생물정보학(Bioinformatics)의 이해

서문

정보통신이 20세기 말을 장식하였다면 21세기를 주도할 과학과 산업의 화두는 유전체(Genome)일 것이라고 말해도 이의를 제기할 사람은 없을 것이다. 불과 10여년 전만 해도 불가능하리라 여겨졌던 유전체 산업이 이렇게 빨리 성공하게 된 가장 큰 이유 중의 하나는 생물정보학(Bioinformatics)의 발전이라 할 수 있다.

생물정보학(Bioinformatics)은 유전체(Genome)에 대한 총체적인 연구를 하는 학문으로, 유전체(Genome)를 문자로 나타낸 서열(sequence)에 대한 연구뿐만 아니라, 단백체를 연구하는 프로테오믹스(proteomics), 단백체의 2D, 3D 구조 연구, 가장 핵심적인 기능 유전체학(functional genomics), 계통 연구, 정상 세포와 병든 세포의 비교 연구 등 많은 다양한 분야들을 포함하는 학문이다. 생물정보학(Bioinformatics)는 90년대 후반부에 들어오면서 인간 유전체 사업(Human Genome Project)의 활성화와 DNA 칩(chip)등 제반 기술들의 발달로 인해 더욱더 풍부한 데이터들을 활용할 수 있게 되었고, 이러한 데이터들로부터 유용한 정보를 캐내고자 하는 여러 시도들이 행해져 왔으며, 지금도 행해지고, 앞으로 더 발전된 시도들이 행하여질 것이다.

생물정보학은 유용한 유전자 및 신약의 개발 분야에서 효율성을 증가시키고 비용을 절감해주는 역할을 할 뿐 아니라 앞으로는 이 분야 자체가 새로운 과학과 산업의 중추적 역할을 맡게 될 것이며 이로부터 많은 산업이 파생될 것이다. 예를 들어 DNA 칩을 이용한 진단 정보가 인터넷을 통해 데이터베이스에 통합, 축척, 분석되어 이를 바탕으로 예방, 진단, 개인별 맞춤 의학과 같은 새로운 의료 산업이 부각될 수 있다.

이 책은 생물정보학의 전부를 알려주기 위한 책이 아니다. 단지 생물정보학의 기본 개념을 이해하고 생물정보학에서 나오는 수학적 문제들을 소개하고

그 중의 일부를 수학적인 관점에서 접근하는 데 역점을 두었다. 다시 말해서 생물정보학을 처음 접하는 사람들을 위한 입문서인 셈이다.

1절에서는 생물정보학의 이해에 필요한 DNA, RNA, 아미노산, 단백질 등과 같은 생물학적 요소를 간단히 설명하고 이러한 생물정보학의 연구에서 대두되는 몇 가지 문제들을 소개한다.

2절과 3절에서는 1절에서 소개한 문제 중 서열 정렬(Sequence Alignment)문제를 집중적으로 다룬다. 서열 정렬 문제와 이를 해결하기 위한 알고리즘 그리고 여러가지 소프트웨어들을 설명할 것이다.

4절에서는 생물정보학의 현재와 미래에 대하여 이야기할 것이다.

이 책의 내용은 주로 Ron Shamir의 강의 노트(Algorithm for Molecular Biology[1])를 참고하여 쓰여졌다. 또, 이 책에 나오는 그림들은 T.A. Brown의 *Genomics*와 Karl Drlica의 *Understanding DNA and gene cloning, a guide for the curious*라는 책에서 발췌하였다.

필자의 바램이 있다면 단지 이 책이 처음 생물정보학을 접하는 사람들의 학습에 조금이나마 보탬이 되었으면 하는 것 뿐이다.

2001년 11월

최형인, 기호삼, 강성수

[1] 이 강의 노트는 웹사이트(http://www.math.tau.ac.il/~rshamir)에서 얻을 수 있다.

1 생물학정보학(Bioinformatics)의 개요

1장에서는 먼저 생물학의 기본지식에 대하여 알아보자. 공부할 주요 내용은 DNA(Deoxy-Ribonucleic acid), RNA(Ribonucleic acid), 유전자(Gene), 염색체(Chromosome), 그리고 DNA로부터 단백질(Protein)으로의 유전 정보의 전달에 관한 것 등이다. 그리고 나서 생물정보학에서 대두되는 여러가지 문제들과 이 문제들을 해결하는 방법들을 개괄적으로 살펴보자.

1.1 유전법칙

유전법칙은 1866년에 폴란드 브룬의 수도원 사제 그레고르 멘델(Gregor Mendel)에 의해 발견되었다. 그는 8년동안 3만여 그루의 다양한 식물을 재배하면서 얻은 결과를 브룬의 자연과학회지에 기고하였다. 그는 생명체가 소유하고 또 후손에게 전달하는 유전적 기본 단위를 유전자(Gene)라고 정의하였다. 그러나 그의 연구 성과를 알아주는 사람은 아무도 없었다. 멘델의 유전법칙 발견으로부터 DNA의 생물학적 역할이 밝혀지기까지는 75년도 더 걸렸다. DNA가 생물체의 유전물질의 주 매개체로 알려지게 된 것이다.

1.2 DNA

DNA의 구조는 Watson과 Crick가 1953년에 발견하였다. 이 때를 분자생물학의 시작이라고 한다. DNA는 뉴클레오티드(Nucleotide)라고 불리우는 기본 단위로 이루어진 커다란 분자이다. 뉴클레오티드는 인산(Phosphate), 당(Sugar), 염기(Base)로 구성되어 있다. 염기는 4가지로 구분되는데 그 이름은 아데닌(Adenine), 구아닌(Guanine), 시토신(Cytosine), 티민(Thymine)이라 불리고 각각 A, G, T, C 라고 표기한다. DNA는 이중 나선 구조(Double helix)로 이루어져 있다. 각 나선은 인산 결합으로 단단히 묶인 뉴클레오티드의 고분자체이고, 두 나선은 수소 결합으로 연결되어 있다. 이 수소 결합은 염기의 쌍으로 이루어진다. 염기에는 퓨린(Purine)계 염기와 피리미딘(Pyrimidine)계 염기가 있는데, 퓨린계 염기에는 A와 G, 피리미딘계 염기에는 T와 C가 있다. A는 G와, T는 C와 수소 결합을 한다. A와 T는 두 쌍의 수소 결합으로 연결되고, T와 C는 세 쌍의 수소 결합으로 연결된다. DNA분자는 방향성을 가지고 있다. 이는 분자의 핵심이 되는 당(Sugar)이 비대칭적이기 때문이다.

사람의 DNA는 총 3.2×10^9 bp(base pair: 염기쌍)로 되어 있다. DNA의 총 염기쌍의 개수는 생물체마다 다르다. 예를 들어 단세포 생물인 아메바는 사람보다 200배나 더 많이 가지고 있다.

