01

Chapter

정보학의 측면에서 바라본 바이오인포매틱스

학습목표 바이오인포매틱스의 의미를 이해하고, 바이오인포매틱스 관련 주요 데이터베이스에 대해서 알아본다.

Bioinformatics

- 바이오인포매틱스의 정의
- 인간게놈프로젝트 이해하기
- 바이오디지털 데이터베이스
- 바이오디지털 데이터의 분석방법

2001년 인간 유전체 초안이 완성되었으며, 이 기념비적인 업적으로 30억 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 역사적 인사건과 인간게놈프로젝트를 비유하며 가벼운 마음으로 이장을 시작해 보도록 하겠다.

1.1 로제타스톤과 인간게놈지도

1799년 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 0.7m인 검은 돌로, 석비를 본 순간 학자들은 돌에 새겨진 문자가 신의 메시지라고 생각하였으며, 4,000년이나 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 23년 만에 밝혀지게 되었던 것이다.

[그림 1.1] 로제타스톤(Rosetta Stone)
(사진출처 : http://www.ancientegypt.co.uk/writing/rosetta.html)

2000년 6월26일 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 30억 개의 정보를 저장하고 있고 이 데이터를 컴퓨터로 옮겨저장하거나, 그 의미를분석하는 일이 가능해졌다.

[그림 1.2] 게놈: 인간의 손에 넘어온 신의 언어

1.2 바이오인포매틱스란?

미지의 암호를 풀어낸다는 개념에서, 인간게놈지도 초안은 21세기의 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 뿐이라는점을 강조하려 한 것이다.

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 수있는 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 부각될것이 자명하다.

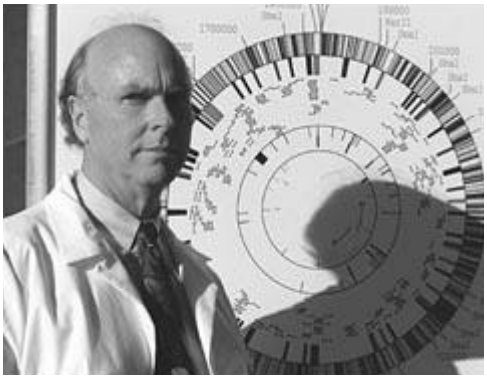
한편, 바이오인포매틱스는 크게는 게놈프로젝트에 관련된 **계산분자생물학**(Computational Molecular Biology)을 말하기도 하지만 최근 이 분야의 연구규모가 방대해짐에 따라 점차 계산분자생물학과는 분리되어 독립된 학문으로 떨어져 나오고 있으며, 대량의 서열 데이터를 분석하는 게놈정보 학문으로 구체화되는 경향도 있는 듯하다. 어쨌든, 2000년 발표된 인간게놈프로젝트의 완성을 기점으로 바이오인포매틱스의 중요성이 더욱 강조된 것만은 분명하다. 다음 절에서는 현재의 게놈 혁명이 일어나기까지의 역사적인 사건에 대해서 정리해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

1.3 인간게놈프로젝트의 완성까지..

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 에너지성이 제안했을 때는 그 중요성에 대해많은 논쟁이 있었다.

이 프로젝트는 인간이라는 생물학적 정보를 디지털화하는 것으로 비유할 수 있었다. 컴퓨터 프로그램이0과 1로 이루어진 것처럼 인간의 게놈은A, C, G, T라는 기호로 표기되는 네 종류의 염기가 특정한 순서로 배열되어 이루어진 신이 만든 프로그램 코드이며 이러한 서열을 완전 구명하는 것을목표로 한 방대한 계획이었던 것이다. 하지만15년에 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 경쟁으로변질되었다.

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 Project) 진영과 셀레라지노믹스사는 서로 악수하며, 유전자 염기서열 초안을 완료했다고 공동발표하게 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 수있을뿐이다. 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 찾기까지수많은시행착오를 겪을 것이다.

