생물정보 분야의 개괄 및 전망

생물정보학연구소 원세연

1. 서 론

지난 몇 년간 생물체와 관련된 여러 분야들, 즉 기초 생물학, 생명공학, 기초 의학, 약학, 농학, 그리고 환경과 관련된 분야 등에 매우 큰 변혁이 있었다. 이 변혁의 핵심 요인은 크게 두 가지로 대별할 수 있는데, 하나는 다양한 형태의 자동화된 high-throughput tool들을 이용하여 대량의 데이터를 생물체로부터 얻어내는 것이고, 다른 하나는 이 대량의 데이터를 컴퓨터로 처리하여 여러 가지 유용한 지식을 얻어내는 것이다. 지난 수십 년간 생명체를 대상으로 하는 연구 방식의 주류를 이루어 왔던 소위 "분자생물학적 연구방식"은 생물체를 이루고 있는 개개의 요소들을 하나 씩 분리하여 그 각각의 성질을 밝히는 식이었다면, 소위 "genomics"적인 연구방식에서는 대상이 되는 생물체의 모든 요소들 또는 최대한 많은 수의 요소들에 대한 것을 한꺼번에 다루어보자는 식이다. 생명현상은 대표적인 창발적(emergent)인 현상으로서, 이와 같이 전체를 살펴보고 다룰 수 있을 때에만 진정으로 그 현상을 이해할 수 있고, 나아가 우리가 이를 조작하여 여러 가지 유용한 결과를 얻어내거나 잘못된 부분을 고치거나 하는 일들을 해낼 수가 있을 것이다. 이러한 시도는 이전까지는 불가능했던 것으로, 최근 들어 여러 가지 바탕 기술들의 급속한 발전으로 말미암아 비약적으로 가능하게 된 것이며, 단순히 이에 수반되는 데이터의 양만을 고려할 때에도 컴퓨터는 이를 위한 필수불가결한 도구이다.

이러한 시도의 출발점을 제공한 것은 소위 말하는 인간 유전체 프로젝트(Human Genome Project)로서, 90년도부터 선진국들을 중심으로 본격적인 출발을 하여 이제 그 첫 단계 목표의 달성, 즉 인간의 유전정보 전체를 밝히는 작업의 마감을 몇 개월 앞에 두고 있는 시점이다. 현재 미국과 영국을 중심으로 2000년 봄까지 인간 유전정보 전체의 첫 번째 판을 완성하려고 막바지 물량공세를 퍼붓고 있는 중이다. 이러한 유전체 프로젝트는 여러 분야에 실로 지대한 영향을 주었는데, 국내에서는 소수의 직접적인 관련이 있는 분야의 사람들을 제외하고는 이 영향에 대한 인식이 극히 저조한 것으로 보인다. 그 영향은 단순히 학문적인 목적의 데이터를 대량으로 얻었다는 데 있는 것이 결코 아니다. 그 가장 큰 영향은, 21세기에는 마치 20세기의 전자공학이나 전산학과 같은 위치를 점하는 분야로서 가장 큰 산업을 형성하게 되리라 많은 사람들이 주저함 없이 예상하고 있는 분야가 태동했다는 데에 있다. 유전체 프로젝트에서 사용된 생물체 시료에 대한 자동화된 high-throughput tool과 컴퓨터의 결합이라는 새로운 패러다임의 연구방식은, 유전정보를 밝히는 작업(즉, 유전체 프로젝트)이 비록 그 출발점을 제공해주기는 했지만 이제는 극히 일부를 차지하는 하나의 예에 불과한 것이 되었으며, 생물체와 관련된 다른 무수한 연구개발들의 방식을 통째로 바꾸고 있는 중이다.