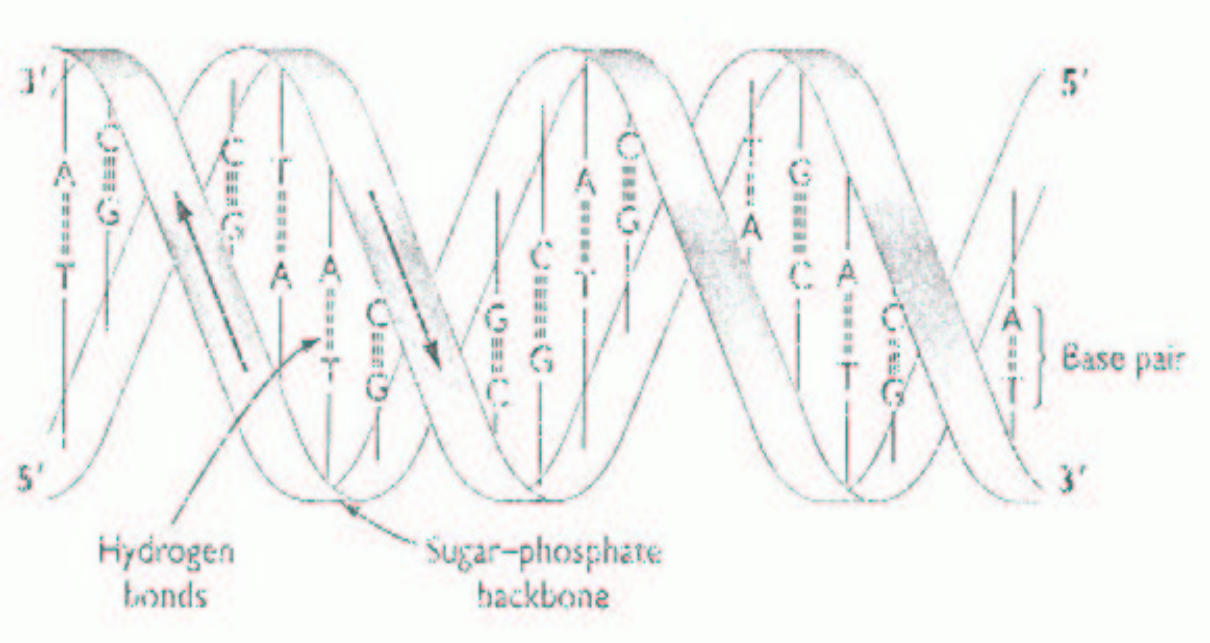

Figure 1: DNA 나선 구조

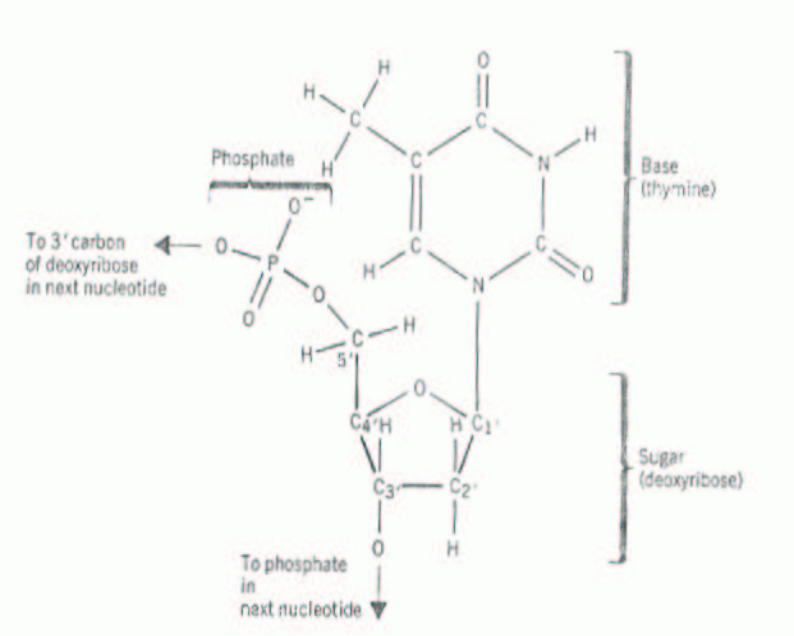

Figure 2: 뉴클레오티드의 구조

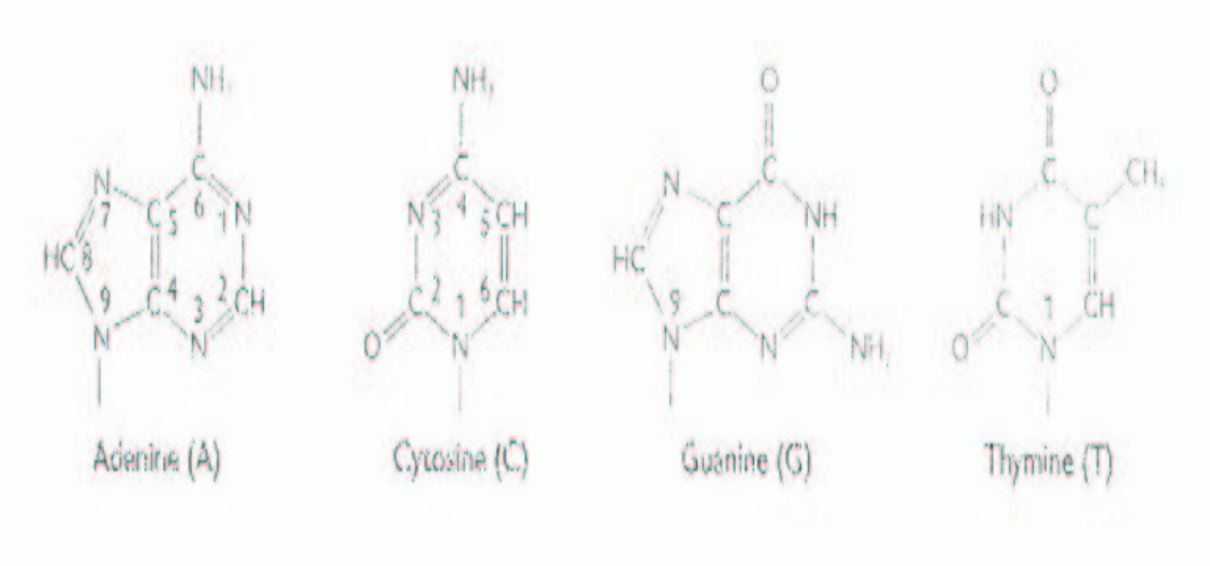

Figure 3: 염기의 종류

1.3 DNA 복제

DNA에 각인된 생물정보는 더 많은 DNA로 복제될 수 있다. DNA가 복제될 때 두 나선은 분리되어 각각이 나선은 새로 생기는 나선을 위한 주형(Template)으로 쓰인다. 주형의 각각의 염기는 새로 생기는 나선의 염기를 자신과 결합할 수 있는 것으로 만들어지도록 한다. DNA 복제는 DNA 폴리메라제(Polymerase)라는 효소(Enzyme)에 의해서 촉진된다.

