Bioinformatics의 소개와 이용

김 광 수 | E-mail:kim@postech.ac.kr
포항공과대학교 화학과 교수

생물정보학은 넓은 의미로 컴퓨터를 이용하여 모든 생화학정보를 광범위하게 연구·활용하는 학문이라고 할 수 있다. 하지만 최근의 인간유전체사업(Human Genome Project)의 인기에 의해 생물정보학이 DNA나 단백질의 서열정보 분석의 분야로만 편중되어 의미되는 경향이 많은 것이 사실이다.

Ⅰ. 서 론

인간을 구성하고 있는 총30억 쌍의 유전자 염기 서열을 분석·규명하고자 하는 다국적인 연구사업이 최근 완료됐다. 인간의 DNA뿐만 아니라 수많은 지구상의 생물과 질병에 대한 유전자 분석이 계획·추진중에 있다. 분자생물학의 급격한 발전은 인체의 서열정보를 밝힐 수 있게 하였고, 엄청난 양의 디지털 유전정보가 공개되고 있다. 이러한 서열정보의 다양함과 방대함은 컴퓨터를 이용하여 생명체 관련 자료를 체계적으로 분석하고 효율적으로 활용할 수 있게끔 전산학의 유용한 기술의 사용은 거의 필수가 되었다. 이런 필요성에 의해 생겨난 학문이 바로 Bioinformatics(생물정보학)이다. 사실 생물정보학은 넓은 의미로 컴퓨터를 이용하여 모든 생화학정보를 광범위하게 연구·활용하는 학문이라고 할 수 있다. 하지만 최근의 인간유전체사업(Human Genome Project)의 인기에 의해 생물정보학이 DNA나 단백질의 서열정보 분석의 분야로만 편중되어 의미되는 경향이 많은 것이 사실이다. 학문적으로 시작단계에 위치해 있다고 볼 수 있을 정도로 해야 할 일이 많은 생물정보학에서 우리 나라가 새로운 분야를 개척하여 우위를 선점하는 것은 매우 의의있는 일이라 하겠다.

그러나 무엇보다도 중요한 것은 이러한 유전정보를 이용함으로써 앞으로 생명학적 변혁에 따른 그 파급효과는 상상을 초월하는 가공할 만한 과학적 혁명을 불러다줄 것으로 기대되는 점이다. 엄청난 유전정보 데이터로부터 생명체의 분자적 구조 및 생리현상, 생물체 변이, 인간의 생각, 감정, 인지에 관한 생명현상의 이해와 발전에 이르는 새로운 생명과학의 창출 등 실로 무궁무진한 과학정보를 창출할 수 있다는 점에서 생물정보학의 체계적인 연구가 시급히 요청되고 있다.

Ⅱ. Bioinformatics의 소개와 경향

분자생물학과 유전공학의 발전은 컴퓨터를 통한 엄청난 크기의 생명정보를 공급하기 시작했고 이를 이용하여 과학적이고 체계적인 데이터베이스를 구축하여 활용하려는 시도가 이루어졌다. 또한 인터넷과 컴퓨터의 보급으로 정보에 대한 공급과 취득이 용이해지면서, 과학자들은 생물정보학을 통해 생명현상과 관련된 새로운 법칙을 찾아내고자 한다. 즉, 신체의 다양한 유전

자 조직이 언제 어디에서 발현해서 단백질을 만드는가 또는 단백질은 어떻게 서로 상호작용을 하고 그러한 상호작용이 질병에는 어떤 역할을 하는가에 관한 데이터베이스를 구축하게 되는 것이다.

