생물정보학(Bioinformatics) I. - 응용분야

바이오칩은 작은 고형 기판 위에 DNA, 단백질 등의 생물분자들을 집적시켜 DNA 결함, 단백질 분포, 반응 양상 등을 분석해낼 수 있는 생물학적 마이크로칩을 말한다. DNA를 집적시켜 놓으면 DNA칩, 단백질을 집적시켜 놓으면 단백질칩이라 명명한다. 바이오칩은 크게 마이크로어레이(microarray)와 마이크로플루이틱스(microfluidics)칩으로 나눌 수 있다. 마이크로어레이는 수천 혹은 수만 개 이상의 DNA나 단백질 등을 일정 간격으로 배열하여 붙이고, 분석 대상 물질을 처리하여 그 결합 양상을 분석할 수 있는 바이오칩을 말하며, 마이크로플루이딕스칩은 미량의 분석 대상 물질을 흘려보내면서 칩에 집적되어 있는 각종 생물분자 혹은 센서와 반응하는 양상을 분석할 수 있는 바이오칩으로 'Lab on a Chip (LOC)'이라고도 불리운다. 생물정보학(bioinformatics)은 생명현상 연구에 필요한 다양한 전산학/통계학/수학적인 것들로 위에서 간단하게 언급한 바이오칩을 이용하여 수행되는 연구분야인 게놈믹스(genomics), 프로테옴믹스(proteomics)을 공통적으로 지원하는 분야로 그 태동은 Frederic Sanger에 의해 단백질 서열결정 방법이 개발된 이후인 1960대부터 시작되었다고 할 수 있으며, 최초로 이를 인식하고 이 분야를 실질적으로 출발을 시킨 사람은 Monte Carlo 방법의 발명으로 유명한 Stanislaw Ulam이란 수학자이다.

생물체는 많은 수의 구성요소들이 복잡한 상호작용을 하는 시스템으로서 생물체를 조작하거나 질병을 치료할 수 있는 방법을 찾아내기 위해서는 이러한 복잡한 생물 시스템에 대해 최대한 관찰을 하고, 그 관찰 결과 얻어진 데이터를 통해서 이 시스템의 성질을 최대한 반영하는 모델을 만들어 가는 과정이 그 기반에 있게 된다. 즉 DNA 칩과 같은 바이오칩을 사용하여 생물체로부터 대량의 데이터를 빠른 속도로 얻어내고, 이 데이터를 통해서 유용한 지식을 얻어내는 것이라 할 수 있다. 이를 위한 모든 과정에서 결국 컴퓨터가 필수적이며 핵심적인 역할을 하게 되며, 바로 이 부분을 구분을 해서 부를 때 사용되는 용어가 생물정보학이다. 생물정보학은 DNA칩과 같은 바이오칩을 이용한 실험결과로부터 생산되는 대량의 데이터를 관리하고 일차적으로 처리할 전산인프라로서 뿐만 아니라 바이오칩으로부터 얻어지는 데이터가 노이즈를 가진 다량의 수치 데이터라는 전형적인 통계학적인 처리대상이기 때문에 이로부터 신뢰할 수 있는 결과를 유추하기 위해서 반드시 필요한 수단이다. 생물정보학의 궁극적인 목표는 바이오칩을 이용한 실험결과로부터 얻어진 데이터와 기존에 알려진 생물체에 대한 다양한 정보를 조합하여, 생명현상에 대한 이해와 실질적인 이용 등 가능한 유용한 지식을 얻어내는 것으로 그 연구대상은 DNA, DNA에 코딩된 정보로부터 만들어지는 단백질의 서열해석, 단백질의 기능 분석, 단백질의 3D 형태를 밝혀 그 기능의 예측, 주어진 단백질에 들어맞는 작은 유기화합물을 디자인 해내는 것, 주어진 두 개의 단백질이 입체적으로 어떻게 상호작용을 하는지를 알아내는 것, 단백질이 이러한 여러 가지 상호작용에 대한 동적인 변화 예측 등 매우 다양하다. 이에 생물정보학의 연구 개발 및 응용 분야에 대해 간단히

살펴보고자 한다.