이러한 새로운 연구의 패러다임은 이미 상당한 효율성을 갖추기 시작했으며, 드디어 헤아

릴 수 없이 다양한 형태의 직접적이고 단기적인 산업화를 내다볼 수 있는 단계에까지 이르게 되었다. 이로 인해 지금까지 전자공학과 전산학의 메카라 할 수 있는 미국 실리콘 밸리의 회사들의 절반에 가까운 수가 이미 이 새로운 패러다임을 바탕으로 하는 소위 바이오텍 회사들로 대체가 되었으며, 머지않아 실리콘 밸리라는 이름 대신에 바이오 밸리로 바뀌게되지 않을까 예상하는 말도 들을 수 있는 상황이다. 이들 바이오텍 회사들은 다양한 목적을 가지고 있으며, R&D가 그 주된 활동인 소위 말하는 "research driven company"들인데, 그 대표적인 모양을 몇 가지 살펴보면 다음과 같다. 우선 가장 큰 부분을 차지하는 것은 인간의 질병을 치료할 수 있는 방법을 개발하고자 하는 목적의 회사들이다. 이들은 보통 세계적인 대형 제약회사들과 결합이 되어 있으며, 연구결과를 이들 대형 회사들이 사후에 구매를 하거나 사전에 미리 투자를 하는 방식을 취하고 있다. 이들의 연구개발에 소요되는 자금의 또 다른 소스로는 선진국들에서 활발한 벤처 캐피탈이 있으며, 주식시장으로부터 직접 조달되는 경우도 흔히 볼 수 있다. 이러한 제약시장은 그 규모가 반도체 시장보다 현재도 훨씬 큰 것으로, 앞으로 이러한 새로운 방법을 통해 지금까지 존재하지 않았던 새로운 형태의 치료약들이 개발되면 그 시장은 더욱 팽창하리라 예상되고 있다. 그 다음은 농업에 관련된 것으로, 전세계적인 식량문제의 해결에 획기적인 전기를 가져오리라 예상되고 있다. 또한, 단순히 식량의 생산량을 늘리는 것만이 아니라, 약품성분이 들어 있는 식품과 같이 지금까지 자연계에는 존재하지 않았던 새로운 형태의 식품을 만드는 것 또한 시도되고 있다. 그 밖에 환경문제 해결에 관한 것들을 비롯하여, 생물체와 관련된 모든 것들이 이들 새로운 연구개발의 패러다임으로 무장한 바이오텍 회사들의 대상이 되고 있다.

이 글에서는 이러한 새로운 연구의 패러다임의 두 기둥중의 하나를 차지하는 컴퓨터 도구들에 대해 개괄적으로 살펴보고자 한다. 21세기의 세계경제의 주역이 이러한 바이오텍 산업이 될 것이라고 흔히들 예상을 하고 있으므로, 21세기에 우리가 선진국의 대열에 동참을 할 수 있기 위해서는 이 바이오텍 그 자체에 대해, 그리고 이를 바쳐주는 중요한 기둥인 전산적인 도구에 대한 준비와 대책이 있어야만 할 것이다. 먼저 이러한 바이오텍에 있어서 왜 컴퓨터가 중요한 도구가 될 수밖에 없는지, 그 이유부터 살펴보고자 한다.

2. 생명체가 가진 정보

생명체는 긴 선형의 기록장치인 DNA에 코딩된 유전정보라는 설계도를 가지고 있다. 이 DNA 속에는 유전자라는 단위로 정보들이 기록되어 있다. DNA는 4가지 염기의 조합을 통해 정보를 기록하며, 이 DNA는 염색체라는 물리적으로 구분되는 단위로 존재한다. 사람의 경우에는 한쪽 부모로부터 각기 23개의 염색체를 물려받으며, 전체 46개인 약간의 차이를 가진 두벌로 구성되어 있다. 하나의 염색체는 하나의 DNA 분자. 즉 일련의 하나의 기록 매체에 해당한다. 이 23쌍의 염색체는 합계 약 60억 개의 염기를 가지고 있으며, 따라서 하나의 염색체는 약 1억 3천만개의 염기로 구성되어 있다. 이 23쌍의 기록장치에 코딩되어 있는 사람의 유전자의 수는 약 10만개의 쌍 정도가 되리라 예상되고 있으며, 훨씬 더 정확한 수(궁극적으로 완벽한 수를 알아내기 위해서는 더 오랜 세월이 걸릴 것이지만)가 사람의 전체 유전정보가 밝혀지는 2000년 봄에는 알려지게 될 것이다. 이 10만 가지 유전자 각각으로부터 (상대적으로 짧은 길이의) RNA 분자가 복사되어 세포 속으로 방출이 되고, 이 RNA 분자에 담긴 정보로부터 다시 단백질이 만들어진다. 이 단백질은 다시 다양한 형태로 수정이 되고, 또한 같은 종류끼리 또는 같은 공정의 일부를 수행하는 것들끼리 뭉쳐서 complex를 이루게 된다. 결국 생명현상의 주류는 이러한 수정된 단백질, 또는 그들의 complex가 복잡한 상호작용을 하는 것이라 할 수 있다.

사람의 몸 속에서는 수천 가지 이상의 각기 다양한 차이점들을 가진 서로 다른 종류의 세포들이 존재하며, 이들 각각은 발현되는 유전자의 종류가 다르다. 일반적으로 한 종류의 세

포에서는 약 2만 가지의 유전자들이 발현된다고 생각되고 있다. 그리고, 특별한 예외를 제외하고는 이들 서로 다른 세포들이 가지고 있는 원본 DNA는 동일하다. 즉, 세포들은 그들이 가진 유전정보의 약 1/5만을 꺼내서 사용한다. 이중에서는 에너지 대사에 해당하는 것과 같은 모든 세포가 반드시 필요로 하는 유전정보가 있고, 헤모글로빈을 만드는 정보와 같이 단 한 종류의 세포(즉, 적혈구)에서만 사용되는 유전정보도 있다. 이처럼 하나의 세포에서는 약 2만 가지의 서로 다른 단백질들이 만들어지게 되고, 다시 이들 단백질들은 그 각각이 몇 가지 또는 수십 가지의 다른 형태로 수정이 되며, 이들은 독자적으로 작용을 하거나 다시 다양한 형태의 complex를 이루어 작용을 하게 된다. 따라서, 하나의 세포는 약 2만 가지의 기본요소가 다시 약간씩 다른 형태로 변형이 된 것들이 다양한 형태로 상호작용을 하는 시스템이라 할 수 있다. 그리고, 단백질과 함께 RNA도, 훨씬 적은 종류이지만 역시 이러한 상호작용에 참여를 하고 있으며, 단백질이 만들어낸 비단백질 산물들이 다시 단백질과 상호작용을 하고 있으므로, 상황이 더욱 더 복잡해진다고 할 수 있다.

