# What is the bioinformatics ?
## (From the chemical engineering perspective)

**권성우, 이병우*, 한종훈**

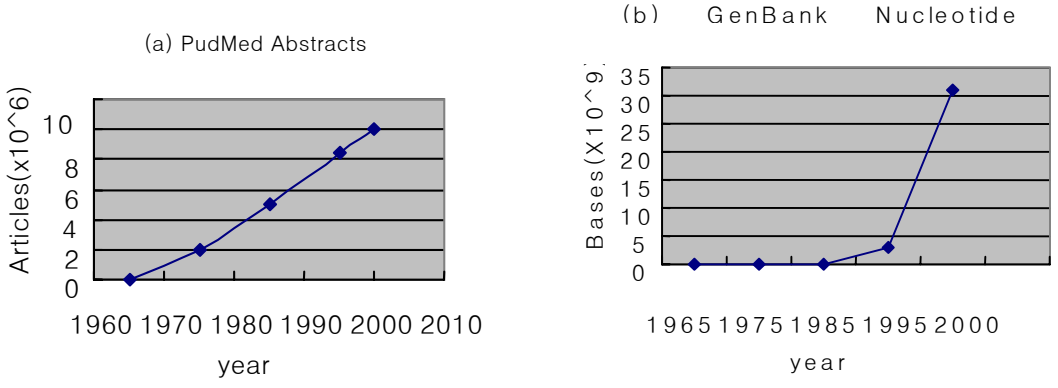
**포항공과대학교 화학공학과, sw74@postech.ac.kr**

***삼성 SDS**

최근 들어 21세기는 바이오텍의 세기라는 말을 흔히 들을 수가 있다. 그리고, 이와 자주 함께 자주 등장하는 것으로는 생물정보학(bioinformatics)라는 용어가 있었을 것이다. 이 글에서는 생물정보학에 대한 소개와, 기존의 바이오텍에 많은 기여를 했던 화학공학자들이 어떤 식으로 접근할 수 있을 지에 대해서 논의를 하기로 한다.

생물체와 이들이 이루는 생물계는 다른 무생물적인 대상에 비해 훨씬 더 정보집약적이라고 할 수 있다. 하나의 생물체를 예로 들자면, DNA속에 담겨진 정보에 의해서 여러 종류의 단백질들이 만들어지고, 이 단백질들은 그 자체의 구조에 대한 정보와 함께 어떤 조건에서 무엇이 어떤 식으로 상호작용을 할 것인가에 대한 정보를 가지게 된다. 여기에 추가로 다세포 생물은 세포, 조직, 기관들 사이의 상호작용부터 시작하여, 개개의 개체 사이의 상호작용, 무생물 환경과의 상호작용, 집단과의 상호작용, 진화에 이르기까지 다양한 계층의 정보를 가지게 된다. 이에 수반되는 정보의 양은 실로 막대하며, 매우 복잡하다. 따라서 생명체에 대한 연구는 본질적으로 컴퓨터를 이용한 정보학적인 도구가 그 핵심을 차지할 수밖에 없는 것이다. 그런데 이전의 생물학에서의 문제는 대규모 정보를 생물체들에게서 얻는 일들을 하는 것이었다. 그 점에서 인간 유전체 프로젝트가 90년대 초부터 시작이 되어 성공적으로 완료가 되었다. 인간 유전체 프로젝트는 여러 가지 파급 효과를 가져왔는데 그 중 하나가 high-throughput 기술 및 기기(DNA chip, Megabase1000, ABI3700, MALDI-TOF 질량분석기 등)의 발명이다. 이러한 기술과 기기가 발명되고 발전함으로 인해 아래 그림에서와 같이 생물학적인 데이터들이 급속도로 증가하게 되었고, 많은 양의 데이터를 다루고 분석 하기 위해서 컴퓨터와 정보학적 방법(informatics technique)을 사용하게 되었다. 바로 이 것이 생물정보학이라고 불리는 것이다.

생물정보학은 biology(생물학)와 informatics (정보학)를 결합하여 만든 단어이다. 현재 그 정의



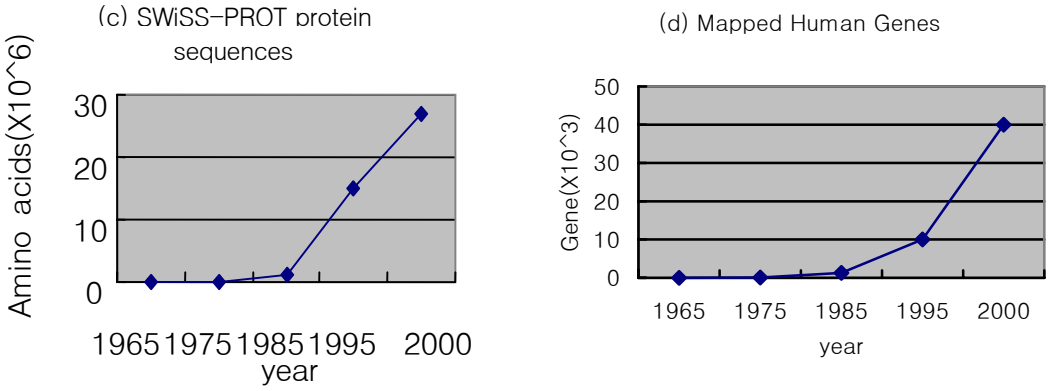



**그림 1. 생물학 데이터의 증가**

| **Data source** | **Data size** | **Bioinformatics topic** |
|---|---|---|
| Raw DNA sequence | 8.2X10$^7$ sequence | 번역부위와 기타 부위의 분류<br>Exon과 intron의 동정<br>gene product 예측 |
| 단백질 서열 | 300,000 sequences | 서열 비교 알고리즘<br>Multiple sequence alignment 알고리즘<br>Motif 서열의 동정과 예측 |
| 거대 분자 구조 | 13,000개 구조 | 2차, 3차 구조의 예측<br>분자간의 상호작용<br>분자의 모사 (force-field 계산, 분자의 운동, docking 예측) |
| 유전체 | 40개의 개체 | Phylogenetic tree 작성 및 분석<br>유전체 규모의 조사(대사회로 분석)<br>질병과 연관된 유전자의 linkage 분석 |
| 유전자 발현 | 약 20회 이상으로 6,000개 이상의 유전자를 측정 | DNA chip을 통해 유전자 발현 양상의 연관성 분석 |
| 대사회로 | | 대사회로의 모사 |
| 생명공학 관련 논문 | 11X10$^7$ 편 | 자연어처리를 통한 논문의 검색<br>논문 검색을 통한 data mining |

