1단원. 게놈에 대한 최초의 보고

서열 통계

1.1 게놈의 시대, 0년

1995년에 메릴랜드에 위치한 게놈연구센터(TIGR)의 크레이그 벤터(Craig Venter)에 의한 한 과학자 무리가 Science 저널에 눈에 띠는 한 논문을 출판하였다. 이 논문은 독립 유기체인 *Haemophilus influenzae* (또는 짧게 *H. influenzae*) 박테리아의 완전한 DNA 서열(게놈)을 보고하였다. 이 시점까지는 오직 작은 바이러스 게놈이나 다른 게놈의 작은 부분만이 서열화되었다. 첫번째 바이러스 게놈 서열(파지 *phiX174*)은 프레드 생거(Fred Sanger)의 그룹에 의해 1978년에, 몇 년 후에는 같은 그룹에 의해 인간의 미토콘드리아 DNA의 서열이 밝혀졌다. 영국 캠브리지에서 연구하는 생거는 1958년에 단백질 서열화 기술을 개발한데 대해 첫번째, 1980년에 DNA 서열화 기술을 개발한데 대해 2번째로 2개의 노벨상을 수상했다. 그러나 박테리아 서열은 바이러스보다 훨씬 거대하므로 *H. influenzae* 논문을 진정한 획기적 사건으로 만든다. TIGR의 그룹에서 서열화한 게놈 크기의 증가 정도는 게놈 시대가 1995년에 시작했다고 말할 수 있게 만들었다.

TIGR의 같은 그룹이 몇 개월 후에 다른 박테리아로서 요도염의 원인인 *Mycoplasma genitalium*의 완전한 게놈의 분석을 출판하였고 그후 첫번째 진핵생물인 곰팡이 *Saccharomyces cerevisiae* (또는 *S. cerevisiae*, 제빵 효모)의 서열이 다른 그룹에 의해서 출판되었다. TIGR 그룹에서 게놈서열을 얻고 조립하는데 만든 방법은 그 자신이 분수령으로 그 방법은 컴퓨터 기술에 크게 의존하였으나 여기서 논의하는 주제에 넘어선다. 다음의 몇 년은 출판된 완전한 게놈의 수가 거대하게 증가하였으며 속도도 여전히 증가하는 중이다. 게놈 시대의 시작전에 일반적으로 이용가능한 DNA 서열 정보들은 과학자들 사이에서 마그네틱 테이프로, CD로, 최종적으로는 인터넷으로 배포되어졌다. 지금 전체 게놈들은 여러 공용 데이터베이스로부터 빠르게 다운로드 받을 수 있다.

*H. influenzae*의 서열화를 끝낸 후에 벤터의 그룹은 어떤 게놈 프로젝트의 다음 단계인 게놈 주석(genome annotation)으로 이동했다. 주석은 다양한 단계를 포함하여 결코 진정으로 완성되지 않는다. 그러나 대부분의 서열화 프로젝트들은 최소한 두 단계를 시행한다. 보통 간단한 첫번째 분석으로 게놈의 주요 구조와 특성을 확인하는 것을 목표로 한다. 두번째 더 복잡한 단계로 이 구조들의 생물학적인 기능들을 예측하는 것을 목표로 한다. 이 책의 첫번째 단원은 서열 주석을 수행하는데 우리에게 기본적인 도구들을 제시한다. 우리는 서열 조립(어떠한 분석이 시작되기 전에 해야만 되는 전체 게놈 서열을 구조화하는 초기 작업)의 더 발전된 주제인 생명정보학의 더 진화된 단계로 떠난다.

지금 우리는 다양한 종과 같은 종의 다양한 개별의 완전한 게놈 서열을 가지면서 과학자들은 유기체들 사이의 전체 게놈 비교와 차이점을 분석하기 시작 할 수 있다. 물론 2001년의 매력적인 헤드라인인 인간체놈 서열의 초안의 완성은 곧 그후 마우스, 쥐, 개, 침팬지, 모기, 다른 유기체로 따르게 되는 게놈 시대의 많은 획기적인 사건 중의 하나일 뿐이다. 표 1.1은 이 책을 통해서 예로 사용되는 유기체이면서 중요한 모델 유기체를 게놈 길이(길이의 단위는 다음 부분에서 정의될 것이다)과 그것의 완성 날짜와 함께 목록으로 만들었다. 우리는 여기서 우리가 생각하는 많은 새로운 분석에 대해 논의하기 전에 해결되어야 하는 정보의 저장, 공유, 관리에 대한 많은 시도들이 있었다는 것을 강조해야만 한다.

이 단원의 나머지에서는 우리는 이전의 게놈 논문의 분석들의 몇 개를 재현하기 위해 게놈 정보의 분석을 시작한다. 우리는 다음의 단원들에서도 이 목표를 계속하여 독자들에게 이 획기적인 분석들을 반복하는데 필요한 개념과 정보, 도구를 제공할 것이다. 우리가 우리의 완전한 게놈의 첫번째 통계적인 실험을 시작하기 전에, 그러나 우리는 어떻게 DNA가 세포 안에서 구조물을 이루는지와 같은 몇 가지 중요 생물학적 사실과 DNA의 분석에 포함된 중요 통계적인 이슈를 요약할 필요가 있을 것이다.

이 시점에서 단지 DNA 서열 정보보다는 더 많은 것을 포함하는 게놈 정보를 강조할 필요가 있다. 1995년에 스탠포드 대학의 패트 브라운(Pat Brown)에 의한 과학자들의 그룹은 한 단일의 실험에서 한 유기체의 모든 유전자들의 활성도를 측정하는 높은 처리량의 기술을 도입하였다. 이 많은 정보들의 분석은 단원 9와 부분적으로 단원 10에서 논의될 것이다.

1.2 게놈의 해부

첫번째 정의로서 우리는 게놈은 세포 안에 포함된 모든 DNA의 한 세트라고 말할 수 있다. 짧게 우리는 어떻게 몇 유기체가 실제적으로 단일 세포 안에 여러 개의 게놈을 가지는지를 설명할 것이다. 게놈은 하나 또는 더 많은 긴 DNA 서열이 함께 모여 염색체들로서 형성된다. 이 염색체들은 선형이거나 원형일 수 있으며 세포가 분열할 때 정확하게 복제된다. 세포 안에서 염색체의 완전한 구획은 단백질을 합성하는데 필요한 DNA와 생존하는데 필요한 다른 분자들과 그들의 합성을 정확하게 조절하는데 필요한 정보들 또한 포함한다. 프롤로그에서 언급했던 것처럼 우리는 각각의 단백질은 하나의 특이적인 유전자에 의해 코딩된다. 이 목적에 필요한 정보들은 DNA에 포함되어 있다.

DNA 분자들은 염기(base)라고 불리는 화학적인 부분이 서로 다른 뉴클레오타이드(nucleotides)라고 불리는 작은 분자들의 체인으로 구성된다. 생화학적인 이유로 DNA 서열은 방향성을 가진다. 각 염색체나 유전자를 읽는데 특이적인 방향을 구별할 수 있다. 세포의 효소적인 기작은 DNA를 5'에서 3'end(이것은 DNA를 만드는 핵산의 화학적인 관례이다)으로 읽는다. 이것은 종종 서열의 각 왼쪽과 오른쪽 끝으로서 제시된다.