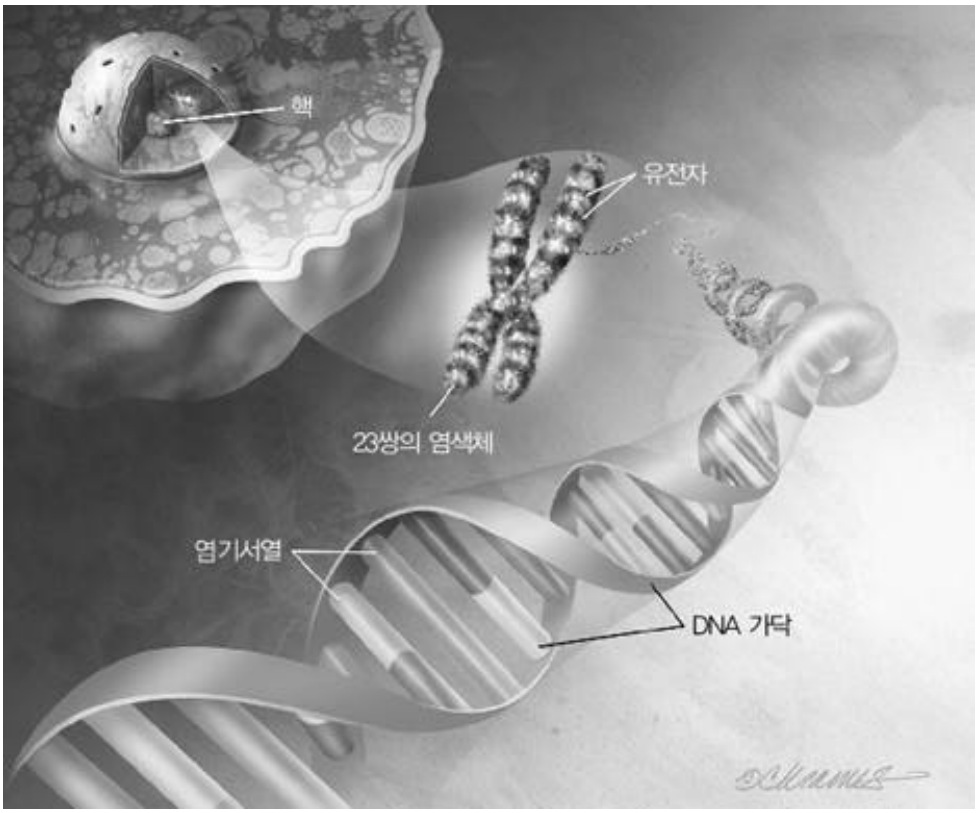
1.4 바이오디지털정보들

앞절에서는 게놈의 역사에 대해서 대강 살펴보았는데, 이번에는 대표적인 바이오디지털 정보들은 어떤 것이 있을지 설명해 보겠다. 이를 위해서는 먼저 게놈이란 무엇인지에 대한 올바른 이해가다시 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 있

듯이 핵 속에 1번부터 22번, 그리고 성을 결정하는 염색체 X, Y까지 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)

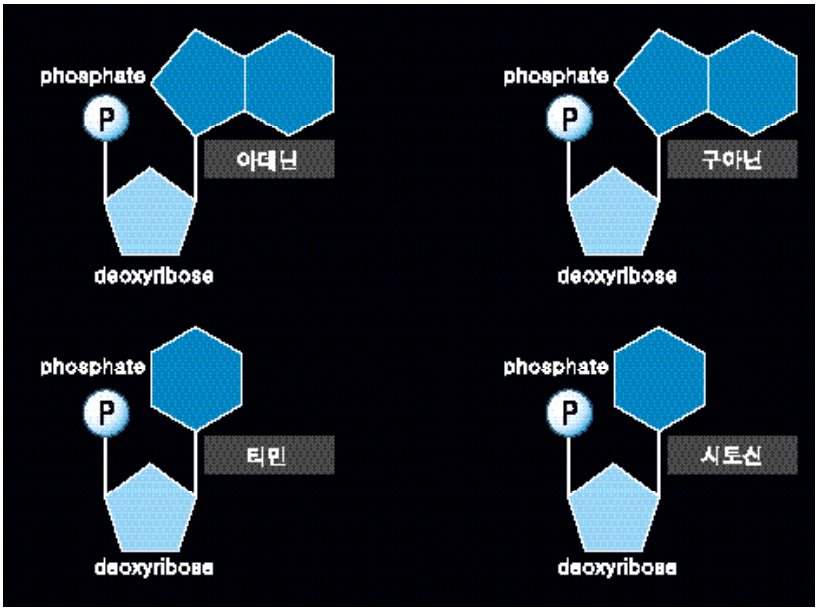
1) 세포는 크게 진핵세포와 원핵세포로 나눌 수 있는데, 일반적으로 인간을 포함한 대부분의 다세포 생물은 진핵세포로 이루어져 있고, 미생물 등의 단세포 생물은 원핵세포로 이루어져 있다. 진핵세포와 원핵세포의 대표적인 차이점은 핵의 유무이다. 핵이 있는 진핵세포는 핵막으로 둘러싸인 핵 안에 유전물질을 갖는 반면에, 핵이 없는 원핵세포는 핵이 따로 존재하지 않고 세포질에 유전물질이 그대로 존재한다.