생물정보학의 가장 기초적인 분야는 새롭게 밝혀진 유전자 서열과 이전에 밝혀진 유전자 서열 사이의 유사성(similarity)이나 상동성(homology)을 찾는 것이다. 비슷한 쌍들은 연구자들에게 새롭게 서열된 단백질의 타입을 예측할 수 있게 해주고 이전에 잘못된 것을 알린다. 이것은 의약품 개발에 있어서 그 결과를 미리 예측·선별함으로써 엄청난 비용의 개발비용을 투입하고도 결과가 좋지 않을 때 발생할 미래의 손실을 방지해 준다. 이런 결과들로 이전에 수없이 반복하던 실험의 상당 부분은 컴퓨터를 이용한 시뮬레이션으로 개발비용을 대폭 절약할 수 있다. 이런 생물학 및 제약학에서 혁명을 실현시킬 수 있는 필수적인 도구는 아마도 데이터베이스 기술일 것이다. 많은 데이터베이스 기술들이 개발·사용되고 있으며, 이에 대해서는 추후 알아보도록 하겠다. 현재에 와서는 다양한 프로그램과 데이터베이스들을 통해 연구자들은 많은 정보를 얻을 수 있게 되었다. 생물정보학의 경제적인 측면에 대하여 전문가들은 5년 이내에 생물정보학의 시장규모가 약 20억 달러에 이를 것으로 내다보고 있다.[1]

현재도 많은 회사들이 경쟁적으로 온라인상에서 다양한 데이터베이스의 접근과 데이터 조작을 지원하고 있다. 또한 연구소나 제약업체들은 그러한 정보에 대해 값비싼 사용료를 지불하더라도 기꺼이 사용하려고 하고 있다. 생물정보는 효율적으로 약물개발의 타겟을 찾게 하며 개발에 대한 비용과 시간을 대폭 감소시키기 때문에 많은 비용을 감수하고도 그러한 정보를 이용하고 있는 추세다.

> 생물정보 데이터베이스의 종류로는 DNA 데이터베이스(Nucleotide Sequence), 단백질 데이터베이스 (Amino Acid Sequence) 등이 있다. 현재 사용 가능한 DNA의 데이터베이스는 대형 데이터베이스, Genomic 데이터베이스, 특수 데이터베이스로 나눌 수 있다.

III. 생물정보 데이터베이스

1. 생물정보 데이터베이스

최근 생물정보 데이터베이스는 굉장한 발전을 해오면서 생물학자들의 연구에 있어서 필수적인 도구가 되었다. 생물학자들이 이러한 전산적 방법을 사용하게 된 이유를 알아보자. 새로운 유전자 서열을 얻을 때 학자는 그것이 이미 데이터베이스에 저장되어 있는지, 아니면 어떤 homologous 서열들을 담고 있는지의 여부를 조사할 필요가 있다. 또, 추정된 부호화된 ORF가 주어지면 이를 이용해서 동종의 단백질을 찾을 수 있다. 그리고 비슷한 non-coding DNA stretch들을 데이터베이스에서 찾을 수도 있고, PCR oligonucleotide의 집합에 대한 잘못된 priming site들을 알아내는 것과 같은 특별한 목적을 위한 다른 용도들 때문에 사용할 수도 있다.

이렇게 사용되는 생물정보 데이터베이스의 종류로는 DNA 데이터베이스(Nucleotide Sequences), 단백질 데이터베이스(Amino Acid Sequence) 등이 있다. 현재 사용 가능한 DNA의 데이터베이스는 대형 데이터베이스, Genomic 데이터베이스, 특수 데이터베이스로 나눌 수 있다. 대형 데이터베이스로는 미국의 Genbank, 유럽의 Embl, 일본의 DDBJ 등이 있으며, 이들은 모두 비슷한 내용을 지니며 협조체제하에 자료의 확충과 갱신이 이루어지고 있다. Genomic 데이터베이스에는 Human(GDB), mouse(MGB), yeast(SGB)등이 있고 특수 datbase에는 ESTs(expressed sequence tags), STSs(squence-tagged sites), EPD(eukaryotic

promotor 데이터베이스), REPBASE(repetitive sequence 데이터베이스) 등이 있다. 단백질을 검색하는 대형 데이터베이스는 Swiss-Prot(고수준 annotiation), PIR(단백질 identification resource) 등이 있고 DNA 데이터베이스를 변환한 것으로는 SPTREMBL(EMBL을 변환), GenPept(GenBank에서 coding regions의 변환) 등이 있다.