지금까지는 하나의 세포라는 시스템을 구성하는 요소들을 살펴보았는데, 이 구성요소들의 상호작용은 각 요소 그 자체에서는 볼 수 없었던 새로운 성질, 즉 창발적 성질(emergent property)을 나타내게 되어 이것이 결국 하나의 세포가 나타내는 성질이 된다. 이 성질은 물론 정적(static)인 것이 아니며, 외부의 입력 또는 그 자체의 필요에 의해 다양한 반응과 변화를 보여주는 형태의 것이다. 그 다음 이러한 세포들은 독자적으로 존재하는 것이 아니라, 같은 종류 또는 다른 종류가 다수 결합된 조직(tissue)의 형태를 이루고 있으며, 이러한 조직들이 모여서 다시 어떤 한가지 또는 몇 가지의 작업을 담당하는 기관(organ)을 이루고, 나아가 하나의 독립된 개체를 형성하게 된다. 그리고, 다시 더 상위의 것으로 이러한 개체들은 같은 종의 다른 개체, 그리고 다른 종의 개체들, 더 넓게는 환경과 상호작용을 하게 된다. 이러한 각 계층들은 각기 새로운 창발적 성질을 가지게 된다.

생명현상에 대해서 우리가 해야 할 일은 크게 "해독"과 "조작"이라는 두 가지로 나눌 수 있다. 궁극적으로 우리가 해독하고 조작할 수 있어야 하는 것은 당연히 위에서 언급한 모든 것이라 할 수 있다. 물론 출발은 세포의 기본 구성요소들에 대한 카탈로그를 만드는 작업(유전체 프로젝트가 바로 이를 위한 대표적인 것이라 할 수 있다)이 되어야 할 것이다. 일반적으로 시스템의 분석을 위해서는 그 시스템을 구성하는 요소들을 파악하고, 그 다음에는 이 요소들 사이의 연결들을 찾아낸 다음, 이 연결로 인해 발생하는 창발적 성질을 이해하는 것이 그 순서이다. 여기에는 다시 연결들 중에서 특별히 중추적 역할을 하는 부분을 찾아내는 것과 같은 이 해독과 조작가능성에 특별히 도움을 주는 여러 가지 중요한 방법들이 있을 것이다.

요소들을 찾아낸 다음에는, 동적인 존재인 생명체로부터 요소들의 연결과 그 창발적 성질에 대한 정보를 해독해내야 한다. 위에서 살펴본 바와 같이 하나의 세포가 가지는 요소들은 기본적인 것이 수만의 단위에 달하게 되고, 다시 이들의 변형까지 고려를 하면 다시 그 order가 하나 또는 그 이상 높아지게 된다. 이 요소들이 보여주는 연결과 창발적 성질을 해독하기 위해서는 여러 가지 high-throughput tool을 사용하게 되는데, 이들은 일반적으로 각 요소들의 움직임을 살펴볼 수 있는 형태의 것들이다. 이 움직임이란 이 요소 각각의 양적(amplitude), 시간적(temporal), 공간적(spacial) 변화를 뜻한다. 이렇게 하여 얻어지는 정보는 실로 막대한 양이 되는데, 유용한 정보가 되기 위해서는 여러 가지 다른 조건에 대한 일정 시간에 걸친 샘플링이 필요하고, 다시 이러한 작업 전체를 다수의 다른 개체로부터 얻어진 것에 대해 반복하는 형태가 되어야 하기 때문이다. 일반적으로 한 가지 목적의 일련의 실험에서 얻어지는 데이터 포인트 수는 쉽사리 수십 억을 헤아리게 된다.

이와 같은 방식을 주된 패러다임으로 하는 생물체에 대한 새로운 연구방식을 소위 "genomics"라 부른다. 이 용어는 일면 임시방편적

인 면이 있으며, 종국에는 더 일반적이고 엄격한 과학적인 용어로 대체될 것으로 보이지만, 그 출처를 한번 살펴보면 다음과 같다. 이 용어는 genome이란 단어에서 왔으며, 원래는 80년대 말에 창간된 한 학술지의 제목에 최초로 쓰인 것으로, 그 당시에는 이렇게 유명한 용어가 될 지 전혀 예상하지 못했을 것이다. genome은 어느 한 생물체가 가지는 유전정보의 "총체"를 뜻하며, 따라서 genomics란 어떤 생물체에 대해 그 생물체가 가지는 유전정보(그리고 이로부터 발현되는 것들) 전부를 총체적으로 연구하는 것을 뜻한다.