표 1 생물정보학에 사용되는 데이터들

는 생물학을 분자 수준에서 보는 것인데, 이들 생물체의 분자들이 가진 데이터 들에 informatics technique을 적용 시켜 대량의 새로운 정보를 얻어내고, 많은 생체 분자들 사이의 연관된 정보를 유기적인 구성을 하는 것이다. 즉, 생물학과 관련된 정보를 향상시키기 위한 모든 전산, 수학, 통계적인 방법이나 접근 방식에 대한 연구라고 할 수 있다.
생물정보학의 목적은 크게 세 가지인데, 첫째는 연구자들이 쉽게 검색을 할 수 있도록 기존에 있는 데이터를 database화 시키는 일이나 기존의 database를 유기적으로 연결시키고 연구자들이 만들어낸 새로운 데이터를 여기에 추가하기 쉽게 하는 일이다. 예를 들자면 ENTREZ나 SRS(Sequence retrieval system)가 여기에 해당이 된다. 둘째로는 데이터들을 분석하는데 있어서 도움이 되는 방법과 수단을 개발하는 것이다. 예를 들자면 FASTA나 PSP-BLAST같은 프로그램을 이용하여 상동성을 검색하는 것이 해당된다. 셋째로는 위의 데이터 분석 방법들을 이용하여 생물학적으로 의미가 있는 결과를 만들어 내는 것이다. 예를 들자면 유전체학(genomics)나 단백질학(proteomics)이 여기에 해당이 된다. 결국 고전적인 생물학에서는 개별 시스템에 대해서 혹은 아주 소수의 관련된 것들을 비교하는 방식으로 연구가 진행되었지만, 생물 정보학에서는 생물체의 데이터를 포괄적으로 분석(global analysis)하여 모든 생물 시스템 혹은 특정 생물 시스템이 가진 원리를 밝힐 수가 있게 된 것이다.

현재 생물정보학에는 다음과 같은 분야가 있다.

## 1. 자료 처리 및 분석

생물학적 정보의 처리량을 높이기 위해서는 사람의 조작을 거치지 않고 아날로그로 들어오는 생물 정보를 자동적으로 감지해서 컴퓨터에서 처리할 수 있는 디지털 정보로 바꿔야 한다. 일반적인 자료의 획득과 이것을 분석하는, 예를 들자면, DNA sequencing 장치에서 뉴클레오티드(nucleotide)가 gel속을 흘러가면서 내는 형광 신호를 해석하여 영상 자료를 분석하는 것을 들 수 있다. 현재의 DNA sequencing 시스템의 효율을 높이려면 향상된 정확도, read-length, 계측의 신뢰도를 주는 염기 판독(base-calling) 알고리즘이 필요하다. 이러한 향상은 다시 DNA 염기 서열의 조합과 마무리를 자동화하는데 도움이 된다. 실험 자료를 환경이 다른 여러 연구소가 서로 공유하면서 여러 종류의 소프트웨어로 분석하려면 자료 표현 기준이 특히 중요하게 된다.

## 2. 염기 서열의 조합

DNA의 염기 서열 판독은 DNA sequencing 시스템의 한계—1000bp이상의 DNA조각을 기계에 넣으면 700bp이후의 DNA의 시그널은 약해져서 시그널이 ACGT중 어떤 것의 시그널인지 제대로 판별할 수 없다—로 인해 DNA를 잘라서 클로닝(cloning)한 후에 각각에 대해 여러 번의 sequencing을 하게 된다. 그 후에 각각의 염기 서열을 결정한 결과를 최종적으로 하나로 합쳐야 하는데, 이때 염기 서열 조합 알고리즘(sequence assembly)이 필요하다. 그러나 서열 판독 기기로 읽은 각각의 염기는 실험 오차의 영향을 받으므로, 대규모의 염기 서열을

판독하려면 다음과 같이 여러 면으로 향상된 염기 서열 조합 알고리즘이 필요하다.
(1) 각 염기 판독의 신뢰도에 대한 정보 사용 (2) 염기 서열 오류의 자동 처리 (3) 최종 결과로 나온 염기 서열에서의 각 염기에 대해 신뢰도 (확률) 예측치를 부여 (4) 추가적인 보충 정보(clone 길이 등)의 사용 (5) sequence closure를 위한 실험 방법의 자동화 지원. 궁극적으로는 실질적으로 사람의 조작이 없이 염기 판독(base calling)으로부터 최종 염기 서열의 조합 및 분석까지의 모든 소프트웨어를 자동화하는 것이 바람직하다.

