제 1 장 생물정보학 들어가기

1.1 생물정보학이란 무엇인가?
생물정보연구소 원세연

생물정보학은 영어로는 **bioinformatics, computational biology**, 또는 **computational molecular biology** 등의 용어로 불리는 분야이며, 국내에서는 영문 bioinformatics가 생물정보학이라는 용어로 번역되어 어느 정도 확립이 되어가고 있는 상황이며, 전세계적으로도 이 용어가 현재 가장 널리 쓰이고 있다. 생물정보학은 매우 다양한 분야를 담고 있는 폭넓은 것이며, 한 문장으로 제대로 모든 것을 담고 있는 표현을 만들기가 쉽지가 않으나, 굳이 적자면 "생명현상 연구에 필요한 다양한 전산학/통계학/수학적인 것들"이라는 표현이 그나마 본질에 어느 정도 접근을 하는 것이라 할 수 있다. 또한 생물정보학은 현재 급격한 변화를 겪고 있는 중이며, 최근에 새롭게 생겨난 분야들이 오히려 그 규모나 중요성으로 볼 때 기존에 존재해오던 분야를 쉽게 능가해버리고 있는 상황이기도 하다.

생물정보학이라 부를 수 있는 분야의 **태동**은 **Frederic Sanger**에 의해 **단백질 서열결정 방법**(이 발명으로 1958년 노벨 화학상 수상)이 개발된 이후인 1960대부터 시작되었다고 할 수 있다. 단백질은 20가지 아미노산이 수백 개 정도가 길게 선형으로 연결된 형태의 고분자로서, 예를 들면 "MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQ"처럼(이것은 사람의 유전자 발현을 조절하는 중요한 단백질 한 가지의 서열의 앞부분 일부이다) 사람의 눈으로는 다루는 것이 불가능한 형태의 데이터가 얻어지게 된다. 만약 쥐, 또는 침팬지로부터 동일한 기능을 하는 단백질들의 아미노산 서열을 얻어냈다고 하면 이들을 상호비교 하려면 어떻게 해야 할까? 또한 이 상호비교는 "유독 유사성이 강하게 유지된 영역들이 따로 존재를 하는가?", "상호간의 거리는 얼마나 되는가?" 등과 같은 질문에 대해 정량적인 답을 할 수 있는 것이어야 한다. 따라서, 무엇인가 **수학적이고 전산적인 도구가 필요**하게 되리라는 것은 쉽게 직감을 할 수 있을 것이다. 최초로 이를 인식하고 이 분야의 실질적인 출발을 시킨 사람은 Monte Carlo 방법의 발명으로 유명한 **Stanislaw Ulam**이란 수학자로서, 자신이 속한 미국 로스 알라모스 국립연구소에서 이러한 방향의 연구를 주위의 젊은 연구원들에게 독려를 했고, 초기의 많은 일들이 이들에 의해서 이루어졌다. 또한, Frederic Sanger는 생물체의 또 다른 중요한 고분자인 **DNA의 염기서열을 밝히는 방법**까지 발명하여 1980년에 **두 번째 노벨 화학상**을 탔으며, 최근에 인간 유전체 전체의 염기서열을 밝히는데 큰 역할을 한 영국의 **Sanger 센터**의 명칭이 바로 그의 이름을 기념하여 붙여진 것이기도 하다.

위에서 두 가지 중요한 생물정보학의 대상이 되는 데이터를 언급을 하였는데, DNA와 이에 코딩된 정보로부터 만들어지는 **단백질의 서열정보**이다. 그 밖에도 다양한 형태의 데이터들이 존재하게 되는데, 하나 씩 차례대로 설명을 해보고자 한다. 우선, 단백질은 다시 **3차원적인 구조**를 만들게 되고, 이 3차원적인 구조를 가진 고분자의

상호작용에 의해서 결국 거의 대부분의 생명현상이 일어나게 된다. 여기에서 다시 또 한 가지 생물정보학의 대상이 되는 데이터가 생겨나게 되는데, 바로 **단백질의 3차원적인 형태**에 대한 것이다. 이에는 여러 가지 문제들이 수반되는데, 우선 1차원적인 아미노산 서열로부터 3차원적인 단백질을 예측해 내는 것(이를 folding 문제라고 한다), 단백질의 3차원적인 형태를 상호 비교하는 것, 주어진 단백질에 들어맞는 작은 유기화합물을 디자인해내는 것, 주어진 두 개의 단백질이 입체적으로 어떻게 상호작용을 하는지를 알아내는 것, 단백질이 이러한 여러 가지 상호작용을 할 때 어떤 동적인 변화가 일어나는지 등 매우 다양한 문제들이 있게 된다.

