시스템 공학 입장에서 본 생물 정보학

권성우, 이병우*, 이희석**, 한종훈
포항공과대학교 화학 공학과, *삼성.SDS, **포항공대 환경 공학과

1. 서론

생물체와 이들로 이루어지는 생물계는 다른 무생물적인 대상에 비해 훨씬 더 정보집약적이라고 할 수 있다. 하나의 생물체를 예로 들면, 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 상호 작용부터 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 정보학적 접근이 그 핵심을 차지할 수 밖에 없지만, 과거 생명공학은 대규모 정보를 생물체들로부터 얻어 내는 실험 방법과 도구가 부족했다. 따라서 생물 데이터를 얻을 수 있는 실험 도구의 문제 였다. 이러한 문제는 고효율 기술 및 기기 (DNA chip, Megabase1000, MALDI-TOF 등)의 발명으로써 해결이 되었다. 고효율 기술과 기기의 발전은 그림 1에서와 같이 생물학적인 데이터들이 급속도로 증가하게 했고, 이로 인해 많은 양의 데이터를 다루고 분석하기 위해서 컴퓨터와 정보학적 방법을 사용하기 시작 했다.

생물정보학은 생물학와 정보학를 결합하여 만든 단어이다. 현재 그 정의는 생물학을 분자 수준에서 보는 것인데 이들 생물체의 분자들이 가진 데이터 들에 정보기술을 적용 시켜 대량의 새로운 정보(DNA, 단백질의 기능, 질병과의 연관성 등)를 얻어내고 많은 생체 분자(DNA, mRNA, 단백질)들 사이의 연관된 정보(DNA-단백질, 단백질-단백질, mRNA-DNA와 mRNA-단백질 상호 작용 등)를 유기적으로 재구성 하여 생명체에 대해 이해 하는 것이다. 즉, 생물 정보학 이란 생물학과 관련된 정보를 얻기 위한 모든 전산, 수학, 통계적인 방법이나 접근 방식에 대한 학문이라 할 수 있다[그림 2].

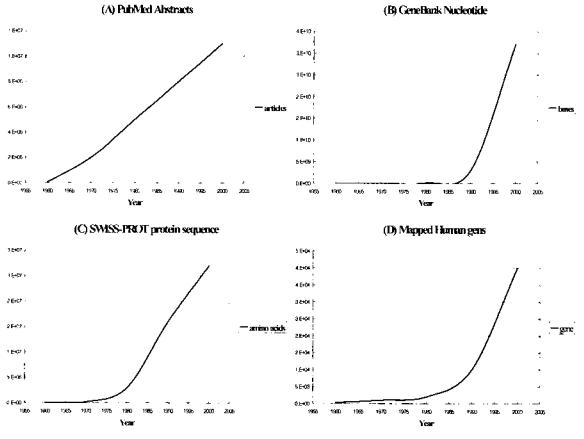

그림 1. 생물학 데이터의 증가

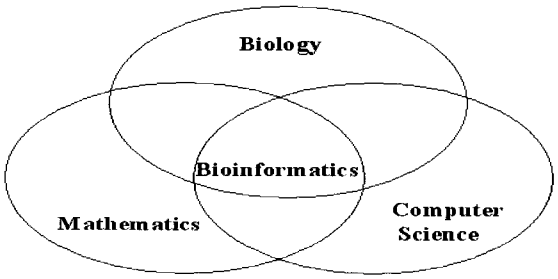

그림 2. 다 학제 분야인 생물 정보학

생물 정보학의 목적은 크게 세 가지로 나눌 수 있는데, 첫째는 연구자들이 쉽게 검색을 할 수 있도록 기존에 있는 데이터를 데이터베이스화 시키는 일이나 기존의 데이터베이스를 유기적으로 연결시키고 연구자들이 만들어낸 새로운 데이터를 여기에 추가하기 쉽게 하는 일이다. 예를 들자면 ENTREZ(http://www.ncbi.nlm.nih.gov/Entrez/)가 여기에 해당이 된다. 둘째로는 데이터들을 분석하는데 있어서 도움이 되는 방법과 수단을 개발하는 것이다. 예를 들자면 bio chip에서 이미지 분석 하는 것이 해당된다. 셋째로는 위의 데이터 분석 방법들을 이용하여 생물학적으로 의미가 있는 결과를 추론해 내는 것이다. 예를 들자면 bio chip의 결과를 가지고 패턴 인식 작업을 통해 유전체 및 단백질체에 대하여 연구 하는 것이 여기에 해당이 된다. 결국 고전적 생물학