궁극적으로 이 책에서 다루고자 하는 것도 정보학의 측면에서 데이터마이닝의 대상인 게놈 정보, 예를 들면, 그림 1.6과 같은 형태로 존재하는 무수한 서열(sequence) 데이터에 포커스를 두고자 한다. 대표적인 바이오디지털 데이터로는A, T, G, C 네 개의 기호로 이루어진 DNA **염기서열**(base sequence)과 A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y 등 20 종류의 아미노산 **잔기서열**(residue sequence)로 이루어진 단백질 서열 데이터가 있을 것이다. 여러분들이 이미 익히 알고 있을 상식적인 내용이기는 하지만, DNA와 아미노산 서열에 대해서 간단히 정리하고 넘어가기로 하자.

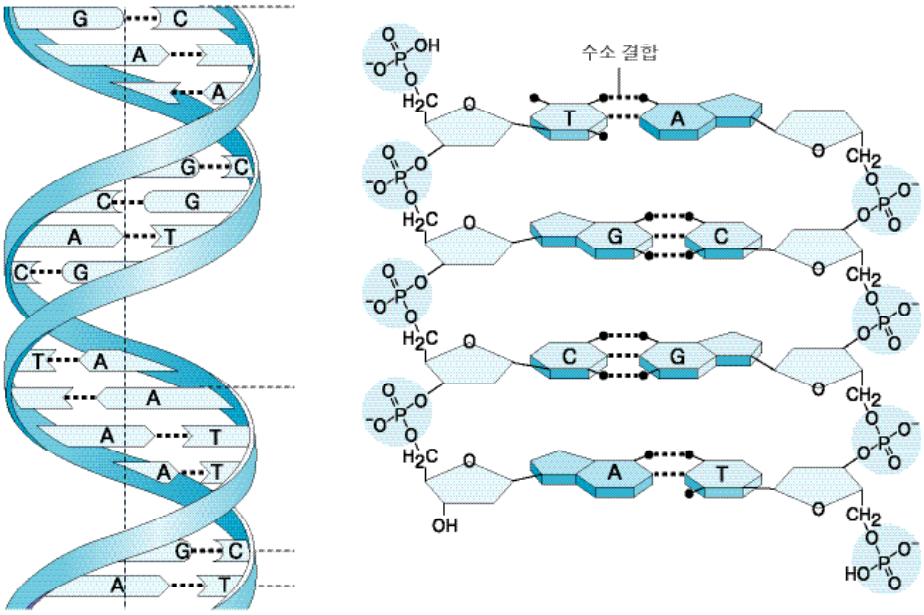
[그림 1.6] 대표적인 바이오디지털 데이터

DNA : 생명현상에 관한 정보를 저장하는 최소 단위는 바로 **뉴클리오타이드**(nucleotide)이다. 뉴클리오타이드는 보통 세포 안에서 DNA(deoxyribonucleic acid) 형태로 존재하며 DNA에는 그림 1.7에서 볼 수 있듯이 아데닌(adenine), 시토신(cytosine), 구아닌(guanine), 티민(thymine) 등 4종류가 있다.

이들은 세포내에서 당-인산 골격에 의해 길게 연결되면서 상보적인 DNA와는 그림 1.8에서와 같이 수소 결합을 이룬다. 이렇게 함으로써DNA 서열은 긴 이중나선구조를 가지면서 한 개체가 가지는 모든 유전정보를 포함한다. 하나의 개체를 이루는 모든 세포는 포함된 조직에 관계없이모두 똑같은 게놈정보를 가진다.