2. 생물정보 데이터베이스 검색 시스템

생물정보 데이터의 양은 굉장히 빠르게 증가하고 있고 그러한 정보들을 접근하고 탐색하는 방법의 획득은 필수적인 것이 되었다. 최근에는 많은 훌륭한 데이터베이스들을 웹을 통하여 사용할 수 있다. 여기에서는 분자생물학에 특별히 관련된 세 가지 종류의 대표적인 데이터 검색 시스템인 SRS(Sequence Retrieval System), Entrez, DBGET에 대해 알아보겠다. 이러한 시스템들은 다용도 분자생물학 데이터베이스에서 텍스트 기반의 검색을 제공하고 검색 표준에 맞는 엔트리들에 관한 링크를 제공한다. 이 세 가지 시스템들은 그들이 탐색하는 데이터베이스와 그들이 담고 있는 서로 다른 정보의 링크들에서 차이점을 보인다.

> 일단 생물 데이터베이스를 검색하려는 사용자는 가장 최신의 데이터베이스를 사용해야 할 것이다. 많은 학자들은 BLAST를 먼저 검색하고 그것의 결과에 따라서 더 나은 툴들(FASTA, ssearch, SW검색, block)을 사용하기를 권장하고 있다. 가능할때는 언제라도 변환된 서열을 사용한다.

SRS(Sequence Retrieval System)는 영국에 위치한 European Bioinformatics Institute에서 개발된 80개 이상의 생물정보 데이터베이스에 연결된 단일 인터페이스다. SRS는 서열, 대사과정, 유전정보의 transcription factor, BLAST, SSEARCH, FASTA 등의 응용프로그램들에 대한 결과, 단백질의 3D 구조 모델, 유전체정보, mapping, 돌연변이 등 엄청난 양의 정보들을 제공한다. 모든 데이터베이스들이 나열된 웹페이지는 마지막으로 업데이트된 일자가 포함된 링크가 담겨져 있다. 사용자는 단순히 질의어나 검색할 단어를 입력하기 전에 하나 이상의 데이터베이스들을 선택한다. 결과를 얻은 후에 CLUSTALW, PHYLIP 같은 정렬알고리즘을 선택하고 실행하게 된다. SRS는 많은 학자들에 의해서 추천되고 있는 우수한 데이터베이스다. Entrez는 분자생물학 데이터베이스 검색 시스템이다. Entrez는 미국의 NCBI(National Center for Biotechnology Information)에서 제작했다. 대표적인 세 가지 텍스트 기반의 데이터베이스 시스템 중 Entrez는 가장 사용하기가 쉽지만 상대적으로 검색하기에 약간 제한적인 정보를 제공한다. DBGET은 일본의 동경대학교에서 개발한 통합 검색 데이터베이스 시스템이다. 동시에 20개 이상의 데이터베이스로의 접근을 제공한다. 더욱 제한된 옵션을 가지기 때문에 DBGET은 위에 언급한 두 개의 데이터베이스보다 기능 면에서 상대적으로 떨어지는 것이 사실이다.

3. 생물정보 데이터베이스 검색 프로그램과 알고리즘

일반적으로 생물정보 데이터베이스 검색 프로그램의 성질을 말할 때 민감성(sensitivity)과 특이성(specifity), 두 가지의 성질에 대해 생각해볼 수 있다. 민감성(sensitivity)은 true positive match를 검출하는 능력을 말한다. 가장 정밀한 검색은 모든 정확한 match들을 찾지만 또한 많은 false positives(잘못 찾은 것)들을 찾을 가능성이 있다. 그리고 특이성(specifity)은 false positive match들을 가려내는 능력을 말한다. 가장 특이성을 만족하는 검색은 true match들만 결과로 나타날 것이다. 하지만 많은 false negative(빠트린 것)들을 가질 수 있다. 연구에 필요한 어떤 알고리즘을 선택할 때는 이 두 가지 성질을 잘 고

려해야 한다.