에서는 개별 시스템에 대해서 혹은 아주 소수의 관련된 것들을 비교하는 방식으로 연구가 진행 되었지만 생물정보학에서는 생물체의 데이터를 포괄적으로 분석하여 모든 생물 시스템 혹은 특정 생물 시스템이 가진 원리를 밝힐 수가 있게 된 것이다.

2. 생물 정보학의 핵심 분야

2.1. 염기 서열의 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. 예를 들자면, DNA 염기 순서 결정 장치에서 뉴클레오티드가 겔 속을 흘러가면서 내는 형광 신호를 해석하여 영상 자료를 분석하는 것을 들 수가 있다. 현재의 시스템의 효율을 높이려면 향상된 정확도, 인식할 수 있는 염기 서열의 길이, 계측의 신뢰도를 주는 염기 판독 알고리즘이 필요하다. 이로써 다시 DNA 염기 서열의 조합과 마무리를 자동화하는데 도움이 된다. 특히, 실험 자료를 환경이 다른 여러 연구소가 서로 공유하면서 여러 종류의 소프트 웨어로 분석하려면, 자료 표현 기준이 중요하다. DNA의 염기 서열 판독은 DNA 염기 순서 결정 시스템의 한계1000bp이상의 DNA조각을 기계에 넣으면 700bp이후의 DNA의 시그널은 약해져서 시그널이 ACGT중 어떤 것의 시그널인지 제대로 판별할 수 없다로 인해 DNA를 잘라서 클로닝한 후에 각각에 대해 여러 번의 염기 순서 결정 작업을 하게 된다. 그 후에 각각의 염기 서열을 결정한 결과를 최종적으로 하나로 합쳐야 하는데, 이때 염기 서열 조합 알고리즘이 필요하다. 그러나 서열 판독 기기로 읽은 각각의 염기는 실험 오차의 영향을 받으므로 대규모의 염기 서열을 판독하려면 다음과 같이 여러 면에서 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 대한 신뢰도 (확률) 예측치 부여 (4)추가적인 보충 정보(clone 길이 등)의 사용 (5)궁극적으로는 사람의 조작이 없이 염기 판독(base calling)으로부터 최종 염기 서열의 조합 및 분석까지의 모든 공정을 자동화하는 것이 바람직하다.

2.2. 유전체의 염기 서열에서의 기능 예측 및 정렬

유전자의 염기 서열은 전사과 번역을 통해 단백질을 만들어 낸다. 이들 단백질은 다른 단백질이나 DNA등과 상호 작용을 통해 세포 내에서 혹은 세포 간의 생명 작용을 결정하게 된다. 그러나 한 개 유전자의 기능을 실험을 통해서 밝히는데 드는 시간이 보통 10년이 걸리므로 수많은 염기 서열에서의 상세한 기능을 다양한 실험을 통하여 결정한다는 것은 현실적으로 불가능하다. 따라서 실험을 하기 이전에 유전체에 대해 관심 있는 기능을 가진 후보 유전자들을 선택한 후 실험을 통해 확인하는 것이 효율적이다.

이렇게 염기 서열에서의 기능을 예측하는 방법으로 많이 사용되는 FASTA (http://fasta.bioch virginia.edu/)는 염기 서열 데이터베이스에 있는 서열간의 유사성을 검색하는 방법이다. FASTA 이외에도 특정한 기능을 한다고 알려져 있는 서열과 기능을 포함하지 않은 서열을 인공 신경망으로 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 유사성 점수(similarity score)는 데이터베이스 크기의 로그에 비례한다. 하지만 두 가지의 염기 서열을 비교함으로써 생성되는 유사성 점수(similarity score)는 이들이 발견된 데이터베이스 크기와는 무관하다. 그러므로 데이터베이스가 커짐에 따라 생물학적으로 의미를 가지지만 관련성이 약한 염기 서열은 무작위로 이루어진 정합(match)보다 작은 유사성 점수(similarity score)를 가지게 될 수 있으며 이로 인해서 노이즈에 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.

