메타지노믹스 – 빅데이터 시대의 새로운 도전

한국생명공학연구원 ▌오 정 수
충북대학교 ▌조 완 섭*

1. 서 론

2010년 전세계에서 생성된 데이터는 zettabyte를 넘어섰고, 데이터의 양은 매년 40%씩 증가하고 있다[1]. 이러한 데이터의 원천을 보면 비즈니스 데이터, 멀티미디어 컨텐츠, 스마트폰과 SNS 데이터, M2M 스트림 데이터 등으로 구분해 볼 수 있으나 그 중에서도 의생명 관련 데이터는 크기나 복잡도 측면에서 다른 어떤 종류의 데이터보다 크고 복잡한 빅데이터로 간주되고 있다. 이러한 빅데이터 시대를 맞이하여 컴퓨팅 기술도 전반적으로 빅데이터를 처리하기 용이한 방향으로 하드웨어와 소프트웨어가 진화하고 있다. 본 고에서는 생명공학 분야의 빅데이터로 한정하여 그 현황과 대응방안을 논의하고자 한다.

차세대 시퀀싱 기술(NGS: Next Generation Sequencing)의 등장으로 기존보다 적은 비용으로 빠르게 많은 양의 서열을 생산하는 것이 가능해 짐에 따라 생물학자들에게 빅데이터에 대한 이슈는 더 이상 새로운 것이 아니다. 지난 10년간 메가베이스(Megabase) 당 시퀀싱 비용은 18개월 마다 컴퓨터의 성능이 2배씩 증가한다는 무어의 법칙(Moore's law)을 이미 크게 넘어섰으며, 시퀀서 한 대가 몇시간 만에 수십 기가바이트(Gigabyte)의 데이터를 생성하는 시대가 되었다.

과거에는 1000 Genome 프로젝트 등과 같이 인간과 관련되거나 동물이나 식물들과 같이 비교적 큰 유전체를 가진 것들에 대한 생물학 빅데이터가 주로 생성되었다. 그러나 최근 환경내의 미생물을 직접 분석하기 위한 접근법인 메타지노믹스(Metagenomics)가 각광받음에 따라 생물학에서의 빅데이터는 더욱더 큰 이슈가 되고 있다. 지구상의 자연환경 어디에나 존재하고 있는 모든 미생물을 대상으로 데이터가 생산되기 때문에 기존 생물학 분야에서도 경험해 보지 못했던 방대한 양의 데이터가 생산되고 있다.

이에 따라 차세대 시퀀싱 데이터를 분석, 관리, 저장, 전송하는 것에 대한 전산학적 문제가 꾸준히 발생되어 왔다. 이러한 문제의 가장 큰 요인은 다른 분야의 빅데이터와 다르게 데이터를 생산, 분석, 관리하는 주체가 일반 소규모 연구실일 수 있기 때문이다. 즉, 차세대 시퀀싱 기술의 발전으로 큰 연구소나 기업이 아닐 지라도 일반 소규모의 대학 실험실에서 조차 적은 비용으로 많은 양의 메타지놈(Metagenome) 서열 데이터를 생산하는 것이 가능하게 되었지만, 이렇게 생산된 데이터를 감당할 만한 고가의 전산장비를 갖추지 못해 분석 및 관리에 어려움을 겪고 있다.

그러나 메타지놈 빅데이터가 연구자들에게 좌절만을 주는 것은 아니다. 이러한 어려움에도 불구하고 메타지놈 빅데이터를 잘 활용한다면, 아직까지 많은 부분이 밝혀지지 않은 환경 내에서의 미생물들의 신비를 풀어갈 수 있는 중요한 실마리를 찾을 수 있을 뿐만 아니라 그 속에서 새로운 지식을 추출할 수 있는 기회가 될 수 있다. 또한 빙산의 일각만 보고 있는 우리의 시야를 획기적으로 넓혀줄 원동력이 될 수 있다.

본 고에서는 메타지노믹스 영역에서의 빅데이터의 상황 및 그것의 장점과 문제점을 고찰해보고 지금까지 이를 해결하기 위한 접근 방법에 대해 논의 해보고자 한다. 본 논문의 구성은 다음과 같다. 제 2장에서는 메타지노믹스의 개념과 필요성에 대해 요약 정리한다. 제 3장에서는 메타지놈 빅데이터를 다룰 때 발생하는 문제점에 대해 알아본다. 제 4장에서는 메타지놈 빅데이터 문제에 대해 현재까지 주로 적용된 빅데이터 기술은 무엇인지 알아본다. 이를 통해 메타지놈 빅데이터에 대해 이해하고 더 나은 해결 방안을 도출하는데 도움이 되고자 한다.

2. 메타지노믹스

메타지노믹스는 자연 환경에서 존재하는 미생물들

* 종신회원

을 배양하지 않고 직접적으로 분석하기 위한 연구 분야이다. 메타지놈 빅데이터에 대한 이해를 돕기 위해 메타지노믹스의 개념에 대해 살펴보기로 하자.