그러면 이제는 생물정보 데이터베이스에 가장 대표적인 3가지의 검색 프로그램을 알아보자. 그 세가지에는 FASTA, BLAST, SW-search가 있다. FASTA는 Pearson과 Lipman의 방법을 사용하는 서열 비교 소프트웨어다. 이 프로그램은 DNA 서열을 DNA 데이터베이스와 비교하거나 단백질 서열을 단백질 데이터 베이스와 비교하게 된다. 실질적으로 FASTA는 FASTA, TFASTA, Ssearch를 포함하는 여러 가지 프로그램들의 집합체다.

BLAST(Basic Local Alignment Search Tool)는 주어진 문제의 해를 구하기 위해 시행착오를 통해서 경험적 지식을 축적하고 이를 문제 해결에 이용하는 방법인 휴리스틱(heuristic) 검색 알고리즘을 사용한 검색 프로그램이다. 프로그램들은 Karlin과 Altschul의 통계적 기법을 사용하여 검색하게 된다. BLAST 프로그램은 산재된 관련 서열들에 대해 민감성(sensitivity)의 최소한의 희생으로 신속한 데이터베이스 검색을 위해 디자인되었다. BLAST는 특별히 압축된 포맷으로 데이터베이스를 검색한다. 사용자 자신의 개인 데이터베이스를 BLAST로 사용하기 위해 사용자는 데이터베이스를 BLAST Format으로 변환시켜야 한다.

Smith-Waterman 검색방법(이하 SW검색)은 pairwise 비교에 대한 full Smith-Waterman algorithm을 사용하는 데이터베이스에서 query를 각 서열에 비교하는 방법이다. 그것은 또한 통계량을 생성하기 위해 검색결과를 사용한다.

SW 검색은 철저히 수행하기 때문에 가장 느린 방법 중 하나다. 우리는 SW algorithm을 실행시키기 위해 특별한 hardware와 software를 사용하게 된다. 세 가지 검색도구를 비교해보면, 개념에 관해서는 SW검색, BLAST는 부분 정렬 도구인 반면 FASTA는 전역 정렬 도구다. 속도는 BLAST, FASTA, SW순으로 빠르다. BLAST는 확률을 계산하고 만일 몇몇의 사용되는 몇몇의 가정이 무효가 되면 때때로 전체적으로 실패하게 되고, FASTA는 주어진 데이터 집합으로부터 신뢰성 있는 계산을 하기는 하지만 만일 데이터 집합이 작으면 문제가 발생할 가능성이 높다.

일단 생물 데이터베이스를 검색하려는 사용자는 가장 최신의 데이터베이스를 사용해야 할 것이다. 많은 학자들은 BLAST를 먼저 검색하고 그것의 결과에 따라서 더 나은 툴들(FASTA, Ssearch, SW검색, block)

을 사용하기를 권장하고 있다. 가능할 때는 언제라도 변환된 서열을 사용한다. 오차의 범위가 0.05이하인 것이 통계적으로 의미가 있으며 보통 생물학적으로도 신뢰할 만하다. 질의하는 서열들의 특이한 조합에 주의해야 한다. 일반적으로 특이한 조합은 편차가 있는 scoring을 발생하게 할 수 있다.

큰 query 서열들은 나누는 것이 좋다. 만일 query하는 서열들이 segments들을 반복하면 그것들을 제거하고 검색을 반복한다.