## 3. 유전체(genome)의 염기 서열에서의 기능 예측 및 정렬(alignment)

유전자의 염기 서열은 전사(transcription)과 번역(translation)을 통해 단백질을 만들어 낸다. 이들 단백질은 다른 단백질이나 DNA등과 상호 작용을 통해, 세포 내에서 혹은 세포 간의 생명 작용을 결정하게 된다. 그러나 수많은 염기 서열에서의 상세한 기능을 다양한 실험을 통하여 결정한다는 것은 실질적으로 불가능하다. 현실적으로 말하면 한 개 유전자의 기능을 실험을 통해서 밝히는 데 드는 시간이 보통 10년이 걸린다. 그러므로 실험을 하기 이전에 유전체에 대해 관심 있는 기능을 가진 유전자를 예측하고, 그 예측이 맞는지를 실험을 통해 확인하는 것이 보다 효율적이다.
이렇게 염기 서열에서의 기능을 예측하는 방법으로 많이 사용되는 FASTA는 염기 서열 데이터베이스에 있는 서열과의 유사성을 검색하는 방법이다. FASTA 이외에도 특정한 기능을 한다고 알려져 있는 서열과 기능을 포함하지 않은 서열을 신경망(neural net)으로 학습시킴으로써 염기 서열의 기능을 예측하는 방법도 있다.
BLAST(Basic Local Alignment Search Tool)는 FASTA와 비슷한 단백질 서열 및 염기 서열의 유사성 검색 프로그램으로서, FASTA보다 훨씬 처리 속도가 빠르지만 별도의 pre-formatted search database가 필요하며, 일치성은 없지만 전반적으로 유사성을 보일 경우에는 검색이 약하다는 등의 단점을 가지고 있다. 또한 BLAST 프로그램에서 사용하는 확률 이론에 의하면, 주어진 수준의 통계적 의미를 위해 필요한 similarity score는 데이터베이스 크기의 logarithm에 비례한다. 하지만 두 가지의 염기 서열을 비교함으로써 생성되는 similarity score는 이들이 발견된 데이터베이스 크기와는 무관하다. 그러므로 데이터베이스가 커짐에 따라 생물학적으로 의미를 가지지만 관련성이 약한 염기 서열은 무작위로 이루어진 match보다 작은 similarity score를 가지게 될 수 있으며, 이로 인해서 ‘noise’에 묻혀 버릴 수도 있다. 이 문제를 해결하기 위해서는 데이터베이스를 간단하게 하거나, 향상된 새 염기 서열 정열 알고리즘을 개발해서 데이터베이스 검색에 사용해야 한다.
HMM(Hidden Markov Model)은 각 시간에 따라 개별적인 상태로 표시가 가능한 시스템에서 그 구조를 명확히 알 수 없을 경우에 시스템을 확률적으로 설명하기 위해 사용할 수 있다. 최근에는 서열의 정열 확률을 계산하기 위해 이 HMM을 도입하기도 한다.

## 4. Microarray technology

많은 양의 genome data를 분석하기 위해서는 새로운 방식의 분석 기술이 필요하다. 이러한

기술들 중에서 대표적인 것이 바로 microarray technology이다. microarray technology 에는 DNA chip, protein chip, lab-on a chip 등이 있다. DNA chip은 solid support에 고정 시킨 DNA와 mRNA 나 다른 DNA를 hybridization 시키는 것으로, 돌연변이(mutation)이나 polymorphism 검출을 위한 resequencing과 특정 상태의 gene expression profile을 연구 할 때 사용하고 있다. 한편 protein chip의 경우 solid support에 protein을 고정화 시킨 것으로 protein간의 상호작용을 연구할 때 사용 할 수가 있고, lab-on a chip의 경우 protein chip 이나 DNA chip 과는 달리 chip 실험을 하기 위한 전처리 과정이 chip 위에서 한꺼번에 이루어져 실험이 편리하고 정확하다는 장점이 있다. 현재 protein chip과 lab-on a chip은 개발 중이며, 많이 사용이 되는 DNA chip의 경우에 주요 연구 분야를 세가지로 나누자면, 첫째 micro fabricate 와 image analysis, 둘째 data analysis 와 mining, 셋째로 application이다.

Micro fabricate 분야의 경우 DNA chip 제작 방식을 말하는 것으로 크게 pin 이나 ink-jet을 이용하여 DNA를 solid support 위에 dotting하는 printing 방법과 photolithography를 이용하여 oligonucleotide를 solid support 위에 합성하는 방법이 있다. Printing 방식의 경우 고정화 시키는 DNA의 양이 적다는 단점이 있지만 chip 제작 단가가 저렴하다는 장점이 있다. 한편 photolithography 방식의 경우 고밀도로 DNA를 고정화 시킬 수가 있는 장점이 있지만 chip 제작 과정에서 사용이 되는 photo-mask의 단가가 비싼 단점이 있다. DNA chip을 hybridization 하면 여러 spot들이 나타나는데 이 것의 색깔과 크기를 정량화 시키는 것이 바로 image analysis에 해당이 된다. hybridization 결과로도 수 백 개에서 수천개의 spot이 나타나므로 사람의 눈을 통해서 하는 것은 불 가능하다. 따라서 반드시 컴퓨터 프로그램을 이용하여 spot을 분석한다. 현재 image analysis의 방식에는 크게 target detection 방식과 grid overlay 방식이 있다. 수백 개에서 수천 개를 이루는 spot으로부터 얻어지는 Data를 분석하는 부분이 바로 Data analysis와 mining에 해당이 된다. 많은 경우에 분석을 보다 간단히 하기 위해서 클러스터링 기법을 이용하여 비슷한 패턴을 보이는 유전자 데이터를 분류한다. 이렇게 많은 양의 데이터를 통계적으로 처리하는 방법으로는 PCA(Principal Component Analysis), decision tree, hierarchical clustering, *k*-means clustering, SOM(Self-Organizing Map) 등이 있다. 한편 DNA chip을 토대로 각각의 유전자가 아닌 인간의 유전자 전체, 유전체 수준에서 유전자와 생명현상과의 관계를 규명하고 유전자의 이상이 질병의 발생과 현상 발현에 어떠한 영향을 미치는지를 규명하는 일이 가능하며, 이를 통하여 질병의 진단과 치료 등에 application을 할 수가 있다.

