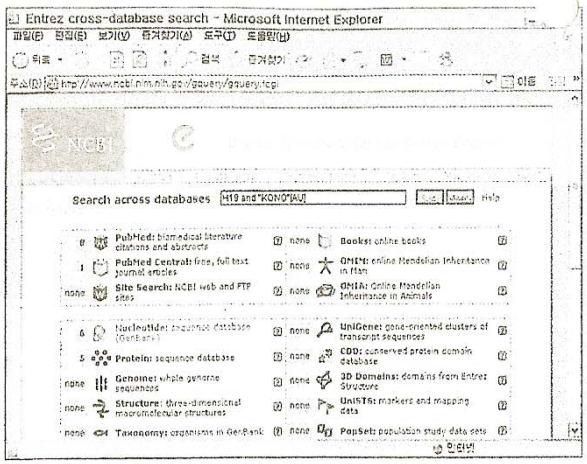
한편, 이러한 사이트에서는, 각종 정보를 검색하기 위한 검색엔진도 제공되게 마련이다. 그림 1.15에서 보다시피 GenBank, EMBL, DDBJ 각각에 대하여, Entrez, SRS, getentry 등의 검색엔진이 제공된다. 예를 들어, NCBI에서는 그림 1.16에서처럼 어떤 질의를 엔트레즈(Entrez) 검색창에 입력하여 유전자 이름, 관련된 논문, 단백질 서열, 3차원 구조 등의 정보를 한꺼번에 검색할 수 있다. 엔트레즈 블리언 검색의 형식은 다음과 같다.

```
키워드[태그] 연산자 키워드[태그] 연산자 키워드[태그] ...
```

예를 들면 H19이라는 유전자에 관련된 검색결과를 찾으려고 하고, 검색결과를 kono라는 일본 저자가 쓴 논문으로 한정지으려고 할 때는

```
H19 and "Kono"[AU]
```

와 같은 형식으로 검색하면 된다. 그림 1.17에 의하면, 펍메드(PubMED)라는 논문 데이터베이스에서는 8개의 결과가, 그리고 뉴클리오타이드(nucleotide) 서열은 6개의 검색결과가, 단백질(protein) 서열은 5개의 검색결과가 나타난다. 각각을 클릭하면 관련 정보를 좀 더 상세히 볼 수 있다.



[그림 1.16] Entrez 검색 (http://www.ncbi.nlm.nih.gov/gquery/gquery.fgci)

1.6 바이오디지털 데이터 분석

마지막으로 화두를 바이오 데이터의 관리가 아닌 바이오 데이터의 분석으로 돌려보기로 하자.

바이오인포매틱스 전문가들이 해야 하는 일은 크게 두 가지로 나누어볼 수 있을 것이다. 하나는 데이터 관리에 관한 것이고, 다른 하나는 데이터 분석일 것이다. 먼저, 속도를 다루는 바이오인포매틱스 연구자들은 모든 데이터를 통합해서 분석하기를 원하기 때문에, 어떤 데이터가 존재하는지를 정확하게 알기를 원하고, 항시 새로운 데이터들이 수정되거나 업데이트되는 순간에 정확하게 이런 일들을 인지하기를 원한다. 바이오인포매티션이 두 번째로 하기를 원하는 작업은 바이오디지털 데이터들의 의미를 여러 가지 알고리즘을 사용하여 예측해 내는 일일 것이다. 쉽게는 연구자들이 새로이 규명된 DNA 정보가 기존에 존재하던 어떤 서열과 가장 유사한지를 찾아내는 일을 할 수 있을 것이다. 하지만, 바이오인포매티션은 늘 좀더 섬세한 질문들에 대해서 대답할 수 있는 애플리케이션을 만들기를 원하고 있다.

서열 분석이라고 하는 것은 DNA에서 유전자를 찾고 이것이 RNA와 단백질로 변환되는 과정에 대한 연구를 포괄하는 용어이다. 서열 분석의 원리는 기본 알고리즘에 대한 이해뿐 아니라 실제적인 응용에 대한 광범위한 이해를 필요로 하며, 관련 데이터는 DNA, RNA, 단백질 서열 등이 있다. 생물분자 서열 분석의 문제를 DNA, RNA, 단백질, 패턴 발견의 네

가지 레벨로 분류해 설명하면 다음과 같다.