IV. 생물정보학의 화학적 이용

생물정보학 전반에 대해 대략적으로 살펴보면 최근의 경향은 대부분 유전체학에 집중되고 있다. 사실 우리나라는 90년대 16개국 350여 개의 연구소가 참여한 인간유전체사업에 참여하지 못했다. 컨소시엄에 우리 나라의 연구실이 한 개도 참여할 수 없었던 것에 반성을 하면서 앞으로 우리가 선진국과 격차를 줄여가려면 어떠한 전략으로 나아가야 하는지를 생각해보는 것은 매우 중요하다. 인간의 유전자 염기서열이 점차 밝혀지고 있는 상황에서 지금은 그러한 유전자 서열 자체의 발견보다는 이러한 유전정보가 어떠한 역할을 하는가 하는 유전자의 기능을 밝히는 유전체 기능분석학이 더욱더 중요해지고 있다. 이 유전체 기능분석학은 연구하는 대상에 따라 크게 DNA나 RNA를 대상으로 하는 유전체학(genomics)과 단백질을 대상으로 하는 단백질체학(proteomics)으로 분류된다. 최근 유전체학에서도 간암이나 위암 같은 질병의 진단을 위한 DNA 칩을 연구하는 등 실용적인 측면을 보다 중시하고 있다. 단백질체학의 발달로 하나의 단백질을 분석하여 다양한 정보를 얻는 데에 걸리는 시간을 획기적으로 단축할 수 있게 되었다. 생물학적 정보가 실제 생체 내에서 영향을 미치기 위해서는 유전자 정보에 상응하는 단백질로 발현되는 것이 바람직하다. DNA 염기서열은 RNA로 전사되고 이에 의해 단백질 1차 구조가 결정된다. DNA 3쌍이 하나의 단백질 잔기(residue)를 결정하며 이렇게 조합된 20개의 단백질 잔기 배열은 공간적 정보를 가지는 2·3차의 구조를 가지게 되고, 때로는 3차 구조들의 집합체인 4차 구조를 이루면서 비로소 생체 내에서 고유의 역할을 수행하게 된다. 그러므로 단백질 염기서열을 밝히는 만큼이나 중요한 것이 생체 내에서 발현된 단백질이 어떤 구조를 가지고 어떤 기능을 수행하느냐 하는 것이다. 이 분야는 구조생물학(structural biology) 또는 분자생물학(molecular biology)으로 불렸지만 현재는 생물정보학의 한 분야(structural bioinformatics & molecular bioinformatics)로 자리잡고 있다.

> 단백질 배열은 DNA 염기 배열과는 다르게 공간적 정보를 가지며 단백질은 분자간 상호작용을 통해 직접적으로 생체 대사를 조절하므로 어떠한 배열, 혹은 어떠한 구조가 생체 내에서 특정한 기능을 나타내는가에 대한 정보는 중요하다. 이러한 구조생물학/분자생물학적 정보를 전산처리하여 체계화하고 데이터 베이스를 구축하는 것은 새로운 기능성 분자의 디자인이나 신약 개발에 매우 중요한 역할을 할 것이다.

단백질 배열은 DNA 염기 배열과는 다르게 공간적 정보를 가지며 단백질은 분자간 상호작용을 통해 직접적으로 생체 대사를 조절하므로 어떠한 배열, 혹은 어떠한 구조가 생체 내에서 특정한 기능을 나타내는가에 대한 정보가 중요하다. 이러한 구조생물학·분자생물학적 정보를 전산처리하여 체계화하고 데이터베이스를 구축하는 것은 새로운 기능성 분자의 디자인이나 신약 개발에 매우 중요한 역할을 할 것이다. 이러한 과정을 거쳐 거시적인 생물정보학적 데이터로부터 미시 분자세계의 정보를 얻어낼 수 있으며, 이로써 생체 시스템을 조절하거나 모방한 새로운 기능성 분자를 만들어낼 수도 있다. 이해를 돕기 위하여 생물학적 정보의 데이터로부터 특정 분자구조에 대한 정보를 얻은 좋은 간단한 예를 소개하겠다. 1950년대 DNA 나선구조에서 방