2.1 메타지노믹스의 개념

미생물은 어디에나 존재하며 전체 생물 중 가장 많은 개수를 점하고 있다. 현재까지 지구상에 있는 미생물의 개수를 추정하면 1030개로 가늠하기 힘들 정도의 어마어마한 양이다[2]. 이러한 미생물들은 그들이 속한 환경과 끊임없는 상호작용을 통해 영향을 미치고 있다. 따라서 “어떤 미생물들이 존재하고 있는가?”, ”왜 그 환경에 존재하고 있는가?” 그리고 “그들이 그 환경 내에서의 어떤 역할을 하고 있으며 어떻게, 어떤 것들과 상호작용을 하고 있는가?”를 이해하는 것은 우리가 살고 있는 환경을 이해하는데 매우 중요하다. 또한 이러한 미생물들의 특정 환경 내에서의 역할을 규명함으로써 산업, 환경, 에너지, 농업, 의료 등 인간의 삶과 관련된 전반적인 모든 분야에 도움을 줄 수 있다(미생물을 이용한 폐수처리, 화학적 화합물 생산, 의약품 생산, 바이오 에너지 생산 등에 이용)[3].

미생물 생태학(Microbial ecology)은 지구상의 생물계(Biosphere)에 존재하는 미생물의 역할과 그들과 환경간의 상호작용을 이해하기 위한 분야로서, 전통적인 미생물 생태학은 실험실에서 미생물을 배양하고 이것을 분석하는 것에 기반하고 있다. 그러나 표 1에서 보듯 환경에 존재하고 있는 미생물들의 99% 이상은 실험실에서 배양할 수 없다[4].

또한, 일반적으로 미생물은 그 주위의 미생물, 또는 환경과 상호작용을 하며 존재하기 때문에 비록 실험실에서 배양할 수 있는 것이라 할지라도 그들의 생육지 내에서의 역할과 미생물의 군집(Community)에 대해 정확히 반영하지 못한다.

표 1 자연 환경에 존재하는 배양 가능한 미생물의 비율[4]

Habitat	Clulturability(%)
Seawater	0.001-0.1
Freshwater	0.25
Mesotrophic lake	01.-1
Unpolluted esturaine waters	0.1-3
Activated sludge	1-15
Sediments	0.25
Soil	0.3

메타지노믹스는 그림 1에서 보듯 자연 환경에서 존재하는 미생물들을 실험실에서 배양 가능하거나 혹은 가능하지 않거나 상관없이 환경에서 직접적으로 유전물질을 추출하여 어떤 것들이 존재하는지 또는 그들이 어떤 역할을 하는지 분석하기 위한 연구 분야로, 기존의 전통적 미생물 생태학의 한계를 뛰어 넘는 접근법이다.

메타지노믹스란 용어는 “Complex”를 뜻하는 그리스어 “Meta”와 유전체학을 뜻하는 영어 “Genomics”가 합성된 것으로서 Jo Handelsman(1998)에 의해 새로운 용어로 1998년 처음 차용되었다[5]. 메타지노믹스는 분류하는 기준에 따라 여러 가지로 나뉠 수 있지만 흔히 시행되고 있는 일반적인 접근 방법에 따라 나누자면 단일 유전자 접근법(single gene approach, target gene approach)과 홀 지놈 샷건 시퀀싱 접근법(whole genome shotgun sequencing)으로 나눌 수 있다.

- **단일 유전자 접근법**

단일 유전자 접근법은 환경 샘플로부터 추출한 16S rRNA(small subunit of the ribosomal RNA)와 같은 특정 유전자를 활용 미생물의 다양성 및 분포를 연구하

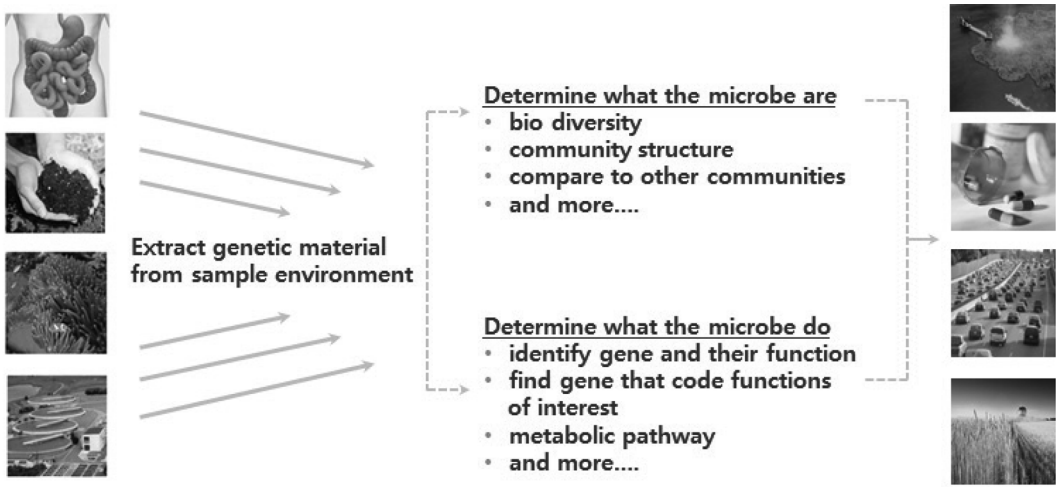

그림 1 메타지노믹스 모식도

기 위한 것으로서, 메타지노믹스가 도입되기 이전부터 미생물 생태학에서 다양성 분석하기 위해 사용된 방법이다. 이 접근법에서는 454 파이로시퀀싱(Pyrosequencing) 플랫폼이 주로 사용된다. 아직까지 SSU rRNA 전체 영역을 커버 할 수 있는 차세대 시퀀싱 플랫폼은 나오지 않고 있으나 다른 플랫폼에 비해 긴 서열(600에서 1000bp까지)을 생산할 수 있는 454 파이로시퀀싱 플랫폼이 다른 것들에 비해 높은 다양성을 갖는 영역을 식별하기에 유용하기 때문이다.