DNA서열은 단일가닥이거나 이중가닥일 수 있다. 이 DNA의 이중가닥의 특성은 염기의 짝지음으로서 생겨난다. 이중가닥일때는 2가닥이 반대방향이며 서로 상보적이다. 이 상보성은 각 A,C,G,T가 한 가닥일 경우 다른 가닥에는 각각 T,G,C,A가 있게 되는 것을 의미한다. 염색체는 이중가닥이며(그러므로 이중 나선, "double helix") 유전자에 대한 정보는 각 가닥에 포함될 수 있다. 중요하게 이 짝지음은 세포가 한 가닥으로부터 전체 게놈을 다시 구성하는 인코딩에서 완전한 반복성을 가져오며 정확한 복제를 가능하게 한다. 그러나 간단하게 우리는 보통 5'에서 3' 방향으로 관심있는 DNA 서열의 한 가닥만을 쓴다.

예 1.1

서열의 방향성과 상보성. 서열 5'-ATGCATGC-3'은 서열 3'-TACGTACG-5'과 상보적이며 종종 다른 방향성이 제공되지 않는다면 간단하게 ATGCATGC라고 쓴다.

우리는 DNA 분자들이 포함하는 염기의 특성에 따라 다른 뉴클레오타이드의 체인으로 구성된다는 것을 보았다. 그 결과, 이중가닥 DNA에 대한 DNA 알파벳 글자는 다양하게 뉴클레오타이드 nucleotides (nt), 염기 base, 염기쌍 base pair(bp)로 불린다. DNA 서열의 길이는 base로서 킬로베이스(1000bp 또는 Kb) 또는 메가베이스(1 000 000 bp 또는 Mb)로 특정될 수 있다. 게놈들은 다른 유기체들에서 사이즈로 킬로베이스에서 메가베이스를 가지며 종종 매우 다른 생물학적인 속성들을 가진다.

원핵생물의 게놈. 이 책을 쓸 때 TIGR의 종합적인 미생물 자원 웹사이트는 217개의 박테리아와 21개의 고세균(1995년에 논의된 2개의 박테리아 게놈을 포함하여)로 239개의 완전한 게놈 서열을 포함한다. 아 숫자는 그 때 이후로 증가했을 것이다. 진정세균(eubacteria)와 고세균(archaea)는 세포 안에 그들의 게놈을 구분하는 진핵생물의 구조체인 핵이 없는 독립 유기체인 원핵생물의 두 주요 그룹이다. 진핵생물의 유기체들은 일반적으로 단일이며 원형의 0.5에서 1.3 메가베이스 사이의 길이의 게놈을 가진다. *M. genitalium*은 알려진 가장 작은 원핵생물 게놈을 가지며 단지 580 074 베이스를 가진다. 상대적으로 작은 게놈을 가지는 데다가 또한 보다 간단한 유전자들과 유전적 조절 서열들을 가진다. 다음 단원에서 우리는 이 이슈들을 보다 깊게 살피며 유전자의 동정에 어떤 영향을 미칠는지 볼 것이다. 이러한 단순성과 더 복잡한 게놈과의 기본적인 유사성 때문에 우리는 이 책에서 수행하는 분석의 첫번째 예로서 많은 원핵생물의 게놈을 사용할 것이다. 우리는 특히 *H. influenzae*, *M. genitalium, Chlamydia trachomatis*에 초점을 맞출 것이다.

바이러스 게놈. 바이러스가 독립 유기체가 아님에도 불구하고 바이러스 게놈의 분석은 매우 유익하다. '전게놈 시대'로 생각되는 지난 1970년대로부터 시작하여 최소한 천 개의 바이러스 게놈들이 서열화되었다. 이러한 게놈들은 보통 5~50Kb로 매우 짧고 매우 적은 유전자들을 포함한다. 그들의 서열화는 생물학에서 획기적인 사건으로서 과학자들이 더 큰 독립유기체들의 게놈을 분석하는데 필수적인 개념적인 도구를 개발하게 해주었다. 그들은 후에 더 큰 게놈들로 전개되어 사용되는 많은 방법들을 연습하는데 훌륭한 모델로서 우리는 기본적인 원칙들을 설명하는데 사용할 것이다. 그들의 분석은 또한 역학적, 의학적 적용으로 HIV와 SARS(단원 6과 7을 보라)를 포함하는 경우에 예증되면서 관련성이 높다. 특히 바이러스 게놈은 단일가닥이거나 이중가닥, DNA- 또는 RNA- 기반(우리는 다름 단원에서 분자 RNA에 대해서 더 배울 것이다)일 수 있다. 그들의 작은 크기 때문에 우리는 더 많은 수의 바이러스 게놈을 컴퓨터로 동시에

분석할 수 있으며 이 작업은 더 큰 게놈 서열의 경우에 많은 기계들을 요구할 것이다.

진핵생물의 게놈. 진핵생물의 핵의 게놈은 보통 유기체의 게놈으로 생각된다(많은 진핵생물에 포함된 기관적인 게놈들의 설명은 다음 단락을 보라). 이 핵 게놈은 원핵생물의 게놈보다 훨씬 크며 몇 곰팡이에서 8Mb부터 단일세포인 아메바의 몇 종에 대해서는 670기가베이스(십억베이스 또는 Gb)까지 크기를 가진다. 인간은 중간인 3.5Gb의 크기를 가진다. 진핵생물 게놈의 큰 크기 때문에 그들의 서열화는 여전히 보통 실험실 협회에 의해서 큰 노력을 요한다. 이러한 실험실들은 한 작업을 나누어서 같은 게놈의 다른 선형 염색체를 서열화하게 된다. 현재 우리는 진화적인 나무의 다양한 가지를 포함하는 50개의 다른 진핵생물 유기체보다 더 많은 것을 대표하는 서열을 가진다 : 곰팡이, *S.cerevisiae*(제빵 효모); 원형 벌레, *Caenorhabditis elegans*; 해마, *Danio rerio*; 초파리와 같은 중요한 곤충, *Drosophila melanogaster*; 모기, *Anopheles gambiae*; 인간 *Homo sapiens,* 쥐, *Mus musculus*와 같은 포유류; 쌀과 같은 식물, *Oryza sativa.*

그런 게놈의 큰 크기는 일반적으로 유전자의 큰 크기 때문이 아니라 거대한 양의 반복적인 "junk" DNA 때문이다. 인간 게놈의 단지 5%만이 기능적이며(단백질을 코딩하는) 최소한 50%의 게놈은 반복적인 부분과 기생 DNA로 이루어져 있다고 알려져 있다. 다세포 유기체의 각 세포 안에 DNA의 많은 양이 채워지는 포장 문제에 더불어 대부분 진핵생물들은 각 세포 안에 그들의 핵 게놈의 2개 복제를 가진다. 하나를 각 부모로부터 받는다. 우리는 단일 보체를 단수체 haploid 세트라고 하며 반대로 두 게놈의 세트를 복수체 diploid라고 한다.

기관적인 게놈. 이러한 거대한 핵 게놈에 더불어 대부분 진핵생물은 또한 각 세포 안에 하나나 더 많은 작은 게놈을 가진다. 이것들은 세포의 기관안에 포함되며 가장 공통적인 것이 미토콘드리아와 엽록체이다. 이러한 기관적인 게놈들은 진핵생물의 세포 안에 살던 공생 원핵생물 유기체의 잔여물로 여겨진다. 우리는 지금 미토콘드리아와 엽록체의 최소 600종의 게놈서열을 가지며 종종 한 종안에 다양한 개별의 여러 개의 전체 게놈을 가진다. 이 게놈들은 보통 단지 수천 베이스로 길며 원형이며 적은 수의 필수 유전자들을 지닌다. 이것들은 세포 안에 더 많은 복제로 인해 더 많은 생산물을 가지는 이런 기관들의 각각이 수 백 개 일 수 있다. 미토콘드리아 DNA(mtDNA)는 특히 인류학적인 분석으로 중요하며 우리는 이것을 5단원에서 인간들 사이의 전체 게놈 비교에 사용할 것이다.