## 5. 구조 데이터베이스의 탐색 및 거대분자(macromolecule) 구조의 결정

최근 protein engineering, crystallography, 분광기가 발달함에 따라서 최근에 밝혀진 단백질 구조의 양도 빠르게 증가하고 있다. 새롭게 밝혀진 구조는 염기 서열의 유사성이 감지되지 않는 경우에도 이미 밝혀진 구조와 점점 더 구조적으로 유사성을 보이고 있다. 새로운 알고리듬들은 단백질 구조를 기존에 알려진 모든 구조의 데이터베이스와 비교할 수 있게 한다. 구조 데이터베이스 검색은 관심을 끄는 생물학적 관계를 발견하는 도구로서 염기 서열 데이터베이스 검색에 비슷한 수준에 이르렀다.

원자 수준의 해상도에서 고분자 구조를 추정하는 방법으로 주로 사용되는 실험 방법은 X-ray crystallography, 핵자기 공명(NMR)이다. 두 가지 방법은 모두 매우 많은 양의 자료를 제공하며, 이 자료의 해석을 위해 강력한 컴퓨터와 정교한 처리 일고리듬이 있는지의 여부에 따라 전적으로 결정된다.

실험에 의한 단백질 구조 결정 방법의 발전에도 불구하고, 아직은 실험을 통해 단백질 생성물의 3차원 구조를 결정하는 것보다 유전자의 염기 서열을 분석하고 이것이 암호화하는 단백질의 아미노산 서열을 유도하는 것이 훨씬 용이하다. 아미노산 염기 서열(단백질의 1차 구조)로부터 단백질의 3차 구조를 직접 예측하는 기능은 새로운 분야인 단백질 공학 및 설계와 함께 구조 기능 연구에 큰 도움이 될 것이다. 원칙적으로는 단백질의 아미노산 서열이 꼬인(folding) 형태의 단백질 의 3차원 구조를 완전히 명시하므로, 아미노산 염기 서열만으로 단백질의 구조를 계산하는 것은 이론적으로 가능하다. 그러나, 살아있는 세포에서 발견되는 길이의 단백질 구조에 대해서 가능한 구조(conformation)의 수는 천문학적으로 많아 기존의 컴퓨터로는 'conformational space'의 탐색 문제가 실질적으로 불가능하다. 또한 이런 구조들의 energetic를 결정하는 생물물리학은 복잡하고 완전히 알려지지 않았다.

최근의 다른 방법은 단백질 folding 문제를 '역 구조(inverse-structure) 문제'로 돌려놓는데, 이것은 단백질의 구조 문제에 두 가지 방식으로 접근할 수 있게 한다. 즉 특정한 구조가 주어진다면 어떤 서열이 그렇게 접힐 것인지, 또는 주어진 서열이 기존에 알려진 구조로 접힐 것인지를 알아보는 방식이다. 단백질 folding 문제와 밀접하게 관련된 문제에는 기질(substrate) 결합(binding), 효소 반응, 세포막과 세포막 단백질 및 단백질-DNA의 모사 등이 있다.

## 6. 분자 발전: 계통 발생학 (phylogenetic) tree의 구성

지금까지 설명한 내용이 주로 하나의 세포나 하나의 조직에 관련된 것이라면, 이제부터 설명하고자 하는 것은 생명의 또 다른 측면이다. 여기에는 하나의 가정이 들어간다. 즉 모든 생물체는 하나의 공통 조상으로부터 진화를 했다는 것이다. 다시 말하면, 각각의 생물체는 지구의 역사와 비슷한 기간동안 우연한 돌연변이의 발생과 환경의 영향 등에 의해서 점차 다른 개체로 나누어 진 것이다. 따라서 현재에도 각각의 생물체들을 보면 여러 공통점을 가진 그룹으로 분류(taxonomy)를 할 수 있다. 이 분류는 70년대 후반부터 급속히 발전한 분자생물학의 발전으로 생물체를 구성하는 생체 고분자(DNA, RNA, 단백질)를 기준으로 하여 이루어진다. 즉 각각 다른 생물체간의 정보의 공통점을 통계적인 방법으로 수량화하여 서열이 유사한 생물체들끼리 그룹을 지어 만들어진 tree를 phylogenetic tree라고 한다. 이러한 phylogenetic tree를 통해 우리는 다른 생물체들 간의 진화적 관계들의 정보와 관심 있는 단백질이나 유전자가 어떠한 식으로 나누어지며(divergence) 진화를 했는지 이해 할 수가 있다. phylogenetic tree에 사용하는 데이터의 타입은 두 가지가 있는데 DNA나 단백질 서열을 문자로 인식하여 배열하고 이를 비교하는 방법인 character-based 방법과 이들 서열 데이터를 유사성 검색 프로 그램(BLAST)을 이용하여 서열간의 부동성(dissimilarity)을 구한 후 이것을

matrix로 전환시켜 phylogenetic tree를 만드는 distance-based 방법이 있다. 그리고 tree를 만드는 방법에는 크게 optimality approaches와 clustering 알고리즘이 있다. 따라서 tree를 만들 때 '어떤 데이터 타입을 쓸 것 인지'와 '어떤 방법으로 만들 것인가'에 따라서 다음 과 같은 방법들이 있다. parsimony 방법과 maximum likelihood 방법은 데이터 타입이 character이며 optimality 방법으로 tree를 만든 경우이고, minimum evolution 방법과 least squares 방법은 데이터 타입이 distance 이며 optimality 방법으로 만든 tree를 만든 것이다. 그리고 UPGMA방법과 neighbor-joining방법은 데이터 타입은 distance이며 clustering 알고리즘으로 tree를 만든 것이다. 각각의 방법들 중에서 여기서는 가장 많이 사용하는 parsimony 방법에 대해 설명하면 다음과 같다. Parsimony 방법은 하나의 염기 서열을 다른 것으로 변형시키기 위해 필요한 변화의 수, 즉 돌연변이적 거리를 최소화하는 점에 기반을 두고 있다. 유전학적으로 관련된 염기 서열들의 돌연변이적 거리의 set이 주어지면, 염기 서열들 간의 유전학적 관계를 나타내는 phylogenetic tree의 재구성이 가능하다. 이 방법의 단점은 단순하고 직관적이며 잘못된 것을 맞게 할 가능성이 높다는 것이다.