1.5절에서도 설명했듯이 게놈 서열은 유전자 부분과 비유전자 부분으로 나뉘고, 유전자 부분은 다시 전사되는 엑손과 전사되지 않는 인트론 부분으로 나뉜다. 그러므로 DNA 수준의 문제에서는 유전자 부분, 비유전자 부분, 엑손, 인트론, 프로모터, 터미네이터 등의 위치를 컴퓨터로 찾는 것 등에 응용된다. RNA 수준에서의 문제는 염기쌍간의 상호작용, 염기쌍과 그들의 에너지간 상호작용, 염기쌍과 에너지간 상호작용에서의 자유 에너지에 기초해서 mRNA, tRNA, rRNA, snRNA의 서열이나 2차 구조를 계산하고 점수를 산정하는 문제 등을 말한다. 이러한 문제는 평가대상이 되는 구조의 수가 많기 때문에 일반적인 컴퓨터 알고리즘으로 해결하기 어려운 문제이며, 주로 전산학에서 말하는 신경망(neural network) 학습이나 확률문법 등의 기법이 활용된다. 단백질 수준에서의 문제로는 단백질의 구조와 기능예측문제에 응용될 수 있다. 단백질의 3차 구조 예측은 포스트게놈 시대 바이오인포매틱스가 해결해야 할 난제이다. 단백질의 구조는 그림 1.10에서와 같이 펩티드 사슬의 접힘(혹은 폴딩)을 통해 생성되는데, 가능한 폴딩 패턴의 수가 무한대이며 단순한 열역학문제로만 풀 수 없고 단백질 주위의 복잡한 상호작용을 모두 고려해야 풀 수 있다. 단백질의 구조적 특성에 관하여 1차 서열정보를 통해서 3차원 구조를 예측하기를 원할 때, 이를 단백질 접힘(protein folding) 문제라 한다. 단백질에 대한 완벽한 구조예측방법은 현재로서는 알려져 있지 않으나 특별한 경우에 적용할 수 있는 2차원 및 3차원 알고리즘들이 존재하고 있다. 실험을 통하지 않고 단백질 3차원 구조를 예측하는 방법으로는 상동성 모델링(homology modeling), 스레딩(threading), 그리고 ab. initio 방법 등이 있다. 게놈프로젝트에서 나온 많은 1차 서열 데이터들은 새로운 방식의 3차 구조예측방법을 제시하고 있다. 마지막으로 패턴 발견문제인데 이는 서열 데이터의 분석보다는 새로운 생물학적 패턴을 발견하는 것이다. 이러한 패턴 발견문제는 전체 게놈 수준에서 구조적인 조직을 찾는 문제를 포함하는데, 반복되는 영역의 발견, 유사도의 계산, 의미 있는 패턴의 발견 등이 일종의 데이터마이닝 문제에 속한다.

1.7 마무리

결국 생명체내에 복잡하게 존재하는 유전적 네트워크를 이해한다는 것이 게놈프로젝트의 최종 목적일지도 모른다. 기초적인 대사회로 등의 구축은 어느 정도 진행되어 왔으나 여전히 생물학 지식의 한계와, 관련 정보들의 방대함으로 인해 유전정보의 총체적 네트워크 구축은 여전히 초보 단계에 머무르고 있다. 다만 최근에 급부상하고 있는 DNA 칩 기술의 발전은 포스트게놈 시대의 연구를 엄청난 속도로 앞당길 것으로 전망되고 있다.

결과적으로, 인간게놈프로젝트를 계기로 의/생물학의 패러다임이 질 중심의 개인적 연구에서 점차 기계와 컴퓨터의 도움을 받아 전체 유전자의 집단적 움직임을 추적하는 양적인 새로운 방식으로 바뀌고 있다. 이에 바이오인포매틱스의 중요성은 더욱 강조될 것이다. 독자들도 이제 어느 정도 바이오인포매틱스와 게놈 혁명에 대한 최근의 변화에 대해서 이해를 할 수 있었으리라 생각한다.

이제부터 초보적인 것이기는 하지만, 게놈 분석과 생명현상에 대한 모델링 작업을 자바라는 언어를 통해서 실습해 보기로 하자.

인터넷 리소스

바이오인포매틱스 학회

The Bioinformatics Organization: http://www.bioinformatics.org/
iSCB(International Society for Computational Biology): http://www.iscb.org/
KSBI (한국생물정보학회): http://www.ksbi.or.kr

온라인 튜토리얼

NCBI Bookshelf: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books/
NCBI Science Primer(What's in a genome?):
http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html
NCBI Bioinformatics Primer:
http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html
About NCBI: http://www.ncbi.nlm.nih.gov/About/index.html
Entrez Tutorial: http://www.ncbi.nlm.nih.gov/Entrez/tutor.html

바이오 데이터베이스

National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov
European Bioinformatics Institute: http://www.ebi.ac.uk/
National Institute of Genetics: http://www.nig.ac.jp/index-e.html
Sakura: http://sakura.ddbj.nig.ac.jp/
DNA Database of Japan: http://www.ddbj.nig.ac.jp/
EMBL nucleotide sequence database: http://www.ebi.ac.uk/embl/
GenBank: http://www.ncbi.nlm.nih.gov/Web/Genbank/
Swiss-Prot: http://www.expasy.org/sprot/

들로부터 특정서열을 확인하는 데 충분하지 않다. 즉 SWISS-PROT의 'A10234'는 PIR의 'A10234'로부터 구별되어야만 한다. DDBJ/EMBL/GenBank 염기서열 데이터베이스는 공통의 등록번호를 공유한다. 따라서 EMBL의 'A12345'는 젠뱅크 또는 DDBJ의 'A12345'와 동일하다. 더 복잡한 것은 젠뱅크 데이터베이스 기록이 단일 서열을 나타낸다 하더라도, 단백질 데이터은행(PDB)[13] 기록은 한 개 이상의 서열을 가지고 있는 단일 단백질 구조를 포함할 수 있다. 이런 이유로, 단백질 데이터은행 서열 검색번호(PDB Seq-id)는 분자 이름, 단일서열의 검색번호를 포함한다. 자주 이용되는 서열 검색번호의 양식과 타입은 이 장의 다음 부분에 기술하였다.