• 홀 지놈 샷건 시퀀싱 접근법

단일 유전자 접근법은 특정 환경 내에서의 미생물의 다양성 연구에 적합하지만 존재하는 미생물이 어떤 역할을 하는지, 그 미생물에 속한 유전자의 환경 내에서의 기능은 무엇인지는 알기 어렵다. 홀 지놈 샷건 시퀀싱을 통한 접근법은 미생물이 환경 내에서 어떤 역할을 하는지 알아보기 위한 접근법이다. 단일 유전자 접근법과는 다르게 특정 유전자만을 선정하여 분석하는 것이 아니라 환경 시료에서 존재하는 미생물의 모든 유전자를 대상으로 한다. 일반적으로 홀 지놈 샷건 시퀀싱은 일루미나(Illumina) 사의 차세대 시퀀서를 이용한다. 비록 일루미사의 플랫폼은 454 파이로시퀀싱 플랫폼에 에 비해 서열의 길이는 짧지만 적은 비용으로 훨씬 더 많은 서열 데이터를 생산할 수 있기 때문에 많은 유전자의 서열을 생산해 내야 하는 샷건 시퀀싱 방법에 더 적합하다.

2.2 메타지놈 빅데이터

초기의 메타지놈 연구는 생거(Sanger)시퀀싱 및 클로닝(Clonning) 기반으로 수행되었기 때문에 데이터를 생산하는데 기술적, 비용적 한계가 있었다. 이에 따라 환경에서 우점하고 있는 미생물만 추출할 수 있거나 정확한 미생물 군집 상태를 반영하지 못해 메타지놈 연구에 한계가 있었다.

시퀀싱 기술의 발전으로 인해 저비용으로 빠르게 많은 양의 서열을 얻는 것이 가능해 지면서 메타지노믹스에 많은 변화를 가져왔다. 이러한 변화 중 가장 큰 것은 그 동안 시도해 보지 못한 대규모 메타지놈 연구를 수행하는 것이 가능해졌다는 것이다. 미국국립보건원이(NIH) 2007년부터 5년간 세계 80여 개국과 공동으로 진행한 인체 미생물 군집 프로젝트(HMP: Human Microbiome Project)가 대표적이다. 250여명의 건강한 남녀로부터 인체의 각 부분에서 샘플을 채취해 그곳에 존재하고 있는 미생물들을 분석하였고, 이를 통해 우리가 아직 모르고 있던 미생물의 인체에서의 역할에 대한 놀라운 사실들을 밝혀 낼 수 있었다[6]. 이 프로젝트에서 생산된 로우(Raw) 서열 데이터의 양만해도 17테라바이트(Terabyte)가 넘는 양이다. 숫자로만 보면 일반적으로 알고 있는 빅데이터보다 적은 양이라고 생각되지만 1테라바이트는 1조 개의 염기서열이며 우리가 분석해야 할 것은 염기서열이라고 감안했을 때 이는 엄청난 양이다.

이 후 이를 뛰어넘는 대규모 메타지놈 프로젝트인 지구 미생물군집 프로젝트(EMP: Earth Microbiome Project)가 현재 진행 중이며, 이는 지구상의 20만개의 환경 샘플에 대한 메타지놈 연구를 목표로 하고 있기 때문에 여기서 생산될 데이터이 양은 가늠하기 힘들 정도이다[7].

앞으로 3세대, 4세대 시퀀싱 기술 등장할 것으로 예고되고 있고[8], 이러한 기술들이 보편화 될수록 대규모 메타지노믹스 프로젝트는 더 자주 수행될 것이며 메타지놈 빅데이터 처리 문제는 가중될 것이다. 그렇지만 이러한 어려움에도 불구하고 빅데이터가 베일에 가려져 있는 미생물 생태의 신비를 파헤칠 수 있는 중요한 단초를 제공해 준다는 것은 명백한 사실임에 틀림없다.

3. 메타지놈 영역에서의 빅데이터 이슈

메타지놈 빅데이터를 다루는데 있어 나타나는 각종 이슈들은 이미 기존의 생물학 분야에서의 제기되었던 빅데이터 이슈와 별반 다르지 않다. 여기서는 각 이슈별로 메타지노믹스에서 맞닥뜨리게 되는 빅데이터 문제에 대해 요약 정리한다.

3.1 데이터 저장의 이슈

다른 영역에서의 빅데이터도 마찬가지겠지만 메타지노믹스에서도 방대한 양의 데이터로 인해 저장 공간 부족 등의 문제가 나타날 수 있다. 일반적으로 미생물 하나의 지놈 사이즈는 4 메가바이트(Megabyte) 정도로 매우 작지만 메타지놈 연구를 위해 채취한 환경 시료에 있는 전체 미생물을 대상으로 했을 때 생산되는 데이터의 양은 엄청난 크기이다. 예를 들어 토양 1그램에 있는 미생물들을 대상으로 할지라도 약 4,000만개의 미생물이 존재하기 때문에 160테라바이트라는 엄청난 양의 서열 데이터가 생산된다. 따라서 연구의 주제나 범위에 따라 수 테라바이트에서 페타바이트(Petabyte)까지 그 크기를 가늠하기 힘들다. 더군다나 그림 2에서 보듯 이미 시퀀싱 비용의 감소 추세는 2004년을 기점으로 데이터 스토리지 비용 하락 추세를 앞질렀으며, 격차는 시간이 갈수록 더욱 벌어지고 있다.