이제 이에 있어서 컴퓨터의 역할에 대해 살펴보면 다음과 같다. 우선, 이처럼 데이터의 양이 막대하다는 점과 이 데이터를 얻어내는 과정에서 효율적인 워크플로우를 구성하기 위해 컴퓨터가 중요하다는 것은 당연하겠지만, 이보다 더욱 근본적인 이유는 이렇게 얻어진 데이터를 가지고 무엇인가를 해볼 수 있는 우리가 가진 유일한 도구가 바로 컴퓨터라는 점에 있다. 즉, 이 데이터는 결코 명시적이지도 자명하지도 않기 때문이다. 이 데이터는 단지 복잡미묘한 창발적 성질이 개개의 요소들의 움직임에 반영된 희미하고 노이즈가 심한 실루엣일 뿐이다. 아주 적절한 비유는 아닐 수도 있지만, 우리가 해내야 할 일은 다양한 각도의 실루엣들로부터 3차원적인 실체를 유추해내야 하는 정도에 비유할 수 있다. 이러한 대량의 데이터에 대해 이러한 작업을 해낼 수 있는 우리가 가진 유일한 무기는 당연히 컴퓨터이다.

그리고, 위와 같이 실용적인 응용의 각도에서 살펴볼 때만이 아니라, 생명현상을 연구하는 학문은 본질적으로 소위 말하는 "information science"가 될 수밖에 없다. 생물체는 각 개체가 서로 다르며, 이것이 하나의 구분되는 "정보적 대상"이 된다. 그리고 이것은 어떤 정해진 시간에는 유한한 양(물론 막대한 양이기는 하지만)이 존재하며, 시간에 따라 지속적으로 변해 가는 성질을 가진다. 이와는 달리, 생물학을 제외한 대표적인 자연과학이라 할 수 있는 물리학이나 화학에서 다루는 대상들은 이러한 성질을 가지고 있지 않거나, 또는 상대적으로 훨씬 약하거나 그 중요성이 훨씬 적다. 단적인 예를 들면, 이 우주에 존재하는 모든 물분자는 동일하며, 이들은 영원히 그러할 것이다. 생명체란 결코 모두가 같은 어느 한 종류의 분자와 같은 식으로 존재하는 것이 아니라, 하나의 종에 있어서도 각 개체가 보이는 모든 차이점 그 자체가 바로 생명현상의 본질의 일부라는 점을 간과해서는 안 된다. 이것은 지금까지의 생물학적 연구에서는 흔히 무시되어온 경향이 있는 것으로, 이런 점까지 제대로 고려할만한 수단을 우리가 가지고 있지 못했다는 점에 기인한 것이라 할 수 있을 것이다. 왜 이런 식으로 더욱더 복잡해 보이고 일면 대부분의 과학적 연구의 중요한 기저를 이루는 방식, 즉 "우주만물의 보편성에 바탕을 둔 방식"을 거스르는 듯이 보이는 방향을 취하지 않을 수 없는 지에 대한 실용적인 이유의 예를 한 가지 굳이 들면 다음과 같다. 거의 동일한 환경 하에서도 어떤 사람은 병에 걸리며 다른 사람은 걸리지 않고, 그리고 어떤 사람에게는 듣는 약이 어떤 사람에게는 잘 듣지 않는 것을 흔히 볼 수 있다. 이것은 단지 "무시할 수 있는 비본질적인 우연한 노이즈"에 해당하는 것이 아니라, 생명현상의 본질이 바로 이러한 것이기 때문이다. 따라서, 생물체를 연구하고 이에 대해 우리가 원하는 조작과 치료를 가할 수 있기 위해서는 이러한 개체간의 차이에 관한 정보들을 수집해야 하며, 다시 이들의 분석을 통한 이해가 필요한 것이다. 또한, 이러한 방식은 "보편적 성질"에 대한 이해 또한 도와주게 되는 것이라는 점은 굳이 설명할 필요도 없을 것이다. 또한, 이러한 방식을 취하게 되면 자동으로 데이터의 양이 막대해지고, 또한 그 분석에 있어서도 복잡성이 배가 되게 된다. 이것이 바로 컴퓨터가 생명체에 관한 연구에 있어서 중요한 도구가 될 수밖에 없는 좀 더 순수한 학문적이고 본질적인 이유이다. 즉, 생물학은 일종의 정보과학이다.