로커스명

로커스라는 단어는 젠뱅크와 DDBJ 기록의 로커스행(locus line)과, EMBL 기록의 ID 행에 나타난다. 사용 초기에는 로커스가 분리된 젠뱅크 데이터 기록을 확인하는 유일한 수단이었다. 유전자 로커스명(genetic locus name)과 유사하게 로커스는 데이터 고유번호(identifier)로서, 또한 유전자 서열의 기능과 이로부터 유래된 생물체를 부호화하는 연상기호로서 고안되었다. 로커스행은 고정된 양식이기 때문에, 로커스명은 10개 이내의 숫자와 대문자로 제한된다. 젠뱅크에서는 몇 년 동안, 로커스명의 첫 세 문자를 생물체 코드로 하고, 나머지를 유전자 코드로 하였다(예. HUMHBB는 'human/3-globin region'임). 그러나 로커스명은 특정 유전자 서열 분야의 생물학적 기능이 원래 생각했던 것과는 다른 것이 발견됨에 따라 유전자 로커스명과 마찬가지로 변하게 되었다. 이러한 로커스명의 불안정성으로 로커스를 검색번호로 사용하여 데이터 검색은 분명한 문제점이 되었다. 더욱이 젠뱅크에 수록되는 서열과 생물체의 수가 갈수록 기하학적으로 증가함에 따라, 새로운 로커스명을 고안하거나 기존의 이름을 효과적인 방법으로 갱신하는 것이 불가능해졌다. 따라서 비록 로커스명이 기존 양식을 망가뜨리지 않기 위해 데이터베이스의 제일 첫 줄에 나온다고 해도, 현재 로커스명은 젠뱅크에서 점차 의미를 잃어가고 있다.

등록번호

염기서열 기록에서 고유 검색번호(identifier)로서 로커스명의 사용이 어려워짐에 따라, 국제 염기서열 데이터베이스 연합(DDBJ/EMBL/GenBank)에서는 등록번호(accession number)를 도입하였다. 이것은 정보를 (상대적으로) 안정하게 유지하는 목적으로 고안되었기 때문에 생물학적 의미를 전혀 갖지 않는다. 원래의 등록번호는 한 개의 대문자와 다섯 개의 숫자로 표시되었다. 그러나 새로운 등록번호는 두 개의 대문자와 여섯 개의 숫자를 갖는다. 첫번째 문자는 각 데이터베이스에 배당하여 등록번호가 협력 데이터베이스 간에 호환되어도 고유한 기록을 유지할 수 있도록 하였다(예. 'U'로 시작하는 것은 젠뱅크에서 유래되었다).

13) PDB: Protein Data Bank (http://www.rcsb.org/pdb)는 3차원 생물학적 거대분자 구조 데이터를 분석, 제공하는 사이트이다.

그러나 로커스명을 사용하는 방법보다는 훨씬 개선되었지만, 등록번호는 사용할수록 문제점과 결점이 드러나게 되었다. 예를 들어, 등록번호가 오랜 시간 안정적이라고는 하나, 특정 등록번호로 검색하였을 때 유전자 서열이 항상 같지는 않은 경우가 많다. 이것은 등록번호가 전체 데이터베이스 기록을 다 확인하기 때문이다. 그런데 만약 서열이 최근에 갱신되었다면(시작부위에서 1000bp가 삽입되었다고 하자), 등록번호는 변경되지 않은 채 같은 등록번호의 최신 버전이 된다. 만일, 원래 서열을 분석하여 등록번호 U00001의 100 위치에 단백질 결합 부위가 있다고 기록했다면, 갱신 후에는 완전히 다른 서열이 100 위치에서 발견될 것이다.

등록번호는 젠뱅크 기록의 등록행(ACCESSION line)에 나와 있다. 이 라인의 첫 번째 등록번호는 '일차 (primary)' 등록번호라고 부르고, 이 기록을 검색하는 데 주요 열쇠가 된다. 대부분의 데이터는 일차 등록번호만을 가지지만 경우에 따라서는 등록행의 일차 등록번호 다음에 '이차 (secondary)' 등록번호라는 것이 나와서 데이터 기록의 역사적인 정보를 제공할 수도 있다. 예를 들어, U00001과 U00002가 단일기록으로 합쳐진다면, 새로운 통합 정보는 U00001을 일차 등록번호로 가지며, U00002가 이차 등록번호가 된다. 오래된 기록은 쓸모가 없어지므로 U00002 기록은 젠뱅크에서 제거된다. 역사적으로 이차 등록번호는 항상 동일한 것을 의미하지는 않기 때문에 사용자는 해석에 항상 주의하여야만 한다(각 데이터베이스는 다른 정책을 가지며, 한 데이터베이스에서도 시간이 지남에 따라 달라질 수 있다). 이차 등록번호를 사용하여도 어떤 변화가 어떤 이유로 일어났는지에 대한 정확한 정보를 제공하기에는 충분하지 않다. 그럼에도 등록번호 시스템은 DDBJ/EMBL/GenBank의 기록을 모두 검색하는 데 가장 잘 관리되고 신뢰받는 방법으로 인식되고 있다.

유전자정보 검색번호(gi Number)

1992년 NCBI는 엔트레즈에서 분석되는 모든 서열에 대해 유전자정보 검색번호(Geninfo Identifiers, gi)를 부여하기 시작했다. 여기에는 DDBJ/EMBL/GenBank의 염기서열, 코딩영역특징(CDS feature)을 해석한 단백질서열, SWISS-PROT, PIR, PRF[14], PDB, 특허 등의 단백질서열이 포함되었다. 유전자정보 검색번호는 데이터베이스에 의해 제공된 등록번호와는 별도로 제공된다. 서열 검색번호의 경우 제공된 출처에 따라 그 형태나 의미가 변하는 반면, gi는 제공된 출처와 관계없이 그 형태와 의미가 일정

14) PRF: Protein Research Foundation(http://www.prf.or.jp/en/index.html)은 일본 단백질 연구진흥회에서 운영하는 단백질 데이터베이스 사이트이다.