2.3. Bio chip 기술

많은 양의 유전체 데이터를 분석하기 위해서는 새로운 방식의 분석 기술이 필요하다. 이러한 기술들 중에서 대표적인 것이 바로 Microarray 기술이다.

Microarray 기술에는 DNA chip, protein chip, lab-on a chip등이 있다. DNA chip은 고형체에 고정 시킨 DNA와 mRNA나 다른 DNA를 잡종형성(hybridization) 시켜 만들게 되는데 특정 상태의 유전자 발현 양상을 연

구 할 때 사용하고 있다. 한편 protein chip의 경우 고형체에 단백질을 고정화 시킨 것으로 단백질간의 상호작용을 연구할 때 사용 할 수 있으며 lab-on a chip의 경우 protein chip 이나 DNA chip 과는 달리 chip 실험을 하기 위한 전처리 과정이 chip 위에서 일괄적으로 이루어져 실험이 편리하고 정확하다는 장점이 있다.

DNA chip의 주요 연구 분야를 세가지로 나누자면, 첫째 미소 제작(micro fabrication)과 이미지 분석, 둘째 데이터 분석 및 의미 있는 유전자 발굴, 셋째로 생명공학에 응용이다. 미소 제작(micro fabricate) 분야의 경우 DNA chip 제작 방식을 말하는 것으로 크게 핀이나 잉크젯을 이용하여 DNA를 고형체 위에 점재하는 인쇄(printing) 방법과 photolithography를 이용하여 올리고뉴클레오티드(oligonucleotide)를 고형체 위에 합성하는 방법이 있다. 인쇄(printing) 방식의 경우 고정화 시키는 DNA의 양이 적다는 단점이 있지만 chip 제작 단가가 저렴하다는 장점이 있다. 한편 photolithography 방식의 경우 고밀도로 DNA를 고정화 시킬 수 있는 장점이 있지만 bio chip 제작 과정에서 사용 되는 photo-mask의 단가가 비싼 단점이 있다.

위의 방법을 통해 만들어진 DNA chip을 잡종화(hybridization)하면 여러 점들이 나타나는데 이 점의 색의 강도와 크기를 정량화 시키는 것이 바로 이미지 분석에 해당이 된다. 잡종화(hybridization) 결과로 수천 개에서 수 만개의 점이 나타나고 이들 색의 강도 결정을 해야 한다. 또한 결정된 색의 강도에 따라서 수치값으로 정량화 하는 것까지가 이미지 분석 단계이다. 그러나 이러한 작업을 사람이 하기엔 불가능하다. 따라서 반드시 컴퓨터 프로그램을 이용하여 점의 크기 와 점이 가지는 색의 강도를 분석한다.

한편, 데이터 분석 및 의미 있는 유전자 발굴이란, 이미지 분석을 통해 수 천 개에서 수 만개의 점을 정량화 시킨 데이터를 다변량 통계나 패턴 인식을 이용하여 분석하는 것을 말한다. 많은 경우에 분석을 보다 간단히 하기 위해서 클러스터링 기법을 이용하여 비슷한 패턴을 보이는 유전자 데이터를 분류한다. 이렇게 많은 양의 데이터를 통계적으로 처리하는 방법으로는 주성분 분석(PCA), 의사 결정 나무(decision tree), 군집화 분석(hierarchical clustering), 자기 조직화(SOM), 인공 신경망과 유전자 알고리즘 등이 있다.

DNA chip이나 protein chip이 고정된 포맷을 가지고 한번에 한가지 일을 수행하는 것과 달리 lab-on a chip은 미세 가공 기술을 이용하여 실험에 필요한 시료 희석, 혼합, 반응, 분리, 정량등 모든 단계를 하나의 칩 위에서 수행하는 기술을 말한다. 즉 일반적으로 생화학 물질의 분석시 사용되는 자동 분석 장치의 시료 전처리 과정에 필수적인 펌프, 밸브, 반응기, 추출기, 분리시스템 등의 기능과 센서 기술을 같은 칩 위에 접목시킨 것이 lab-on a chip이다.