1.3 게놈 서열의 확률 모델

모든 모델은 틀리지만 어떤 것들은 유용하다.(G.E.P. box)

첫번째 독립유기체의 전체 게놈이 1995년에 서열화되었을 때 이미 세포의 기능의 일반적인 작용과 DNA 서열에 대해 많은 부분들이 알려져 있었다. 1980년대에 이용 가능한 작은 바이러스와 기관들의 완전한 게놈들은 다양한 유기체로부터 각 유전자의 서열로서 있었다. 1995년 이전에 규모적으로 비교하여 심지어 간단한 박테리아 유전자를 분석하는데 해결할 수 있었던 문제들이, 이전의 실험들은 현대의 전체 게놈 분석으로 진화하는데 기본적인 통계적인 기술들을 제공하였다.

전산 게놈학의 연구의 큰 부분은 통계적인 방법으로 구성된다. 수백만 또는 수십억의 데이터 포인트들을 포함하는 어떠한 사업이라도 필수적으로 통계학을 필요로 하지만 그 문제는 특히 DNA 서열의 연구에 중요하다. 이것의 한 이유는 우리는 종종 수백만의 염기의 서열에서 관심있는 구조(유전자 같은)를 찾으려 하며 많은 중요한 경우들에서 그런 서열들의 대부분은 생물학적으로 관련된 정보를 포함하고 있지 않다. 즉, 게놈 서열에서 signal-to-noise ratio는 매우 낮을 것이다. 우리가 보는 바와 같이 관심있는 부분들은 종종 random background에 담겨져 있다. 그것들을 알아내기 위해서는 세련된 통계학적인 알고리즘 도구를 필요로 한다.

첫번째 과정으로서 우리는 DNA 서열의 확률 모델들을 필요로 한다. 이것들은 이 책에서 보이는 분석의 모든 것에 대한 기반을 둘 것이다. 우리는 첫번째로 DNA의 확률 모델에 대한 기본적인 개념을 정의하고 게놈 서열 정보의 분석에 대한 간단한 통계적인 틀을 제시한다. 확률 모델의 유용성을 강조하기 위해 우리는 게놈 서열의 간단한 분석을 수행함으로서 단원을 계속 진행한다. 다음의 단원들은 흥미로운 복잡한 생물적인 질문들과 그들에 대해 대답하는데 필요한 적합한 통계적인 방법들을 소개할 것이다.

알파벳, 서열, 서열 공간. 우리는 DNA가 3차적인 속성을 가지는 복잡한 분자라는 것을 잊지 않음에도 불구하고 종종 알파벳 기호의 서열 {A, C, G, T}의 1차적인 물체로서 모델로 하는 것이 편리하다. 이런 추상적 개념은 굉장히 강력하여 우리가 다수의 통계적인 도구로 전개하게 해준다. 이것은 또한 부정확하며 분자의 3차 구조에 포함된 정보를 무시한다. 이 책에서

우리는 이런 접근을 어떠한 경고 없이 만들어 그것에 기반한 분석의 통계적이며 전산적인 방법들을 개발할 것이다.

정의 1.1

DNA 서열과 게놈 : 정식 모델. "DNA 서열", **s**,는 알파벳 N={A,C,G,T}로 이루어진 한정된 줄이다. "게놈"은 모든 DNA 서열의 한 세트로서 유기체의 기관들과 연관되어 있다.

한 알파벳으로 구성된 줄로서의 게놈의 표현은 매우 일반적이며 우리가 서열 진화, 서열 유사성, 서열 분석의 다양한 형태의 통계적인 모델을 개발하게 해 준다. 그것들 중의 어떤 것은 이 단원에서 논의된다.

정의 1.2

서열의 구성원소. 우리는 서열의 구성원소를 각각의 뉴클레오타이드를 s_i로 표시하여 $s=s_1s_2\cdots s_n$으로 표시한다. 색인 k의 한 세트일 경우 우리는 서열을 그들의 원래 순서에서 **s**의 상응하는 원소로 함께 연관시킴으로서 서열을 생각할 수 있다. K={i, j, k}면 $s(K) = s_1s_2\cdots s_k$. 그 세트가 정수의 닫힌 간격이라면, K=[i, j] , 우리는 또한 K=(I : j)로서 상응하는 서열은 substring s(i : j)로서 표시할 수 있다. 단지 한 원소로 형성되어 있다면 K={ i } 그리고 나서 이것들은 단일의 특이적인 서열의 기호로서 s_i = s(i)로서 표시한다.

예 1.2

서열의 원소들. DNA서열에서 **s**= ATATGTCGTGCA로 우리는 s(3:6) =ATGT와 s(8) =s_8 =G.

주의 1.1

줄과 서열. 우리는 생물학에서 서열을 보통 표준 컴퓨터 과학 용어에서는 줄이라고 부른다. 이 구별은 컴퓨터 과학에서 두 개의 다른 물체인 subsequence와 substring을 정의할 때 관련이 있다. 우리는 이 책에서 subsequence를 긴 서열의 연속된 짧은 서열을 말하며 컴퓨터 과학에서는 substring이라고 부른다. 컴퓨터 과학에서는 긴 서열의 비 연속적인 세트를 subsequence라고 한다.

거의 모든 확률 서열 분석 방법들은 두 개의 간단한 모델, 또는 그것의 변형들의 한 개로 추측한다. 그들은 다항식의 모델로서 Marcov 모델로 아래에 설명할 것이다. 종종 모델링의 경우에 이것들은 모든 면에서 진짜 DNA 서열을 모방할 필요가 없다. 그들의 주요 특징들은 그들이 효율적으로 컴퓨터화하는 동안에 서열의 특성을 충분히 갖는다. 즉, 그들은 정확성과 효율성의 사이의 절충안의 결과이다. 이 단원에서 우리는 DNA 서열을 다룸에도 불구하고 이 책의 나머지에서는 우리는 생물적인 서열의 다양한 다른 형태를 찾을 것이다. 서열은 다른 알파벳으로 정의된다. 우리가 제시하는 모든 알고리즘은 어떠한 형태의 서열에도 유효하며 우리는 종종 그들을 일반성을 유지하기 위해 포괄적인 알파벳 ∑로 정의한다. 가장 일반적인 다른 형태의 생물학 서열은 RNA 서열(또한 4개의 알파벳으로 정의된다, N_{RNA} = {A, C, G, U})과 아미노산 서열(20개의 알파벳에 기초하여 A={A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}이며 2단원에서 논의된다)이다. 종종 또한 알파벳은 뉴클레오타이드 알파벳 N으로부터 3개의 알파벳으로 형성된 코돈의 서열을 정의하는데 유용하며 C={AAA, …, TTT}로서 나타내어질 것이다.