또한 분석을 하지 않은 원시 데이터의 양도 매우 많

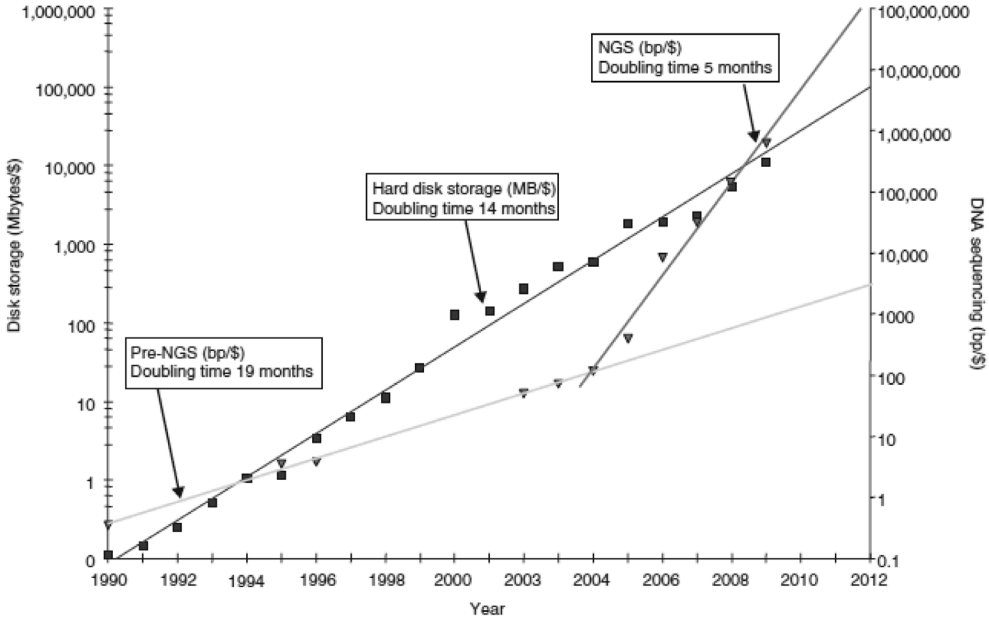

그림 2 연도별 시퀀싱 비용과 디스크 스토리지 가격 추세[9]

지만 분석을 진행할수록 부가적으로 생산되는 데이터의 양은 원시 데이터의 몇 배가 될 수 있다는 사실은 연구자들을 압박하는 요인 중 하나다. 그러나 우리가 진정으로 간과하고 넘어가서는 안 되는 것 중 하나는 하드디스크는 언제라도 고장이 날 수 있기 때문에 소중한 데이터의 일부가 언제라도 없어질 수 있다는 점이다.

3.2 데이터 관리의 이슈

실제 메타지놈 서열데이터와 분석결과를 저장하는 것도 중요하지만 그것의 환경, 시간, 온도 등과 같은 특성 정보도 간과해서는 안 된다. 미생물들은 그들이 속한 환경과 끊임없이 상호작용을 하기 때문에, 이러한 속성들은 환경 내에 있는 미생물의 상태 변화에 영향을 미치는 요소로서 추후 분석에 중요하게 사용될 수 있다. 또한 새로 생성한 데이터와 기존의 데이터에 대해 비교 분석을 수행하는 경우에도 이러한 메타 정보를 관리하는 것이 필요하다. 그러나 대부분 메타지놈 연구를 진행하면서 관리되는 데이터는 일반 파일 형식이거나 혹은 압축된 파일 형식이 많다. 데이터의 양과 메타지놈 샘플의 개수가 많아질수록 연구자들도 모르게 중복된 데이터로 인해 저장 공간의 낭비가 발생할 가능성이 높고 데이터가 어디에 있는지, 어떤 데이터인지 바로 찾기가 힘들다. 또한 분석 결과를 체계적으로 관리하는 것이 쉽지 않다. 이러한 데이터는 다른 빅데이터의 특성과 비슷하게 비정형적인 데이터이거나 방대한 양으로 인해 우리가 일반적으로 생각하는 데이터베이스로는 관리가 힘들 측면이 있다.

3.3 데이터 전송의 이슈

다른 생물학 분야에서도 그렇듯 메타지놈 연구에서도 국제적인 협력을 통해 여러 나라가 동시에 프로젝트를 수행하거나, 혹은 데이터의 생산과 분석을 서로 다른 연구실에서 나누어 수행하는 일이 많다. 특히 전산학적 지식이 부족한 일반 생물학자들은 분석을 위해 회사나 다른 연구자에게 의뢰를 하는 경우가 비일비재하다. 이럴 경우 지역 간 혹은 국가 간 데이터를 서로 주고받아야하는 일이 발생하는데 빅데이터를 안정적으로 먼 곳까지 전송하는 것조차 쉬운 일이 아니다. 가장 확실한 방법은 데이터가 담긴 하드디스크를 배달하는 것이지만 이는 너무나 비효율적인 방법으로 근본적인 해결책이 아니다. 따라서 일반적으로 메타지놈 빅데이터를 전송하기 위해서 가장 많이 쓰이는 방법은 인터넷을 통해 주고 받는 방법이지만 데이터의 크기가 큰 관계로 네트워크에 부하가 많이 걸릴 수 있기 때문에 안정적인 전송을 담보하지 못한다. 또한 분석을 진행하면 할수록 데이터 사이즈가 커지기 때문에 데이터를 주고받는데 어려움이 가중될 수 있다.