젠뱅크 포맷

젠뱅크는 서열 기록을 'DNA-중심'으로 보여준다.[반면 젠펩트(GenPept)는 '단백질-중심'의 서열기록이다.] 이 자료들의 호환성을 유지하기 위하여 일부 자료에 대한 지도화 작업이 수행되고 있는데, 다른 서열들의 특징간, 또는 같은 서열의 중복된 특징에 대한 지도화이다.

젠뱅크 포맷에서 코딩영역 특징의 단백질산물은 번역(/translation) 자격항목으로 표시되며, 자체적인 특징을 가진 서열로 표시되지는 않는다. 단백질(/product) 자격항목은 산물 생물서열에서 가장 규모가 큰 단백질 특징이다. 완성 펩타이드, 신호 펩타이드 등의 특징은 NCBI 데이터 모델에서는 단백질 생물서열에 기능해석되어 있는데, 젠뱅크 포맷에서는 코딩영역 구간을 통하여 DNA 좌표 시스템에 지도로 표시되어 있다.

유전자특징은 염기서열의 한 영역을 명명하며, 유전자의 표현형에 영향을 미친다고 알려진 영역을 포함한다. 이 영역에 포함된 다른 특징들은 유전자특징으로부터 유전자 자격항목(/gene qualifier)을 선택할 것이다. 따라서 다른 특징들에 유전자 자격항목에 대한 기능해석을 따로 첨부할 필요는 없다.

파스타 양식

파스타 양식은 정의행(definition line)과 서열 문자를 포함하며 다양한 분석 프로그램에 입력용으로 사용할 수 있다 (3장 참조). 정의행은 꺾음괄호 >로 시작하며 대개는 다음에 예시된 바와 같이 서로 구분되는 형태의 서열 검색번호가 뒤따른다.

```
>gi | 2352912 | gb | AF012433.1 | HSDDT2
```

정의행의 나머지 부분은 해당 서열의 이름인데, 염기-단백질 생물서열 세트의 특징과 기타 정보에 따라 소프트웨어가 만들어낸다.

세그먼트 생물서열에서는 개개의 초벌 생물서열 일부가 각 세그먼트를 분리하는 대시(dash)와 함께 표시될 수 있다.(정규 블라스트 검색 서비스는 이 방법을 사용하여 검색결과를 얻는데, 결과 히트(hit)가 개개의 젠뱅크 기록에 지도로 표시된다.) 세그먼트 생물서열은 단일 서열로 처리될 수 있으며 이 경우 초벌 생물서열(raw Bioseq)의 부분 서열들이 연결될 것이다.(이러한 형식은 엔트레즈에서 블라스트 관련서열을 만드는 데 이용된다. 7장 참조.)

블라스트(BLAST)

기본 국소정렬 검색도구(the Basic Local Alignment Search Tool, BLAST; Altschul 등, 1990)는 서열의 유사성을 밝히는 데 가장 많이 사용되는 방법이다. 블라스트 프로그램은 사용자들이 제공한 검색 대상 서열에 대하여 NCBI의 전체 데이터베이스를 대상으로 하여 검색을 수행한다. 각각의 히트에 대한 출력(output)이 서열-정렬(Seq-align)이며, 이것들이 서열-기능해석(Seq-annot)으로 합쳐진다(블라스트 검색 수행에 대한 보다 자세한 사항은 8장을 참조).

블라스트 검색 결과 얻어진 서열-기능해석은 전통적인 블라스트 결과 리포트를 만드는 데 사용될 수 있으나 엔트레즈나 세퀸 같은 소프트웨어와 함께 사용하면 훨씬 더 유용하다. 이들 소프트웨어의 디스플레이 프로그램인 뷰어는 정렬에 관한 정보를 편리하게 나타내도록 디자인되었다. 예를 들어, 그래픽 보기(Graphical View)는 단지 질의 서열에 관한 삽입(insertion)과 삭제(deletion)를 보여주는 데 반해, 정렬보기(Alignment View)는 정렬된 부위에서 염기(또는 잔기)가 일치하지 않는 부분을 보여준다. 또한 서열보기(Sequence View)는 자세한 정렬 내용을 염기와 잔기까지 보여준다. 전체적인 것에서 세부적인 사항으로 정보를 좁혀나가는 이들 소프트웨어는 한 개 보고서만을 사용하는 것보다 서열 사이의 관계를 훨씬 더 용이하게 파악하도록 해준다.

마지막으로, 서열-기능해석이나 서열-정렬은 공백정렬 프로그램(banded or gapped alignment program) 등의 다른 프로그램을 이용하여 질을 개선할 수도 있다. 이 결과는 디스플레이 프로그램으로 다시 보내서 출력해 볼 수 있다.

엔트레즈

엔트레즈(Entrez) 서열 검색 프로그램(Schuler 등, 1996; 7장 참조)은 NCBI 데이터 모델이 간직하고 있는 연관성을 이용하기 위하여 디자인되었다. 예를 들어 서열 기록에 기재된 인용문헌은 메드라인 UID 또는 펍메드 ID를 포함할 수도 있는데 이를 이용해 펍메드 논문에 직접 링크될 수 있고 그를 엔트레즈가 검색해낸다. 또한 코딩영역 특징의 산물서열-위치는 단백질 산물의 생물서열을 제시하는데, 이것도 엔트레즈가 검색할 수 있다. 관련 기록들의 검색은 데이터 모델에서 실행 버튼을 누르기만 하면 된다. 엔트레즈의 유전체 분과(genome division)는 이 데이터 모델을 한층 더 잘 활용하여 대규모 유전체의 특정부위를 바로 보여준다. 웹엔트레즈에서 프롯테이블(ProtTable) 버튼을 누르는 경우와 마찬가지이다.