다항식의 서열 모델. 가장 간단한 DNA 서열의 모델은 뉴클레오타이드가 독립적이며 동일하게 분포되어 있다는 것이다("i.i.d" 추정) : 추측통계학 과정에 의해 생성된 서열은 알파벳 N으로 같은 분포로 그들로부터 독립적이며 무작위로 i 위치에 4개의 뉴클레오타이드를 생성한다. 이것은 다항성 서열 모델로 불리며 알파벳 $p=(p_A, p_C, p_G, p_T)$로 가능한 분포를 선택함으로서 간단히 정의되며 4개의 뉴클레오타이드들을 서열 s의 i 위치에서 관찰할 확률은 $p_x=p(\mathbf{s}(i)=x)$로 표시한다. 이 모델은 또한 일반적인 기호인 $x\in\sum$ (x는 어떤 알파벳이라도 가능하다)로 관찰할 확률을 계산하는데 사용된다.

이런 모델로 우리는 모든 4개의 뉴클레오타이드가 동등한 빈도($p_A=p_C=p_G=p_T$)로 또는 그들이 어떤 데이터세트로부터 관찰되는 빈도인지를 추측한다. 우리의 유일한 필수조건은 분포가 일정하게 $p_A+p_C+p_G+p_T=1$ 로 표준화를 만족할 필요가 있다는 것이다.

다항식 모델은 우리가 쉽게 주어진 서열(P로 표시된)의 확률을 계산하도록 하며 또한 주어진 모델(L로 표시된)의 데이터의 가능성으로 불린다. 주어진 서열은 $s=s_1s_2\cdots s_n$의 확률은

$$P(s) = \prod_{i=1}^{n} p(\mathbf{s(i)})$$

이것은 각 뉴클레오타이드들의 확률을 모두 곱한 값과 동일하다.

물론 우리는 DNA 서열이 정말로 무작위라고는 기대하지 않으나 무작위로 생성된 서열에 대한 예측을 설명하는 모델은 유용할 수 있다. 게다가 이런 간단한 모델은 이미 특정 응용

에서 유용한 서열의 행동을 충분히 설명하며 수학적으로 다루는 데도 매우 쉽게 된다. 심지어 이런 간단한 모델에서의 위반을 찾는 것은 우리에게 게놈의 흥미로운 부분을 가리킬 것이다. 우리는 쉽게 이 모델이 현실적인지 실제 데이터를 그것의 추정과 맞는지 확인해봄으로서 테스트할 수 있다. 우리는 이것을 서열의 다양한 부분의 기호들의 빈도 - 서열의 독립적이고 동일하게 분포하는(i.i.d) 확률 분포의 정상성의 추정을 확인하기 위해 - 를 측정함으로서 또는 이웃하는 뉴클레오타이드들 사이의 관계를 찾아 독립적인 추정의 오류를 테스트 함으로서 할 수 있다. A, C, G, T의 빈도가 변화하며 가까운 기호들 사이의 강한 연관성이 있는 지역은 꽤 흥미로울 수 있으며 우리는 그것을 나중에 살펴볼 것이다.

Marcov 서열 모델. 더 복잡한 DNA 서열 모델은 Marcov 체인 모델 이론에 의해서 제공된다. Marcov 체인에서 한 기호를 관찰할 확률은 서열 안의 그 기호 이전의 기호에 의존한다. 그렇게 하면 Marcov 체인들은 뉴클레오타이드들 사이의 지역적인 연관성을 모델로 할 수 있다. Marcov 체인은 각 기호의 확률이 오직 그 바로 앞의 하나에만 의존한다면 order 1이라고 말할 수 있으며 지난 기호들에 대한 의존도가 더 긴 길이까지 확장될수록 order는 증가하게 된다. 우리는 이전의 기호에 대한 의존도가 없기 때문에 다항성 모델을 order 0의 Marcov 체인모델로서 생각할 수 있다. 어떻게 우리가 어떤 모델을 고르고 어떤 order의 Marcov 모델을 선택할 것인가? 이것은 가설 테스트의 문제로 2단원에서 논의할 것이다. 지금은 우리는 단순히 Marcov 모델의 기본적인 측면을 논의할 것이다. (Marcov 체인은 생명정보학에서 중요 도구로서 우리는 단락 5.4.1과 5.4.2에서 그것들을 다시 마주할 것이다).

간략히 Marcov 체인은 상태(이 경우에는 알파벳의 기호들)의 한 세트와 한 상태로부터 다른 상태로 전환 확률로 정의되는 과정이다. 전환 확률은 전환 매트릭스 T로 정리된다. 상태 공간을 통한 과정의 궤도는 서열로 정의한다. 이것은 그림 1.1에 나와있다.

그림 1.1 Marcov 체인 과정의 궤도는 기호의 서열을 형성한다. 4개 뉴클레오타이드의 하나로 시작하여 서열에서 다음 뉴클레오타이드의 확률은 현재 상태에 따라 결정된다.

그림은 4개의 뉴클레오타이드 중에 어느 하나로부터 시작하여 서열에서 다음 뉴틀레오타이드의 확률은 현재 상태에 의해서 결정되는 것을 보여준다. 우리가 G로 시작한 경우에 4개의 다른 뉴클레오타이드의 어느 것이 다음에 나타난 확률은 p_{GA}, p_{GC}, p_{GG}, p_{GT}로 주어진다. 다른 뉴클레오타이드 상태로의 이동은 다른 전환 확률을 의미한다. 우리가 상태 A로 이동했다면 이것은 p_{AA}, p_{AC}, p_{AG}, p_{AT}일 것이다. 이런 형식으로 Marcov 체인은 DNA 서열을 정의한다. 전환 확률의 모두는 매트릭스 T로 주어지며 시작 상태의 확률은 π = ($\pi A, \pi C, \pi G, \pi T$)로 주어진다. 다시 확실한 표준화 변수로 보면

$$T = \begin{matrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ P_{GA} & p_{GC} & p_{GG} & p_{GT} \\ P_{TA} & p_{TC} & p_{TG} & p_{TT} \end{matrix}$$

$$\pi = \pi_A \quad \pi_C \quad \pi_G \quad \pi_T$$

Marcov 모델은 그러므로 더 이상 기호들이 독립적이라고 추정하지 않으며 짧은 단위의 연관성이 나타날 수 있다. (전환 확률이 모두 0.25라면 우리는 다시 다항성 모델로 돌아간다.)

예 1.3

Marcov DNA 서열. Marcov 체인을 사용하여 일정한 시작 확률(π)과 전환 매트릭스가 정의된다.

		to A	to C	to G	to T
	from A	0.6	0.2	0.1	0.1
T =	from C	0.1	0.1	0.8	0
	from G	0.2	0.2	0.3	0.3
	from T	0.1	0.8	0	0.1

우리는 다음의 서열을 만든다.

ACGCGTAATCAAAAAATCGGTCGTCGGAAAAAAAAAATCG

볼 수 있듯이 As 뒤에 많은 As가 나오며 Ts는 Cs 뒤에, Cs는 Gs 뒤에 나오며 우리의 전환 매트릭스로 설명된다.

전환 매트릭스의 진입은 형식적으로 다음과 같이 정의된다.

$P_{xy} = p(s_{i+1} = y \mid s_i = x)$.

이것은 상태 x에서 상태 y로 가는 확률은 이전에 상태 x일 때 상태 y로 가는 조건부 확

률과 같다는 것을 말한다. 우리는 일반적으로 전체 서열의 확률을 이음(joint) 확률이라고 정의한다.

$P(s) = P(s_1 s_2 \cdots s_n)$.

이 확률의 계산은 우리가 그것을 인자화(factorization)할 수 있을 때 매우 간단해 질 수 있다. 다항식 모델의 경우에 인자화는 명확하다. $P(s) = p(s_1)p(s_2) \cdots p(s_n)$. Marcov 체인(order 1)의 경우 단지 조금 더 복잡하다.