최근 lab-on a chip의 개발 동향은 미세유체 역학과 관련된 MEMS(Micro Electro Mechanical System) 기술을 기존의 분석 기술에 접목시킴으로써 수 나노 리터에 해당하는 적은 양의 액체 시료를 단위 칩 상에서 다룰 수 있도록, 시료 분석에 필요한 모든 구성 요소를 소형화와 결합을 사용하여 하나의 칩 위에 올리려 하는 추세이다.

위의 bio chip 제작 및 데이터 분석 기술을 토대로 bio chip은 수천 개에서 수 만 개의 유전자 발현 양상을 연구할 수가 있다. 이것은 생명체의 특정 상태(예를 들어 질병)와 1:1 대응을 가지는 유전자의 발견할 수 있다. 더욱이, 수 천개에서 수 만개의 유전자들의 특정 상태에 대한 연관을 동시에 bio chip을 이용해서 연구함으로써 현재까지 밝혀 내지 못한 특정 상태에 관여하는 유전자들의 움직임을 한 눈에 알아 볼 수 있게 된다. 한편 이때 많은 유전자들이 동시에 변화하므로 이들 변화하는 유전자 중 어떠한 것이 특정 상태와 관련하여 상관성 있게 변화하고, 어떤 유전자들이 특정 상태와는 상관없이 변화하는가 하는 것을 정확히 분류하여 알아 낼 수 있는 생물정보학 기술이 중요한 역할을 하게 되며 이를 통하여 질병의 진단과 치료, 대사 공학에 응용과 신약 개발등에 bio chip을 응용할 수가 있다.

2.4. 구조 데이터 베이스의 탐색 및 거대 분자 구조의 결정

최근 단백질 공학, 결정학, 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 알고리즘들은 단백질 구조를 기존에 알려진 모든 구조의 데이터 베이스와 비교할 수 있게 한다. 구조 데이터베이스 검색은 관심을 끄는 생물학적 관계를 발견하는 도구로서 염기 서열 데이터 베이스 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 알고리즘이 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 단백질의 3차원 구조를 완전히 명시하므로, 아미노산 염기 서열만으로 단백질의 구조를 계산하는 것은 이론적으로 가능하다. 그러나 살아있는 세포에서 발견되는 길이의 단백질 구조에 대해서 가능한 구조의 수는 천문학적으로 많아 기존의 컴퓨터로는 'conformational space'의 탐색 문제가 실질적으로 불가능하다.

최근의 다른 방법은 단백질 꼬인 (folding) 문제를 '역구조(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 것 인지를 알아보는 방식이다. 단백질 구조화(folding) 문제와 밀접하게 관련된 문제에는 기질 결합(binding), 효소 반응, 세포막과 세포막 단백질 및 단백질-DNA의 모사 등이 있다.

2.5. 분자 발전: 계통 발생학의 phylogenetic tree 구성

지금까지 설명한 내용이 주로 하나의 세포나 하나의 조직에 관련된 것이라면, 계통 발생학에는 하나의 가정이 들어간다. 즉 모든 생물체는 하나의 공통 조상으로부터 진화를 했다. 다시 말하면, 각각의 생물체는 지구의 역사와 비슷한 기간동안 우연한 돌연변이의 발생과 환경의 영향 등에 의해서 점차 다른 개체로 나누어 진다. 따라서 현재에도 각각의 생물체들을 보면 여러 공통점을 가진 그룹으로 분류를 할 수 있는데 이 분류는 70년대 후반부터 급속히 발전한 분자 생물학의 발전으로 생물체를 구성하는 생체 고분자 (DNA, RNA, 단백질)를 기준으로 이루어 졌다. 즉 각각 다른 생물체간의 정보의 공통점을 통계적인 방법으로 수량화하여 서열이 유사한 생물체들끼리 그룹을 지어 만들어진 트리를 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 데이터 형태의 타입은 두 가지가 있는데 DNA나 단백질 서열을 문자로 인식하여 배열하고 이를 비교하는 방법인 문자 기초 방법과 이들 서열 데이터를 유사성 검색 프로그램(BLAST)을 이용하여 서열간의 부동성(dissimilarity)을 구한 후, 군집화(clustering) 알고리즘을 통해 phylogenetic tree를 만드는 거리 기초 방법이 있다.