## 7. SNP(single nucleotide polymorphism)

Phylogenetic tree와 같이 하나의 세포나 하나의 조직, 또는 하나의 개체가 보여주는 데이터가 아닌, 서로 얽혀있는 집단이 시간에 따라 변해 가는 데이터가 또 있다. 사람의 경우 부모가 그 다음 자식에게 유전 정보를 물려줄 때는 부모와 같은 것을 물려 주는 것이 아니라 자신이 가진 두벌의 염색체(상동 염색체)를 뒤섞어서(cross-over) 물려 주는데, 이것이 일어나는 상동 염색체의 위치는 매우 우연한 위치에서 일어나며, 보통 한 세대 당 한 번 정도의 매우 적은 횟수로 일어난다. 즉, 부분적으로 원래의 연결 상태가 여러 세대가 지나도 그대로 유지가 된다. 만약 이 것을 카드에 비유를 하자면, 새로 산 카드는 스페이드와 하트와 클로버와 다이아몬드 각각이 순서대로 섞이지 않은 상태이다. 만약 이것을 극히 적은 횟수로 위에서 아래로 넣는 방식으로 섞는다고 생각 하자. 그 후에 중간의 어떤 위치의 카드를 빼낸다면 그리고 빼낸 카드가 하트라면 그 빼낸 카드의 섞은 후의 위치에서 바로 위 카드나 바로 아래 카드는 하트일 가능성이 높다. 즉 여기서 카드 전체는 유전체에 해당되고 스페이드와 하트와 클로버와 다이아몬드는 특정 유전자라고 볼 수가 있고 카드를 섞는 것은 상동 염색체 간의 뒤섞임(cross-over)에 해당하며 임의의 카드는 바로 genetic marker에 해당한다고 볼 수 있다. Genetic marker는 길 안내 표시 판에 해당 되는 것으로 우리가 관심 있는 유전자가 염색체상에 어디에 있는 것인지를 안내 해주는 표시 판 역할을 한다. Marker는 여러 가지 조건이 있는데 우선 너무 자주 돌연변이(mutation)가 생기거나 너무 돌연 변이가 생기면 안 되며 사람의 경우 5-10 세대당 바뀌는 것이 적당하다. 그리고 이들 marker들은 유전체(genome)상에 많을수록 그리고 bialleric 할수록 좋다. 이러한 조건들과 가장 잘 일치하는 것이 바로 SNP이다. SNP이란 DNA상의 서열의 개인적인 차이점들 중에서 하나의 염기의 차이를 말하는 용어이며, 현재 개발된 high-throughput sequencing으로 찾아 낼 수 있다. 이렇게 각각 개인별로 자기가 가진 유전 정보 중에서 SNP을 찾아 SNP에 대한 개인 목록을 만들고 이 SNP

목록을 가능한 한 모든 다른 여러 사람들에 대해서 만들고, 이들 목록을 서로 비교하여 특정 SNP에 대한 연관 관계가 가까운 사람들끼리 분류를 한다고 생각하자. 그러면 특정 유전자 SNP의 연관 관계가 비슷한 사람들끼리는 특정 유전자 조각에 대해서는 일란성 쌍둥이가 될 것이다. 그 다음은 이 개개의 SNP들과 질병들 사이의 관계를 찾아야 하는데 이것은 각 개인의 의료 기록을 이용하면 된다. 즉, 특정 SNP을 가진 사람들이 특정 질병에 걸린다면, 거기에 속하는 사람은 특정 질병이 발병을 안 했더라도 발병할 확률이 높은 것으로 추측할 수 있는 것이다. 이러한 방법을 통한 진단을 맞춤 의학(personalized medicine)이라고 하며, 현재 연구중인 방법들 중에서는 SNP을 marker로 이용한 방법 뿐이다. 이 방법을 위해서는 개개인의 의료기록이며 이것을 얼마나 정확하고 효율적으로 데이터베이스화 하는 일과 그리고 연관관계를 정확하게 찾는 방법들이 매우 중요하다. 현재 DNA 구조를 밝힌 제임스 왓슨이 나서서 개인의 의료 기록과 DNA를 기부하는 Human Gene Trust라 명명된 프로젝트를 진행중이다. 그리고 질병과 SNP간의 연관관계를 찾는 일도 DNA chip 데이터 분석의 경우와 같이 단순한 일대일 대응 관계가 아니고 매우 복잡하다. 마지막으로 맞춤의학의 한가지 예를 들면, 이러한 유전정보 조각별 일란성 쌍둥이들이 보여주는 약에 대한 반응들을 토대로 하는 것이다. 즉, '나와 이 부분이 유전정보 조각별 일란성 쌍둥이들인 사람들은 거의 모두 이 약에 거부 반응을 보인다'와 같은 것을 찾아낼 수 있게 되는 것이다.

## 8. 데이터베이스와 데이터베이스의 통합

유전자의 염기 서열을 포함한 다양한 생물학적 정보가 급격히 증가하고, 이에 대한 연구가 활발히 이루어짐에 따라 생물학적 정보에 대한 수요도 전 세계적으로 발생하고 있다. 최근에는 DNA 및 RNA 염기 서열뿐만 아니라 단백질, 거대분자 구조 등의 데이터베이스도 인터넷 상에서 찾을 수 있다.