$P(s) = P(s_n | s_{n-1})\, P(s_{n-1} | s_{n-2}) \cdots P(s_2 | s_1)\pi(s_1)$

또는

$P(s) = \pi(s_1)\prod_{i=2}^{n} p(s_i | s_{i-1}) = \pi(s_1)\prod_{i=2}^{n} P_{s_{i-1}s_i}$

$\pi(s_1=x)$는 s_1위치에서 기호 x가 나올 확률을 말한다. 즉, 우리는 한 기호의 확률이 전체 서열의 이음 확률의 계산을 간단히 하기 위해 오직 이전의 기호에 의존한다는 사실을 이용한다.

표 1.2 *H. influenzae* 게놈의 기본적인 통계 : 각 뉴클레오타이드의 수와 그것의 상대적인 빈도. 서열의 총 길이는 1 830 138bp 이다.

1.4 게놈에 주석달기 : 통계적인 서열 분석

게놈 서열에서 흥미로운 다양한 인자들이 있으며 많은 부분이 이 책의 다양한 단원에서 논의될 것이다. 예를 들어, 2단원은 유전자의 구조와 그들을 어떻게 찾는지 그들이 어떻게 조절되는지를 논의할 것이다. 여기서 우리는 뉴클레오타이드, 디뉴클레오타이드(베이스의 쌍),다른 짧은 DNA단위들의 빈도와 같은 DNA 서열의 더 간단한 통계적인 특징들을 검사한다. 우리는 이런 게놈 주석과 다음의 분석을 위한 다음 단계로서 예비 설명의 기본적인 중요성을 볼 것이다.

염기 조성. 게놈 서열의 가장 기본적인 특징 중 하나는 염기 조성으로 A,G,C,T 뉴클레오타이드가 존재하는 비율이다. *H. influenzae*의 경우 우리는 쉽게 염기의 각 타입의 수를 세어서 각 염기의 빈도를 알기 위해 게놈의 총 길이(2작업은 DNA서열의 오직 한 가닥만에서 행해진다.)로 나눌 수 있다. 표 1.2는 게놈에서 각 염기가 나타나는 숫자와 그것의 상대적인 빈도를 보여준다.(총 길이, L, *H. influenzae*의 게놈은 1 830 138bp이다)

우리는 4개 뉴클레오타이드가 게놈에서 동일한 빈도로 사용되지 않았다는 것을 알 수 있다. A와 T는 G와 C보다 훨씬 더 흔하다. 사실, 모든 염기들이 어떤 게놈에서든지 동일한 빈도로 사용되는 것은 확실이 이상하다. 우리는 단지 DNA 분자의 한 가닥에 있는 염기의 수를 세지만 이중 나선의 상보성 때문에 다른 가닥의 모든 염기의 빈도도 정확히 알 수 있다는 것을 알아야 한다. 상보적 서열의 빈도는 T=0.3102, G=0.1916, C=0.1898, A=0.3082일 것이다.

한 게놈의 공통적인 염기 조성에 더하여 서열에서 뉴클레오타이드의 빈도에서 지역적인 변동을 고려하는 것은 흥미롭다. 우리는 염색체를 따라 크기 k의 움직이는 창으로 각 창의 각 염기의 빈도를 측정하고 창의 중심 위치에 이 수치를 할당함으로서 지역적인 염기 조성을 측정할 수 있다. 이것은 길이 $L - k + 1$의 벡터로서 나타날 수 있으며 그림 1.2와 1.3에서 보여진다.(창 크기는 각각 90 000bp와 20 000bp이다)

그림 1.2 *H. influenzae*의 서열을 따라 뉴클레오타이드 빈도의 지역적인 변동을 보여주는 움직이는 창 그림.(90 000bp의 창 크기)

그림 1.3 *H. influenzae*의 서열을 따라 뉴클레오타이드 빈도의 지역적인 변동을 보여주는 움직이는 창 그림.(20 000bp의 창 크기)

물론 더 작은 창 크기는 염기 조성의 높은 다양성을 보이며 큰 창은 다른 염기 조성을 가진 작은 지역을 놓치므로 중요 조건은 k를 선택하는 것에 있다. k를 다양하게 하는 것은 그림 1.3에서 보여지는 것처럼 다른 크기에서 패턴들을 보여준다. 두 경우는 흥미로운 패턴을 감출 수 있으므로 k를 선택할 때 몇 가지 검사가 필요하다.

표 1.3 3개의 다른 유기체에서의 전체 GC 비율의 비교(이 양은 종간에서 매우 특이적으로 다를 수 있다)

*H. influenzae*의 게놈의 염기 조성에는 꽤 많이 다양성이 있다는 것은 명확하다. 이러한 다양성은 i.i.d. 확률 분산으로부터 그려지는 뉴클레오타이드의 다항성 모델의 추정에 첫번째 중요한 오류를 보여준다. 이 분포는 염색체를 따라 확실히 변한다.

이런 통계적인 분석에 기초하여 다양한 다른 분석들이 수행될 수 있다. 예를 들어 우리는 서열에서 염기 조성이 변화거나 뉴클레오타이드의 몇 부류(C+G vs. A+T)의 이음 빈도의 변화한 위치를 확인하는데 관심이 있다. 이것은 다음의 몇 단락에서의 주제이다. 우리는 또한 빈도('AT'와 같은 문자의 쌍)과 다른 짧은 DNA 단위들을 검사할 것이다.

GC 함량. 위에서 수행된 4개 뉴클레오타이드의 빈도의 분석은 꽤 자연스러우나 사실은 대부분의 생물학적인 요구는 몹시 복잡하다. 대부분의 논문들이 보고하는 것은(일반적으로 필요한 것) A와 T의 모임 빈도(AT 함량) 대 C와 G의 모임 빈도(GC 함량)이다. 이런 두 양들은 합계가 항상 1이 되어야 하며 단지 GC 함량 만을 보고한다. 단순히 GC함량을 보고하는 이유는 -화학적인 이유 때문- 게놈에서 A와 T의 함량과 같이 G와 C의 함량이 종종 매우 유사하기 때문이다. 이런 방식으로 오직 한 값이 4개를 대신하여 보고되는데 필요하다. 표 1.2를 되돌아보면 우리는 *H. influenzae*의 게놈에서 이 경우를 볼 수 있다.

GC 함량의 간단 분석은 확연한 생물학적 정보를 보여줄 수 있다. 예를 들어 박테리아에서 이런 빈도들은 유기체의 복제 기관에 의존하며 종에 따라 매우 달라질 수 있다(표 1.3을 예로 보라). 이것 때문에 GC 함량은 게놈 평균으로부터 달라진 함량이 있는 위치를 확인함으로서 외부의 유전 물질이 게놈에 삽입되었는지를 찾는데 사용될 수 있다. Horizontal gene transfer와 같은 현상에서 한 종이 다른 유기체-바이러스와 같은-로부터 subsequence를 얻었는지를 잘 알려진다. 각 박테리아 게놈 서열을 환성한 후에 비전형적인 염기서열의 지역을 찾음으로서 연구자들은 평행적으로 획득한 유전자들을 찾는다.