1.4 유전자와 염색체

모든 생물은 두 가지 그룹으로 나누어진다. 하나는 프로케리오트(Prokaryote)이라는 핵이 없는 단세포 생물이고 다른 하나는 유케리오트(Eukaryote)라는 상위의 유기체인데 그것의 세포에는 핵이 있다.

유케리오트의 핵 안에는 염색체가 있고 염색체 속에 DNA가 저장되어 있다. 모든 세포는 쌍으로 된 염색체를 가지고 있고 각 쌍은 똑같은 정보를 가지고 있다(예외적으로 수컷은 X, Y 성염색체를 한 쌍 가지고 있다). 생식 세포의 경우는 각 염색체를 한 벌씩만 가지고 있다.

염색체의 수는 종(Species)마다 다르다. 사람은 22쌍의 염색체에 추가로 한 쌍의 성염색체를 갖는다. 사람의 염색체의 크기는 대략 3×10^7에서 3×10^8이다.

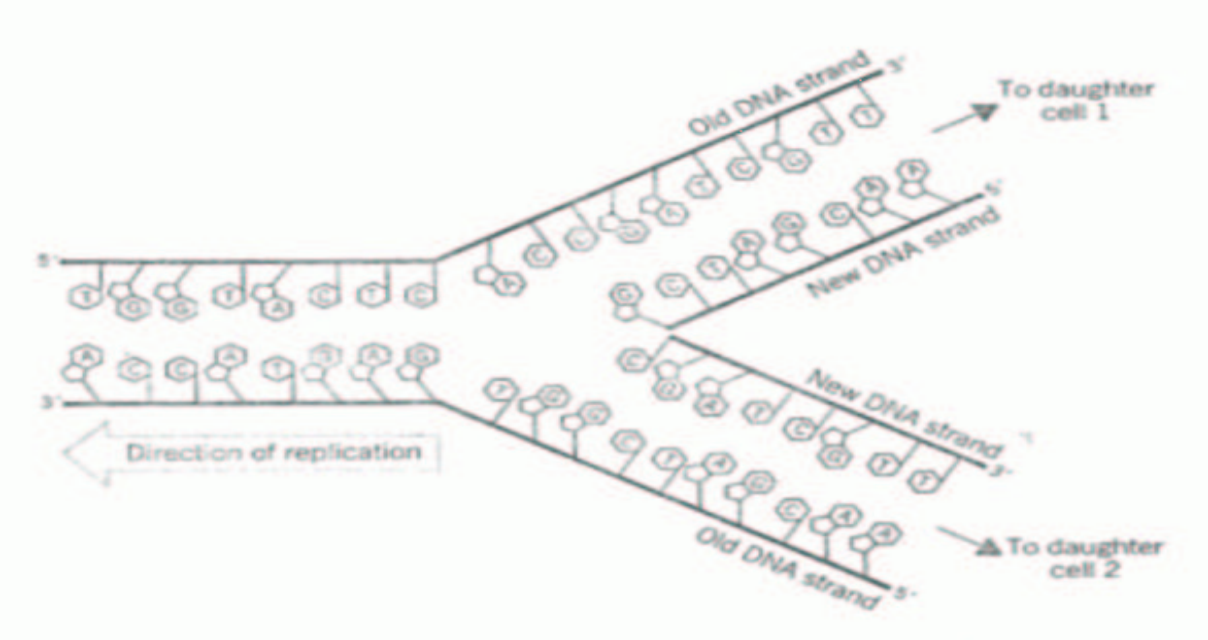

Figure 4: DNA 복제

유전자(Gene)는 염색체 속에 있는 뉴클레오티드 염기의 특정 배열로서 단백질(Protein)을 구성하는 정보를 갖고 있는 부분이다. 다시말해서 유전자는 염색체의 일부분이다. 염색체를 구성하는 DNA의 $5-10\%$가 유전자이고 나머지 $90-95\%$는 유전자가 아니다(이 부분을 junk-DNA라 한다). 그러나 이 부분이 매우 중요한 부분이라고 추측되고 있지만 아직까지는 그 역할을 잘 알지 못하고 있다. 사람의 유전자의 총 개수는 대략 3만에서 8만개로 추정된다.

1.5 RNA와 복사

세포는 RNA(Ribonucleic acid)라는 핵산(Nucleic acid)를 가지고 있는 데, 이것은 유전정보를 운반하는 역할을 한다. DNA가 주로 핵 속에 있는 것과는 달리 RNA는 핵 바깥의 세포질(Cytoplasm) 속에서도 발견된다. DNA처럼 RNA도 퓨린계와 피리미딘계의 염기를 갖는 데 티민(T) 대신 우라실(U: Uracil)을 갖는다. RNA는 이중나선구조가 아니고 한 가닥으로 되어 있는 것이 DNA와 다른 점이다.

RNA에는 mRNA(messenger RNA)와 tRNA(transfer RNA)가 있다. mRNA는 유전정보를 핵 속의 DNA에서 리보솜(Ribosome)으로 운반하는 역할을 한다. 리보솜은 세포속에서 RNA를 이용하여 단백질(Protein)을 만드는 역할을 한다. mRNA는 DNA 한 가닥에서 유전자(Gene)가 있는 곳을 복사(Transcription)한다. 우선 DNA의 이중나선이 갈라지