이러한 단백질의 3차구조에 관한 여러 가지 문제들에 관련된 분야는 소위 **구조 생물학**이라 불리는 것으로, 이미 오래 전부터 상당히 큰 규모의 분야를 형성해 오고 있었다. 오늘날에는 생물정보학이라는 일종의 umbrella term 아래에 일부로 포함되기도 하며, 더 본질적으로는 생명현상이 결국은 3차원적인 실체를 가지는 단백질들의 상호작용에 의해서 일어나는 것이므로, 따로 명확하게 구분을 하기 힘든 점이 있는 것이기도 하다. 또한, 특히 최근에는 구조 생물학 고유의 연구 분야들에서도 대량으로 얻어진 DNA와 단백질의 서열 데이터를 종합적으로 분석함으로서 3차원적인 구조 그 자체의 연구에도 큰 도움을 얻고 있기도 하므로, 결국 상호간에 강하게 융합이 일어나고 있다고 할 수 있다.

그 다음은 이러한 단백질들과 염색체 상에 DNA 형태로 들어있는 유전자, 그리고 이들 사이의 중간 단계라 할 수 있는 RNA가 어떤 것들끼리 어떻게 상호작용을 하며, 어디에 얼마나 존재하고, 어떤 환경이나 조건에서 어떻게 양이나 구조 등이 변하는지에 대한 데이터가 있다. 이들을 밝혀내는 현재 가장 각광을 받고 있는 도구가 바로 소위 **DNA chip이라 불리는 것과 proteomics 도구**라 불리는 것들이다. 이 도구들은 결국 이들에 대해 일종의 스냅사진을 제공해주게 된다. 당연히 스냅사진도 여러 장을 연속으로 찍게 되면 움직임을 알 수 있게 되듯이, DNA chip 또는 proteomics 도구도 다양한 변이를 준 상황에서 가능한 많은 양의 데이터를 얻어내고, 이를 분석함으로써 생물체내에서 실제로 일어나고 있는 현상을 최대한 밝혀내고자 하는 것이 결국 오늘날 소위 genomics 또는 proteomics 등으로 불리는 분야들의 목표인 것이다.

이에 대해서는 상당히 긴 설명이 필요로 하게 되는데, 최대한 시도를 해보고자 한다. 우선 DNA chip(오늘날 주로 쓰이고 있는 형태에 대해서는 더 엄밀한 학술적인 용어로는 DNA microarray라고 해야 한다)이나 proteomics 도구들의 특징은 소위 "global"이라는 단어로 요약을 할 수가 있다. 기존의 동일한 목적, 즉 생물체 내부의 분자적인 메커니즘 규명을 위해 사용되던 방법들은, 주로 하나 또는 몇 개의 RNA나 단백질을 추적하는 식인 것에 반해, 소위 genomics적인 방법에서는 대상이 되는 세포나 조직 속에 들어 있는 **"모든" RNA나 단백질을 추적해보는 것인 점**이 바로 본질적인 차이이다. 하나 또는 몇 개를 추적할 경우에는 대부분의 경우 복잡한 수학이나 대용량의 데이터 처리나 고급스러운 통계적인 처리가 굳이 필요가 없다는 것은 당연히 짐작할 수 있을 것이다. 반면에, 이러한 소위 genomics적인 방식에서는 일단 데이터의 양부터 우리가 직접 펜을 들고 실험 노트에 기록이 가능한 양을 훨씬 초월하는 것이다.