3.4 데이터 분석의 이슈

연구자들이 메타지놈 연구를 수행하는데 있어 가장 어려움을 겪는 부분은 분석하는데 있다. 방대한 데이터의 양으로 인해 많은 전산자원을 필요로 하지만 일반 연구자들이 고가의 전산자원을 갖추기란 쉽지 않다. 더군다나 그림 3에서 보듯이 시퀀싱 기술의 발전 속도가 무어의 법칙으로 대변되는 컴퓨터의 성능의 발전 속도

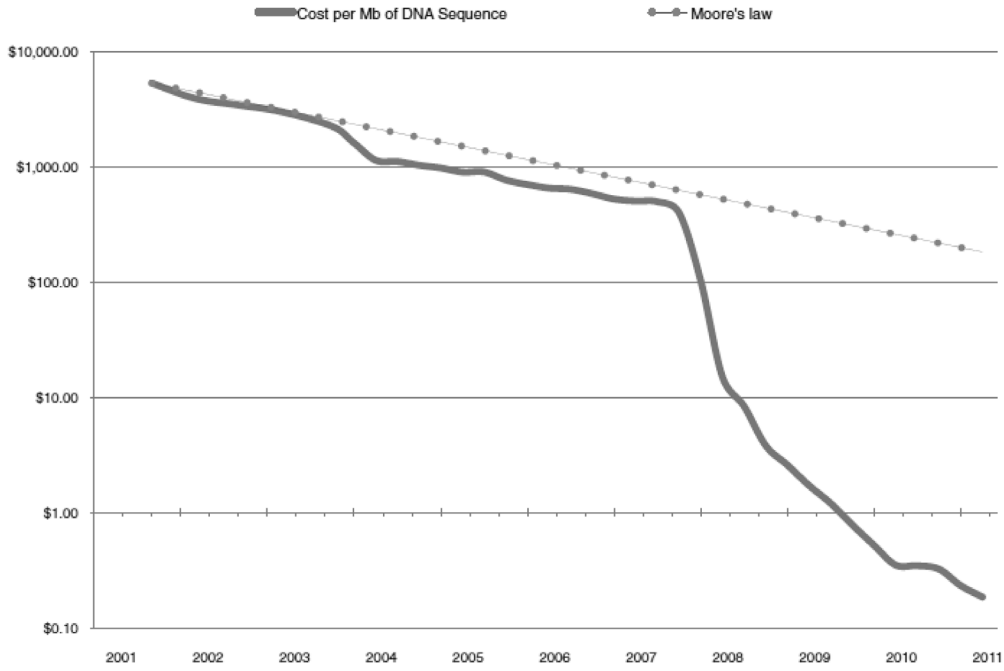

그림 3 MB당 DNA 시퀀싱 비용[10]

보다 더 빠르고 그 격차는 점점 벌어지고 있어 전산자원의 보충만으로 해결이 그리 쉽지 않다.

보통 메타지놈 데이터를 분석하기 위해서는 분석 목적에 따라 몇 단계를 거쳐야 하는데 그중 가장 많은 부하를 야기하는 것은 Binning과 어셈블리(Assembly) 과정이다.

Binning이란 미생물 다양성 분석 시 필수적인 작업으로 각 서열들을 OTU(Operational Taxonomic Unit)에 할당하는 작업을 뜻한다. 정확한 결과를 도출하기 위해서는 각 서열간 유사성 비교를 해야 하며 최대 2n-1개의 서열 정렬을 수행해야 하기 때문에 시간도 오래 걸리고 많은 CPU자원이 필요하다.

차세대 시퀀서로부터 나온 서열은 전통적인 생거기법에 비해 길이가 짧다. 특히 휼지놈 샷건 시퀀싱에 주로 쓰이는 일루미나의 제품의 경우 데이터 산출량은 매우 많지만 짧은 서열 길이로 인해 전체 유전자를 커버하지 못하는 단점이 있다. 이를 해결하기 위해 짧은 서열들을 길 서열로 구성하는 서열 어셈블리 작업이 수반되어야 한다. 그러나 휼 지놈 샷건 시퀀싱 같이 많은 종이 섞여 있는 메타지놈 샘플에서 모든 유전자를 대상으로 추출한 서열을 어셈블리 하는 것은, 서로 다른 종에 있는 유사한 서열 영역들이 잘못 어셈블리 될 수 있기 때문에 매우 어려운 작업이다. 특히 미생물들은 수평적 유전자 이동(HGT: Horizontal gene transfer)을 통해 다른 종간 유전 물질을 주고받아 Mosaicism이 발생할 수 있다. 따라서 다른 종의 서열인데 같은 종의 서열로 어셈블리 되어, 이 후 분석에서 잘못된 결과를 도출할 수 있다. 또한 어셈블리 과정 시 메모리상에서 서열을 로딩한 후 각 서열들을 모두 비교해야 하기 때문에 많은 메모리 자원이 필요하다. 일례로 인간 장내 미생물 메타지놈 데이터에 대해 어셈블리 수행 시 512 기가바이트(GB)라는 엄청난 양의 메모리가 필요로 하였다[11]. 이러한 빅 메타지놈에 대한 분석문제를 해결하기 위해서는 정확도를 높일 수 있는 새로운 알고리즘 개발과 함께 분산 및 병렬처리 기술이 필수적으로 적용되어야 한다.