이를 위해서는 실험 결과를 분석하기 위해 사용되는 소프트웨어의 표준화도 고려해야 한다. 이러한 데이터베이스의 중요도는 이들이 가지고 있는 데이터의 양이 얼마나 빨리 증가 하는가로 판단할 수 있다. 생물정보학에서 주로 사용되는 실험 방법이 분석 및 모사 알고리즘이라는 것을 볼 때, 실험실에서의 실험 결과 뿐만 아니라 인터넷 데이터베이스의 자료가 실험 재료로 사용된다. 따라서 인터넷 데이터베이스와 이것을 지원하는 데이터 서비스는 이제 생물정보학에서 없어서는 안 되는 도구이며, 모든 분자 생물 과학에서도 꼭 필요한 존재가 될 것이다. 그러나 세계 여러 곳에서의 실험 결과는 실험 조건 및 방법이 다르므로 원하는 정보를 빠른 시간 내에 찾는 것도 하나의 중요한 연구 과제로 볼 수 있으며, 데이터베이스의 표준화도 필요하다.

기존의 자동화된 생물학 데이터베이스는 분리되어 있을 때보다 상호 연결되어 있을 때 더 유용하다. 이러한 이유는 앞에서 말한 바와 같이, 생물체의 정보들은 각각이 여러 계층으로 나누어지는데, 각각의 계층은 각각의 데이터를 가지며, 또한 이들 각 계층이 유기적으로 연관이 되어 있기 때문이다. 따라서 생물체에 대한 데이터들은 당연히 각각의 계층들간에 서로 유기적으로 연관이 되어 있어야만 한다. 그러나 생물학 데이터베이스를 구축하고 자료를

넣는 전문적 기술은 연구소 한 곳에 있는 것이 아니다. 그러므로 생물학 데이터베이스는 다양한 연구팀들이 여러 곳에서 다양한 목적으로 서로 다른 데이터 모델과 지원 데이터베이스 관리 시스템을 사용해서 구축된다. 이로 인해서 각 데이터베이스가 가지는 단어의 개념이 다르며 각 데이터베이스의 구성 또한 flat text file이나 관계형이나 객체 지향 방법 등으로 다양하게 만들어져 있다. 그리고 각 데이터베이스의 query도 다르며 semantics도 다르다. 그 결과 이들이 가지고 있는 관련된 자료를 연결하는 것은 수월하지 않다.

데이터베이스 통합을 위한 접근 방식에는 두 가지가 있는데, 하나는 '데이터 창고'라고도 하는 다양한 주요 데이터베이스의 복합적인 자료를 포함한 거대한 데이터베이스의 구축(consolidation)이고 또 다른 하나는 기존의 독립적인 데이터 베이스와 연결(link)하는 방법(federation)이 있다. 거대한 데이터베이스의 구축의 경우, 다른 데이터베이스로부터 앞으로 어떻게 수정될지도 모르는 자료를 복사하는 식의 통합 시도는 매우 어려운 일이다. 이 방법의 대안으로는 데이터 통합에 필요한 지식을 데이터베이스 query 도구에 넣는 것에 달려있다. 이 도구는 관련된 데이터베이스로 자동적이나 반자동적으로 적절히 형성된 query를 보내며, 회수한 자료를 사용자에게 논리적인 보고서로 통합하는 능력이 있어야 한다. 기존의 독립적인 데이터베이스간을 연결시키려면 각각의 데이터베이스에 공통적인 query가 있어야 하고 공유한 데이터들간의 semantics도 유사해야만 한다. 그리고 각 데이터베이스 간의 연관된 개념에 대한 계통도(thesaurus)를 만들어야 하고 각각의 레코드(record)들 요소 중 중요한 필드(field)부분은 같아야 한다. 현재 가장 많이 사용되는 데이터베이스 중의 하나인 ENTRZ나 SRS의 경우 federation 방식의 초기 형태로 구성되어 있다. 그러나 federation방식의 경우에는 n개의 데이터 베이스를 통합하기 위해서 $O(n^2)$이 필요하다. 따라서 이 문제의 해결을 위해 현재 global schema를 통하여 O(n)으로 감소 시키려는 노력이 진행이 되고 있다. 한편으로 최근에는 데이터 자체를 표준화 시키기 위해서 XML(Extensible Markup Language)을 이용한다. XML은 어떤 종류와 내용을 가진 데이터라도 사용자가 정의한 markup language를 이용해서 저장이 가능하도록 특별히 설계되어있고, 이것으로 데이터를 만들 경우에는 단순한 주제어 검색이 아닌 hierarchical 구조를 제공하여 완벽한 parsing을 가능하게 한다. 따라서 현재 이것을 이용하여 internet DB의 개발이 이루어지고 있다.

Internet에 존재하는 sequence 데이터베이스로는 NCBI, EBI, DDBJ의 세 가지의 대형 데이터베이스 센터가 있으며, 이외에도 SWISS-PROT 등의 여러 가지가 있다. 인터넷의 분자생물학 DB의 최근 목록은 Nucleic Acids Research의 2001년 1월호에서 찾아볼 수 있다("The Molecular Biology Database Collection: an online compilation of relevant database resources", *Nucleic Acids Research*, 2001, **29**(1), http://nar.oupjournals.org/cgi/content/full/29/1/1).