예를 들어 *H. influenzae*의 게놈은 비이상적인 GC 함량의 30Kb 지역을 게놈 안으로 1.56Mb에 포함한다(원형 염색체에서의 위치는 일반적으로 세포가 분열할 때 DNA 복제가 시작하는 위치와 관련이 있다). 우리는 이 지역을 GC 함량의 위치를 *H. influenzae*의 게놈에 표시함으로서 그림 1.2에서 보는 것처럼 그림의 오른쪽 끝을 향하는 것을 알 수 있다. 그림을 검사함으로서 우리는 대량 1.56Mb에서 1.59Mb에서 DNA의 짧은 부분이 나머지 게놈의 부분보다 높은 GC 함량을 가진다는 것을 할 수 있다. 이 부분은 바이러스 DNA가 *H. influenzae*의 게놈에 고대에 삽입되었기 때문이다. 더 많은 분석(단원 2와 3에서 개발된 방식을 사용하여)은 이런 외부 서열의 동일성과 그것의 가장 가능한 유래를 확인할 수 있다.

이런 종류의 분석을 위한 다음 단계는 평균과는 많이 다르거나 매우 다른 지역들 사이의 경계를 정의할 수 있는 통계적인 특징들을 자동적으로 찾는 방법을 가지는 것이다. 우리는 이 방법은 변화 지점 분석 change point analysis라고 부른다.

변화 지점 분석. 우리는 GC 함량과 같은 통계적인 특징이 변하는 서열의 지역들을 확인하는 방법을 가질 것이다. 이 지역들은 변화 지점이라고 불리는데 게놈을 대략 동일한 통계 양상의 지역으로 나누며 중요한 생물학적인 신호들을 확인하는 데 도움을 줄 수 있다. 변화 지점 분석은 다양한 방식으로 행해질 수 있다. 하나는 특히 효과적인 접근으로 숨겨진 Marcov 모델에 기초하여 단원 4에서 세세하게 논의될 것이다. 지금은 우리의 기술을 변화 지점 분석을 수행하는데 매우 기초적인 접근으로서 제한한다.

다른 통계적인 양상의 지역을 찾는 대부분의 간단한 전략은 2개의 지역을 구별하는 역치를 정하는 것을 포함한다. 2개의 창 사이에 이 역치를 지나면 우리는 변화 지점을 확인하게 된다. 물론 사용되는 창의 크기처럼 이 역치를 정하는 것은 통계적인 문제이다. 둘 다 가설 테스트(이것은 다음 단원에서 논의된다)의 문제에서 무작위 데이터 안의 변화를 찾는 확률을 가지고 해야만 한다. 지금은 GC 함량이 변하는 지점이 통계적인 분석에 필요 없이 확실한 지점을 간단한 예로 주어야 할 것이다.

예 1.4

λ-phage 의 GC 함량의 변화. 박테리오파지 람다(λ)는 1982년에 완전히 서열화된 첫번째 바이러스 게놈 중의 하나로 48 502 bp의 길이를 가진다. 파지는 박테리아를 감염시기는 바이러스로 박테리오파지 람다는 박테리아 잘 알려진 모델 시스템인 *E.coli*를 감염시킨다. 파지 게놈의 분석은 완전히 다른 GC 함량을 가진 2개의 부분으로 구성된다. 첫번째 G+C rich와 두번째 A+T rich부분이다. 이것은 게놈에서 염기 조성의 균등한 지역을 확실하게 나누는 변화 지점의 간단한 예이다(그림 1.4를 보라).

주의 1.2

GC 함량과 관련된 물리적인 특성. AT-와 GC-rich 지역들 사이의 흥미로운 차이점은 2개의 DNA 가닥이 떨어지기 위해 필요한 에너지이다. AT-rich 지역은 더 낮은 온도에서 분리된다. 이 사실에 미루어 대양의 구멍에서 발견되는 높은 온도에서 사는 호열성 유기체들은 매우 GC-쪽으로 치우친 게놈을 가지는 것은 놀랍지 않다. 박테리오파지 람다 게놈에서 GC 함량의 차이는 박테리아 세포에 감염되어 들어갈 때 DNA를 빠르게 분리하기 위해 필요하기 때문일 것이다.

k-mer빈도와 모티프-기울기 motif-bias. 측정되는 게놈의 또 다른 간단하고 중요한 특

징은 길이 2(디뉴클레오타이드 또는 2분자체 dimer)또는 더 긴(3분자체 trimers, k-mers) 뉴클레오타이드 부분의 빈도이다. (길이 k의 부분은 컴퓨터 과학에서 k-grams 또는 k-tuples로 불린다. 우리는 생물학적인 표기로 k-mers라고 부르는 것을 고수한다) 우리는 또한 k(1과 15 사이의)의 값들을 완화하기 위해 적게는 비이상적인 k-mers를 찾는데 관련된 알고리즘적이고 통계적인 문제들을 생각할 것이다. 우리는 "비이상적인" k-mers를 기대되는 것보다 종종 더 많이나 적게 나타나는 것으로 정의한다. 이런 부분들의 위치나 숫자에서의 기울기는 그들의 기능에 대한 중요한 정보를 밝힐 수 있다.

우리는 k-mers를 게놈의 오직 한 가닥을 다시 살펴서 센다. 크기 k의 창의 경우에 우리는 서열을 따라 한 번에 한 염기로 이동하며-한 겹치는 방식으로- 우리가 관찰하는 모든 k-mers를 기록한다. 이것은 L -k + 1 은 길이L의 서열에서 가능한 k-mers이다.(길이 L의 서열에서 2분자체는 L-1)

예 1.5

*H. influenzae*에서 2-mer 빈도. *H. influenzae에서* 디뉴클레오타이드 빈도는 여기서 보여진다.

	*A	*C	*G	*T
*A	0.1202	0.0505	0.0483	0.0912
*C	0.0665	0.0372	0.0396	0.0484
*G	0.0514	0.0522	0.0363	0.0499
*T	0.0721	0.0518	0.0656	0.1189

표의 왼편에는 2분자체의 첫번째 뉴클레오타이드이며 두번째 뉴클레오타이드는 위에 있다.

우리는 게놈을 따라 흥미있는 특정 2-mers의 빈도를 *H. influenzae*의 2분자체 AT에 대하여 그림 1.5와 같이 그려 볼 수 있다(이것은 AT단위의 빈도를 표시하며 A+T의 이음 빈도가 아니다; 유사하게 CG 선도 2-mer CG의 빈도를 나타낸다).

호열성 미생물과 척추동물과 같은 특정 유기체 등에서 디뉴클레오타이드 TA가 드물게 나타나는 것부터 CGs(종종 CpGs라고 불리는)의 낮은 빈도까지 뉴클레오타이드 사용에서 독특한 통계적인 기울기의 많은 예들이 있다. 이런 기울기들을 더 쉽게 보기 위해서 – 카오스 게임 설명 chaos game representation 또는 게놈 사인 genome signature – 다른 k-mers의 관찰되는 빈도들의 색깔 코드로 간단한 그림을 보는 것이 편리하다. 이러한 표현들은 우리가 더 큰 값들일수록 k:가 1-mers인 경우 4개의 고유한 모티프, 2-mers인 경우 16, 4인 경우는 k번째 power를 생각하여 다른 부분의 빈도에서 패턴을 더 쉽게 볼 수 있게 한다(책의 웹사이트에 카오스 게임 설명 뒤의 알고리즘에 대한 설명을 보라).

예 1.6

*H. influenzae*의 빈도 단위. 더 나아간 예로서, 우리는 *H. influenzae*의 게놈 서열에서 가장 빈번한 10-mers를 찾을 수 있다. 이 경우에는 우리는 부분 AAAGTGCGG와 ACCGCACTTT가 가장 빈번하며 둘다 500번이상 나타난다는 것을 알 것이다. 그러나 이런 발견의 의미는 통계적으로 생물학적인 의미로 설립되는 것이 필요하다.

