리눅스 클러스터기반 유전자서열분석 병렬처리 모형 개발 및 성능 검증

박미화[0*], 김재우*, 박춘구**, 유승식*
*포스데이타 솔루션개발연구소
**생물학전문연구정보센터
(bfpark[0], jaewookim, ssyoo)@posdata.co.kr
madreach@bric.postech.ac.kr

Linux Cluster based Biological Sequence Parallel Processing Model Development and Efficiency Verification

Mi Hwa Park[0*], Jae Woo Kim*, Chungoo Park**, Seung Sik Yoo*
*POSDATA Solution Development Institute
**Biological Research Information Center

요 약

Human Genome Project와 같은 대형 Sequencing 프로젝트와 High-throughput Sequencing 기술의 발전으로 현재 Expressed Sequence Tag (EST)와 같은 대량의 DNA 서열들이 생산되고 있다. 이를 효과적이고 효율적으로 분석해야 할 필요성이 증대되고 있다. 대부분의 실험자들이 서열 분석을 위해 우선적으로 BLAST 검색을 이용하고 있다. 하지만 대량의 서열, 검색 DB의 크기, BLAST 검색 결과의 복잡성에 의해 어려움을 겪고 있다. 이에 빠르고 정리된 결과를 보여줄 수 있는 BLAST 검색 시스템의 필요성이 커지고 있다. 이에 본 논문은 미국생명공학연구소(NCBI)에서 제공하는 유전자 서열 검색 툴인 BLAST(Basic Logical Alignment Tool)를 클러스터 수퍼 컴퓨터 구축 기술을 기반으로 한 병렬처리와 Gene Ontology를 이용하여 방대한 양의 서열 검색 결과를 요약하는 모형을 제시한다. 이것은 신약개발 및 유전자 발굴 등의 연구기간을 획기적으로 단축시켜 신약 개발, 농업, 화학, 의료, 환경 등 생명공학 연구에 핵심적인 역할을 할 수 있다. 또한 성능 실험을 통하여 분석결과 대기시간을 최소화하는 병렬처리모형의 효율성을 검증하였다.

1. 서론

최근 의학, 약학, 농학, 화학, 환경학 등을 아우르는 생물학에 대한 연구가 급진전되면서, 유전자서열, 예컨대 DNA 서열들의 생산량 또한 급증하고 있다[1]. 이러한 유전자서열의 생산량 증가에 비례하여, 신규 생산된 일련의 유전자서열들을 효율적으로 분석해야 할 필요성 또한 점차 증대되고 있다[2]. 이러한 생명공학분야의 급속한 발전은 대용량 고성능의 처리가 가능한 컴퓨터를 요구하고 있으나 IT인프라에 대한 막대한 투자비용이 따르기 때문에 관련 연구기관 및 기업에서 쉽게 컴퓨터를 도입하지 못하고 있는 실정이다. 이러한 상황에서 저가격 고성능의 서버와 고성능의 네트워크 기술을 접목하여 기존의 서버를 대체할 수 있는 범용적이며 시스템의 안정성과 신뢰성을 보장하는 리눅스 클러스터 컴퓨터를 개발하게 되었고, 이를 기반으로 대용량 유전자 서열 정보를 초고속으로 분석 가능한 병렬처리 시스템을 구축하여 대량의 서열 검색 결과를 생명공학분야 지식체계인 Gene Ontology (GO)를 이용하여 서열의 기능에 대한 Global View를 제공하는 클러스터기반 유전자서열 분석 병렬처리 모형을 개발하게 되었으며 그 성능을 검증하여 시스템의 효율성을 제시하고자 한다.