4. 메타지놈 분야에서의 빅데이터 기술 활용

전 장에서 살펴보았듯 메타지노믹스에서 빅데이터를 다루는 데 있어 많은 어려움과 문제가 있다. 빅데이터 기술을 중심으로 이러한 문제를 해결할 수 있는 방안에 대해 정리해 보고자 한다.

4.1 하둡과 맵리듀스

클라우드 상의 하둡(Hadoop)은 대량의 데이터를 처리하기 위한 분산처리 프레임웍으로 구글 맵리듀스(MapReduce)와 구글 파일시스템(Google File System)의 자바 오픈소스 구현체이다. 자바 오픈소스 검색엔진 라이브러리로 유명한 루씬의 창시자 더그커팅이 야후의 지원을 받아 구현하였다. 현재는 빅데이터의 산업에서 없어서는 안 될 핵심 요소로서 자리 잡고 있으며 메타지노믹스 분야에서 대용량 데이터 문제의 해결책

으로서 점점 주목받고 있다.

하둡을 통해 얻을 수 있는 가장 큰 장점은 맵리듀스를 통한 빅데이터의 분산처리를 들 수 있다. 이미 생물학분야에서는 하둡의 맵리듀스를 기반으로 차세대 시퀀서로 부터 나온 빅데이터를 분석하기 위한 프로그램들이 꾸준히 등장해 왔으며, 빅데이터 처리를 난감해하던 연구자들에게 많은 도움을 주었다. 메타지노믹스 분야에서도 아직 많이 나오진 않았지만 최근 하둡 맵리듀스 기반의 어셈블러나 클러스터링 도구들이 등장하기 시작했다[12,13]. 이러한 흐름은 분석시 부하가 많이 걸리던 연산이나 메모리 부족 등의 어려움을 해소하는데 도움을 주며, 연구자들이 기존에 시도해보지 못하던 큰 규모의 메타지놈 데이터 분석을 할 수 있게 되어 아직도 베일에 싸여있는 다양한 환경에서의 미생물의 생태와 역할에 대한 이해를 높이는데 중요한 역할을 할 것이다.

하둡을 적용함으로서 얻을 수 있는 또 하나의 이점은 데이터의 저장과 관리가 용이해 진다는 것이다. 하둡의 분산파일 시스템인 HDFS(Hadoop file system)를 사용하면 값비싼 스토리지가 없더라도 일반 리눅스 장비에 있는 하드디스크들을 묶어서 하나의 스토리지로서 사용할 수 있게 되어 기존보다 저비용으로 큰 데이터를 저장할 수 있다. 또한 저장 공간이 부족할 때 마다 바로 추가가 가능할 뿐만 아니라 파일을 분산, 복제하여 저장하기 때문에 하드디스크 장애가 일어난다 할지라도 복구가 가능하다. 또한 Hbase, MongoDB, Cassandra와 같은 NoSql 도구들을 활용함으로서 일반 관계형 데이터베이스 시스템(RDBMS)이 처리하기 힘든 빅 메타지놈 데이터에 대한 메타정보나 분석결과 등의 정보들을 신속하게 검색하고 추출할 수 있게 되어 관리가 용이해진다.

최근들어 비교적 복잡한 맵리듀스 과정을 간단하게 처리할 수 있도록 하는 도구들이 속속 나타나고 있으며, Hive나 Pig 등이 대표적인 예이다. 또한, 수천가지의 고도의 통계 분석과 마이닝 알고리즘을 라이브러리 형태로 제공하는 R도 빅데이터 분석의 필수 도구로 활용되고 있다.

4.2 클라우드 컴퓨팅

일반 연구자가 자신의 실험실에 전산장비를 갖추고 빅 메타지놈 데이터를 다룰 수 있다면 아주 이상적인 상황일 것이다. 그러나 우리가 알고 있는 현실은 이것이 쉽지 않다는 것이다. 가장 큰 이유는 빅데이터 분석을 위한 고가의 전산장비를 갖추기가 힘들다는 것과 일단 한번 갖춘다고 하더라도 데이터사이즈가 점점 늘어남에 따른 추가적인 비용을 부담하기가 만만치 않다는데 있다. 또한 전산학적 지식이 부족한 일반 생물학자들이 서버나 클러스터와 같은 전산장비를 다루고 관리하는 것이 쉽지는 않다.

생물학 분야에서 이에 대한 요즘 떠오르는 대안 중 하나는 클라우드 컴퓨팅(Cloud computing)이다. 클라우드 컴퓨팅은 대용량 서버나 스토리지를 갖춘 클라우드 서비스 업체에 일정 금액을 내고 온라인 상에서 전산자원을 빌려서 사용하는 것을 말한다. 현재 표 2에서 보듯 Amazon을 비롯한 다양한 업체들이 클라우드 서비스를 제공하고 있다.