이상한 DNA 부분을 찾기. 간단한 통계적인 분석은 모티프의 적거나 많은 출현을 찾아서 관찰된 기울기가 의미가 있는지를 결정(우리는 2단원에서 패턴의 의미에 대해서 논의할 것이다)하는데 도움을 주는데 사용될 수 있다. 우리는 2-mers의 경우에 이 아이디어를 설명한다. 주요점은 주어진 k-mer의 관찰된 빈도 N와 전형적인 다항식 모델인 백그라운드 모델 하에서 예측된 것을 비교하는 것이다. 두 양 사이의 비율은 얼마나 많이 특정 부분이 백그라운드 모델로부터 벗어나는지를 보여주며 adds ratio라고 불린다.

$$\text{odds ratio} \approx \frac{N(xy)}{N(x)N(y)}$$

다항식 모델에서 생성된 서열이라면 이 비율은 대략 1이어야 한다. 1로부터 확연히 벗어난다면 돌연변이 과정이나 자연 선택에 의한 어던 효과라는 것을 나타낸다. 물론 1로부터 벗어남이 특정 역치보다 커야하며 이것은 각 범주에서 관찰된 모티프의 숫자와 서열의 길이에 의존한다. 10단원에서 우리는 9-mer 모티프의 의미성을 평가할 때 "참고" 서열이 주어진 더 현실적인 백그라운드 모델을 사용할 것이다.

예 1.7

*H. influenzae*에서 비이상적 2분자체찾기. *H. influenzae*의 게놈에서 디뉴틀레오타이드

빈도의 관찰된/예측된 비율은 다음과 같다.

	*A	*C	*G	*T
*A	1.2491	0.8496	0.8210	0.9535
*C	1.1182	1.0121	1.0894	0.8190
*G	0.8736	1.4349	1.0076	0.8526
*T	0.7541	0.8763	1.1204	1.2505

1에서 벗어난 값을 가지는 2분자체는 비이상적으로 나타나며 벗어남의 양이 특정 패턴인지를 고려하기 위해서는 2단원에서 논의되는 도구로 분석될 필요가 있다. 예 1.5에서의 표와 차이점을 쓰면 2분자체 CC는 전의 표에서는 극히 드물게 나오는 것처럼 보이지만 이 분석에서는 이것은 뉴클레오타이드 C가 시작할 때부터 낮은 빈도이기 때문에 의미있는 기울기가 아닌 것처럼 나타난다.

비이상적인 모티프의 생물적인 관련성. 뉴클레오타이드 모티프의 더 적거나 많은 출현은 생물적인 변수들을 반영하여 돌연변이적이거나 선택적이다. 빈번한 단위들은 반복적인 부분들(특정 게놈에서 매우 일반적인 특징), 유전자 조절 특징, 또는 다른 생물적인 기능을 가지는 서열들로 인한 것이다. 드문 모티프들은 전사 인자(2단원을 보라)들을 위한 결합 부분과 바라지 않는 구조적인 특징(DNA의 꼬임(kinking)을 유도하기 때문에)을 가지는 CTAG와 같은 단위들 또는 박테리아의 내부적 면역 시스템과 부합하지 않는 단위들을 포함한다. 박테리아 세포들은 바이러스에 의해 감염될 수 있으며 그들은 반응하여 restriction sites로 알려진 특정 뉴클레오타이드 단위에서 DNA를 자를 수 있는 단백질인 제한 효소를 생산한다. 이런 발견은 거대한 기술적인 응용을 가져와 1978년에는 노벨상을 수상했다.(Werner Arber, Dan Nathans, Hamilton Smith)

이 뉴클레오타이드 모티프들은 제한 효소에 의해 인식되며 박테리아 숙주의 제한 효소를 피하기 위해서 많은 바이러스 게놈에서 under-represented 된다.

그림 1.6

이 그림(게놈 사인으로 불리는)의 색깔 코드는 *H. influenzae*에서 나타나는 다른 k-mers의 관찰되는 빈도들이다. 그림의 사각형 안에 나타나는 4개의 k-mers가 있다.(같은 접두사가진 k-mer들은 같은 4분원에 위치한다.) 이런 표현은 서열에서 비이상적인 통계 패턴을 측정할 수 있게 도와준다.

제한 위치의 흥미로운 특징은 그들은 종종 ABBA와 같이 2 방향에서 같은 방향으로 읽을 수 있는 서열인 palindrome을 형성하는 것이다. 그러나 DNA 서열에서 우리는 palindrome을 그들에 상보적이라고 부른다. 예를 들어 서열 5'-AACGCGTT-3'은 상보적인 서열 3'-TTGCGCAA-3'이기 때문에 palindrome이며 세포에 의해 5'-AACGCGTT-3'라고 읽는다.

주의 1.3

패턴 일치 대 패턴 발견. 컴퓨터 과학에서 패턴 일치와 패턴 발견의 작업들을 구분하는 것은 관습적이라는 것을 인식하라. 첫번째 경우에 하나는 주어진 특이적인 단위(또는 다른 패턴)이며 서열에서 그들의 발생을 찾는 것이 요청된다. 두번째 경우는 흥미로운 특정 특징(가장 빈번하며 가장 놀라운 단위)을 가지는 서열에서 패턴을 찾는 것을 필요로 한다. 생명정보학에서 첫번째 경우는 예를 들어 서열의 특정 분류를 특성화하는 새로운 모티프를 찾는 것과는 반대되게 주어진 DNA 모티프의 발생을 찾는데 상응한다. 알고리즘적이며 통계적인 이슈들은 2가지 작업에서 매우 다를 수 있다.

1.5 데이터 찾기 : GenBank, EMBL, DDB

이 책을 통하여 우리는 우리가 새롭게 얻은 지식을 실제 DNA(와 단백질) 서열을 분석하는데 어떻게 적용하는지를 보여줄 것이다. 이것은 우리는 게놈 서열 데이터를 접근하여 다루고 가공랑 수 있어야만 하는 것을 의미한다. 포함된 대부분의 단계는 지금 표준화되어 모든 생명정보학 연구자들의 일반적인 도구의 부분이다.

온라인 데이터베이스. 어떤 분석의 첫번째 단계는 서열을 얻는 것이다. 모든 출판된 게놈 서열은 인터넷을 통하여 이용가능하며 모든 출판되는 DNA 서열은 모든 공용 데이터 베이스에 수록되어야 하는 것이 관련된 과학적인 저널의 요구사항이다. 서열 분포의 주요 수단들은 International Nucleotide Sequence Database Collaboration의 구성원들이다. 이것은 3가지 큰 데이터 베이스의 협회이다: DNA Database of Japan(DDBJ), the European Molecular Biology Laboratories(EMBL), US National Institute of Health에 의해 지원되는 GenBank. 이런 데이터 베이스들은 모든 공용의 이용가능한 DNA 서열 데이터를 모으고 교환하여 공짜로 이용가능하

게 만든다. 2005년 8월에 3개의 데이터베이스에 거의 DNA 서열의 1000억 베이스가 모아져서 저장되었다(이것은 은하수의 별들의 숫자보다 조금 적다).

주요 데이터베이스의 웹 주소는 다음과 같다.