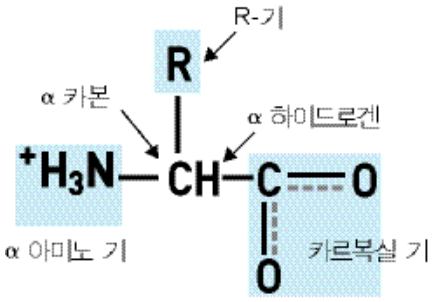
[그림 1.7] DNA 분자의 구조



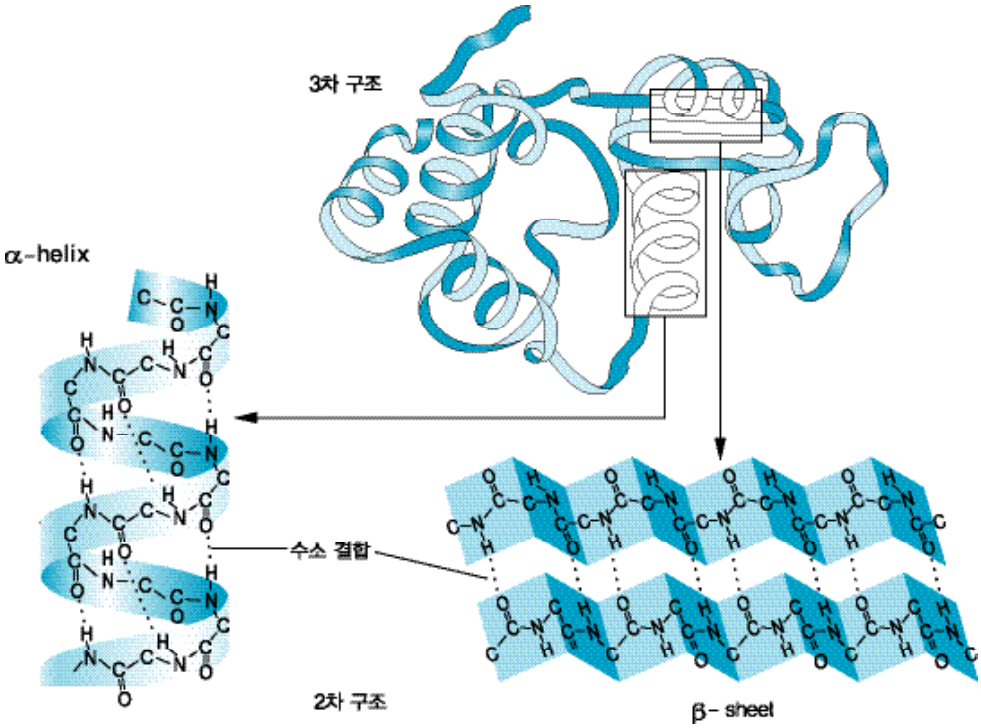
(a) DNA 이중나선구조 (b) 수소결합

[그림 1.8] DNA 이중나선 (사진출처 : http://www.anselm.edu)

단백질 : 생명현상에서 단백질은 매우 중요한 역할을 한다. 이토록 중요한 단백질은 아미노산으로 구성되는데, 아미노산은 그림 1.9와 같이 가운데의 탄소 분자에 수소, 아미노 기(amino group), 카르복실 기(carboxyl group), 그리고 여러 가지 형태의 곁가지(side chain)가 결합하는 구조를 가진다.



[그림 1.9] 아미노산의 기본적인 분자구조

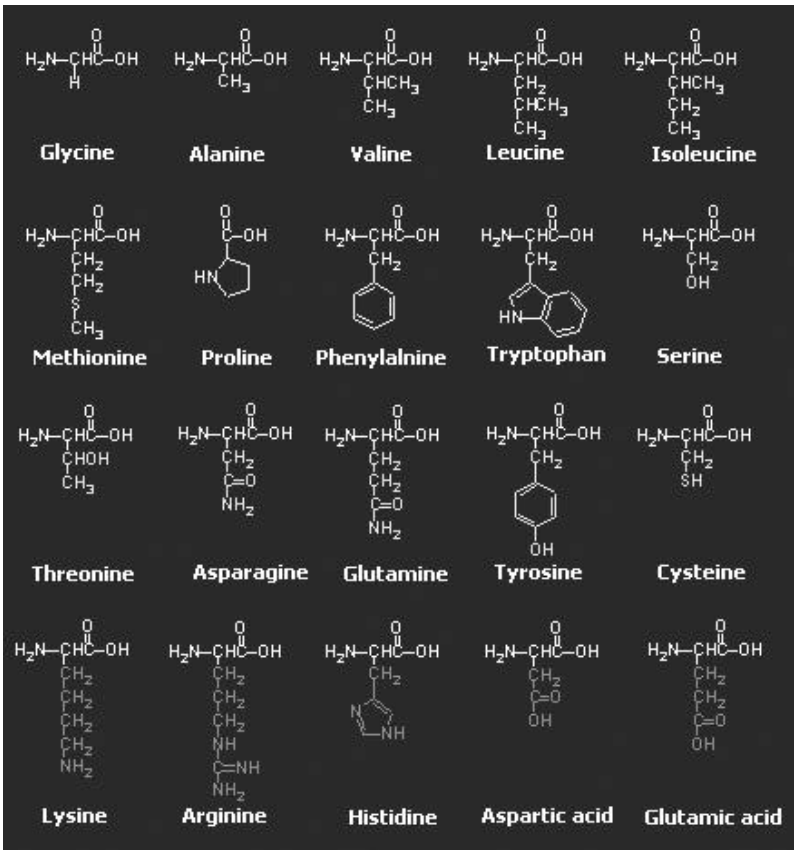


[그림 1.10] 단백질의 2차 구조(α helix와 β sheet)와 3차 구조

아미노산 분자는 다른 아미노산 분자와 펩티드 결합을 형성하면서 중합체(polymer)를 이루는데, 이것을 특별히 폴리펩티드(polypeptide)라고 부른다. 아미노산이 펩티드 결합을

형성할 때 두 개의 수소 분자와 한 개의 산소 분자가 떨어져 나가므로 폴리펩티드를 이루는 각 아미노산들을 **잔기**(residue)라고 부르는 것이다.