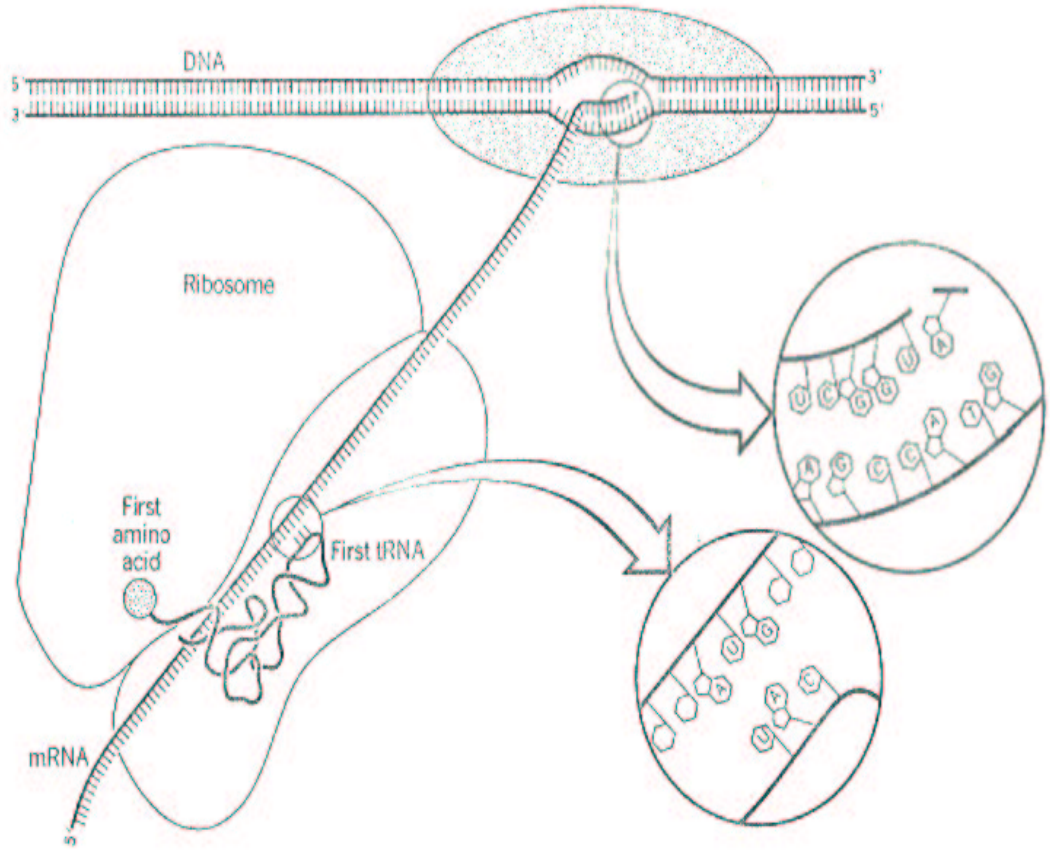

Figure 5: 단백질 합성 과정: (a) DNA가 분리되면서 mRNA가 염기 서열을 복사(Transcription)한다 (b) mRNA에서 인트론을 떼어내고 코돈을 만든다(Splicing) (c) 리보솜에서 tRNA가 mRNA의 코돈에 대응되는 anti-코돈을 만든다 (d) anti-코돈으로부터 아미노산을 만든다(Translation) (e) tRNA는 아미노산을 폴리펩타이드에 연결한다

면서 한 쪽 가닥에서 복사를 시작한다. 복사는 반대쪽 염기를 만들면서 이루어지기 때문에 복사 후에는 자기와 같은 염기가 아니라 복사가 이루어지지 않은 쪽과 같은 염기 배열을 이루게 된다. 대부분의 유전자(Gene) 주변에는 프로모터(Promoter)라고 불리우는 특별한 염기 배열이 있다. 프로모터는 RNA 폴리메라제(Polymerase)라는 효소(Enzyme)에게 복사가 시작되는 곳을 알려주는 역할을 한다. 복사가 끝나는 지점에는 수 십 개의 아데닌(A)이 mRNA 끝에 덧붙여진다. mRNA는 엑손(Exon)과 인트론(Intron)으로 구성된다. 핵 안에서 일어나는 접착(Splicing)이라는 과정에서 mRNA로부터 인트론이 떨어져 나간다. 인트론이 떨어져 나간 후에 엑손을 이어 붙여 최종적으로 성숙된 mRNA 분자가 만들어진다. 평균적으로 유케리오트의 엑손은 200bp 정도 되고 인트론은 10000bp 쯤 된다. 많은 경우에 접착(Splicing)의 패턴은 복사가 이루어지는 조직에 따라 달라질 수 있다. 예를 들어, 간(Liver)에서 복사된 어떤 유전자의 mRNA로부터 떨어져 나간 인트론은 뇌(Brain)에서 복사될 때 같은 mRNA로부터 떨어져 나가지 않을 수도 있다. 이러한 변이(Variation)을 선택적 접착(Alternative splicing)이라고 하는 데, 이것은 조직에서의 전체적인 단백질 변화의 원인이 된다. 한 DNA 나선으로부터 합성되고 수정된(edited) 후에 mRNA 분자는 핵에서 나와 리보솜을 향해간다. 그 곳에서 단백질로 전이(Translation)된다.

1.6 변이

변이(Mutation)는 부정확한 복제에 의해 DNA 정보가 부분적으로 바뀌는 것이다. 이러한 복제 오류는 방사선이나 다른 주위 환경에 의해 일어날 수 있다. 변이는 여러가지 종류가 있다. 옳은 염기 대신에 잘못된 염기가 병합될 때 치환(Substitution)이 일어난다. 치환이 발생하는 위치에 따라 단백질 서열(Sequence)이 바뀔 수도 바뀌지 않을 수도 있다. 왜냐하면 바뀐 치환이 발생하는 위치가 인트론에 속하면 이 부분은 접착(Splicing) 과정에서 떨어져 나가기 때문이다. 삽입(Insertion)과 삭제(Deletion)는 하나 이상의 염기를 더하거나 제거하는 것이다. 또 재정렬(Rearrangement)은 염색체의 전 영역에 걸쳐 염기 순서를 바꾸는 것이다.

변이는 여러가지 이유로 중요하다. 유전병 또는 암과 같은 다른 질병을 일으킬 수 있는 데 이는 유전자 변질의 원인이 되기도 한다. 동시에 변이는 자연 선택(Natural Selection)이 작용하는 표현형(Phenotype: 유전자의 작용과 환경에 의해 외부에 나타나는 성질) 변이와 종(Species) 창조 그리고 종의 변화의 근원이다. 예를 들어, 사람과 쥐의 유전체(Genome)은 매우 비슷하다. 그들 사이의 주요한 차이점은 DNA의 내부적인 순서가 다르다는 것이다.

변이로 생긴 DNA 서열(Sequence)의 차이가 있을 때, 이 변이가 어떤 것인가를 알아내는 것이 서열정렬문제(Sequence Alignment Problem)이다.

문제 1.1 서열 정렬 문제(Sequence Alignment Problem) 뉴클레오타이드(또는 아미노산)의 일치와 불일치에 대한 점수(score)들이 주어질 때, 두 DNA(또는 단백질) 서열에 대하여 적당한 공백(Opening gap)을 포함하여 그 점수(score)의 합이 최대가 되는 맞춤(Match)을 찾아라.

예를 들어, 뉴클레오타이드가 일치하면 점수(score)가 2, 불일치하면 점수(score)가 -1라고 주어졌다고 하자. 그리고 두 DNA 서열 AAGGTAATTCCC와 AAGGGGTTCTCCC가 있다고 하자.

1. 두 서열의 맞춤(Match)을
   A A G G T A A T - T C C C
   A A G G G G T T C T C C C
   라 하자. 위 맞춤에서 −는 공백을 나타낸다. 뉴클레오타이드가 일치하는 위치는 1, 2, 3, 4, 8, 10, 11, 12, 13이고 나머지 5, 6, 7, 9는 불일치한다. 따라서 점수(score)의 합은 $9 \times 2 + 4 \times (-1) = 14$이다.