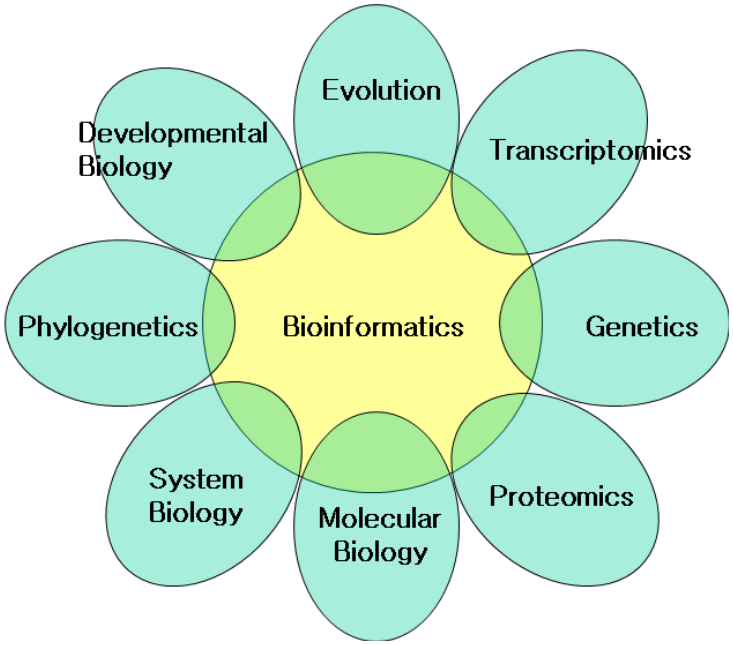

그림 1 생물학의 여러 관련 분야와 생물정보학과의 관계

또한, 실험에 수반되는 시료들의 개수가 기존의 소위 분자생물학적인 방식에 비해서 훨씬 크다는 점도 염두에 두어야 할 중요한 점이다. 한 장의 DNA chip에는 보통 수천 개에서 수만 개의 DNA 단편들이 고착이 되며, 이들 개개의 단편들은 DNA chip 제작을 위해서는 모두 하나하나 따로 다루어줘야만 하는 것들이다. 또한 한 장의 DNA chip을 사용한 실험에서는 이 수천 개 또는 수만 개의 단편들에 대응하는 양이 모두 측정치로 얻어지게 되며, 이들은 각기 따로 따져주어야만 하는 데이터 포인트들이 된다. 그리고, 일반적인 DNA chip 실험에서는 단 한 장만 사용해서는 의미 있는 결과를 얻을 수가 없고, 가능한 한 많은 장수의 chip을 사용해야만 한다는 점으로 인해 더욱 많은 양의 데이터가 얻어지게 된다.

그 다음은 이렇게 얻어진 데이터의 복잡성에 대한 점이 있다. 이렇게 얻어진 데이터는 결국 복잡한 요소들이 상호작용 하는 복잡한 현상의 한 단면이다. 이러한 현상은 결코 단순한 관계로는 표현이 되지 않는 종류의 것이며, 결국 통계적인 처리나 복잡한 모델링 방법을 사용할 경우에만 실용성 있는 데이터의 해석이 얻어질 수 있는 성질의 것이다. 바로 이 점이 오늘날 생물정보학의 가장 큰 부분을 이루고 있는 핵심적인 내용이다.

위에서 genomics적인 실험 그 자체에 수반되는 데이터 처리, 그리고 실험 결과 얻어진 데이터의 분석에 대한 것들을 극도로 간략하게 살펴보았다. 또 한 가지 이해를 해야 할 중요한 점은, 과학적인 질문 그 자체를 만드는 것에 수반되는 것과 실험의 디자인에 수반되는 것들이다. 이를 위해서도 질문을 만들게 된 출발점이 "단순비교"인 것이어서는 시작부터 잘못된 것이 된다. 즉, 실험 그 자체, 그리고 이로부터 얻어지는 데이터가 이처럼 복잡하고 많은 양이 수반되는 것이므로, 과학적인 질문의 출발 그 자체부터, 그리고 실험 디자인을 위한 것에서도 반드시 통계학적/전산학적인 지식을 기반으로 하여 출발할 수 있어야만 한다.

따라서, 자연스럽게 다음과 같은 결론이 나오게 된다. "앞으로 생명현상을

연구하려면, 기존에 생물학과 등에서 배울 수 있었던 것과는 완전히 다른 지식들이 필요로 하게 되겠군." 정도로 표현을 할 수 있을 것이다. 이것은 이미 우리를 휩쓸고 지나간 현실이며, 선진국들에서는 정신없이 이에 대한 적응을 해나가고 있는 중이기도 하다. 국내에서도 이를 대비하고, 위에서 언급한 점들을 제대로 갖춘 새로운 세대를 키워내지 않는다면, 21세기에는 가장 큰 산업이 된다고 하는 바이오텍의 앞날도 어두울 수밖에 없을 것이다.