향족 고리(aromatic ring)들이 겹쳐서 쌓인 구조가 중요한 역할을 하는 것이 발견되면서 방향족 고리간 상호작용(aromatic-aromatic interaction)이 주목을 받기 시작했다. 이후로 화학자들은 방향족 고리들은 쌓인(stacking) 구조가 가장 안정한 것으로 인정해 왔지만 1985년에 이르러 Burley와 Petsko에 의해 다른 가능성이 제시되었다. 이들은 Cambridge Crystallographic Data Base로부터 34개의 단백질 구조를 선택하여 이에 포함된 방향족 잔기(Phe, Tyr, Trp)들 중 고리 사이의 중심간 거리가 7 이내인 580여 쌍의 방향족 고리들에 대하여 짝지어진 형태에 대한 빈도, 분리 정도와 이면각과 같은 상호작용 구조, 비결합 상호작용 에너지, 상호작용을 하는 잔기들의 이차 구조적 위치 등을 분석하였다. 이러한 분석결과, 단백질에 존재하는 방향족 고리 중 약 60%가 방향족 고리 쌍으로 존재를 하며, 이 중 80%는 세 개 혹은 그 이상의 방향족 고리가 서로 상호작용을 하는 네트워크를 형성하고 있다는 것이 밝혀졌다. 그러나 이러한 방향족 고리들은 이전에 안정하다고 인식되어온 서로 겹쳐져 쌓인 구조보다는 T자 모양이나 비스듬한 구조로 더욱 많이 존재한다는 것이 발견되었다. 이러한 사실은 방향족 고리들이 서로 층층이 쌓인 면-면 상호작용(Aromatic-aromatic face-to-face interaction)보다는 고리들의 이면각이 90도 정도가 되는 방향족 고리간 끝-면 상호작용(Aromatic-aromatic edge-to-face interaction)이 더욱 안정하다는 결과를 보여주는 것이다. 이러한 사실은 1990년에 Sanders와 Hunter에 의해 수행된 계산화학적 방법을 통한 연구에 의해 뒷받침되었다. 그들은 방향족 분자들 사이의 상호작용에 기하학적 구조가 아주 강한 영향을 미친다는 사실을 설명하기 위해서 간단한 모델을 디자인하여 이들의 상호작용을 계산·분석한 결과, 쌓인 구조보다는 T 모양이나 비껴 쌓인(displaced stacked) 구조가 안정함을 증명하였다. 최근 McGaughey는 방대한 양의 단백질 구조 데이터베이스를 검색 30,444 쌍의 방향족 잔기들의 구조를 분석하여 이전의 결과들을 뒷받침하였다. Burley와 Petsko에 의해 제시된 구조분석 데이터는 생체 혹은 분자세계에서 방향족 고리간 상호작용을 이해하는데 결정적인 역할을 했으며 이후 신약 같은 새로운 분자를 디자인하는 데 있어서 중요한 역할을 할 것으로 기대되고 있다. 이렇듯 어떠한 생물적·화학적 실험도 거치지 않고 단순히 구조정보 분석만을 가지고도 훌륭한 연구결과를 낼 수 있다는 점에 우리는 주목해야 한다. Protein Data Bank(www.pdb.org)를 통해 제공되는 12,600개 가량의 단백질 구조들 속에는 우리가 알지 못하는 귀중한 정보들이 있다. 그러나 이들 구조 정보를 체계화하고 공통점을 분석하여 새로운 사실을 발견하는 방향의 연구는 아직 초보적인 단계이며, 새로운 아이디어와 빠르고 정확한 분석방법의 개발이 절실한 상황이다.

V. 결 론

지금까지 Bioinformatics에 대한 최근 경향, 생물정보학에 필수적인 생물정보 데이터베이스, 그리고 생물정보학의 화학적 이용에 대해서 알아보았다. 결론적으로, 생물정보학의 개념은 구조생물학·분자생물학적 데이터 분석 및 응용에도 확장할 수 있으며, 이러한 구조 데이터분석은 DNA 염기서열의 정보를 직접적으로 이용하는 것과 더불어 생체연구, 의술발전, 신약개발, 특히 생명현상의 이해 및 생체변이 등에 혁신적 발전이 이루어져 실로 가까운 시일 내에 새로운 미지의 생명과학을 창출할 것으로 기대한다.

■ 주

1) Scietific American : Feature Article : The Bioinformatics Gold Rush : July 2000