세퀸

세퀸(Sequin)은 DDBJ/EMBL/GenBank 데이터베이스에 데이터를 투고할 때 이용되는 프로그램으로, 가공하지 않은(raw) 서열 데이터와 기타 생물학적 정보를 취합하여 투고 가능한 데이터 형태로 만들어 준다(4장 참조). 세퀸은 NCBI 데이터 모델을 활용하여, 반복 투고되는 서열과 비교함으로써 입력되는 사항의 유효성을 검증한다. 예를 들면, 서열 투고자가 염기서열과 단백질서열을 모두 제공하기 때문에 세퀸은 코딩영역 위치를 용-

http://genome10k.soe.ucsc.edu/

http://www.onekp.com/project.html

The 1KP Project

Links

Home

What is the 1KP Project?

Why Sequence 1000 plants?

Transcriptomes not Genomes

Essential Plant Phylogeny

Media

Contact Us

What is the 1KP Project?

A new initiative launched in November 2008 will acquire gene sequence information for 1000 plant species. Our mandate includes everything from algae to land or aquatic plants, with a particular focus on plants that make valuable bioproducts. The project is led from Alberta by Gane Ka-Shu Wong and Michael Deyholos, and the sequencing will be done at BGI-Shenzhen. An international multidisciplinary consortium has been formed to participate in this research. All of our sequence data will be released to the public upon publication, specifically through GenBank and other open access websites. This project will begin what we hope is a longer term effort by the research community to study the vast biodiversity that to date has barely been touched by genomics. Not only will this lead to great science, but also, we believe it will lead to commercialization opportunities.

● **Entrez** search:
ex)
Magnolia and complete and "Soltis"[AU]

TABLE 7.1. Entrez Boolean Search Statements

General syntax:

`search term [tag] boolean operator search term [tag] . . .`

where `[tag]` =

[AD]	Affiliation
[ALL]	All fields
[AU]	Author name O'Brien J [AU] *yields all of* O'Brien JA, O'Brien JB, etc. ''O'Brien J'' [AU] *yields only* O'Brien J
[RN]	Enzyme Commission or Chemical Abstract Service numbers
[EDAT]	Entrez date YYYY/MM/DD, YYYY/MM, or YYYY
[IP]	Issue of journal
[TA]	Journal title, official abbreviation, or ISSN number Journal of Biological Chemistry J Biol Chem 0021-9258
[LA]	Language
[MAJR]	MeSH major ropic *One of the major topics discussed in the article*
[MH]	MeSH terms *Controlled vocabulary of biomedical terms (subject)*
[PS]	Personal name as subject *Use when name is subject of article*, e.g., Varmus H [PS]
[DP]	Publication date YYYY/MM/DD, YYYY/MM, or YYYY
[PT]	Publication type Review Clinical Trial Lectures Letter Technical Publication
[SH]	Subheading *Used to modify MeSH Terms* hypertension [MH] AND toxicity [SH]
[NM]	Substance name *Name of chemical discussed in article*
[TW]	Text words *All words and numbers in the title and abstract, MeSH terms, subheadings, chemical substance names, personal name as subject, and MEDLINE secondary sources*
[UID]	Unique identifiers (PMID/MEDLINE numbers)
[VI]	Volume of journal