또한, 아미노산 잔기가 길게 연결되어 있는 것을 단백질의 1차 구조(primary structure)라고 한다. 단백질의 1차 구조는 다시 αhelix, βstrand(혹은 βsheet) 등의 2차 구조를 형성한다.

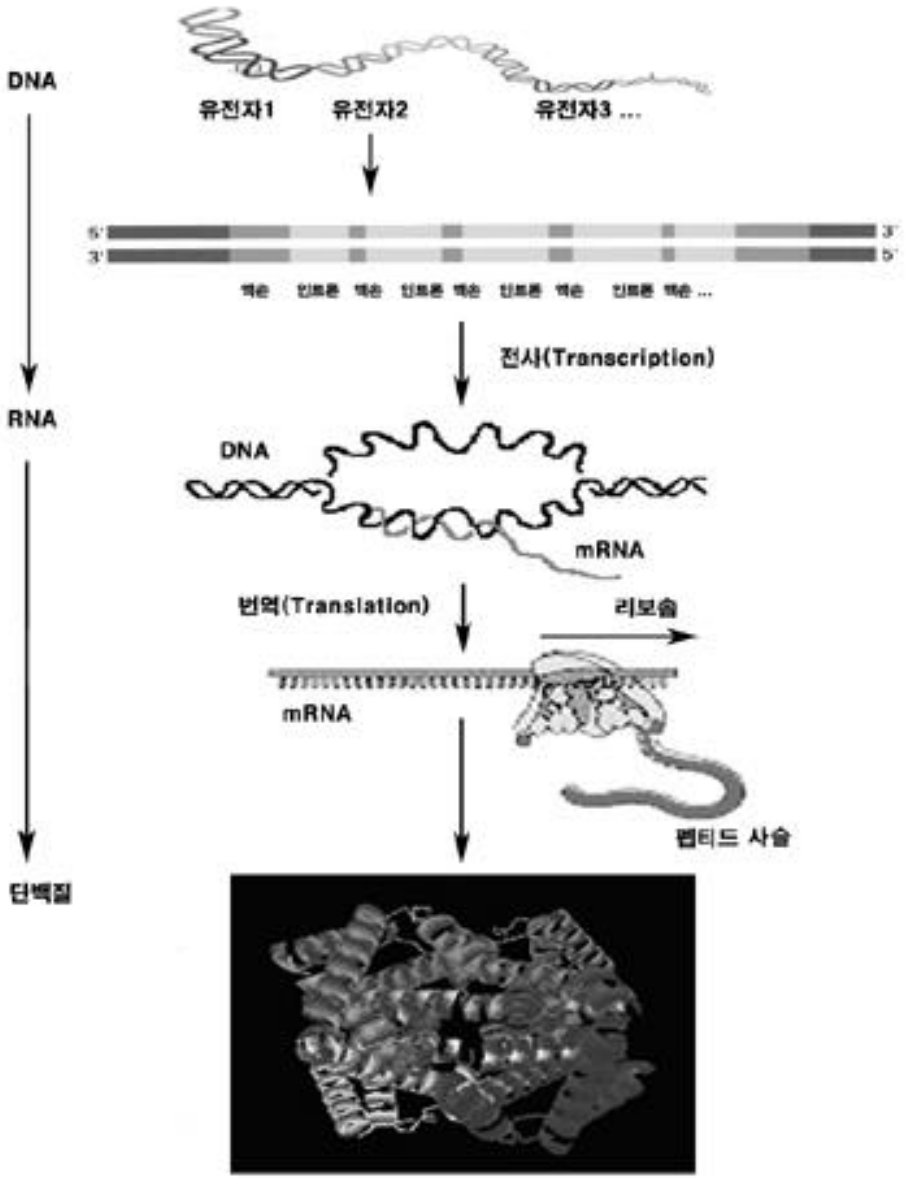


[그림 1.11] 20가지 아미노산의 분자구조

폴리펩티드의 2차 구조는 특정 부위에서 접힘(folding) 현상이 일어나면서 더욱 복잡한 3차원적인 구조를 이루는데, 이를 3차 구조라고 한다. 또한 여러 개의 3차 구조가 결합하여 더 큰 형태의 단백질을 이루기도 하는데, 이러한 형태를 4차 구조라고 한다. 이처럼 단백질은 각각 고유한 3차원 구조를 형성함으로써 생명현상에 관여한다. 따라서, 단백질 구조를 파악하는 것은 생명현상을 연구하는 것에서 대단히 중요한 일이다. 단백질의 구조는 X-ray 결정학(chrystallography) 혹은 핵자기공명법(NMR: Nuclear Magnetic Resonance) 등으로 규명하고 있다. 단백질을 이루는 아미노산은 곁가지(side chain 혹은 R-group)의 종