또한, 이러한 복잡한 일을 해내는 데 있어서 마치 신의 선물과 같은 점이 있는데, 그것은 바로 생명체가 가지는 정보의 기저를 이루는 것이 "디지털" 포맷으로 되어 있다는 점이다. 생물체를 이루는 핵심 요소라 할 수 있는 DNA, RNA, 그리고 단백질은 모두 디지털화

된 정보를 가지고 있다. 물론 이들의 상호작용으로 나타나게 되는 생명현상 그 자체는 결코 단순한 디지털화 형태의 것이 아니며 디지털 포맷으로 쉽게 변환시킬 수 있는 형태의 것도 아니지만, 그 기본요소가 디지털 포맷으로 되어 있다는 점이 여러 모로 큰 도움이 되리라는 것은 쉽게 상상할 수가 있을 것이다. 예를 들면, 이러한 점 덕분에 우리는 변해 가는 유전정보를 전혀 정보의 손실이 없이 그대로 컴퓨터 안에서 다룰 수가 있으며, 컴퓨터 안에서의 디지털화된 정보에 가한 조작을, 복잡하고 손실이 따르게 되는 역변환과정을 그치지 않고 그대로를 다시 실세계에 가할 수가 있는 것이다.

이상에서 생명체가 가진 정보가 무엇이며, 그 정보를 다루기 위해서 컴퓨터가 왜 중요한지를 개괄적으로 살펴보았다. 아래에서는 좀 더 구체적인 내용들을 살펴보고자 한다.

3. DNA 칩의 예

DNA 칩은 생물체로부터 정보를 얻어내는데 쓰이는 현재 가장 각광을 받고 있는 high-throughput tool이다. 이는 수 cm 또는 그 이하의 유리 또는 실리콘 등으로 된 칩 위에 수만에서 수십만 가지에 달하는 서로 다른 DNA 분자를 2차원 매트릭스 형태로 심어놓은 것이다. 개개의 점에는 만드는 방식에 따라 다르지만, 한 가지 종류의 (즉, 같은 염기서열을 가진) DNA 분자가 pico mole 또는 femto mole (즉, 일억에서 천억 개) 정도가 심어지게 된다. 만드는 방식은 칩 위에서 직접 합성을 하는 방식과 따로 합성을 한 것을 칩에 붙이는 식의 두 가지 방식으로 크게 나눌 수 있다. 전자의 경우에 현재 쓰이는 방식은 일반적인 반도체 칩 제작 기술을 변형한 것이며, 자세한 것을 알기를 원하는 사람은 아래에 소개한 참고문헌에서 살펴보기 바란다. 이 기술을 사용할 경우에는 하나의 DNA 분자의 길이는 염기 숫자 20개 내외를 넘을 수가 없다는 한계를 가지고 있지만, 대신 훨씬 고밀도로 만들 수가 있다는 장점을 가지고 있다. 미리 합성할 경우에는 효소를 이용하여 합성하는 방식과 화학적인 합성을 하는 방식이 있는데, 후자는 역시 수십 개 정도라는 길이의 한계를 가지게 된다. 붙이는 방식은 만년필과 같은 원리로 찍는 식인 것, 잉크젯 프린터와 같은 원리를 사용하는 것 등 여러 가지가 있다.

이렇게 만들어진 DNA 칩은 생물체로부터 얻어진 시료와 반응을 시키게 되는데, 이때 사용되는 시료는 DNA 또는 RNA이며, 보통 형광물질로 표지가 되어 있다. 즉, 어떤 생물체 시료에 약물의 투여와 같은 조작을 가하거나, 또는 암조직과 같은 것에서 DNA나 RNA를 추출하여, 이에 형광물질 표지를 붙이면 된다. 그 다음에는 수용액 상태의 적절한 조건에서 이 칩 표면에 앞에서 준비한 시료가 반응을 하도록 한다. 이 반응은 소위 hybridization이라 부르는 것으로, DNA(또는 RNA)가 서로 염기서열이 상보적인 것끼리 (즉, 아데닌과 티민, 그리고 구아닌과 시토신) 결합하려고 하는 성질을 이용하는 것이다. (두 가지 종류의 핵산들, 즉 DNA와 RNA는 서로 같은 종류끼리, 그리고 다른 종류끼리의 잡종 형태의 결합이 모두 가능하다.) 이들의 결합정도가 바로 얻고자 하는 데이터가 되는데, 이는 결합에 참여하지 못한 것을 씻어낸 다음에 남아 있는 형광물질의 량을 측정하여 얻어낸다. 이것이 DNA 칩의 대략적인 작동 방식이다.