and *boolean operator* = AND, OR, or NOT

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help
Search Clear

Display Settings: GenBank

Amborella trichopoda PISTILLATA-like protein gene, partial cds

GenBank: AY337760.1
FASTA Graphics

Go to:

```
LOCUS       AY337760               1682 bp    DNA     linear   PLN 23-FEB-2005
DEFINITION  Amborella trichopoda PISTILLATA-like protein gene, partial cds.
ACCESSION   AY337760
VERSION     AY337760.1  GI:37992966
KEYWORDS    .
SOURCE      Amborella trichopoda
  ORGANISM  Amborella trichopoda
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; basal Magnoliophyta; Amborellales;
            Amborellaceae; Amborella.
REFERENCE   1  (bases 1 to 1682)
  AUTHORS   Kim,S., Yoo,M.J., Albert,V.A., Farris,J.S., Soltis,P.S. and
            Soltis,D.E.
  TITLE     Phylogeny and diversification of B-functiion MADS-box genes in
            angiosperms: Evolutionary and functional implications of a
            260-million-year-old duplication
  JOURNAL   Am. J. Bot. 91 (12), 2102-2118 (2004)
REFERENCE   2  (bases 1 to 1682)
  AUTHORS   Kim,S., Soltis,D.E. and Soltis,P.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-JUL-2003) Dept. of Botany, University of Florida,
            P.O. Box 118526, Gainesville, FL 32611, USA
FEATURES             Location/Qualifiers
     source          1..1682
                     /organism="Amborella trichopoda"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:13333"
     mRNA            join(<1..112,342..408,485..546,671..770,862..903,
                     1107..1151,1279..>1410)
                     /product="PISTILLATA-like protein"
     CDS             join(<1..112,342..408,485..546,671..770,862..903,
                     1107..1151,1279..1410)
                     /codon_start=3
                     /product="PISTILLATA-like protein"
```

```
                     /codon_start=3
                     /product="PISTILLATA-like protein"
                     /protein_id="AAR06649.1"
                     /db_xref="GI:37992967"
                     /translation="GILKKAKEISVLCDAKVSLVIFSSAGKMSEFCSPSIELKNMLEE
                     YQRTSGKKLWDSRHEYLSTEVDRMKKDNEQMRIELRHLMGEDLNSLTPHELNRIEDSL
                     QMGLSSVRAKQMEHIRTRTEMLKNNERILEDQNKQLKYIMHQIEGGDEAERRYQNQQN
                     GRDYPQQALTAFRVQPIQPNLQQNK"
ORIGIN
        1 gcggaatact gaagaaggcc aaggagattt cggttctatg cgatgccaag gtctccctcg
       61 tcatattctc cagtgctggc aaaatgtccg agttttgcag tccatccatc gagtatctct
      121 ctctcttctt tgggtttcct attgtttttg ggtttagtct ctctctgttt tttgtttcat
      181 ttgtctctgg gttctcatta tttcctgttt agattctttt gtttgttggg ttttcctctc
      241 tctctctctc tctctctctc tctctctctc tctctctctc tctcctctgt tttaagaatt
      301 aagcttttcg ggtgtggatt tcctctgttt cgggatagaa ggttgaagaa tatgctagaa
      361 gagtaccaga ggacttcagg gaaaaagtta tgggattccc gtcatgaggt aactcctctc
      421 tctctctctc tctctctcta cattgctgtt tccggctgtt tgcgagtaat gggtggtggt
      481 gtagtatttg agcacagagg tagataggat gaagaaagac aacgaacaga tgaggattga
      541 gttgaggtaa atttcaagtt ttttttttgt gaccagagtt aaatttcaag tgattgggat
      601 tttaagttgt ctgaaatctt ttgaattact tactgatgca tatgggtttg atgtgtaatt
      661 tgttttgaag gcacttgatg ggagaagatc tcaactcatt gacgccccat gaactcaata
      721 ggattgagga ctccctgcaa atgggcctct ccagtgttcg tgctaaacag gtctgctaaa
      781 gtttttcttg aattggatca gagacattat ctcttatatc caggttcctc aacctttttg
      841 cctctctctc tggctttaca gatggaacac attcgcacca ggactgagat gctaaaaaac
      901 aacgtaatgt tatctctgtc tctctgatgc tcgaaaataa tgtagtttca ccttttctct
      961 cagctgttaa aaaatatcta gtattcacct ctcccttaaa tgttgaacca ataatgtaat
     1021 aaccccaaca tccatctctc tcttttatgg aaataattca tgggcgactg catcagcaat
     1081 ggatgtgtga gattactctt tttcaggaaa ggattctcga agatcagaac aaacaattga
     1141 agtacataat ggtaactcct tttgcatgtt ttagtctttc aaattaagat taagacacca
     1201 agagatgaga ccaacgctta ctctcatcat cattggagcc ttatcccgac tgaccaacac
     1261 ttcgtttcac taaaccagca tcagattgaa ggtggtgatg aagcggagcg caggtatcaa
     1321 aaccaacaga acggaaggga ttatcctcag caagctctca ctgcatttcg tgtgcaaccc
     1381 atccaaccca atcttcagca gaataaatag acacctaaaa agtgatgcac taccttgaac
     1441 aaaatctagt agatctgttt gttagacctg aaatagtttg gtttggggag aagactcttt
     1501 atatgcatgc accttggact tagttttaat ttgttgacaa ggaacaagct tgtttatgtt
     1561 ttcgggagct tttgggcaat agaaagaggc atctctttta ttagcatcag ttgaatgcta
     1621 gatcatggat agagttctca tatatgtaag acatggtttg ctgaagcaag attattatcc
     1681 ct
//
```

Arabidopsis thaliana PI (PISTILLATA); DNA binding / transcription factor (PI) mRNA, complete cds

NCBI Reference Sequence: NM_122031.3

FASTA Graphics

Go to:

```
LOCUS       NM_122031               1095 bp    mRNA    linear   PLN 21-AUG-2009
DEFINITION  Arabidopsis thaliana PI (PISTILLATA); DNA binding / transcription
            factor (PI) mRNA, complete cds.
ACCESSION   NM_122031
VERSION     NM_122031.3  GI:145358258
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; malvids; Brassicales; Brassicaceae; Camelineae;
            Arabidopsis.
COMMENT     REVIEWED REFSEQ: This record has been curated by TAIR. The
            reference sequence was derived from AT5G20240.1.
            On Apr 18, 2007 this sequence version replaced gi:30687709.
FEATURES             Location/Qualifiers
     source          1..1095
                     /organism="Arabidopsis thaliana"
                     /mol_type="mRNA"
                     /db_xref="taxon:3702"
                     /chromosome="5"
                     /ecotype="Columbia"
     gene            1..1095
                     /gene="PI"
                     /locus_tag="AT5G20240"
                     /gene_synonym="F5O24.130; F5O24_130; FLORAL HOMEOTIC
                     PROTEIN PISTILLATA; PI; PISTILLATA"
                     /function="Floral homeotic gene encoding a MADS domain
                     transcription factor. Required for the specification of
                     petal and stamen identities."
                     /db_xref="GeneID:832146"
                     /db_xref="TAIR:AT5G20240"
     CDS             162..788
                     /gene="PI"
                     /locus_tag="AT5G20240"
                     /gene_synonym="F5O24.130; F5O24_130; FLORAL HOMEOTIC
                     PROTEIN PISTILLATA; PI; PISTILLATA"
                     /note="PISTILLATA (PI); FUNCTIONS IN: transcription factor
                     activity, DNA binding; INVOLVED IN: regulation of
                     transcription, DNA-dependent; LOCATED IN: nucleus,
                     cytoplasm; EXPRESSED IN: 20 plant structures; EXPRESSED
                     DURING: 8 growth stages; CONTAINS InterPro DOMAIN/s:
```