표 2 클라우드 컴퓨팅 및 이기종 컴퓨팅 환경 제공 사업자 목록[14]

Environment	URL
Cloud computing	
Amazon Elastic Compute Cloud	http://aws.amazon.com/ec2
Bionimbus	http://www.bionimbus.org
NSF CluE	http://www.nsf.gov/cise/clue/index.jsp
Rackspace	http://www.rackspacecloud.com
Science Clouds	http://www.sciencecloud.org
Heterogeneous computing	
NVIDIA GPUs	http://www.nvidia.com
AMD/ATI GPUs	http://www.amd.com
Heterogeneous cloud computing	
SGI Cyclone Cloud	http://www.sgi.com/products/hpc_cloud/cyclone
Penguin Computing On Demand	http://www.penguincomputing.com/POD/Summary

심지어 차세대 시퀀싱기술의 선두 주자 중 하나인 일루미나사 조차 BaseSpace라는 차세대 시퀀싱에 특화된 클라우드 컴퓨팅 서비스를 제공한다. 클라우드 컴퓨팅은 더욱 빠르게 메타지놈 데이터를 분석할 수 있게 도와 줄 뿐만 아니라 연구자들이 전산자원에 대한 비용 문제를 경감시켜줄 대안으로서 그 활용가치가 매우 높다.

4.3 이기종 컴퓨팅

몇 년 전부터 과학 및 엔지니어링 분야에서 그래픽 처리장치(GPU) 활용하여 처리 속도를 높이려는 시도가 있어왔다. GPU는 CPU에 비해 가격도 저렴할 뿐만 아니라 태생적으로 여러 개의 작업을 동시에 처리하도록 설계된 병렬 프로세서이기 때문에 대량의 데이터 분석에 CPU보다 빠르다는 장점이 있다. 이기종 컴퓨팅(heterogeneous computing) 시스템은 이런 GPU의 장점을 활용해 어플리케이션의 처리 속도를 높이기 위해 CPU와 GPU로 이루어진 시스템을 말한다. CPU는 어플리케이션의 순차적 실행에 사용되고 GPU는 데이터 처리에 사용하여 보다 빠른 성능을 이끌어 낼 수 있도록 한다.

그림 4에서 보듯 대량의 염기 서열데이터를 다뤄야 하는 생물학분야에서도 이런 GPU의 특징을 활용하면 엄청난 처리 성능 향상을 이룰 수 있다는 것을 알게 되면서 GPU 기반의 분석 도구들이 속속 개발되었다.

최근 이러한 흐름에 맞춰 메타지노믹스 영역에서도 GPU를 활용하여 대량의 데이터를 빠르게 처리하려는 움직임이 있으며 MetaBinG[16], Parallel-META[17], GHOSTM[18] 등과 같은 도구들이 선보여졌다. 또한 앞으로 메타지노믹스에 많이 쓰일 것으로 예상되는 반도체 기반 차세대 시퀀서 중 하나인 아이온 프로톤(Ion Proton)은 처리 속도 향상 및 비용절감을 위해 GPU를 채택하였다.

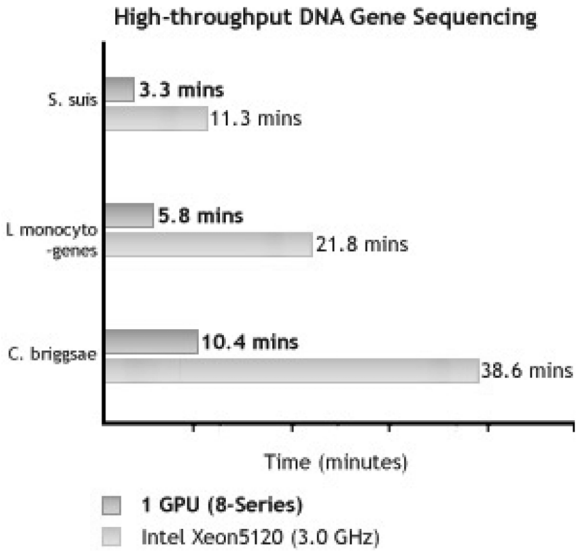

그림 4 High-through DNA sequence alignment using GPUs [15]

아직까지는 메타지노믹스 영역에서 이기종 컴퓨팅의 활용이 그리 많지는 않지만 앞으로 생산될 메타지놈 데이터의 양을 고려해 볼 때 이기종 컴퓨팅을 활용한 대용량 메타지놈 데이터 처리는 더 각광을 받을 것으로 예상된다.

4.4 빅데이터 전송 솔루션

앞서 빅데이터의 전송 문제에 대해 살펴보았듯 방대한 양의 데이터를 인터넷상에서 전송하는 것은 연구자들에게 또 다른 이슈가 되고 있다. 연구자들끼리 데이터를 공유하는데도 문제가 될 수 있지만 클라우드 서비스를 이용하는데 있어 걸림돌이 되고 있다. 아무리 고성능의 전산장비를 제공해 준다 하더라도, 먼저 원격에 있는 클라우드 서버로 빅데이터를 전송하지 못한다면 연구자들에겐 클라우드 서비스 업체에서 제공하는 전산장비조차 무용지물인 것이다.

그러나 다행히 이 문제에 대한 솔루션이 우리에게 존재하고 있고 성공적으로 이 문제를 극복하고 있다. 현재 까지 가장 잘 알려져 있고, 가장 먼저 생물학 분야에서 적용된 솔루션은 아스페라(Asprea)사의 fasp기술이다. 아스페라의 대용량 데이터 전송기술은 영화 반지의 제왕이나 아바타 제작시 영화를 편집하는 미국과 컴퓨터 그래픽을 담당하는 뉴질랜드 간 수테라 바이트의 데이터를 네트워크를 통해 대륙 간을 넘어 주고받는데 사용되어 유명해졌다. 이미 몇 년 전부터 아스페라의 전송 솔루션은 SRA(Sequence Read Archive)와 같은 차세대시퀀싱 서열을 저장하기 위한 공용 저장소에도 사용되어 왔으며, 현재 아마존과 같은 대형 클라우드 서비스 업체 뿐만 아니라 중국의 BGI(Beijing Genomics Institute)에서 운영하는 클라우드 서비스에까지 적용되어 활용되고 있다. 이 외 Globalspace 사에서 제공하는 Enhanced File Transfer 기술이나 삼성 SDS의 Rapidant같은 빅데이터 전송 기술도 있다. 우리가 이러한 기술을 잘만 활용한다면 빅데이터 전송 시의 어려움은 상당 부분 해결 할 수 있을 것이다.