류에 의해 그림 1.11과 같이, 20 가지로 나누어지는데, 곁가지의 종류는 아미노산 분자를 코딩하는 세개의 연속적인 뉴클리오타이드 분자(코돈, codon)에 의해결정된다.

중심이론 : 이러한 대량의 바이오디지털 정보간의 상호관계는, 분자생물학에서 말하는 중심이론(central dogma)이라고 하는가설하에서대략적으로 정리될 수 있다.



[그림 1.12] DNA로 이루어진 유전자 서열에서 전사과정을 통해 mRNA가 생성되고 다시 번역과정을 통해 아미노산 서열을 만든다. 아미노산 서열은 다시 α-helix와 β-sheet 등의 2차 구조를 이루면서 고유한 3차원 구조를 형성한다

중심이론(central dogma) 또는 중심명제란 아주 간단히 말하면, 유전정보가 물 흐르듯이 DNA 에서 RNA 로(또는RNA 에서 DNA 로), 다시 단백질로 최종 전달된다는 이론이다. 그리고, DNA 에 담긴 정보가 **전사**(transcription)와 **번역**(translation)이라는 과정을 거쳐 단백질을 만들고, 이 단백질이 생명현상을 주도하는 역할을 한다는 것이다.[2] 사실, 전사와 번역은 언어학에서 쓰는 말로 이해하면쉽다.

게놈에 존재하는 유전자의 DNA 서열은 그림 1.12 에서와 같이 먼저 전사(transcription)라는 과정을 거치면서 RNA 서열로 바뀐다. 물론 DNA 가닥 자체가 변하는 것은 아니고, RNA 로 이루어진 DNA 원본 서열의 복사본이 생성되는 것이다. 그런데 DNA 와 RNA 의 변환은, 같은 종류의 문자로 바꾸는 과정이므로transcribe(" 베끼다," " 전사하다")라는 말을 쓴다.

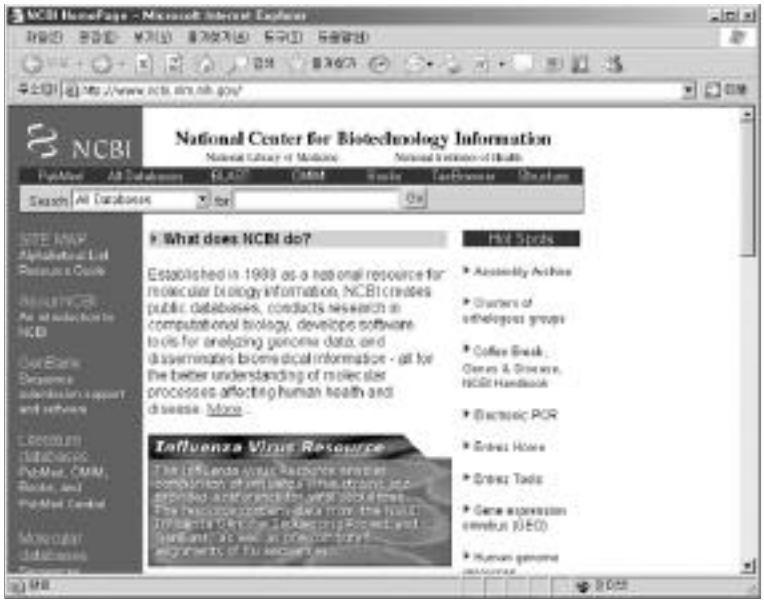
하지만, 유전자 서열의 RNA 복사본을 아미노산 서열로 변형하는 과정에서는 번역(translation)이라는 용어를 쓴다. 이것도 역시 세포내의 리보솜이라는 곳에서 RNA 서열이 코딩하는 아미노산 분자를 순서대로 가져다 붙이는 것일 뿐 RNA 서열 자체가 단백질 서열로 바뀌는 것은 아니다. 어쨌든, RNA 로부터 단백질을 만드는 과정은 전사와는 달리, 뉴클리오타이드(nucleotide)를 아미노산(amino acid)이라는 다른 언어로 바꾸는 과정이므로 translate(" 해독하다," " 번역하다")라는 말을 쓰는것이다.