## 9. 화학공학자의 역할

단세포나 여러 세포로 이루어진 조직이나 여러 조직으로 이루어진 기관들은 화학공학적인 관점에서 보면 물리화학적 공정이 직렬(serial)이나 병렬(parallel)으로 연결되어 있다고 생각을 할 수가 있다. 이러한 관점을 통해 많은 화학공학자들은 생명공학의 많은 문제들을 풀어왔

다. 최근 들어 전자공학, 기계공학, 분석 기술, 생화학, 나노 기술, 고분자 화학, 재료 과학의 발달에 의해서 high-throughput 기술이 발전이 되었다. 이 기술에 의해서 생명 공학은 정보 혁명이 일어나고 있다. 즉 앞서 말한 세포 내의 여러 고분자 물질들(DNA, RNA, 단백질)에 대한 막대한 양의 정보를 매우 손쉽게 얻을 수가 있게 된 것이다. 이러한 막대한 정보를 가지고 현재 생명 공학은 생물정보학이라는 방법을 이용하여 신약 개발과 맞춤의약 등 여러 가지 분야에 도전을 할 수 있게 한다. 생명 공학의 정보 혁명은 화학공학자들에게 두 가지 분야에 도전의 기회를 제공한다. 하나는 high-throughput 기술의 개발이고 나머지 하나는 high-throughput 기술을 이용하여 생물 정보를 만들고 이를 이용하는 분야이다. 먼저 high-throughput 기술 개발에 있어서는 DNA chip을 제작하는 경우를 생각해 보자. 이 경우 DNA를 부착시키기 위한 소재 개발에 대한 문제가 있다. 현재로는 유리나 고분자 재료를 사용하고 있으며, DNA가 보다 잘 결합하는 소재의 개발이 필요하다. 소재의 개발 역시 화학 공학의 분야에 포함이 된다. 그리고 현재 가장 많이 사용하고 있는 DNA chip인 Affymetrix사의 GeneChip의 경우에는 반도체 가공 공정과 유사한 photolithography방법을 이용하여 chip을 만들고 있다. 이러한 표면 반응 역시 화학공학의 분야에 포함된다. 또한 시료를 chip 위에 올려 놓는 경우에 일부 DNA chip을 제외하면 자연적으로 유체 내의 이동에 의해 DNA간의 결합이 이루어지도록 기다리게 된다. 이렇게 기다리는 시간을 단축하는 문제는 이동현상의 문제가 된다. 현재 사용되고 있는 DNA chip에서 발전한 개념이며, 실용화를 위해 활발히 연구하고 있는 lab-on-a-chip에서는 하나의 실험장치를 이용해 실험전 처리부터 결과 분석까지를 수행하려고 한다. 여기에는 microvalve 등이 사용되며, 이를 위해서는 역시 미세한 구조에서의 이동현상, 그리고 현재 반도체 가공 공정에서 사용되고 있는 것과 유사한 표면 반응 등의 처리가 필요하다. 그리고 실험에서 검색하고자 하는 DNA와 결합하는 DNA의 쌍을 찾고 그것을 복제 등을 통해 제조하는 것은 생물화공 분야에서 해결해야 할 문제다. DNA chip 및 microarray의 결과를 통계적으로 처리하는 방식은 화학공정에서의 다변량 데이터 처리 방식을 확장하는 방식으로 해결할 수 있다. 한편 나머지 하나는 high-throughput기술을 이용하여 생물 정보를 만들고 이를 이용하는 분야에 대해 어떤 것이 있을지 생각해 보자. 우선 mRNA와 단백질 발현 데이터를 가지고 여러 유전자의 기능을 알아내는 것을 보면 이것은 화학 공학에서 공정 확인과 비슷하다. 즉, 세포가 받는 환경을 입력 변수라고 생각을 하고 이로 인하여 유전자의 발현이 달라지게 되는데 이것은 단백질이나 mRNA의 양 변화로써 알 수가 있다. 이러한 양의 변화를 출력 변수라고 본다면 우리는 세포에 대한 전달 함수를 구할 수가 있을 것이다. 즉, 세포를 하나의 공정으로 보고 공정에 대한 전달 함수를 구해서 수학적 모델링을 할 수가 있는 것이다. 모델이 성공적으로 구해 진다면 특정 입력변수를 가지고 우리는 전달 함수를 통해 출력 변수를 알아 낼 수가 있다. 따라서 입력 변수인 특정 환경에 의해 발현되는 유전자와 출력 변수인 세포의 상태를 알 수가 있게 되는 것이다. 그리고 생물체 내에서 일어나는 전사와 번역 기작의 조절은 매우 정교하게 이루어지는데, 이것의 경우 화학공학에서 공정 제어의 원리와 공정 제어 방법을 바탕으로 조절 기작을 수학적 모델링에 바탕을 둔 분석을 할 수가 있다. 한편 전사와 번역의 과정을 보면 이것은 화학 공