소위 proteomics라 불리는 분야에 대해서 간략하게 언급을 하고자 한다. proteomics도 위에 설명한 소위 genomics적인 것들과 본질적으로는 아무런 다를 바가 없다. 단지 RNA의 양들을 추적하는 대신에 좀 더 복잡한 도구로 더 복잡한 성질의 단백질들에 대한 데이터를 얻어내는 정도의 차이가 있을 뿐이다. 그리고, 이러한 모든 것들에는 자동화가 매우 중요한데, 사람의 손에 일일이 의존하지 않고 자동화된 장치를 사용하기 때문에 바로 오늘날의 소위 genomics 혁명이 가능하게 된 것이라는 점은 굳이 언급할 필요도 없을 것이다. 그리고, 단순히 기계의 자동화된 조작만이 아니라, **각 단계에서의 의사결정의 자동화도 매우 중요한 점** 중의 하나이다. 아무리 각 단계에서 자동화된 기계로 대량의 데이터를 얻어낸다 해도, 그들 사이의 연결 부분에서 사람이 일일이 판단을 해주어야 한다면 결국 병목현상이 일어나게 되고, 이로 인해 전체적인 생산성은 크게 떨어질 수밖에 없을 것이다. 이 의사결정도 생물정보학의 큰 분야를 형성하고 있으며, 핵심적인 것이기도 하다. 또한, genomics나 proteomics나 모두 가장 최초로 얻어지는 데이터는 결국 이미지 형태의 데이터인 경우가 대부분이다. 따라서, 이미지 분석 분야도 매우 중요하리라는 것을 짐작할 수 있을 것이다. 이처럼 소위 생물정보학이라 불리는 분야는, 온갖 다양한 전산학적/통계학적/수학적인 것들을 담고 있다.

그 다음은 소위 복잡계 현상에 대한 것을 매우 간략하게 언급을 하고자 한다. 복잡계에 대한 과학은 대략 70년대부터 시작된 비교적 최신의 학문이다. 복잡계는 다수의 요소들이 상호작용을 할 때 일어나는 현상에 대한 학문으로, 이 세상의 대부분의 것들이 결국 이러한 복잡계적인 현상이다. 생명현상이 이러한 복잡계적인 현상의 대표적인 것이라 할 수 있을 것이다. 오늘날 genomics 혁명이 가져다준 가장 큰 변혁중의 하나가 바로, 생명현상을 드디어 복잡계적인 현상의 하나로서 제대로 들여다 볼 수 있는 방법이 생겼다는 것이다. 이전에는 이렇게 접근을 하기를 원해도 적절한 데이터를 얻을 수가 없었으므로, 현실성 없는 수식의 장난에 불과하기 십상이었는데, 드디어 이 한계를 넘어설 수 있게 된 것이다. 최근의 생물정보학의 주요한 학회들을 보면, 바로 이 접근방법을 가장 중요한 세션으로 내세우고 있는 것도 볼 수 있다.

그 다음, 또 다른 차원의 데이터에 대해서 설명하고자 한다. 지금까지 설명한 데이터는 주로 하나의 세포, 하나의 조직, 또는 하나의 개체가 보여주는 것들에 대한 것이었다면, 이제부터 설명하고자 하는 것은 생명의 또 다른 본질, 즉 서로 얽혀있는 집단이 시간에 따라 변해 가는 것이라는 점에 대한 것이다. 우리 자신은 결국 부모로부터 물려받은 유전자들의 조합이며, 다시 또 부모들은 조부모로부터 물려받은 유전자들의 조합이다. 그리고 이 유전자들은 조금 씩 변해가면서 환경과 상호작용을 해나가는 것이다. 즉, 하나의 생물체란 시간에 따라 동적으로 변해 가는 유전자들의 조합 한 세트가 잠시

모여 있는 그릇과 같은 것이라 할 수 있을 것이다. 이 분야 역시 위에서 설명한 구조 생물학의 경우와 마찬가지로 이미 오랜 기간동안 잘 정립된 분야를 형성해 오고 있었던 분야이기도 하다. 또한, 현대 통계학의 발전의 많은 부분이 이 분야에서 영향을 받은 것이기도 하다. 그리고 구조 생물학과 마찬가지로 이 분야 역시 최근의 genomics 붐의 영향으로 상당한 변혁을 겪고 있으며, 또한 이전에 비해 훨씬 더 실용적이고 직접적인 응용을 가지게 되는 분야로 탈바꿈하고 있는 상황이기도 하다. 이 분야가 바로, 최근의 genomics 붐과 함께 생겨난 인기 있는 대중적인 용어 중의 하나인 **맞춤 의학(영어로는 보통 personalized medicine이라고 한다)**의 학문적인 세부의 핵심이기도 하다.