IF YOU ARE PREPARED FOR SUBMITTING A SEQUENCE (SEQIN FILE OUTPUT), SEND A MESSAGE TO GENBANK MAMAGER: gb-sub@ncbi.nlm.nih.gov

Dear GenBank manager:
I would like to submit seven sequences to the GenBank.
Please see attached file: Fossil2.seq

Thank you very much,

Sangtae Kim

****

Sangtae Kim Ph.D.
Postdoctoral Research Associate
Dept. of Botany
University of Florida
Gainesville, FL
TEL: 352-392-7924

****

Dear GenBank Submitter:

Thank you for your direct submission of sequence data to GenBank. We have provided GenBank accession numbers for your nucleotide sequences:

Pam	AY337727
Pbo	AY337728
Aba	AY337729
Ppa	AY337730
Lno	AY337731
Lbe	AY337732
Mri	AY337733
Lcu	AY337734
Pps	AY337735
Nu.ad.PI	AY337736
Nu.va.PI	AY337737
Pe.am.PI	AY337738
As.lo.PI	AY337739
Eu.be.PI	AY337740
Eu.la.PI	AY337741
Ri.sa.PI	AY337742
Am.tr.AP3-1	AY337743
Am.tr.AP3-2	AY337744
Nu.va.AP3-1	AY337745
Nu.va.AP3-2	AY337746
Il.pa.AP3	AY337747
Pe.am.AP3	AY337748
As.lo.AP3	AY337749
Eu.be.AP3-1	AY337750
Eu.be.AP3-2	AY337751
Ma.gr.AP3	AY337752
Gu.ti.AP3-1	AY337753
Gu.ti.AP3-2	AY337754
Gu.ti.AP3-3	AY337755
Gu.ti.AP3-4	AY337756
Gu.ti.AP3-5	AY337757
Ri.sa.AP3-1	AY337758
Ri.sa.AP3-2	AY337759
Am.tr.PI	AY337760

We strongly recommend that these GenBank accession numbers appear in any publication that reports or discusses these data, as they give the community unique labels with which they may retrieve your data from our on-line servers.

We are now processing your submissions and will mail you copies for your review prior to their release to the public database.

You have requested that your data are to be held confidential until:

Jul 31, 2004

They will not be released to the public database until this date, or until the data or accession numbers appear in print, whichever is first.

Please send any revisions, including bibliographic information (e.g., conversion from unpublished to published), biological data (e.g., new features), or sequence data as text in the body of an email to:

gb-admin@ncbi.nlm.nih.gov

Since the flatfile record is a display format only and is not an editable format of the data, do not make changes directly to a flatfile. In addition, please do not use bold, highlighted, or colored edits and do not use attached files or spreadsheets for additional or revised data.

You may be interested to know that there are two GenBank submission tools available, BankIt and Sequin. BankIt is available through the World Wide Web (WWW), and Sequin, a stand-alone program, can be downloaded from NCBI's anonymous ftp site. You can access these submission tools through the NCBI Home Page (http://www.ncbi.nlm.nih.gov/).

For more information about the submission process or the available submission tools, please contact GenBank User Support at info@ncbi.nlm.nih.gov or at (301) 496-2475.

Respectfully,

Lawrence Chlumsky

The GenBank Direct Submission Staff
Bethesda, Maryland USA
****

(301) 496-2475
(301) 480-2918 fax
gb-admin@ncbi.nlm.nih.gov (for updates/replies to GenBank entries)
info@ncbi.nlm.nih.gov (for general questions regarding GenBank)
****

Dear GenBank Submitter:

Thank you for your submission.

Based on the data submitted to us, the scheduled release date for your submission is:

Dec 31, 2005

However, if the accession number is published prior to that date, the sequence will be released upon publication.

Note that the entire sequence will be released when the article citing this accession number is published or on the above release date, whichever comes first. If this is not what you intended, please notify us immediately so that we can discuss your submission in more detail.

If this date is not correct, please get in touch with us as soon as possible (gb-admin@ncbi.nlm.nih.gov, telephone at (301) 496-2475, or fax at (301) 480-2918), otherwise this submission will be released on the date indicated above. The data would then be available over the network data servers which provide daily updates of GenBank data. The data are simultaneously made available to EMBL in Europe and the DNA Data Bank of Japan.

Minor changes may have been made in your original submission in order to conform to database annotation conventions. You can greatly assist us in presenting your data in as accurate a manner as possible by paying specific attention to the following in your review:

- Spelling (particularly author names)
- Citation data (author order, page span, etc.)
- Nomenclature ('official' gene names, product labels, etc.)
- Taxonomic and source data
- Feature spans and descriptions (particularly non-coding regions)

Please send any revisions, including bibliographic information (e.g., conversion from unpublished to published), biological data (e.g., new features), or sequence data as text in the body of an email to:

gb-admin@ncbi.nlm.nih.gov

Since the flatfile record is a display format only and is not an editable format of the data, do not make changes directly to a flatfile. For complete information about different methods to update a sequence record, see: http://www.ncbi.nlm.nih.gov/Genbank/update.html

An accession number has been assigned to each nucleotide sequence and was provided to you at the time receipt of your submission was acknowledged. Note that during the processing of your records, we have assigned protein identifiers to any proteins that are in your records. This is fielded as /protein_id.