2. 두 서열의 맞춤(Match)을
   A A G G - - T A A T - T C C C
   A A G G G G T - - T C T C C C
   라 하면 뉴클레오타이드가 일치하는 위치는 1, 2, 3, 4, 7, 10, 12, 13, 14, 15이고 나머지 5, 6, 8, 9, 11은 불일치한다. 따라서 점수(score)의 합은 $10 \times 2 + 5 \times (-1) = 15$이다.

위의 예에서 두 번째 맞춤의 점수(score)의 합이 최대인지에 대해서는 나중에 자세히 배우기로 하자.

1.7 유전체 재정렬

재정렬(Rearrangement)은 한 염기의 치환, 삽입, 삭제보다 훨씬 드물게 일어난다. 예를 들어, 어떤 조직에서 치환은 한 세대에 10번 정도 일어나지만 치명적이지 않은 재정렬은 500만년에서 1000만년에 한 번 정도 일어난다. 특히 사람과 쥐는 8000만년 전부터 달라지기 시작했는 데 그 동안 140-150번 정도의 재정렬이 발생했을 뿐이다. 또, 한 번 재정렬이 일어난 곳에서 다시 일어날 가능성이 희박하기 때문에 변이의 해석이 애매모호하

지 않다. 따라서, 발생하는 재정렬들의 종류와 그들의 발생 순서를 알면 진화 과정을 더 잘 이해할 수 있다.

문제 1.2 유전체 재정렬 문제(Genome Rearrangement Problem) 한 유전 정보의 주어진 두 순열(Permutation)에 대하여 한 순열로부터 다른 순열로 변환하는 데 필요한 최소 수의 작용(Operation)들을 구하여라.

DNA 서열에서 일어나는 작용(Operation)들은 다음과 같다.

- 삭제(Deletion): $abc \to ac$
- 삽입(Insertion): $ac \to abc$
- 중복(Duplication): $abc \to abbc$, not $abc \to abcb$
- 역(Reversal): $abc_1c_2 \cdots c_nd \to abc_n \cdots c_2c_1d$

예를 들어 서열 AGT를 GTA로 변환하기 위해서는 다음과 같은 과정을 거치면 된다.

1. AGT → ATG → GTA (역 2번)
2. AGT → GAT → GTA (역 2번)
3. AGT → GT → GTA (삭제와 삽입)

이 외에도 여러가지 방법이 더 있을 수도 있다. 그렇다면 한 번의 작용(Operation)만으로 변환할 수 있을까? 결론은 없다. 따라서, 작용의 최소 수는 2이다.

연습 1.3 DNA 서열 AAGTC와 그 서열의 한 순열인 AGTCA가 있다고 하자. AAGTC에서 AGTCA로 변환하는 데 필요한 최소 수의 작용(Operation)들을 각자 구해보자.

연습 1.4 작용(Operation)이 역(Reversal) 뿐이라고 할 때, 다음 순열 12345이 41523으로 바뀌는 데 필요한 작용(Operation)들을 순서대로 나열해 보고, 필요한 작용들의 최소 수를 추측해보자.

1.8 mRNA의 전이

연속된 mRNA 염기 세 개를 코돈(Codon)이라 부르는 데, 이것은 전이(Translation) 과정에서 아미노산(Amino acid)을 결정한다. 뉴클레오타이드가 DNA와 RNA분자의 단량체(Monomer↔Polymer)인 것처럼 아미노산은 단백질의 단량체이다.

유케리오트에서는 mRNA는 Non-Coding 영역(Coding 영역이 아닌 부분)이 옆에 붙어있는 Coding 영역으로 구성된다. 유전자의 DNA에서 Coding 영역은 단백질 생성을 위해 사용되는 엑손 또는 엑손의 일부분이다.

RNA 분자로부터 단백질을 합성하는 과정은 리보솜이라는 세포구조에서 이루어진다. mRNA가 리보솜을 만나면 리보솜이 mRNA를 따라 가면서 코돈을 특별한 문자로 해독해낸다. 이 문자가 바로 20개의 아미노산이다. mRNA의 코돈이 특정 아미노산으로 해독되면 이에 해당하는 아미노산이 tRNA에 의해 운반되어 순서대로 연결된다. 즉 tRNA는 mRNA와 아미노산 사이의 어뎁터(Adapter) 역할을 한다. 따라서 이 tRNA는 두 부분으로 구성된다. 한 쪽은 세 개의 RNA 염기로 구성된 anti-코돈[2]들을 붙들고 있고 다른 쪽은 아미노산의 사슬을 붙들고 있다. tRNA는 anti-코돈에 해당되는 아미노산을 리보솜으로 가져와서 폴리펩타이드(Polypeptide)라고 불리는 아미노산의 사슬에 연결한다. 리보솜은 mRNA를 따라 움직이면서 이 과정을 계속해 나간다. 여러개의 리보솜이 한 mRNA 분자에서 동시에 작업할 수도 있다.

1.9 유전암호

아미노산을 생성된 아미노산이 얽혀서 삼차원 입체구조가 이루어지는 데 이것이 단백질이다. 코돈의 종류는 $4^3 = 64$가지이므로 아미노산 20개에 대응되고도 남는다. 따라서 코돈에서 아미노산으로의 대응은 다대일 대응이다. 이 대응관계를 유전암호(Universal genetic code)라고 한다.

따라서 유전암호는 mRNA에 저장된 유전정보로부터 단백질 서열(Protein sequence)을 결정하는 논리적 대응관계라고 말할 수 있다.

유전암호는 지구상에 생존하는 거의 모든 생물체에 공통으로 적용된다는 사실은 왓슨과 크리크에 의해서 발견되었다. Table 1는 mRNA의 코돈과 아미노산의 관계를 나타낸다.

Table 1를 살펴보면 UAA, UAG, UGA와 같은 세 개의 코돈은 단백질 합성을 종료하는 정지코돈(stop codon)으로 작용한다. 마치 문장의 마침표(.)와 같은 역할을 하는 코돈이다. 다대일 대응이므로 예를 들어 GGU, GGC, GGA, GGG는 모두 글리신이라는 같은

[2]코돈은 mRNA에 붙어있는 3개의 염기를 말하고, anti-코돈은 tRNA에 붙어있는 3개의 염기를 말한다.