이에 대해서 기본적인 이해가 가능하도록 설명을 시도해보고자 한다. 특히 사람의 경우에는 직접 실험을 해보거나 조작된 교배를 해볼 수가 없으므로, 결국 있는 그대로의 데이터를 이용할 수밖에 없다. 한 개인이 가지는 유전자들의 조합을 추적을 하려면 다른 개인과 차이를 보이는 DNA 상의 서열의 차이에 대한 목록을 가지고 있어야만 할 것이다. 이 목록으로 쓰일 수 있는 것으로 현재 가장 각광을 받고 있는 것이 바로 소위 SNP(Single Nucleotide Polymorphism)이라 불리는 것이다. 즉, DNA 상의 서열의 개인적인 차이점 중에서 단일 염기의 차이를 일컫는 용어이다. SNP 이외에도 다양한 형태의 변이들이 있으나, SNP이 주목을 받는 이유는 다른 변이들에 비해 훨씬 풍부하기 때문이다. 사람은 두 벌의 각기 약 30억 개의 염기로 된 유전정보를 각 세포마다 가지고 있으며, 양친으로부터 각기 한 벌씩을 물려받고, 다시 자식에게 절반씩을 물려준다는 점은 잘 알고 있을 것이다. SNP은 이 30억 개의 긴 염기서열 상에 대략 수백 염기당 하나 씩 존재하고 있다.

그 다음 자식에게 물려줄 때는 있는 그대로를 물려주는 것이 아니라, 두 벌 사이를 새롭게 뒤섞어서 물려주게 되는데, 유전정보의 양이 매우 많으므로 부분적으로 볼 때는 이처럼 뒤섞이는 것은 드물게 일어나는 현상이 된다. 즉, 부분적으로는 원래의 연결 상태가 여러 세대가 지나도 그대로 유지가 되게 된다. 그 다음은 유전자들의 조합이 동일한 두 사람, 즉 일란성 쌍둥이를 생각해봐야 할 차례이다. 이들은 외모부터 상당히 닮았으며, 여러 가지 질병에 관련된 것에서도 서로 유사한 점들을 보여주기도 한다. 이제 SNP이 어떻게 질병의 정복과 관련이 있는지의 이해를 시도해보자. 많은 사람들로 이루어진 어느 한 집단에 대해서 SNP의 목록을 추적할 수 있다면, 결국 유전정보의 작은 한 조각에 국한해서는 자신과 일란성 쌍둥이인 사람들을 찾아낼 수 있을 것이다. 물론, 이러한 조각들은 많은 수이고, 또한 조각별로 일란성 쌍둥이에 해당하는 사람들도 많은 수가 될 것이다. 그리고, 가능하면 많은 수의 SNP 목록을 사용하고, 동시에 가급적 큰 수의 집단에 대해서 이와 같은 데이터를 얻어낼수록 더욱 좋을 것이라는 점도 쉽게 이해를 할 수 있을 것이다. 또한, 서로 연관관계가 가까운 사람들이 모인 집단일수록 더욱 "유전정보 조각별 일란성 쌍둥이"를 찾아내기가 용이하리라는 점도 이해할 수 있을 것이다. 바로 이 점으로 인해서, 불과 수십 명이 두 차례에 걸쳐 이주를 해서 형성된 민족인 아이슬란드 전체 국민에 대해 이와 같은 작업을 수행하고 있는 deCODE genetics라는 회사가 최근에 지속적으로 국제적인 큰 뉴스거리를 제공해주고 있는 바로 그 이유이다. 그 다음은 이 개개의 조각들과 형질들 사이의 관계를 찾아내야 하는 일이 남아 있게 된다. 이 때 정확한 의료기록이 매우 중요한 역할을 하게 되며, 우리가 해결하고자 하는 대부분의 질병들은

노인이 되어서 나타나게 되는 것들인 점도 중요하게 고려해야 할 점 중의 하나이다. 이 연관관계를 찾아내는 과정은 매우 복잡한 데이터 분석을 수반하게 된다. 그리고, 우리 나라의 경우에도 당연히 이러한 의료기록의 정확한 데이터베이스화를 해내는 일이 중요한 관건이 된다. 마지막으로, 맞춤 의학의 한 가지 예를 들면, 이러한 유전정보 조각별 일란성 쌍둥이들이 보여주는 약에 대한 반응들을 토대로 하는 것이 있다. 즉, "나하고 이 부분이 유전정보 조각별 일란성 쌍둥이들인 사람들은 거의 모두 이 약에 거부반응을 보이더라"라는 것과 같은 것을 찾아낼 수 있게 되는 것이다.