이와는 달리 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 세트가 주어지면, 염기 서열들 간의 유전학적 관계를 나타내는 phylogenetic tree의 재구성이 가능하다. 이 방법의 단점은 단순하고 직관적이며 잘못된 것을 맞게 할 가능성이 높다는 것이다.

2.6. 데이터 베이스와 데이터 베이스의 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 서열뿐만 아니라 단백질 등의 데이터 베이스도 인터넷 상에서 찾을 수 있다.

이를 위해서는 실험 결과를 분석하기 위해 사용되는 소프트웨어의 표준화를 고려해야 한다. 이러한 데이터베이스의 중요도는 이들이 가지고 있는 데이터의 양이 얼마나 빨리 증가하는 가로 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 안되는 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 데이터 베이스는 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 텍스트 파일 이나 관계형이나 객체 지향 방법 등으로 다양하게 만들어져 있다. 그리고 각 데이터베이스의 질의도 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 데이터 베이스의 구축이고 또 다른 하나는 기존의 독립적인 데이터 베이스와 연결하는 방법이 있다. 거대한 데이터 베이스의 구축의 경우 다른 데이터베이스로부터 앞으로 어떻게 수정될지도 모르는 자료를 복사하는 식의 통합 시도는 매우 어려운 일이다. 이 방법의 대안으로는 데이터 통합에 필요한 지식을 데이터베이스 질의 도구에 넣는 것에 달려있다. 이 도구는 관련된 데이터 베이스로 자동적이나 반자동적으로 적절히 형성된 질의를 보내며, 회수한 자료를 사용자에게 논리적인 보고서로 통합하는 능력이 있어야 한다. 기존의 독립적인 데이터 베이스간을 연결시키려면 각각의 데이터베이스에 공통적인 질의가 있어야 하고 공유한 데이터들간의 semantics도 유사해야만 한다. 그리고 각 데이터베이스 간의 연관된 개념에 대한 계통도(thesaurus)를 만들어야 하고 각각의 레코드들 요소 중 중요한 필드부분은 갖아야 하다. 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.

한편, 최근 데이터 자체를 표준화 시키기 위해서 XML(Extensible Markup Language)을 이용한다. XML은 어떤 종류와 내용을 가진 데이터라도 사용자가 정의한 markup language를 이용해서 저장이 가능하도록 설계되어있고, 이 것으로 데이터를 만들 경우에는 단순한 주제어 검색이 아닌 계층적 구조를 제공하여 완벽한 파싱을 가능하게 한다. 따라서 현재 이것을 이용하여 인터넷 DB의 개발이 이루어지고 있다.

3. 시스템 공학자의 새로운 도전과 기회

단세포나 여러 세포로 이루어진 조직이나 여러 조직으로 이루어진 기관들은 시스템 공학자 관점에서 보면 물리화학적 시스템이 직렬이나 병렬로 연결되어 있다고 생각을 할 수가 있다. 이러한 관점을 통해 많은 시스템 공학자들은 생명공학의 많은 문제들을 풀어왔다. 최근들어 전자공학, 기계공학, 분석 기술, 생화학, 나노 기술, 고분자 화학, 재료 과학의 발달에 의해서 고효율 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 제어, 자동화 그리고 시스템 공학자들에게 두 가지 분야에 도전의 기회를 제공한다. 하나는 고효율 기술의 개발이고 나머지 하나는 고효율 기술을 이용하여 생물 정보를 만들고 이를 이용하는 분야이다. 먼저 고효율 기술 개발이란 생물학적 데이터를 대량으로 얻기 위한 하나의 실험 도구 및 관련 기술 개발을 말한다. 이러한 도구들은 단기간으로 대량의 데이터를 손 쉽게 얻기 위한 목적이므로 반드시 자동화가 이루어져야 한다. Bio chip을 제작하는 경우를 예를 들자면, 현재 사용되고 있는 bio chip에서 microarray에 기반 한 DNA chip이나 protein chip의 경우 대량 생산과 실험의 재현성이 필수적이다. 하지만 아직 자동화가 부족하여 대량 생산이 어려우며, 실험자의 수작업이 많기에 재현성 떨어지는 단점이 있다. 한편 lab-on-a-chip에서는 마이크로 벨브등 여러 요소들이 필수적인데, 이를 위해서는 미세한 구조에서의 이동현상, 그리고 전과정(시료 희석, 혼합, 반응, 분리, 정량등)의 자동화 및 제어 기술이 필요하다. 그리고 Bio chip의 이미지 결과를 통계적으로 처리하는 방식은 컴퓨터 공학이나 전기, 전자 공학에서의 비전 문제에 해당하며, 많은 양의 노이즈를 제거하는 것이 핵심이다. 또한 이미지 분석 한 결과에서 필요한 생물학적 정보를 추출하는 과정의 경우 패턴 인식 문제에 해당 되며, 생명체 데이터에 적합한 패턴 인식 알고리즘 개발이 필수적 이다.