한편, 게놈에 있는 각각의 DNA 정보는 의미가 다르므로 주석(annotation)을 달아놓아야 한다. 예를 들면, 전체 게놈 중 유전자인 부분을 따로 표시해 주어야 할 것이고, 이 유전자에서도 의미가 서로 다른 부분들이 복합적으로 섞여 있어서 또다시 구분해 주어야 한다. 즉, 유전자는 서로 크기가 다른 엑손(exon)과 인트론(intron)이란 조각들이 교대로 반복해서 연결되어 만들어진다. 이 중 엑손에는 단백질을 만드는 정보가 들어 있는 것이며 유전자의 DNA 가 RNA 로 바뀔 때 인트론은 떨어져나가 버리고 엑손 정보들만 고스란히 RNA 로 옮겨오는 것이다(스플라이싱 splicing). 그런데 이 엑손이 정보를 어떻게 저장하고 있는가 알아보니 연이어진 염기 세 개가 한 개의 아미노산을 암호화하고 있던 것이고 이것을 코돈(codon)이라 부른다. 이 염기 세 개의 조합에 의해 어떤 아미노산이 만들어질지 결정되는 것이고 아미노산이 길다랗게 사슬처럼 연결되면 단백질이 되는 것이다. 한편 DNA 의 네 가지 염기로 많은 수의 조합을 만들 수 있는데 ACT, ATT, CTG 등 총 가능한 조합의 수는 4×4×4=64 개가 된다(자세한 과정은 5 장 등의 예제를 통해서 살펴볼 것이다). 이 외에도 게놈정보는 프로모터(promoter)와 같이 전사를조절하는특정 영역으로 구분될 수있다.

2) 유전자 서열로부터 단백질이 만들어지는 것을 유전자 발현(gene expression)이라고도 한다.

1.5 대표적인게놈포털사이트

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 대부분은분석결과와 함께 전 세계의 연구자들이공유할수 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

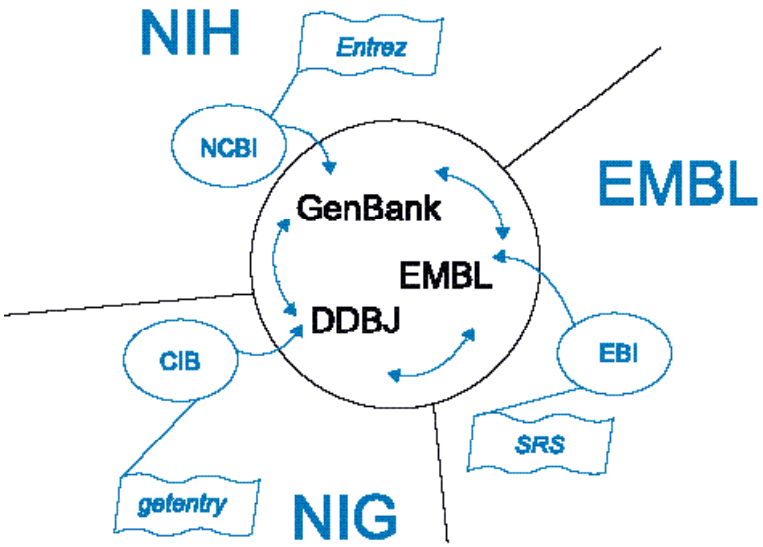
예를 들면 앞절에서 언급하였던, 인간게놈프로젝트의 결과물, 즉 30억 개의 DNA 서열이나 분석자료 및 연구내용들도 바로 이 사이트를 통해서 여러분들이 직접 다운로드받을 수도 있을 것이다. NCBI의 http://www.ncbi.nlm.nih.gov/genome/guide/human/를 직접 방문하여인간게놈데이터를 확인해 보기 바란다.

[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 다양한알고리즘들이세포 안에 있는 바이오디지털 데이터에도 그대로적용될 수 있다는 점은 놀라운 사실이다. 이들 알고리즘에 의한 서열분석방법에 대해서는1.6 절에서 간단히 소개하기로하자.

한편, NCBI 사이트에 등록되는 서열 데이터들은 사이트내의GenBank 라는 서열 데이터베이스에 저장된다. GenBank 에 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 같은정보도 제공하고, 이외에도 연구자들이 직접 간단한 분석작업을할 수 있도록 몇 가지 소프트웨어도 서비스하고 있다. GenBank 데이터 형식에 관해서는7 장에서 자세히 학습하게될 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 국제협약에의해 서로 동일한 데이터를 유지하고관리한다.