위에서 설명한 바와 같은 유전정보와 질병 사이의 관계를 찾아내는 작업은 당연히 대규모의 데이터를 상호 비교하는 복잡한 일이며, 여기에서도 역시 DNA chip 데이터의 분석의 경우에서와 같이 이들 사이의 관계가 단순한 일대일 대응 관계가 아니라는 점에서 근본적인 복잡성을 가지게 된다. 이에 대한 연구가 현재 생물정보학 연구의 또 다른 큰 분야를 차지하고 있다. SNP과 연관이 있는 이야기를 한 가지 덧붙이면, DNA 구조를 밝힌 제임스 왓슨이 나서서 각자의 의료기록과 DNA(혈액을 제공하면 혈액 속의 백혈구에서 DNA를 얻어낼 수 있게 된다)를 기부하는 Human Gene Trust라 명명된 프로젝트를 최근에 출범시켰다. 관심이 있는 사람들은 www.dna.com을 방문해보면 더 많은 정보를 얻을 수 있을 것이다. 우리 나라의 경우에도 이와 같은 형태의 것이 반드시 필요할 것이다.

지금까지는 주로 데이터를 분석하는 것들에 대해 설명을 했지만, 이와 더불어 **데이터 그 자체를 다루고 관리하고 상호연결을 하는 일** 또한 상당히 까다로운 전문적인 일이 됨을 쉽게 상상할 수 있을 것이다. 특히 생물체로부터 얻어지는 데이터는 그 양이 막대할 뿐 아니라, 생명현상은 결국 서로 복잡하게 얽힌 것이므로, 얻어진 데이터 또한 이 연결을 반영하는 상태로 상호 연결을 한 상태에서 들여다보지 않을 수가 없다. 또한, 상대적으로 단순한 비즈니스 데이터 등에 비해서는 훨씬 복잡한 성질을 가진 데이터를 어떻게 컴퓨터 상에 모델링해서 넣어놓을 수 있는가에 대한 것 또한 현재 활발하게 연구가 진행되고 있는 분야이다. 마지막으로, 이러한 **복잡한 데이터와 컴퓨터 앞에 앉은 사람 사이의 인터페이스를 어떻게 만들 것인가에 대한 것 또한 중요한 분야 중의 하나**이다.

이상으로 현재 중요한 위치를 차지하고 있는 생물정보학의 여러 분야들을 간략하게나마 살펴보았다. 이러한 생물정보학은 오늘날의 바이오텍 붐의 핵심을 차지하고 있는 것이고, 필수 불가결한 것이기도 하며, 또한 필요한 인력의 수가 상당히 크다는 점도 쉽게 짐작할 수가 있을 것이다. 그리고, 이 분야는 인력양성이 유독 힘들다는 점이 현재 전세계적으로 나타나고 있는 특징이기도 하다. 우선 기존의 생물학 분야의 교육에서는 위에서 설명한 것과 같은 수학/전산학/통계학 등이 크게 결핍되어 있었다. 지금까지 대부분의 소위 분자생물학자들은 이러한 지식들에 대해서 대학학부를 지나면서 거의 담을 쌓고 지내게 되고, 또한 이러한 수리적인 지식과 사고능력은 젊은 시절부터 차근히 쌓아가지 않으면 체득하게 되기가 상당히 힘든 것인 점도 있다. 그 다음 문제는 기존의 생물학 교육과는 상당히 이질적인 것들을 포함한 교육을 해낼 수 있는 시스템을 실제로 만들어내는 것이 힘들다는 현실적인 점도 있다. 반면에 생물정보학의 한쪽 절반을 이미 가지고 있는 사람이라 할 수 있는 전산학자, 통계학자, 수학자들은 또한, 소위 "분자 또는 물질"에 대한 것들을 배우는 분야에 속해 있지 않다는 점이 있다. 전산학자 혹은

통계학자들은 이 세상에 많지만, 이와 동시에 생물학자이기도 한 사람이 극히 드문 것이 바로 오늘날 전세계적으로 생물정보학 전문가 부족이 극심한 이유의 핵심의 하나이다.

따라서, 국내에서도 위에서 설명한 바와 같은 생물정보학의 특징들을 잘 이해를 하고, 제대로 된 기초 지식을 가지고서 생물정보학 인력의 양성을 위한 계획들을 세워나가야만 할 것이다. 이것은 21세기의 가장 큰 산업이 된다고 하는 바이오텍을 위한 핵심적인 준비이므로, 결코 늦출 수도 또한 성급하게 졸속으로 준비를 해서도 안 되는 그런 까다롭고 힘든 일일 것이다.