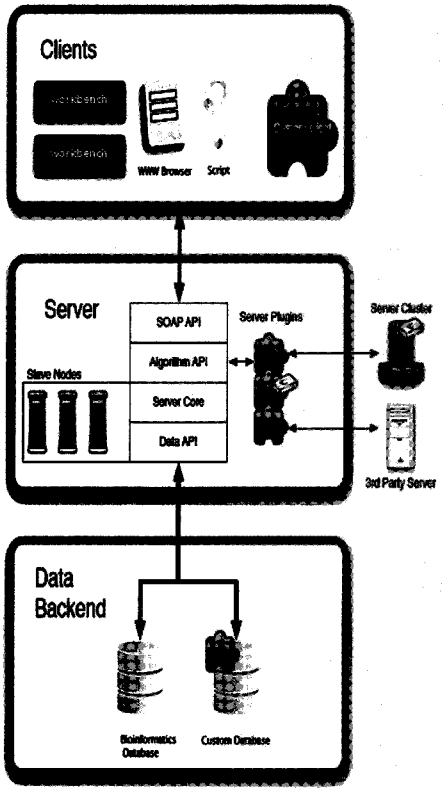
NGS 시대의 생물정보학 프레임워크에 대한 제언

강병철 ((주)인실리코젠 연구소장)

"우리 아들은 누굴 닮아서 이렇게 키가 크지?" 라는 말을 이제는 생물정보학적 분석 결과를 바탕으로 유전자의 서열정보를 보며 설명하는 시대가 왔다. 개인의 유전적 질환을 유전자 서열에 기초하여 말하고, 일상적으로 먹는 먹거리의 원산지 표시가 유전자 감식을 통해 이뤄지고 있는 등 모든 생명체의 유전 코드를 수집하고 있다 해도 과언이 아니다. 이러한 흐름은 2004년 등장한 차세대 시퀀싱(Next Generation Sequencing, NGS) 장비의 도입으로 증폭되고[1], 곧 출현할 제 3 세대 시퀀싱 장비에 의해 더욱 가속화 될 것이다[2]. 단적으로 최초의 Human Genome Project는 13년에 걸쳐 완료되었지만 현재는 일주일이면 세 사람의 유전체정보를 모두 밝힐 수 있게 되었고, 더 나아가 3 세대 시퀀싱 기술은 개인 유전체 분석에 단 100만원의 비용이면 가능한 것으로 예측된다[3]. 시퀀싱 장비 제조사의 화려한 마케팅 문구를 보면 더 이상의 새로운 도전 과제가 없는 것처럼 보이지만, 적어도 생물정보학의 입장에서는 새로운 이슈에 직면하게 되었다.

NGS 장비에서 쏟아지는 데이터는 규모와 형식 모두에서 이전의 세대와 구분이 된다. 첫 째 규모면에서 이전의 시퀀싱 장비가 1회 운전하여 생산되는 데이터 파일이 수십 메가바이트였지만, NGS는 테라바이트까지 소요되기도 한다. 또한 하나의 리드의 크기가 상대적으로 짧은 특징을 가진다. 또한 1회 분석 시 생산되는 리드 개수도 수 백 개에서 수 십 만개로 증가하였다[4]. 이러한 NGS 데이터 특징들 때문에 기존의 생물정보학 아키텍처로는 생산성과 유연성을 확보하기 어려운 측면이 있다. 즉, 증가하는 데이터 용량을 따라가기 위해 스토리지를 추가하고 네트워크상에서 연결시키기 위해서 시스템 엔지니어가 투입되고, 계산 능력을 증설하기 위해서 새로운 서버를 추가할 때 분산 처리를 위한 복잡한 그리드 엔진을 설정하고, 분산 처리를 위한 자원 관리 등 생물정보학자의 입장에서는 실제 유전체학적인 문제를 해결하기 위해서 들이는 비용보다 환경 정비를 위해서 더 많은 노력을 해야 되는 경우도 있다. 따라서 시스템 관리자, 개발자, 생물정보학자, 생물학자 모두에게 생산성을 가져다 줄 수 있는 소프트웨어 프레임워크에 대해서 생각해볼 시점이 되었다.

먼저 결론부터 제시하면 그림1과 같은 3 계층의 프레임워크를 제안해본다. 이러한 구조는 유수의 연구 기관에서도 이미 채택한 방법으로 데이터 백엔드(Data backend), 분석서버, 클라이언트(client)로 구성된다. 데이터 백엔드는 단순히 스토리지 서버와 데이터베이스 서버의 조합을 말하는 것이 아니다. 실제 현장에서 생물정보학적 분석을 수행하는데 사용되는 다양한 저장 방식에 대해서 일관된 접근 방식을 제공하고 이를 쉽게 관리할 수 있는 data persistence layer를 구성하는



[그림 1] NGS시대의 분석을 위한 3단계 시스템 아키텍처.
데이터 관리를 위한 데이터베이스, 데이터분석용 연산 서버, 생물학자를 위한 데스크탑용 워크벤치를 통해 NGS시대의 대용량 데이터 처리를 수행한다.