일단 여기까지 과정에서 두 가지 종류의 중요한 전산적인 도구가 존재하는데, 하나는 당연히 신호(그리고 영상)처리에 관한 것이라는 점은 이미 간파할 수 있었을 것이다. 다른 하나는 DNA 칩의 특징을 가만히 살펴보면 알 수 있는 것으로, 단 한 종류의 칩에도 수만 가지 또는 수십만 가지의 서로 다른 DNA 염기열이 붙여야 하는 것이므로, 이점으로 인한 전산적인 도구가 또 필요하게 된다. 그리고, DNA 칩의 장점, 즉 최근에 주인공들로 등장한 모든 high-throughput tool들의 장점은 위에서 설명한 것과 같은 작업(즉, DNA 칩의 경우 어떤 시료에서 추출한 핵산과 hybridization을 시켜보는 것과 같은)을 하루에도 수천 번이나 수만 번, 또는 그 이상을 해낼 수 있다는 데에 있다. 이에 수반되는 물리적인 리소스의 개수만 해도 엄청난 수가 될 것이며, 이를

관리하고 추적할 수 있는 전산적인 도구 또한 큰 무게를 가진 것임을 쉽게 상상할 수 있을 것이다. 이는 소위 LIMS(Laboratory Information Management System)라 부르며, 이러한 연구방식의 기반을 형성하는 중요한 전산적인 도구이다.

물론 더 흥미로운 것은 이렇게 하여 얻어낸 데이터로부터 정보를, 나아가 유용한 지식을 얻어내는 것일 터인데, 이것이 바로 현재 모두가 갈망하고 있고 오르고자 노력하고 있는 태산에 해당하는 것이다. 일단 이를 헤내기 위해 다시 중요한 인프라가 되는 것은 소위 말하는 데이터 웨어하우스(warehouse)를 구성하는 것이다. 특히 DNA 칩으로 얻어낸 데이터에 대해서는 이 부분은 현재 그 첫 번째 쓸만한 버전을 내놓기 위해 개발이 막 진행되고 있는 중이다. 여기에서 한 가지 언급하고 지나가야 할 것은, 현재 누가 이런 일들을 하고 있는가 하는 점에 대한 것이다. 이 분야의 일들이 주로 일어나고 있는 선진국들에서 볼 때, 이는 크게 상업적인 집단과 정부지원을 받는 집단으로 나눌 수가 있는데, 다른 분야들과는 달리 정부지원 부분이 상당히 큰 역할을 하는 것을 볼 수 있다. 이는 생명체를 다룬다는 점, 그리고 인간의 질병을 정복하고자 하는 일이라는 점에서 그 근본적인 이유를 찾을 수 있다. 물론 바로 이 점으로 인해 돈벌이가 되는 일이므로 상업적인 집단들의 움직임 또한 활발할 수 밖에 없는 것이다. 상업적인 부분과 정부지원 부분은 경쟁과 동시에 협력이라는 특이한 관계를 형성하고 있다. 그리고 경쟁이 있는 이유는, 공공의 이익을 위해서는 이러한 연구개발의 결과물, 즉 치료약이나 치료방법 등이 소수에 의해 장기간 독점이 되는 것을 막아야 한다는 데에 있다. 이것은 매우 특이한 문제를 발생시키는 것으로, 현재 특히 특허권을 어디까지 허용할 것인가 하는 문제를 중심으로 매우 복잡한 법률적 사회적 문제를 야기하고 있는 중이다. 이 글에서는 여기에 대해서는 더 이상 다루지 않을 것이며, 관심이 있는 사람은 참고문헌을 참조하기 바란다.

이러한 데이터 웨어하우스의 구성 문제가 더욱 복잡해지는 이유중의 하나는, 생물체로부터 얻어지는 매우 비균질적인 데이터베이스들을 상호 연결해야 한다는 데 있다. 지금 소개하고 있는 DNA 칩을 사용하여 얻은 데이터도 결국 어떤 생물체의 어떤 시료로부터 온 것이고, 이 생물체로부터는 DNA 칩 이외에도 다양한 방법으로 데이터를 얻어낼 수가 있으며, 이들은 또 나름대로 특징을 가진 데이터베이스를 형성하게 된다. 그리고 이러한 현상은 다시 두 방향의 차원으로 더 넓어지게 되는데, 위에서도 언급한 것처럼 개개인의 차이가 그 한 방향이다. 다른 한 방향은 좀 더 심오한 것으로, 모든 생물체가 결국 하나의 조상을 가지고 있다는 점에 기인한 것이다. 즉, 어느 생물체가 가진 어떤 정보는 일반적으로 그 친척이 되는 정보를 다른 생물체에서 찾을 수가 있다. (이를 "homology"라고 부른다.) 즉, 기본적으로 생물체가 가진 정보는 갑자기 하늘에서 떨어지는 식으로 생겨난 것이 아니라, 그 조상이 가진 정보가 다양하게 조금 씩 변형이 된 것이다. 따라서, 다른 생물체에서 이들 homologous한 정보를 조사하면 여러 가지 유용한 것을 얻어낼 수 있게 된다. 사람의 경우에는 이 점이 중요한 이유가 한 가지 더 생기게 되는데, 그 이유는 사람에게 "실험"을 해보는 것에는 한계가 있기 때문이다. 이 정보들이 서로 닮은 정도는 상식적으로 생각할 수 있는 것보다 상당히 큰데, 사람이 가진 정보의 상당 부분은 맥주나 빵을 만들 때 쓰는 효모와 닮았으며, 대장균과 닮은 것도 무시할 수 없는 양이 존재한다. 그리고, 이러한 닮은 정보들의 연결은 생명현상의 이해에 있어서 핵심적인 부분을 차지하고 있다. 따라서, 생명현상 연구를 위한 데이터 웨어하우스는 이러한 여러 차원에 걸친 복잡한 연결들을 모두 고려해야 하는 순수한 전산학적인 면에서 볼 때에도 큰 도전이 필요한 형태의 것이다.