GenBank www.ncbi.nlm.nih.gov
EMBL www.ebi.ac.uk
DDBJ www.ddbj.nig.ac.jp

GenBank에서 서열들은 접근 숫자(accession number)로 정의되며 한 숫자는 단일 서열과 연관된다. 각 서열은 또한 meta-data(데이터에 대한 데이터, 또는 주석) –서열화된 유기체의 종 이름과 같은-의 특정 양은 매우 유용할 수 있다. 3단원에서 우리는 어떻게 accession number 보다 서열 유사성에 의해 몇 데이터베이스를 검색하는지를 볼 것이다. 여기서 accession number의 몇 가지 예와 그들의 참고 사랑이 같이 있다.

AF254446 *Homo sapiens neanderthalensis* mitochondrial D-loop, hypervariable region I
NC_001416 *Bacteriophage lambda*, complete genome
NC_000907 *Haemopilus influenza* Rd KW20, complete genome
NC_000908 *Mycoplasma genitalium* G-37, complete genome
NC_001807 *Homo sapiens* mitochondrion, complete genome

데이터 형식과 주석. 서열과 그것의 주석이 주어질 때는 여러가지 형식들이 있다. EMBL, GenBank, DDBJ와 다른 곳들은 그들의 고유한 표준을 사용하며 데이터베이스와는 연관이 없는 많은 형식들 또한 표준으로 생각된다(그러나 서열 분석 프로그램과는 관련된다). 이것들 중의 하나가 FASTA(fast "A"로 발음하는)로 불린다. 서열에 대한 정보의 매우 제한적인 양에 따라 서열 정보를 요약하는데 공통적으로 사용된다. 형식은 첫번째 줄이 ">" 기호로 표시되며 주석이 따라오며 이것은 줄이 끊어지지 않는다면 얼마든지 길어질 수 있다. 서열 정보는 다음 줄에서 시작하며 모든 서열 정보는 첫번째 글자로서 또다른 "<"기호가 나올때까지의 줄이다.

예 1.8

FASTA 형식의 서열. 이것은 FASTA 형식의 작은 DNA 서열이다. 이것은 러시아에서 발견된 네안데르탈의 잔여물에 속한다. 우리는 이것을 5단원에서 인간 기원을 공부하는데 사용할 것이다. 한 서열 이상이 표시될 필요가 있을 때 그들은 간단히 다른 것 아래에 같은 형식으로 표시된다.

> Homo sapiens neanderthalensis mitochondrial D-loop, HVR I
CCAA ~ TTGA

FASTA 형식은 대부분의 서열 분석 소프트웨어에서 수용되며 대부분의 온라인 서열 데이터베이스에서 제공된다. 그러나 주석의 양은 제한되어 다른 표준들이 더 많은 주석을 전달하기를 원할 때 사용된다.

GenBank 형식은 다양한 단락을 포함하며 그것 중의 주요 하나는 다음과 같다: LOCUS 서열을 확인하는 것으로 서열의 짧은 DEFINITION과 고유한 ACCESSION number이 다음에 나온다. accession number는 안정한 방식으로 서열을 확인하는 것이다. 서열을 다루는 과학적인 출판물들에서 보고되며 다른 데이터베이스에서 교차 참고(cross-reference)로서 사용될 수 있다. SOURCE와 ORGANISM 영역은 서열의 생물적인 기원을 확인한다. REFERENCE 영역은 서열에 관련된 참고 글들을 포함한다. 보통 한 개 이상의 참고물들이 나열된다. FEATURES 단락은 위치에 대한 기본적인 정보와 서열에서 흥미로운 다양한 부분의 정의를 포함한다(조절 서열과 같은). SEQUENCE단락은 주요한 것으로 뉴클레오타이드을 나열한다. 서열은 각 10개의 염기가 편리하게 숫자가 주어져 있다. 기호 //는 형식의 끝을 의미한다.

예 1.9

GenBank 형식의 서열. 짧은 네안데르탈 서열은 GenBank 형식에서 이것과 같은 FASTA 형식으로 보여진다.

주의 1.4

표준 뉴클레오타이드 알파벳. 모든 DNA 사이트에서 찾아지는 서열들은 표준 뉴클레오타이드 알파벳으로 쓰여진다. 부정확한 뉴클레오타이드에 대한 기호들 또한 존재한다.(서열화 부정확성으로 인해 염기서열에서 한 염기나 다른 것으로 확실하지 않은 경우) 가장 일반적인 기호들은 다음과 같다.

A Adenine N aNy base

C	Cytosine	R	A or G (puRine)
G	Guanine	Y	C ot T (pYrimidine)
T	Thymine	M	A or C (aMino)

1.6 연습문제

(1) GenBank로부터 Bacteriophage lambda, accession number NC_001416 의 완전한 게놈 서열을 다운로드하여 GC 함량을 다양한 창 크기로 선택하여 분석하여라.

(2)인간과 침팬지의 미토콘드리아 DNA의 통계적인 특징을 비교하여라.(각각 NC_001807 과 NC_001643)

(3) 쥐의 미토콘드리아 DNA에서 비이상적인 2분자체를 찾아라.(NC_001665)

1.7 읽을 목록

DNA의 분자 구조는 왓슨과 크릭에의해서 설명되었으며 1953년에 출판되었다.(Watson and Crick, 1953). 게놈 서열의 첫번째 물결은 1980년 초반이었으며 작은 파지와 바이러스들을 포함했으며 미토콘드리아 게놈(Anderson *et al.,* 1981; Bibb *et al.*, 1981)을 포함한 대부분이 프레드 생거의 방법에 기초하였으며(Sanger *et al.,* 1982), (Sanger *et al.*, 1978).

게놈의 두번째 물결은 원핵생물과 진핵생물을 포함하며 첫번째 독립미생물이 서열화되었으며 1990년대 중반에 시작되었다. 논문(Fleischmann *et al.*, 1995)은 *H. influenzae*의 서열화와 게놈의 기본 통계적인 분석을 보고한다. 그리고 이 과정으로 추천되는 읽을거리는 *M.genitilium*의 서열을 발표하는 논문(Fraser *et al.,* 1995)이다. 이 단원에서 논의되는 게놈의 특징들은 위의 논문들에 나와있다. Blattner et al., (1997)은 *E.coli*를, Goffeau *et al.,* (1996)은 첫번째 진핵생물인 곰팡이 *S.cerevisiae*의 게놈의 분석과 서열화를 보고한다.

게놈 서열의 세번째 물결은 1998년에 시작하여 다세포 유기체를 포함한다. *C.elegans*의 완전한 서열이 처음으로 완성되었으며 다음으로 인간의 게놈이 두 개의 경쟁하는 그룹에서 동시에 Science와 Nature(Consortium, 2001; Venter, 2001)에 출판되었다. 마우스, 쥐, 닭, 개, 소, 침팬지의 게놈도 지금 증가하는 속도에 있다.

게놈 서열의 통계적인 특성의 일반적인 논의는 Karlin *et al.*(1998)에서 이 단원에서 나타나는 개념을 포함하여 찾을 수 있다. GenBank의 설명은 글 Benson et al. (2004)에서 찾을 수 있다. 이 단원을 이해하기 위해 필요한 생물학적인 사실들의 논의는 평행적 유전자 이동과 DNA 구조, 일반적인 세포 생물학을 포함하는 Brown(1999)와 Gibson과 Muse(2004)에서 찾을 수 있다.

이것과 연결되어 더 많은 논문들과 모든 예들과 연습문제에 대한 소프트웨어와 데이터는 이 책의 웹사이트에서 찾을 수 있다.

www.computatioal-genomics.net