[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, NCBI 에서는 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 H19 이라는 유전자에 관련된 검색결과를 찾으려고 하고, 검색결과를 kono 라는 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 바이오디지털데이터분석

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

바이오인포매틱스 전문가들이 해야하는 일은 크게 두 가지로 나누어볼 수 있을 것이다. 하나는 데이터 관리에 관한 것이고, 다른 하나는 데이터 분석일 것이다. 먼저, 속도를 다투는 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 늘좀더 섬세한 질문들에 대해서대답할수 있는 애플리케이션을 만들기를원하고있다.

서열 분석이라고 하는 것은 DNA 에서 유전자를 찾고 이것이 RNA 와 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 레벨로분류해 설명하면 다음과 같다.

1.5 절에서도 설명했듯이 게놈 서열은 유전자 부분과 비유전자 부분으로 나뉘고, 유전자 부분은 다시 전사되는 엑손과 전사되지 않는 인트론 부분으로 나뉜다. 그러므로DNA 수준의 문제에서는 유전자 부분, 비유전자 부분, 엑손, 인트론, 프로모터, 터미네이터 등의 위치를 컴퓨터로 찾는것 등에 응용된다. RNA 수준에서의 문제는염기쌍간의상호작용 염기쌍과 그들의 에너지간 상호작용, 염기쌍과 에너지간 상호작용에서의 자유 에너지에 기초해서 mRNA, tRNA, rRNA, snRNA 의 서열이나2 차 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. 단백질의3 차 구조 예측은 포스트게놈 시대 바이오인포매틱스가 해결해야 할 난제이다. 단백질의 구조는 그림 1.10 에서와 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 1 차 서열정보를 통해서 3 차원 구조를 예측하기를 원할 때, 이를 단백질 접힘(protein folding) 문제라 한다. 단백질에 대한 완벽한 구조예측방법은 현재로서는 알려져 있지 않으나 특별한 경우에 적용할 수 있는 2 차원 및 3 차원 알고리즘들이 존재하고 있다. 실험을 통하지 않고 단백질 3 차원 구조를 예측하는 방법으로는 상동성 모델링(homology modeling), 스레딩(threading), 그리고ab. initio 방법 등이 있다. 게놈프로젝트에서 나온 많은 1 차 서열 데이터들은 새로운 방식의 3 차 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 데이터마이닝문제에속한다.

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 있는DNA 칩 기술의 발전은 포스트게놈시대의연구를 엄청난 속도로 앞당길것으로 전망되고 있다.

결과적으로, 인간게놈프로젝트를 계기로 의/생물학의 패러다임이 질 중심의 개인적 연구에

서 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 수있었으리라 생각한다.

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

인터넷리소스

바이오인포매틱스학회

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

온라인튜토리얼

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

바이오데이터베이스

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/

PIR: http://pir.georgetown.edu/
PDB: http://www.rcsb.org/pdb/
wEBIn: http://www.ebi.ac.uk/embl/Submission/webin.html
BankIt: http://www.ncbi.nlm.ni.gov/BankIt/
Entrez: http://www.ncbi.nlm.nih/Entrez
NCBI Human Maps Help:
http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html

연습문제

1. 다음 문장들의 빈칸을 채워라.

(a) _____는 서열의 유사성을 검색하는 데 가장 많이 쓰이는 도구로서 NCBI의 전체 데이터베이스를 대상으로 하여 검색을수행한다.

(b) NCBI에서 생물학 관련 논문들의 요약문을 모아놓은 데이터베이스의 이름은 _____이다.

(c) GenBank의 대표적인 검색엔진은_____이고, EMBL의 대표적인 검색엔진은 _____ 이며, DDBJ의 대표적인 검색엔진은_____이다(그림 1.15를 참조할 것).

(d) NCBI는 유럽의_____, 일본의_____의 두 파트너와 함께 국제협약에 의해 서로 동일한 데이터를유지하고관리한다.

2. 인터넷을 검색하여(예: http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html/), 바이오인포매틱스(bio-informatics)와 계산생물학(computational biology)을 각각 어떻게 정의하고 있는지 알아보자.

***3. 이 책에서는 GenBank(2장), PDB(8장) 등 몇 개의 바이오 데이터베이스에 대해서만 언급할 것이다. 이 외에도, NCBI(http://www.ncbi.nlm.nih.gov/)의 Entrez에서 특정 서열이나 유전자들을 검색하면 그림 1.16과 같이 여러 가지 통합 정보가 검색된다. p53이라는 유전자를 키워드로 검색하여 나타나는 여러 가지 정보들의 연관성을 조사해 보기 바란다. 생명과학 전공자가 아닌 경우, 각 데이터베이스의 특성을 파악하려면 상당한 시간이 걸릴 것이며, 주위의 도움이 필요할지도 모르겠다.**