이제는 좀 더 본격적인 흥미를 끌 수 있는 것으로, 이러한 데이터로부터 무언가를 알아내는 작업이 남아 있다. 아마도 지금까지 논의를 따라오면서 자연스럽게 "데이터 마이닝"이란 용어가 떠오르게 되었을 것이다. 생물체로부터 얻어지는 정보에 관한 이러한 데이터 마이닝 기법은 이제 막 출발이 되었다고 볼 수 있으

며, 예를 들어 DNA 칩으로부터 얻어진 데이터에 대해서는 현재 클러스터링을 해보는 수준 정도에 머무르고 있으나, 본격적으로 DNA 칩 데이터가 얻어지기 시작한 것이 겨우 2년 남짓 밖에 되지 않았으므로 곧 더욱 발전적인 기법들이 등장하리라 당연히 예상할 수 있다. 아직 걸음마 단계라 할 수 있는 이러한 시도들로부터도 기존의 생물분야 연구자들에게는 가히 충격적이라 할 수 있는 결과들이 쏟아져 나오고 있다. 이것은 지금까지 우리가 생명현상을 전체로 보지 않고, 부품들을 하나 씩 뜯어보는 식의 연구가 얼마나 무모한 소위 장님 코끼리 만지기 식의 방식이었는지를 깨닫게 해주는 것이며, 이것은 바로 아직 많은 수의 확실하게 성공한 예가 생겨난 상황이 아닌 데도 너도나도 주저함 없이 이 패러다임 변혁의 대열에 동참하고 있는 이유가 무엇인지를 보여주고 있다.

이러한 DNA 칩 데이터는 위에서 언급한 바와 같이 상업적인 부분에서 생산된 것과 정부지원에 의한 것의 두 가지가 존재하게 되는데, 후자의 경우에는 물론 기본적으로 공개가 되어 있으며, 아래 참고문헌에서 이들을 찾아볼 수 있다. 그 다음 데이터 마이닝과 함께 자연스럽게 따라오게 되는 것은 "데이터 시각화"이다. 대량의 복잡한 데이터를 가지고 무언가를 해보기 위해 우리가 가진 이 두 가지 최고의 무기들은 모두 생물체로부터 얻어진 정보라는 일면 난공불락으로 보이는 성에 이제 막 도달한 상태이다. 이제 이들을 새롭게 갈고 닦아 인류의 이러한 모든 시도들이 의미가 있게 만드는 것은 많은 부분 전산학자들의 몫이다.

4. 결 론

이상에서 생물정보에서 대한 전산분야와 관련된 면들을 큰 줄기만을 따라가는 식으로 살펴보았다. 위의 논의에서 제외된 것중에는, reverse engineering적인 방식으로 생물체 내부에서 일어나는 현상을 모델링 해보고자 하는 시도인 소위 "genetic network"이라 부르는 또 줄기의 큰 흐름을 차지하고 있는 분야도 있다. 그리고, 위에서는 단지 DNA 칩만을 예를 들어 설명을 했으나, 이와 같은 high-throughput tool은 이미 다양한 것들이 나와 있으며, 앞으로도 더욱더 혁신적인 것들이 개발 될 것이다. 그리고 이미 일어나고 있는 현상이 바로 데이터로부터 지식을 뽑아내는 것이 속도결정 단계라는 점이다. 즉, 데이터를 생산해내는 속도는 이미 엄청난 상황이며 점점 더 가속이 되어 가고 있으나, 이 데이터를 가지고 무언가를 해볼 수 있는 전산적인 도구들이 아직은 저 멀리 뒤쳐져 있는 상황이다. 그리 적절한 표현은 아닌 것 같지만, 이러한 모든 시도들에 있어서 맡은 바의 역할을 가장 제대로 해내지 못하고 있는 분야가 바로 전산학이라 할 수도 있을 것이다.