<게놈프로젝트 및 데이터베이스>

게놈프로젝트에 의해 실험실에서 얻어진 결과물은 생물학 서열 그 자체이다. 즉 4개의 알파벳으로 표현되는 핵산 염기서열이나 20개의 알파벳으로 표현되는 아미노산 서열을 얻는 것이다. 이들 문자 조합의 수가 사람의 손으로 직접 다룰 수 있는 한계를 훨씬 벗어나기 때문에 컴퓨터를 이용하여야 하며, 또한 이들 서열을 효과적으로 다루기 위한 가장 기본적인 단계로서 서열정보 데이터베이스를 구축하는 것이 중요하다. 생물학 서열정보 데이터베이스가 중요한 이유는 서열 자체만으로는 가치 있는 정보의 추출이 어렵기 때문에 이들이 가지고 있는 모든 연관된 정보를 데이터베이스를 통해 구조화함으로써 데이터 클러스터링 및 마이닝을 통해 모든 예측 가능한 정보를 얻을 수 있도록 하기 위함이다. 그림. 1에 DNA 서열의 형성에 대해 나타내었다. DNA 서열을 결정하는 기계로부터 짧은 길이의 염기서열을 수십이나 수천만개 구하고 나면, 위와 같은 여러 가지 프로그램을 사용하여 이들을 연결하는 긴 DNA 서열을 구성하게 된다. 그림. 1에서 위의 그림은 각 단편이 연결되는 순서와 위치를 보여주고 있고, 아래 그림은 각 단편의 염기서열과 이들이 연결되어 형성한 긴 염기서열을 보여주고 있다. 인간의 유전체 서열도, 수천만 개로 임의로 조각내어 이들 각 단편의 서열을 기계로 결정한 후, 그림. 1과 같은 프로그램들로 연결하여 얻어지게 된다.

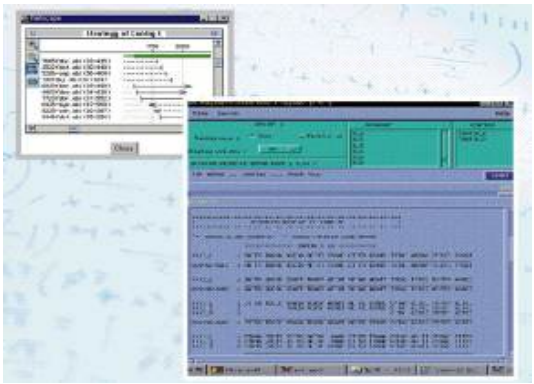
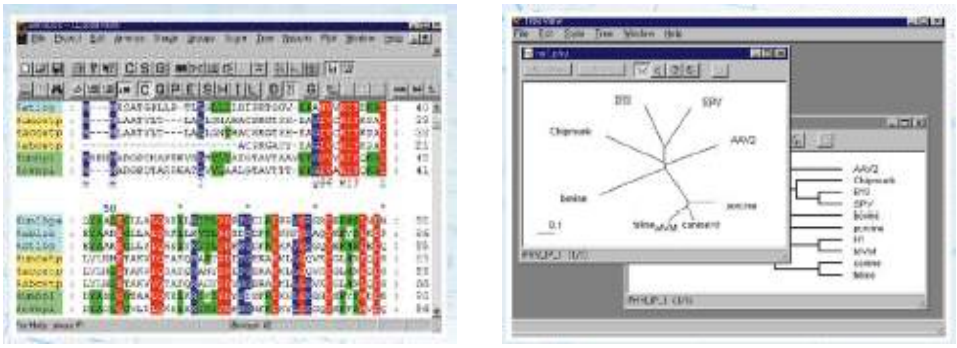


그림.1 DNA 서열의 형성

<서열 분석>

생물정보학은 생물학 데이터(서열정보) 만으로 기존에 알려지지 않은 새로운 현상을 찾는데 그 목적이 있다. 그러나 서열 데이터 자체만으로 가치 있는 특정

의 정보를 유추하기란 쉬운일이 아니기 때문에 다양한 방법으로 서열 분석에 접근하고 있다. 서열 분석을 위해 가장 대표적인 방법이 서열정렬(sequence alignment)이다. 서열정렬이란 단백질서열이나 핵산서열 사이의 상관관계를 나타내는 것으로 가장 생물정보학의 가장 기본이 되는 연구방법 중의 하나이다. '서열간의 관계'는, 서열들이 기능적으로나 진화적으로 어느 정도 연관성이 있고, 서열의 어느 부분들이 그러한 연관성을 가지고 있는가를 나타내는 방법으로 표시될 수 있다. 서열정렬은, 어떤 기준에 의한 계산으로부터 만들어지는데 그 계산 과정이 복잡하고 계산량이 많기 때문에 컴퓨터 프로그램을 사용해야만 한다. 사실상 모든 서열정렬은 프로그램에 의해 이루어지므로, 서열정렬을 하여 서열을 분석하고자 하는 생물학자는 프로그램의 사용법과 해석 방법을 습득해야 한다. 서열정렬은 관심 대상인 서열과 상동성(homology)이 높은 서열들을 알아내어 그 서열의 기능을 유추하거나, 관련있는 서열들간의 정량적인 상관관계(진화적인 연관성과 같은)나 관련 기능 부위 등을 예측하기 위한 목적으로 이용된다.