한편 나머지 하나인 고효율 기술을 이용하여 생물 정보를 만들고 이를 이용하는 분야의 경우, 우선 mRNA와 단백질 발현 데이터를 가지고 여러 유전자의 기능을 알아내는 것을 보면, 시스템 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 수 있다. 모델이 성공적으로 구해 진다면 특정 입력변수를 가지고 우리는 전달 함수를 통해 출력 변수를 알아 낼 수 있다. 따라서 입력 변수인 특정 환경에 의해 발현되는 유전자와 출력 변수인 세포의 상태를 알 수 있다. 그리고 생물체 내에서 일어나는 전사와 번역 기작의 조절은 매우 정교하게 이루어지는데 이것의 경우 시스템 공학에서 공정 제어의 원리와 공정 제어 방법을 바탕으로 조절 기작을 수학적 모델링에 바탕을 둔 분석을 할 수가 있다. 이외에도 대사회로 네트워크나 유전자 네트워크(genetic network)를 규명하는 부분에도 시스템 공학 지식을 접목 한다면 체

계적이고 합리적인 분석을 통해 생명 현상을 조명해 나갈 수 있으리라고 생각된다.

생물정보학은 생물학, 컴퓨터 과학, 응용 수학, 통계, 전산학 사이에 위치한다. 이로 인해서 생기는 세 그룹 사이의 용어와 과학적 접근 방식의 양식의 상당한 차이는 큰 문제로 지적 되고 있다. 그러나, 시스템 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 시스템 공학자들이 도전해 볼 분야로 생각된다.

참고 서적

1. Vassily Hatzimanikatis, "Bioinformatics and functional genomics : Challenges and opportunities", *AIChE journal* 2000; 46; 12: 2340-2343.
2. Won Se-Yeun, "Current trend in bioinformatics", *Recent Advances in Bioprocess Engineering* 1998; 6: 203-217.
3. Nicholas M Luscombe, Dov Greenbaum & Mark Gerstein, "What is bioinformatics? An introduction and",overview. http://bioinfo.mbb.yale.edu/~nick/bioinformatics/ 20001.
4. Andreas D. Baxeveanis, B.F.Francis Ouellette. "Bioinformatics a practical guide to the analysis of genes and proteins." Wiley, New York 1998.

학회발간 논문집 재고정리 안내

회원님 재중

안녕하십니까.

회원 여러분의 발전과 행운을 기원합니다. 우리 학회에서는 학회의 학술사업을 더욱 활성화하고 회원님들께 보다 충실한 서비스를 제공하기 위하여 최선의 노력을 하겠습니다.

이번에 우리 학회는 출판물 재고정리를 하고자 하오니 아래 목록을 보시고 필요하신 논문집이 있으시면 신청하여 주시기 바랍니다.

<목 록>

⊙ 제어·자동화·시스템 공학 논문지 : 권당 2,000원
⊙ Transaction on Control, Automation and Systems Engineering : 권당 2,000원
⊙ ISIE2001(3권) : 권당20,000원(총60,000원)
⊙ KACC('95년-2000년) : 권당 20,000원
⊙ ICCAS CD(2001년) : 개당 50,000원

-자료 요청시에는 먼저 재고가 있는지 확인해 주신 후
1) 수신자명
2) 우편번호
3) 주소
4) 해당자료 권 호(행사명,연도)
5) 연락처를 명기하여 연락해 주십시오.

<연락처>
E-mail: finance@icase.or.kr
Tel: 02-508-5801
Fax: 02-555-4746