생물학 분야의 사람들이 전산학 지식까지 익히는 것이 어렵다는 것은 당연한 것이라 치고, 이에 대해서는 더 이상 논의를 하지 않기로 하자. 그렇다면 그 반대의 경우는 어떠한가? 우선, 이러한 생물체로부터 얻은 데이터를 전산학적으로 다루는 데 있어서 필요한 생물학적인 지식은 어느 정도인가 하는 질문을 할 수 있을 것이다. 이것은 물론 "많을수록 좋다"라는 것이 가장 좋은 답이다. 워크플로우를 다루는 것과 같은 일부 분야에서는 이러한 생물학적 지식이 그리 중요하지 않은 경우도 있으나, 궁극적으로는 자신이 도대체 뭘 가지고 이러고 있는 지 그리고 정확히 뭘 하려고 이러고 있는지는 알고 있어야 하는 것이다. 그런데 문제는 생물체에 대한 데이터인 경우 이것을 제대로 알게 되는 것이 그리 간단한 일이 아니라는 데 있다. 이것이 바로 전산이 주요한 도구로 사용되는 다른 대부분의 분야들과는 달리, 필수 과목의 수가 2배인 별도의 교육과정이 굳이 필요한 이유인 것이다. 우리 나라의 경우는 "인터넷"만으로도 공급이 모자라는 상황이니, 공부를 2배나 해야 하는 이런 이상한 분야까지 신경 쓰게 될 리가 없다는 것이 많은 사람들의 현실적인 전망이기도 하다.

그러나 다시 선진국들로 눈을 돌려보면, 이러한 대규모 데이터와 이에 대한 전산처리라는 방식은 매우 급격하게 발생하였는데, 물론 90년 초부터 예상들은 되어왔으나 본격화된 것은 모두 최근 몇 년간에 일어난 일이라 할 수 있

다. 그리고, 이러한 것은 당연히 "인적 자원의 문제"를 야기한다는 것을 상상할 수 있을 것이다. 즉, 이러한 일을 해낼 사람들이 어디에 있는가 하는 것이다. 미국의 경우에는 국가적인 차원에서의 체계적인 촉진을 하고 있으나, 여전히 극심한 "사람 부족"에 시달리고 있는 중이며 이 추세는 한 동안 계속될 것이다. 이 분야는 현재 거의 대부분 바이오텍 회사들에서 상시 모집 광고가 나와 있는 상태라 하면 크게 틀린 표현이 아니다. 이에 당연히 수반되는 엄청난 고액의 연봉으로 인해, 그나마 얼마 되지 않는 기존의 전문가들을 상업적인 부분에서 흡수해 가버림으로써 새로운 전문가의 교육 자체에 문제가 생기는 등, 다양한 문제들이 발생하고 있는 중이다. 미국의 경우 이제 대부분의 대학들에서 이러한 전문가를 교육시키기 위한 교육과정들이 이미 생겨났거나 생겨나고 있는 중인데, 일반적으로 전산학과를 중심으로 생물분야의 학과들이 결합을 하는 식이며, 수학과 또한 통계학과 등도 중요한 역할을 하고 있다. 국내의 경우에는 물론 당장의 상황은 미국의 경우와는 전혀 다르다 할 수 있는데, 무엇보다도 바이오텍 산업이 아직은 미약하다는 데에 그 이유를 찾을 수 있을 것이다. 그렇지만, 국제화된 세상에서 우리만 특별할 수는 결코 없는 것이며, 이미 국내의 생물분야에서도 이러한 전문가의 필요성을 느끼기 시작하는 경우를 점점 더 흔히 볼 수 있게 되어가고 있다. 우리나라의 경우에도 단지 그 시기만이 문제일 뿐, 머지않아 이 분야 전문가에 대한 수요가 폭발하리라는 예상은 너무나 당연한 것이다. 물론 이는 우리 나라가 주저앉지 않고 선진국을 향해 나아가는 노력을 21세기에도 계속한다는 가정 하에서만 맞는 말일 것이다.

참고문헌

[1] Guide to Bioinformatics at Stanford University (and, Professor Donald Knuth on Bioinformatics), http://www-smi.stanford.edu/people/altman/bioinformatics.html

[2] National Human Genome Research Institute (NHGRI), http://www.nhgri.nih.gov

[3] The International Society for Computational Biology, http://www.iscb.org

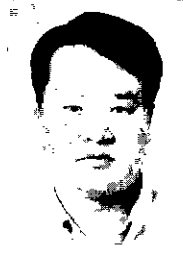
[4] The Chipping Forecast, Nature genetics, Vol.21 supplement, 1999.

[5] The Cancer Genome Anatomy Project (CGAP), http://www.ncbi.nlm.nih.gov/CGAP

[6] Ethical, Legal, and Social Issues (ELSI), http://www.ornl.gov/hgmis/resource/elsi.html

[7] Baldi, P. and Brunak, S., Bioinformatics: The Machine Learning Approach, MIT Press, London, 1998.

[8] Gusfield, D., Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, New York, 1997.



원 세 연

1985 고려대학교, 생물학과, 이학사
1989 Wayne State Univ., Depart. of Bio. Sci., 이학석사
1995 KAIST, 생물과학과, 이학박사
1995 KAIST, 인공지능연구센터, Post Doc.
1995~1998 연구개발정보센터, 선임연구원
1998 (주)바이오니아, DNA 연구소, 책임연구원
E-mail:sywon@bioinfo.kaist.ac.kr