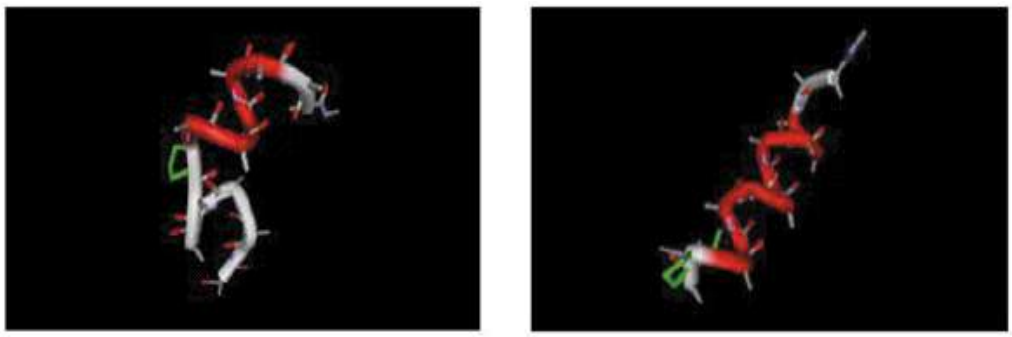
(a) (b)

그림. 2 유전자 분석의 예: (a) 여러 개의 단백질을 서열로 비교하여 기능적으로 관련된 부위와 서로간의 유사한 정도를 분석: (b) 염기서열을 비교하여 진화적인 연관성을 분석

<기능 예측 및 3D 구조 분석>

서열정렬 등과 같은 방법을 통해 그 기능이 밝혀지지 않은 서열 부위의 기능을 유추해 낼 수 있다. 단지 서열 문자정보 자체만을 알고 있을 때 등재된 데이터베이스 등을 이용해 서열 정렬을 실행하고, 비슷하거나 혹은 연관성 있는 특성을 찾아내어 클러스터링을 한 다음 네트워크 상호작용 분석을 통해 그 기능을 유추해 낼 수 있다. 그러나 단백질의 경우는 조금 다르다. 단백질은 1차원적인 서열 분석만으로는 그 기능을 명확히 해석하는데 한계가 있기 때문에 2D구조, 3D구조 분석을 통해야만 어느 정도 예측 할 수 있다. 단백질 아미노산 서열은 서열 유사성이 높더라도

구조적으로 차이가 충분히 날 수 있어서 전혀 다른 단백질로 분류되거나 그 반대로 서열 유사성이 낮더라도 구조적으로 비슷하여 유사한 기능성을 갖는 단백질로 분류되는 경우가 종종 있다. 이는 단백질의 기능이 특정 모티프의 구조 및 특성에 따라 결정되기 때문이다. 생명현상을 직접적으로 조절하는 물질인 단백질의 작용은 주로 효소작용으로 설명되며 가장 대표적인 모델로서 “Key and Rock” 구조로 이해할 수 있다. 이는 단백질의 구조(특정부위)에 의해 그 기능성이 결정되는 것을 명확히 보여준다. 따라서 단백체 연구에 있어서 구조 분석은 매우 중요한 분야이다. 실험실에서 단백질의 3D 구조를 밝혀 내는 가장 대표적인 방법은 X-ray crystalography, NMR 등이 있으며, 이들을 통해 단백질 분자의 3차원 구조를 알아내어 기능을 예측한다. 그러나 컴퓨터 하드웨어 및 소프트웨어의 발달로 단백질의 아미노산 배열로부터 2차 구조를 예측하거나, 컴퓨터 시뮬레이션을 통한 가상의 3차원 분자 모델링을 통해 새로운 단백질을 밝혀내는 예가 급격히 늘고 있다. 그러나 현재 까지 개발된 알고리즘으로는 단백질 서열 정보로부터 실제와 유사한 3차원 구조를 직접 예측하기에는 많은 어려움이 있기 때문에 각각의 아미노산 분자들의 구성과 화학, 물리학적 특성을 이용하여 분석을 하는 방법을 택하고 있다. 단백질 3차원 구조 모델링을 위해서는 컴퓨터 하드웨어 의존도가 높기 때문에 고성능의 컴퓨터 시스템을 갖추는 것도 고려해야 한다.