We strongly recommend that these numbers appear in any publication which reports or discusses these data, as they give the community convenient labels with which they may retrieve your data from our on-line servers.

Thank you once again for your submission.

Sincerely,

Lori Black, PhD
GenBank Direct Submission Staff
gb-admin@ncbi.nlm.nih.gov

GenBank flat file:

```
LOCUS       DQ070749                741 bp    DNA     linear   PLN 22-JUL-2005
DEFINITION  Nuphar advena AP3-like protein gene, partial cds.
ACCESSION   DQ070749
VERSION     DQ070749
KEYWORDS    .
SOURCE      Nuphar advena
  ORGANISM  Nuphar advena
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; basal Magnoliophyta; Nymphaeales;
            Nymphaeaceae; Nuphar.
REFERENCE   1  (bases 1 to 741)
  AUTHORS   Kim,S., Koh,J., Yoo,M.-J., Kong,H., Hu,Y., Ma,H., Soltis,P.S. and
            Soltis,D.E.
  TITLE     Expression of floral MADS-box genes in basal angiosperms:
            Implications for the evolution of floral regulators
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 741)
  AUTHORS   Kim,S., Koh,J., Yoo,M.-J., Kong,H., Hu,Y., Ma,H., Soltis,P.S. and
            Soltis,D.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-MAY-2005) Dept. of Botany, Univ. of Florida, P.O. Box
            118526, Gainesville, FL 32611-8526, USA
FEATURES             Location/Qualifiers
     source          1..741
                     /organism="Nuphar advena"
                     /mol_type="genomic DNA"
                     /specimen_voucher="S.Kim 1140, FLAS"
                     /db_xref="taxon:77108"
```

```
LOCUS                       741 bp    DNA     linear   PLN 22-JUL-2005
DEFINITION  Nuphar advena AP3-like protein gene, partial cds.
ACCESSION   DQ070749
VERSION     DQ070749
KEYWORDS    .
SOURCE      Nuphar advena
  ORGANISM  Nuphar advena
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; basal Magnoliophyta; Nymphaeales;
            Nymphaeaceae; Nuphar.
REFERENCE   1  (bases 1 to 741)
  AUTHORS   Kim,S., Koh,J., Yoo,M.-J., Kong,H., Hu,Y., Ma,H., Soltis,P.S. and
            Soltis,D.E.
  TITLE     Expression of floral MADS-box genes in basal angiosperms:
            Implications for the evolution of floral regulators
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 741)
  AUTHORS   Kim,S., Koh,J., Yoo,M.-J., Kong,H., Hu,Y., Ma,H., Soltis,P.S. and
            Soltis,D.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-MAY-2005) Dept. of Botany, Univ. of Florida, P.O. Box
            118526, Gainesville, FL 32611-8526, USA
FEATURES             Location/Qualifiers
     source          1..741
                     /organism="Nuphar advena"
                     /mol_type="genomic DNA"
                     /specimen_voucher="S.Kim 1140, FLAS"
                     /db_xref="taxon:77108"
     mRNA            join(<1..12,128..227,393..434,501..539,627..>741)
                     /product="AP3-like protein"
     CDS             join(<1..12,128..227,393..434,501..539,627..>741)
                     /note="Nu.ad.AP3.2"
                     /codon_start=2
                     /product="AP3-like protein"
                     /protein_id="AAZ22442"
                     /translation="RSIRQRNGEDLDMLNHSELCGLEQNLSEALKKIRSVLDNKIKRQ
                     IDTYRKKIKAADSIRNIGFMELQELNCSFDGSEENYESMLVMRNGNAQPFPISVQPNH
                     "
ORIGIN
        1 caggagcatc aggttgtgta cctctctaga attttttat gtgctttttg ctctttgttt
       61 ttctaatggg ctgacgaagg atccttttt attatgactg atattttaca agatgggtgt
      121 tgtttaggca aaggaatggc gaggatttag atatgttaaa ccattctgag ctgtgcggtc
      181 ttgagcaaaa tctgagcgaa gcgcttaaga aaatccgatc agtattggta tgtttttcaa
      241 tgtgtatgtc caagttgttt tgaatcatat atttcggttt taggttttaa aatcagggta
      301 gaatttatat gttttatgaa catgcatgta tcaattaagt gataattttc aatcttggat
      361 attcatctta cctgattttc catgtgcatc aggataacaa aatcaagaga cagatagata
      421 cttataggaa aaaggtaaat gggttatcaa aaaacaacat ttattttgta gccatggatt
      481 aacggtgctt cttgatgcag ataaaggcag ccgattccat tagaaacata ggtttcatgg
      541 tatattactt gaaccataat tttagtggtt tatgttattt gcattcttct ctacgccact
      601 taccctcttt tgtttttgcc aaacaggagt tacaagaact caactgtagt tttgatggaa
      661 gtgaagaaaa ctatgaatcc atgctggtga tgaggaatgg caatgcgcaa ccgttcccaa
      721 tcagtgtgca acccaatcac c
//
```