1st position	2nd position				3rd position
	U	C	A	G	
U	페닐알라닌	세린	티로신	시스테인	U
	페닐알라닌	세린	티로신	시스테인	C
	류신	세린	(종료)	(종료)	A
	류신	세린	(종료)	트립토판	G
C	류신	프롤린	히스티딘	아르기닌	U
	류신	프롤린	히스티딘	아르기닌	C
	류신	프롤린	글루타민	아르기닌	A
	류신	프롤린	글루타민	아르기닌	G
A	이소류신	트레오닌	아스파라진	세린	U
	이소류신	트레오닌	아스파라진	세린	C
	이소류신	트레오닌	리신	아르기닌	A
	메티오닌(시작)	트레오닌	리신	아르기닌	G
G	발린	알라닌	아스파르트산	글리신	U
	발린	알라닌	아스파르트산	글리신	C
	발린	알라닌	글루탐산	글리신	A
	발린	알라닌	글루탐산	글리신	G

Table 1: 유전암호(Genetic Code)

아미노산으로 번역된다. 또 한가지 특이한 코돈 AUG가 있는데 이것은 메티오닌이라는 아미노산으로 번역되는 동시에 단백질 합성의 시작신호(start signal)가 된다. 그러므로 리보솜은 mRNA 상의 AUG라는 코돈을 찾아 단백질 합성을 시작하게 된다.

문제 1.5 유전자 탐색 문제(Gene Finding Problem) 주어진 DNA 서열에 대하여 ORF[3] (Open Reading Frame), 엑손, 인트론의 위치를 예측하여라.

가장 쉬운 방법은 서열을 따라 가면서 정지 코돈을 찾는 것이다. 만일 여러 개의 정지 코돈이 한 영역에서 서로 가깝게 위치해 있으면 그것은 종료를 나타내는 부분이므로 Coding 영역이 될 수 없다.비교적 긴 서열이 정지 코돈을 포함하고 있지 않으면 그것은 아마도 Coding 영역을 포함하고 있을 것이다.

유케리오트의 DNA에서 이 문제는 엑손과 인트론이 서로 끼워진 모양으로 존재하기 때문에 더욱 더 복잡하다. 이 경우에 정지 코돈은 그 서열이 유전자(Gene) 안에 있지 않다는 것을 알려주지 못한다. 단지 엑손 안에 있지 않다는 것을 알려 줄 뿐이다. 게다가 더욱 복잡하게 하는 것은 어떤 DNA 서열은 6가지의 다른 방법으로 해석될 수 있다는 사실이다(서열에서 시작 점을 잡는 방법이 3가지이고 방향도 2가지이다). 예를 들어 ACCUUAGCGUA는 RF(Reading frame)를 다음과 같이 세가지 경우로 생각할 수 있다.
(ACC)(UUA)(GCG)UA, A(CCU)(UAG)(CGU)A, AC(CUU)(AGC)(GUA)
또, 방향이 ACCUUAGCGUA일 수도 있고 AUGCGAUUCCA일 수도 있다.

2 DNA Chip

2.1 이종혼합(Hybridization)

앞에서 살펴보았듯이 보통의 상태에서 DNA 분자들은 두 나선이 수소 결합으로 묶인 이중 나선 구조로 이루어져있다.

그런데, DNA가 들어있는 용액을 데우면 수소 결합이 깨져서 두 나선은 분리된다. 이렇게 한 나선으로된 DNA를 변성(denatured) DNA라고 한다. 다시 이 용액을 식히면 두 나선의 서로 맞는 염기끼리 다시 수소 결합을 할려고 한다. 만일 서로 맞는 쌍이 존재하면 재빨리 수소 결합을 해서 다시 이중 나선 구조를 이루지만 항상 그것을 보장할 수는 없다. 경우에 따라서는 다른 DNA 분자들의 나선 사이에 수소 결합이 일어날 수도 있고, 또 길이가 다른 나선 사이에 수소 결합이 일어날 수도 있다.

[3]염기 서열을 3개씩 묶는 방법을 말한다. 예를 들어 ACCUUA가 있다면 (ACC)(UUA) 또는 A)(CCU)(UA 또는 AC)(CUU)(A 중의 하나이다. 이는 3개의 염기가 하나의 아미노산을 만들기 때문이다.

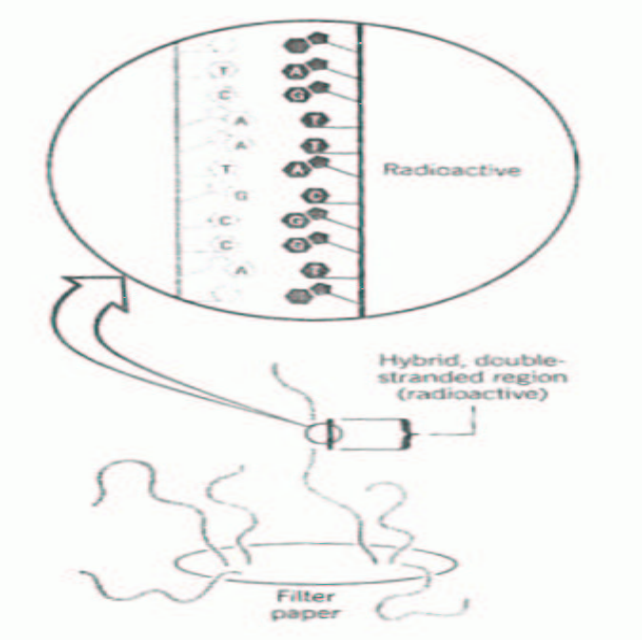

Figure 6: 이종혼합

데워진 타겟(target) DNA 분자를 생각하자. 타겟 DNA의 한 나선에 맞는 올리고뉴클레오타이드(oligonucleotide: 간단히 올리고라 하자)라 불리우는 짧은(10-12개의 염기를 가진) 한 나선을 타겟이 들어있는 용액에 넣고 이 용액을 식히면 올리고는 타겟 DNA의 맞는 부분에 착 달라붙는다. 그 결합된 결과를 하이브리드(hybrid) DNA라 한다. 이런 이유로 올리고는 맞는 염기 서열의 존재 여부를 시험하는 데 사용된다.

많은 생물학적 기술은 이 이종혼합에 기초한다. 예를 들어 사람의 DNA와 쥐의 DNA의 Coding 영역에서 얻은 올리고를 생각하자. 그리고 이 두 DNA를 이종혼합하자. 이 두 DNA 사이에 상동(homology: 외형적으로는 다르게 보이지만 본질적으로는 유사한 성질)이 Coding 영역에 존재하기 때문에 사람의 DNA의 Coding 영역이 어디에 있는지를 추측하기 위해 이종혼합을 이용한다. 올리고에 형광 염료나 방사선 물질을 묻혀서 사람의 DNA와 어디에서 이종혼합을 하는지를 확인한다.

이러한 이종혼합 실험을 과거에는 종이 크기의 필터를 사용하여 했는데 최근 10년 전부터는 기술이 발달하여 칩(Chip) 크기로 소형화 되었다. 따라서, 동시에 여러 개를 행렬처럼 늘여 놓고 병렬적으로 실험할 수 있게 되었는데 이를 DNA 칩(Chip) 또는 DNA Array라고 한다. DNA 칩에는 두 가지 종류가 있는데 하나는 타겟을 칩으로 만들고 올리고는 용액에 넣어서 실험하는 것이고 다른 하나는 반대로 올리고를 칩으로 만들고 타겟을 용액에 넣어서 실험하는 것이다.