5. 결 론

최근 몇 년간 빅데이터는 과학, 의학, 경제를 비롯한 다양한 분야에서 활용되어 왔으며 빅데이터를 다루는

사람들뿐만 아니라 일반인들도 점차 빅데이터가 가진 힘에 주목하기 시작했다. 메타지노믹스도 예외는 아니어서 이러한 흐름에 동참하기 시작하였고 우리 주변 어디에나 존재하고 있지만 그 동안 베일에 싸여 있던 미생물의 신비를 탐색하는데 이러한 빅데이터의 활용은 필수적이게 되었다. 메타지놈 빅데이터는 빙산의 일각만 보던 우리의 시야를 넓혀줄 것이며 농업, 의학, 환경, 에너지 등 우리 인간 삶 전반에 걸쳐 영향을 미칠 수 있는 엄청난 파급력을 가지고 있다.

본 고에서는 메타지놈 빅데이터 대해 알아보고 그것을 다루는데 발생하는 문제점과 이를 해결하기 위한 빅데이터 기술에 대해 알아보았다. 비록 메타지놈 빅데이터를 다루는 것이 쉽지 않은 것은 사실이지만 이미 몇몇 앞선 연구자들은 빅데이터 기술을 활용하여 이것에 맞서고 있고 메타지노믹스를 이끌고 있다. 우리가 어려움을 극복하고 메타지놈 빅데이터를 잘 활용할 수 있다면 그 가치는 우리의 노력만큼 값질 것이다.

참고문헌

[1] McKinsey, Big Data: The next frontier for innovation, competition, and productivity, 2011. 05.

[2] Simon, Carola, and Rolf Daniel, Metagenomic analyses: past and future trends, Applied and environmental microbiology, 77(4), 2011

[3] Prosser, James I., et al.,The role of ecological theory in microbial ecology, Nature Reviews Microbiology, 5(5), 2007

[4] Amann, Rudolf I., Wolfgang Ludwig, and Karl-Heinz Schleifer, Phylogenetic identification and in situ detection of individual microbial cells without cultivation, Microbiological reviews, 59(1), 1995

[5] Handelsman, Jo, et al., Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, Chemistry & biology, 5(10), 1998

[6] Peterson, Jane, et al., The NIH human microbiome project, Genome research, 19(12), 2009

[7] Gilbert, Jack A., et al., Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project, Standards in genomic sciences, 3(3), 2010

[8] Niedringhaus, Thomas P., et al., Landscape of next-generation sequencing technologies, Analytical chemistry, 83(12), 2011

[9] Stein, Lincoln D., The case for cloud computing in genome informatics, Genome Biol, 11(5), 2010

[10] Sboner, Andrea, et al., The real cost of sequencing: higher than you think, Genome Biol, 12(8), 2011

[11] Qin, Junjie, et al., A human gut microbial gene catalogue established by metagenomic sequencing, Nature, 4649 (7285), 2010

[12] Yang, Xiao, Jaroslaw Zola, and Srinivas Aluru, Parallel metagenomic sequence clustering via sketching and maximal quasi-clique enumeration on map-reduce clouds, Parallel & Distributed Processing Symposium, IEEE, 2011

[13] Guo, Xuan, et al, Cloud Computing for De Novo Metagenomic Sequence Assembly, Bioinformatics Research and Applications, Springer Berlin Heidelberg, 2013

[14] Schadt, Eric E., et al., Computational solutions to large-scale data management and analysis, Nature Reviews Genetics, 11(9), 2010

[15] Schatz, Michael C., et al., High-throughput sequence alignment using Graphics Processing Units, BMC bioinformatics, 8(1), 2007 http://www.nvidia.com/object/bio_info_life_sciences.html

[16] Jia, Peng, et al, MetaBinG: Using GPUs to accelerate metagenomic sequence classification, PloS one, 6(11), 2011

[17] Su, Xiaoquan, Jian Xu, and Kang Ning, Parallel-META: efficient metagenomic data analysis based on high-performance computation, BMC systems biology, 2012

[18] Suzuki, Shuji, et al., GHOSTM: a GPU-accelerated homology search tool for metagenomics, PloS one, 7(5), 2012

약 력

오 정 수

2013 충북대학교 공학(박사)
2005~현재 한국생명공학연구원 연구원
관심분야 : 생물정보학, 메타지노믹스, 클라우드 컴퓨팅
E-mail : ofang@kribb.re.kr

조 완 섭

1996 한국과학기술원 전산학(박사)
1987~1990 한국전자통신연구원 (ETRI)
2001~2002 미국 Univ. of Florida Post-Doc. 연구원
1997~현재 충북대학교 경영정보학과 교수
관심분야 : 빅데이터, 데이터베이스, 데이터 다차원, 기업정보화
E-mail : wscho@chungbuk.ac.kr