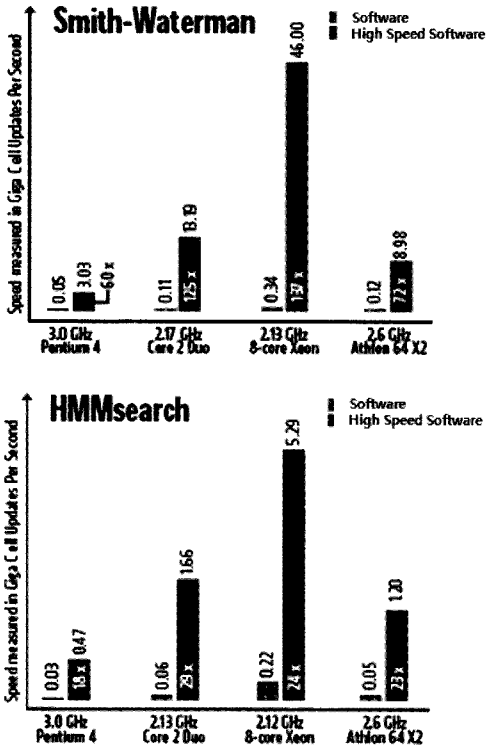
뜻 한다. 즉,분석 서버에서 다루는 데이터 오브젝트를 일반적인 하드디스크(파일시스템), 오라클, MySQL, DB2와 같이 다양한 DBMS과 연결하여 저장하고 읽을 수 있는 환경이다. 시스템 관리자 측면에서는 개별적인 데이터베이스 구조를 이해할 필요 없이 운용 가능한 데이터베이스 서버만 있으면 저장 용량을 확장시킬 수 있고, 개발자는 단일화된 API를 통해서 데이터를 접근하므로 개발이 단순해지는 장점이 있다.

두 번째, 계산용 서버 계층은 NGS 정보를 처리하기 위한 HPC (High-Performance Computing) 기능을 갖추면서 분산처리, 자원관리, 프로세스 큐 관리와 같이 대규모 분석을 위한 관리 기능이 있어야 한다. 또한 이러한 기능과 자원을 일관된 형식으로 접근할 수 있는 API를 제공함으로써 다양한 어플리케이션 개발이 용이해야할 것이다. 특히 웹이나 네트워크 어플리케이션 개발을 위해선 SOAP (Simple Object Access Protocol) API제공하거나 리스소의 접근을

RESTFUL하게 구성해야할 것이다. 계산 서버의 확장도 복잡한 시스템 설정 없이 Unix, Linux, Windows, MacOS 모두를 지원하고 각 서버마다 특정 분석 소프트웨어를 할당하고 이를 배분할 수 있어야 한다. 그러나 이러한 부분들은 Perl이나 Python 스크립트 작성에 익숙한 많은 생물정보학자에게는 여전히 복잡하고, 지난 10여 년간 개발했던 다양한 스크립트를 재활용할 수 없다는 약점이 있다. 이를 해소하기 위해서는 분석서버에서 외부 스크립트를 플러그인 방식으로 연계할 수 있는 아키텍처를 가져야한다. 즉, 생물정보학자에게는 생물학적 문제 해결에 집중하고 검증된 과거 스크립트를 최대한 활용할 수 있는 환경이 된다. HPC 지원도 중요한 이슈중의 하나이다. NGS가 등장한 이후 염기서열 데이터베이스가 10배 증가하는데 걸리는 시간이 고작 18개월에서 24개월 정도라고 한다.

Moore의 법칙에 따르면 CPU의 처리속도는 18개월마다 2배 정도 증가한다는데, 산술적으로 따져보면 서버를 아무리 증설하여 처리능력을 키우더라도 데이터 증가 속도를 따라잡을 수 없다는 결론이다. 하지만 최근 기존의 CPU에서 멀티미디어를 처리하기 위해서 준비된 기능을 생물정보학적으로 응용하여 고속 어셈블리(표1), Smith-Waterman 검색, HMM 계산을 지원하는 기술이 등장하여 계산 병목을 줄이는 데 많은 기여를 하고 있다(그림2). 마지막으로 클라이언트에 대해서 이야기해보자. 웹이 대세인 요즘에 클라이언트를 다시 제기하는 것은 웹으로는 NGS 데이터를 조회하기에는 적절하지만 실제 분석이나 재구성을 위한 인터페이스를 구현하는 데는 많은 한계점이 있다. 물론 RIA와 같은 웹 기술들이 복잡한 인터페이스 구현에 적용되지만 여전히 많은 비용이 발생한다. 따라서 데스크탑 컴퓨터에서 오피스 프로그램처럼 쉽게 사용할 수 있는 클라이언트의필요성을 다시제기한다.