2.2 Oligo-Fingerprinting

이 DNA 칩은 칩 행렬의 각 원소가 타겟 DNA이다. 이 칩을 이미 알려진 올리고가 들어 있는 용액에 넣으면 올리고와 맞는 DNA 사이에 이종혼합이 일어난다. 이 때, 만약 올리고에 형광 염료나 방사선 물질을 묻혀서 넣으면 칩에서 이종혼합이 일어나는 위치를 알 수 있다. 칩을 데워서 올리고를 떼어낸 다음 다른 올리고를 가지고 실험을 계속할 수 있다. m의 타겟 DNA와 n개의 올리고에 대하여 실험을 한 다음 그 결과를 다음과 이진수 행렬 M을 DNA i와 올리고 j사이에 이종혼합이 일어나면 $M_{ij} = 1$, DNA i와 올리고 j사이에 이종혼합이 일어나지 않으면 $M_{ij} = 0$으로 정의한다. 그러면 이 행렬 M만 보면 타겟 DNA에 대하여 알 수 있다.

이러한 칩의 장점은 많은 타겟 DNA에 대해서 한 번에 실험을 할 수 있다느 것이다. 1989년에 이런 종류의 칩에 대한 6개의 특허가 출원되었고 여전히 법정에서 논란이 되고 있다. 따라서, 많은 사람들은 오늘날 다른 종류의 칩을 사용하고 있다.

2.3 Oligonucleotide Array

이 칩에 대한 기본적인 아이디어는 칩에 타겟 DNA 대신 올리고를 넣는다는 것을 제외하고는 앞에서 설명한 칩과 유사하다. 일반적으로 올리고의 길이는 25이하로 타겟 DNA에 비해서 매우 작기 때문에 앞에서 설명한 칩보다 훨씬 고밀도의 칩을 만들 수 있다. $1 \times 1cm^2$의 칩 1개에 10만개 정도의 올리고는 쉽게 넣을 수 있다.

칩을 타겟 DNA가 들어 있는 용액에 넣으면 타겟 DNA와 맞는 올리고가 반응한다. 이 때, 만약 올리고에 형광 염료나 방사선 물질을 묻혀서 넣으면 칩에서 이종혼합이 일어나는 위치를 알 수 있다. 칩을 데워서 타겟 DNA를 떼어낸 다음 다른 타겟 DNA를 가지고 실험을 계속할 수 있다.

이 칩은 타겟 DNA의 개수가 적고 올리고의 개수가 많은 경우에 유용하다. 또 다른 장점은 누구나 사용할 수 있는 표준 올리고 칩을 대량으로 생산할 수 있어서 비용을 많이 절감할 수 있다는 것이다.

2.4 이종혼합을 이용한 정렬(Sequencing)

보통 올리고 칩은 정렬(Sequencing)을 위해 이용될 수 있다. 길이가 k인 서열을 갖는 올리고를 모두 포함하는 올리고 칩을 생각하자. 이 길이가 k인 서열을 k-mer라고 부른다. 만일 이 올리고 칩을 어떤 타겟 DNA가 들어있는 용액에 넣으면 칩의 올리고 중에서 타겟 DNA에 맞는 것들은 이종혼합을 할 것이다. 그러면 타겟 DNA에 길이가 k인 연속되

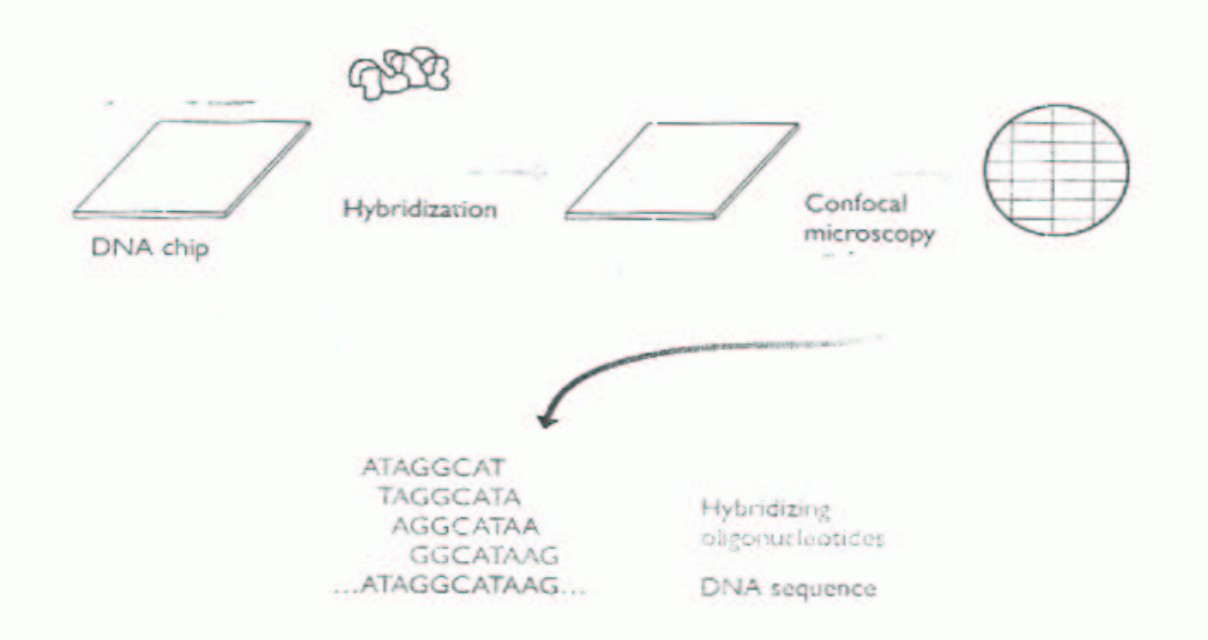

Figure 7: 이종혼합을 이용한 정렬

는 부분서열이 어떤 것들이 있는지를 조사할 수 있다.

정의 2.1 서열 X가 주어질 때, 서열 X의 연속되는 부분서열 중 길이가 k인 모든 서열의 집합을 서열 X의 k-스펙트럼(k-spectrum)이라 한다.

예를 들어, 서열 X가 AAGTTTC라면 서열 X의 3-스펙트럼은 {AAG, AGT, GTT, TTT, TTC }가 된다.

문제 2.2 서열 재구성 문제(Sequence Reconstruction) 길이가 k인 서열들의 집합 S가 있을 때, S를 k-스펙트럼으로 갖는 서열이 존재할까? 존재하면 그 서열을 구하여라.

방향성있는 그래프[4] G=(V,E)를 생각하자. 여기서 V는 S와 같고, V의 두 원소 v_1, v_2에 대하여 v_1의 첫번째 원소를 뺀 길이 k-1의 서열이 v_2의 마지막 원소를 뺀 길이 k-1의 서열과 같을 때, $(v_1, v_2) \in$ E라고 하자. V의 원소를 꼭지점(Vertex), E의 원소를 모서리(Edge)라고 부른다.