(a) (b)

그림. 3 proline이 단백질 α-helix의 중간에 위치하는 경우(a)와 첫 번째에 위치하는 경우(b)에 최적화된 모형 단백의 3-D 구조분석의 예

<신약개발>

단백질 3차원 구조 모델링을 이용하여 직접적으로 생명공학 산업에 이용될 수 있는 분야가 “신약개발”이다. 그러나 이것은 매우 단순화시킨 예이며 생물정보학에서의 “신약 개발”은 Genomics, Proteomics, Cheminformatics, Molecular design, Chemical genomics 등의 다양한 분야로부터 컴퓨터 과학 기술력을 이용하

여 생합성, 생물적 활성, 각종 생물 칩 개발 등의 여러 분야와 결부되어 개발이 이루어진다. 신약개발 구성의 예를 그림 4.에 나타내었다.

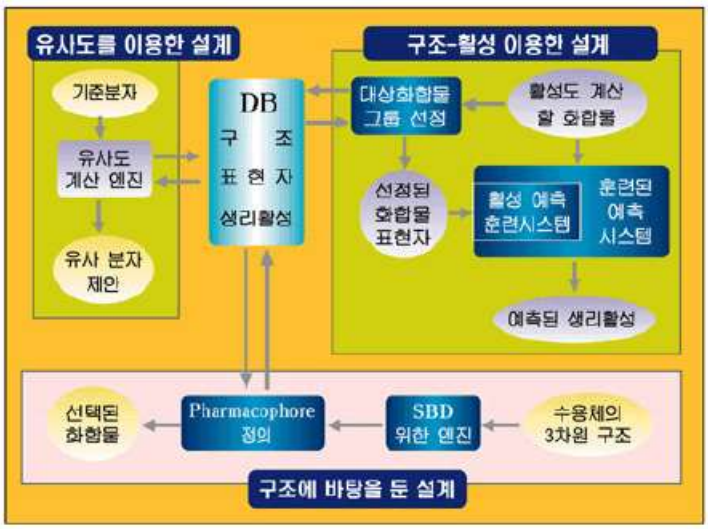


그림. 4 신약개발 구성도의 예

<프로그램 개발>

실제로 생물학 데이터를 이용하여 유용한 정보를 추출해 내기 위해서는 목적에 적합한 컴퓨터 프로그램이 필요하다. 데이터베이스가 매우 잘 구성되어 있고, 에러 없는 정보를 가지고 있다 하더라도 이를 적절히 처리해 줄 소프트웨어가 없으면 연구자는 큰 벽에 부딪히게 된다. 때문에 목적에 맞는 응용프로그램을 개발하는 것은 매우 중요하다. 단순히 핵산 서열을 아미노산 서열로 번역만 해주는 간단한 프로그램으로부터 서열 정렬 프로그램, 분자구조 이미지 가시화 프로그램, 가상 PCR 프로그램, promoter 예측 프로그램, 매우 복잡한 단백질 3차원 구조 예측 프로그램 등 그 개발 범위는 매우 광범위하다. 인간게놈프로젝트의 완료 기간이 단축된 것도 유전자 조각을 자동으로 짜맞추는 프로그램이 적시에 개발되었기 때문이라고 해도 과언이 아니다. 그러나 생물학 관련 분석 툴 및 응용 프로그램, 제어프로그램 등을 개발하기 위해서는 생명현상에 대한 기본적인 이해가 필요할 뿐만 아니라 단백질 구조 분석용 프로그램과 같이 복잡성이 매우 높은 경우에 모든 분야의 전문가들이 모여 서로 조언해 줄 필요가 있다.

생물정보학은 기초생물학, 전산학 그리고 수학, 물리학, 화학공학 등 타 과학 영역간의 연계를 기반으로 하는 연구이므로 생물정보학 연구의 성과는 관련 학문과 산업에 직접적으로 기여할 수 있을 것으로 예상된다.