그 이유는 대다수의 생물학자들은 윈도우나 맥킨토시와 같이 데스크탑이나 랩탑컴퓨터에 익숙하다. 한동안 유닉스 일색이

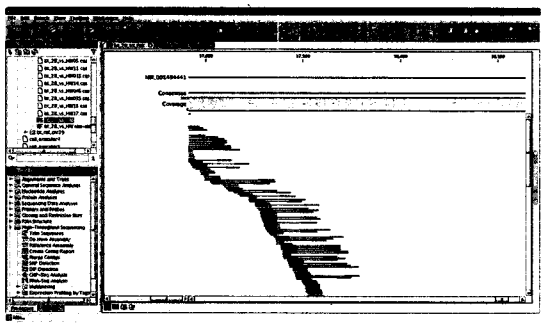


[그림 2] 초고속 연산장치를 이용한 command 방식의 생물정보 분석 프로그램의 속도 향상.
대표적인 분석 프로그램인 Smith-Waterman, BLAST, HMMsearch, HMMpfam, clustalW의 속도를 최대 130배 이상 향상 시킬 수 있다.

던 생물정보학 분야도 이러한 요구를 반영하여, 많은 오픈소스 및 상용 소프트웨어가 이러한 운영체계를 지원하였다. 그러나 NGS 장비의 등장은 다시 과거로 되돌리는 현상을 만들

[표 1] Ilumina 데이터를 이용한 CLC Assembly Cell과 ABySS 벤치마킹 결과.
Human 38X에 해당하는 130Gbp의 데이터를 이용하여 De novo assembly를 수행하였다. ABySS와의 비교로 16배의 메모리 절약과 시간 절약은 SIMD기술을 통한 생물정보학적 소프트웨어가 하드웨어 증축을 대신할 수 있음을 보여준다.

	CLC Assembly Cell	ABySS
Number of computers	1	21
Total RAM usage (GB)	21	336
Time spent (hours)	6	15
Contig >=1000bp		
Number of contigs	638,661	549,522
Mean size (bp)	1,918	1,703
Max size (bp)	18,222	15,911
N50 size (bp)	2,021	1,731
Sum (Gbp)	1.2	0.9



[그림 3] 생물학자를 위한 데스크탑용 워크벤치.
다양한 운영체제에 적용되며, 계산용 서버 및 데이터베이스와 연계하여 복잡하고 다양한 생물데이터의 분석에 전문 생물정보학자의 도움 없이 적극적인 연구 활동을 할 수 있도록 지원하고 있다.

었다. 즉 대용량 정보처리를 위해서 리눅스와 유닉스 운영체제의 서버에서 분석하고 결과를 생물학자에게 제공하는 방식이다. 그러나 결과 파일 자체도 정보량이 많아 생물학자는 파일을 열어 보는 것조차 어려운 경우도 있다. 이를 해결하기 위한 방안으로 대용량 데이터의 연산처리는 두 번째로 언급된 계산용 서버에서 처리하고 분석 결과는 데스크탑 워크벤치를 통해 서비스함으로써 생물학자의 적극적인 피드백을 유도할 수 있는 방안이 필요하다(그림 3).

가끔 고객(생물학자)으로부터 NGS 분석결과를 텍스트 파일이나 엑셀파일로 받았는데 자신의 컴퓨터에서 열어볼 수 없었다는 푸념을 듣는 경우가 있다. 요즘 유행하는 표현으로 생물정보학자와의 "소통"이 어렵다는 것이다. 이 글의 제안이 NGS 데이터 활용을 극대화하고, 생물학자에게는 자신의 데이터에 대한 소유권을, 생물정보학자에게는 문제해결에 집중할 수 있는 IT기반을, 개발자에게는 표준화된 개발환경을, 관리자에게는 확장성과 유연성을 제공하는 프레임워크를 구축하는데 도움이 되었으면 한다.

[참고 문헌]

[1] 김선영. (2009) Personal Genomics 연구 개발 동향. Biochemistry and Molecular Biology News
[2] John Eid. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science Express. 133-138.
[3] 김남신, 추인선 (2009) 차세대 시퀀싱 기술의 활용. 생화학 분자생물학 뉴스 12월호
[4] Mardis ER. (2008) The impact of next-generation sequencing technology on genetics. Trends Genet. 24(3), 133-141.

인실리코젠 Codes팀

Codes란 일련의 정보를 편지, 문장, 단어, 제스처와 같은 다른 형식의 표현으로 변경하는 것을 의미합니다. 디지털 시대에서는 모든 정보를 0과 1의 디지털 방식으로 코드화할 수 있으며, 이와 마찬가지로 모든 생물의 유전 정보는 4가지 DNA 염기(ATGC)로 코드화할 수 있습니다. Codes팀은 이와 같은 생명정보를 담고 있는 DNA 코드를 분석하는 연구자들을 위한 생물정보 컨설팅을 통해 가치 있는 정보를 제공하고 있습니다.