[4]꼭지점(vertex)들의 집합 V와 이 꼭지점들 사이의 관계를 나타내는 모서리(Edge)의 집합 E로 이루어진다. 꼭지점 u, v ∈ V에 대하여 u가 v와 어떤 관계를 가질 때, (u,v) ∈ E라고 한다. 방향성이 있다는 것은 (u,v)와 (v,u)가 다른 것으로 간주 된다는 것을 의미한다. 예를 들어 집합 V = { 1,2,3 }에 대하여 관계를 a<일 때 (a,b) ∈ E라고 하면 E = { (1,2), (2,3), (1,3) }이다. 방향성 있는 그래프는 꼭지점을 모두 써 놓고 두 꼭지점 사이에 관계를 화살표로 그려보면 쉽게 이해할 수 있다.

예를 들어, S = { ATG, AGG, CAG, GCA, GGT, GTC, TCC, TGC, CCA, CAG }가 주어졌을 때 E는 다음과 같다.

$$\begin{aligned} E \;=\; & \{(ATG,TGC),\ (AGG,GGT),\ (CAG,AGG),\ (CCA,CAG),\ (GCA,CAG),\\ & (TGC,GCA),\ (GGT,GTC),\ (GTC,TCC),\ (TCC,CCA)\} \end{aligned}$$

이 때, 문제는 이 그래프의 헤밀톤 경로(Hamiltonian path)를 찾는 것으로 바뀐다. 헤밀톤 경로란 모든 모서리를 반드시 단 한번씩만 모서리의 방향에 따라 순차적으로 통과하는 경로를 말한다. 위의 예에서 헤밀톤 경로를 구하면

ATG → TGC → GCA → CAG → AGG → GGT → GTC → TCC → CCA → CAG

이다. 따라서 S를 3-스펙트럼으로 갖는 서열은

ATGCAGGTC

이다. 그러나, 임의의 방향성있는 그래프에서 헤밀톤 경로를 구하는 문제를 세일즈맨 문제(Travelling salesman problem)라고 하는 데, 이는 NP-hard[5] 문제로 알려져 있다. 따라서, S가 충분히 크면 이 방법으로 서열 재구성 문제를 푸는 것은 쉬운 일이 아니다.

연습 2.3 3-mer의 집합 S가

S = { AAA, AAC, ACA, CAC, CAA, ACG, CGC, GCA, ACT, CTT, TTA, TAA}

로 주어질 때, 그래프 G=(S,E)의 E를 구하고 헤밀톤 경로가 있으면 구하고 S를 3-스펙트럼으로 갖는 서열이 있으면 구하여라.

다행히 서열 재구성 문제를 더 빨리 풀 수 있는 방법이 있다.
방향성있는 그래프 G=(V,E)를 생각하자. 여기서 V는 S의 모든 원소들에 대하여 길이가 k-1인 모든 연속되는 부분서열을 모아놓은 집합이다. 그리고, V의 두 원소 v_1, v_2에 대하여 v_1의 첫번째 원소를 뺀 길이 k-2의 서열이 v_2의 마지막 원소를 뺀 길이 k-2의 서열과 같고 v_1에 v_2의 마지막 원소를 붙인 길이 k의 서열이 S의 원소일 때, $(v_1, v_2) \in$ E라고 하자. 이 그래프를 de-Bruijn 그래프라 하고 모든 모서리를 반드시 단 한번씩만 모서리의 방향에 따라 순차적으로 통과하는 경로를 오일러 경로(Euler path)라 한다.

예를 들어,

S = { AAA, AAC, ACA, CAC, CAA, ACG, CGC, GCA, ACT, CTT, TTA, TAA}

라 하자. 그러면 그래프 G=(V,E)의 V = { AA, AC, TA, CA, CG, GC, CT, TT}이고

$$E \;=\; \{(AA,AC),\ (CA,AA),\ (AC,CG),\ (CA,AC),\ (AC,CA),\ (TT,TA),$$

[5]비결정적 튜링머신에 의해 다항식으로 표시되는 시간복잡도로 풀 수 있는 문제보다 본질적으로 더 어려운 문제를 말한다. 자세한 것은 알고리즘 이론에서 P, NP, NP-complete와 같이 찾아보고 비교해보기 바란다.

(GC,CA), (CG,GC), (AC,CT), (CT,TT), (AA,AA), (TA,AA)}

이다. 따라서, 오일러 경로를 구하면

AC → CA → AA → AA → AC → CG → GC → CA → AC → CT → TT → TA → AA

이므로 S를 3-스펙트럼으로 갖는 서열은

ACAAACGCACTTAA

이다.

이 방법은 수학적으로 좋은 방법이지만 몇 가지 문제가 있다. 오일러 경로가 여러 개 있을 수도 있고, 그래프의 모서리가 몇 번 나오는지는 알 수 없다는 것이다.

연습 2.4 3-mer의 집합 S가

S = { AAA, AAC, ACA, CAC, CAA, ACG, CGC, GCA, ACT, CTT, TTA, TAA}

로 주어질 때, S를 3-스펙트럼으로 갖는 서열 중에서 ACAAACGCACTTAA와 다른 것이 있으면 구하여라.

연습 2.5 4-mer의 집합 S가

S = { CTGT, ACTG, TCAG, TGTT, AACT, GTTC, TTCA}

로 주어질 때, S를 4-스펙트럼으로 갖는 서열을 구하여라.

2.5 이종혼합의 다른 용도

이종혼합은 이미 서열이 알려진 DNA를 이용하여 DNA의 변이를 감지하는 목적으로 이용할 수 있다. 예를 들어 HIV의 DNA 서열을 이미 알고 있다고 할 때 이것을 DNA 칩으로 만들어서 다루고자하는 여러가지 바이러스의 형태를 조사하는데 이용할 수 있다. 바이러스 DNA를 포함하는 용액에 이 칩을 넣어서 이종혼합이 일어나는지를 체크하기만 하면 된다. 여러가지 바이러스에 테스트하면 이종혼합은 서로 다른 올리고들에서 나타날 것이다.

또, 이종혼합은 모두 정상이고 한 뉴클레오타이드만 변이된 SNP(Single nucleotide polymorphism)을 감지하는 데 이용할 수 있다.

References

[1] T. A. Brown. *Genomics*, John Wiley & Sons (Asia) Pte Ltd, 1999

[2] Pavel A. Pevzner. *Computational Molecular Biology - An Algorithmic Approach*, the MIT Press, 2000

[3] Ron Shamir. *Algorithm for Molecular Biology*,
Lecture notes http://www.math.tau.ac.il/~rshamir, 2000

[4] Karl Drlica. *Understanding DNA and gene cloning, a guide for the curious*, John Wiley & Sons, 1992