학에서 고분자의 중합 공정과 유사하다. 차이점은 생물체 내에서 이루어지는 것은 기존의 화학 공학보다 매우 복잡한 형태로 이루어 져있다는 점이다. mRNA는 DNA로부터(전사) 그리고 단백질은 mRNA로부터(번역) 합성이 된다. 이것은 반응기가 직렬로 연결된 경우라고 생각할 수가 있고 여러 종류의 단백질들의 합성은 이러한 직렬로 연결된 반응기들이 많은 수로 병렬로 연결이 된 것이라고 생각을 할 수가 있다. 또한 고분자에서 말하는 단량체는 생물체에서는 핵산(nucleic acid)과 아미노산으로 볼 수가 있다. 특히 고분자 공정과 비슷한 점은 생물체 내에 중합된 단백질이나 mRNA는 그들의 생산물 자체가 다시 중합을 위한 촉매(autocatalytic)로써 쓰인다는 점이다. 즉, 단백질-mRNA, 단백질-DNA, 단백질-단백질 간의 상호 작용을 하면서 단백질과 mRNA의 중합 과정인 전사와 번역을 조절 하는 것이다. 이렇게 상호 작용을 하는 문제는 화학공학에서 열역학에 해당이 된다. 그러나 생체 내에서 전사와 번역 과정은 알려지지 않은 물리 화학적인 여러 요소가 관여된다. 또한 반응 메커니즘은 nonelementary reaction kinetics가 대부분이고, 비선형성을 가지며 시간에 따라 세포 내 공정이 달라지며 반응을 하는 환경도 기존에 배운 화학공학과는 다른 점도 있어 쉽지 만은 않다. 또한 단백질의 3차 구조 예측의 문제 경우에도 화학공학의 고분자 설계(polymer design)분야와 비슷하다. 보통 X선 결정 방법이나, NMR에 기초를 두어서 이미 3차 구조를 아는 단백질의 아미노산 서열과 아직 3차 구조를 모르는 단백질의 아미노산 서열을 서로 비교하여 유사한 것을 찾는 것이다. 이 경우 화학공학에서 다루는 보통의 고분자와는 달리, 생물체 내에 있는 단백질은 20개의 다른 아미노산 단량체로 이루어진 고분자이기에 복잡하며, 특히 2차 구조와 3차 구조의 폴딩(folding)과 모티프(motif)의 구조가 세포 내 환경에 민감하게 변화하므로 예측하기가 매우 어려운 점이 있다. 지금까지 살펴본 내용을 정리하면 아래 표 2와 같으며, 생명공학 분야의 여러 문제는 화학공학의 중요 연구 분야와 매우 유사하다. 따라서 화학 공학자들이 관심을 가진다면 생물정보학에 그리고 궁극적으로 생명 공학 분야에 매우 중추적인 역할을 할 것으로 보인다.

| 화학공학 | 생명공학(생물정보학, 유전체학, 단백질학) |
|---|---|
| 공정 확인(Process identification) | mRNA와 DNA 발현 데이터의 해석 |
| 공정 제어(Process control) | 생체 내 효소의 조절<br>전사 및 번역의 조절 |
| 고분자 공정(Polymer processes) | DNA와 mRNA와 단백질의 합성<br>DNA, 단백질, lab on a chip의 제작 공정 |
| 고분자 설계(Polymer design) | 단백질 구조 분석 및 모사<br>DNA와 RNA의 2차 구조 분석 및 모사<br>lead compound 개발 및 신약 개발<br>docking의 예측 |
| 반응 공학(Chemical reaction networks)<br>환경 화학(Atmospheric chemistry)<br>고분자 화학(Polymer chemistry)<br>촉매 공학(Catalysis) | 단백질, DNA, RNA의 상호작용 분석<br>세포 내 반응 경로 분석<br>세포 내 대사회로의 통합 |
| 유체 역학(Fluid mechanics)<br>이동 현상(Transport phenomena) | 세포간의 신호 전달 체계 분석.<br>DNA, protein, Lad on a chip 개발 |
| 열역학(Thermodynamics) | Cellular energetics<br>세포 구성 물질의 물리화학적 성질 분석 |
| 생물 화학 공학(Biochemical engineering)<br>대사 공학(metabolic engineering)<br>조직 공학(tissue engineering) | 거의 모든 분야를 포괄.<br>High-throughput기술을 이용하여 세포 내 대사 회로의 분석 및 조절<br>조직 배양을 통한 이식, |

**표 2 생명 공학에서 화학공학과 연관성이 많은 분야**

생물정보학은 생물학, 컴퓨터 과학, 응용 수학, 통계, 컴퓨터 및 소프트웨어 프로그래밍의 사이에 위치한다. 생물학자들은 자료 관리 및 분석 프로그램에 대한 빠른 해결책을 원하고, 컴퓨터 과학자와 수학자들은 그들에게 흥미가 있는 기본적인 연구 문제를 찾으며, 소프트웨어 프로그래머들은 필요한 프로그램을 만들기 위해 충분히 잘 정의된 문제를 앞의 두 부류에 요구한다. 이러한 형식의 문제들은 생물정보학에서만 유일하게 나타나는 문제는 아니며, 일반적으로 다분야 전공이 필요한 분야에서 나타나는 문제이다. 이와 같은 문화의 차이는 세 그룹 사이의 용어와 과학적 접근 방식의 양식의 상당한 차이와, 이 차이를 줄이기 위한 노력을 세 그룹이 모두 과소 평가한다는 이유로 비롯된다.

이런 문제는 다분야 유전체(genome) 과학이나 분자생물학을 하고자 하는 수학자와 컴퓨터

전공자들을 위한 특별 장학 프로그램, 일부 대학에서의 생물정보학이나 전산생물학 대학원 프로그램의 설치, 여러 대학에서의 생물정보학 과목 개설 등에 의해서 바뀌고 있다. 그러나 화학 공학에서는 이미 컴퓨터, 생물, 수학의 세 가지 분야에 대한 지식을 사용하고 충분히 활용하고 있다. 따라서 세 분야의 전문가들과 함께 생물정보학을 연구하면서 중간자적 시각에서 문제를 해결하는데 도움을 줄 수 있을 것이며, 앞으로의 화학공학자들이 도전해 볼 분야로 생각된다.

참고 서적

Vassily Hatzimanikatis. Bioinformatics and Functional Genomics: Challenges and Opportunities. *AIChE journal* 2000; 46; 12: 2340-2343.

Won seyeun. Current Trend in Bioinformatics. *Recent Advances in Bioprocess Engineering* 1998; 6: 203-217.

Nicholas M Luscombe, Dov Greenbaum & Mark Gerstein. What is bioinformatics? An introduction and overview. http://bioinfo.mbb.yale.edu/~nick/bioinformatics/ 20001.

Andreas D. Baxeveanis, B.F.Francis Ouellette. Bioinformatics A Practical Guide to The Analysis Of Genes And Proteins. Wiley, New York 1998.