2. 기존 유전자분석 처리 방법과 문제점

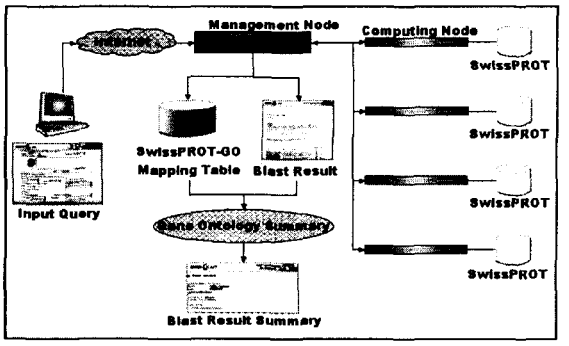
유전자서열분석(Gene Annotation)의 목적은 유전자의 key feature를 밝혀내는 것이고, 특히 유전자와 그 생산물에 대한 것을 알아내는 것이다[1]. 기존에 알려진 다른 종의 유전자서열이나 cDNA등의 similarity에 대한 사전 정보 없이 유전자를 찾는 방법과, 그런 정보를 바탕으로 유전자를 찾는 방법이 있다. 처음 방법을 Ab initio method라 하고, 둘째 방법을 Homology Method라고 한다. Homology Method 중 가장 널리 쓰이는 툴이 BLAST(Basic Logical Alignment Search Tool)이다[3]. 종래의 체제하에서 유전자서열의 주 생산처인 유전자서열 생산업체, 유전자서열 생산 연구소들은 미국생명공학연구소 (NCBI: National Center for Biotechnology Information)[4], 스위스 생물정보학 협회 (Swiss Institute of Bioinformatics)[5], 유럽 생물정보학 협회(European Bioinformatics Institute)[6] 등과 같은 저명 생물학 기관에서 운영하는 웹 시스템, 예컨대, BLAST system(Basic Local Alignment Search Tool system), 스위스-프롯 시스템(SWISS-

PROT system) 등에 온라인 접속한 후, 해당 시스템에서 제공하는 일련의 온라인 분석 서비스를 활용하여, 자가 생산한 유전자서열에 대한 상세 정보를 획득하고 있다. 그러나 BLAST 시스템, 스위스-프롯 시스템에 배치된 각 데이타베이스는 그 보유 정보량이 워낙 방대하기 때문에 유전자서열 생산업체, 유전자서열 생산 연구소 등에서 아무리 상세한 검색을 시행한다 하더라도, BLAST 시스템, 스위스-프롯 시스템 등에서는 어쩔 수 없이, 필요 이상으로 방대한 규모의 유전자서열 분석 결과를 제공할 수밖에 없게 된다. 결국, 별도의 조치가 취해지지 않는 한, 유전자서열 생산업체, 유전자서열 생산 연구소 등에서는 해당 유전자서열 분석 결과를 판독하는데 많은 어려움을 겪게 된다. 이처럼, 유전자서열 생산업체와 유전자서열 생산 연구소에서, 유전자서열 분석 결과를 판독하는데 많은 어려움을 겪는 경우, 유전자서열의 생산, 연구속도는 그만큼 더디어 질 수밖에 없으며, 결국, 의학, 약학, 농학, 화학, 환경학 등을 아우르는 생물학 전반의 발전속도가 필요 이상으로 늦어지는 심각한 문제점이 야기된다. 더욱이, 종래의 체제 하에서, BLAST 시스템, 스위스-프롯 시스템 등은 통상 전 세계에서 접속하는 다수의 유전자서열 생산업체, 유전자서열 생산 연구소 등을 상대로 일련의 유전자서열 분석 서비스를 제공하는 것이 일반적이다. 이러한 상황하에서는 보유 데이타베이스의 정보 저장환경을 아무리 개선한다 하더라도, 시스템 측에서는 각 유전자서열 생산업체, 유전자서열 생산 연구소 등의 요청 정보를 소규모로 장시간 처리할 수밖에 없게 된다. 결국, 유전자서열 생산업체, 유전자서열 생산 연구소 등에게 일련의 유전자서열 분석 결과를 최적화된 속도로 제공할 수 없게 되고, 그 결과 유전자서열 생산업체, 유전자서열 생산 연구소 등에서는 유전자서열 분석이 필요한 매 시기마다 시스템 측의 분석 결과 출력을 장시간 기다려야만 한다.

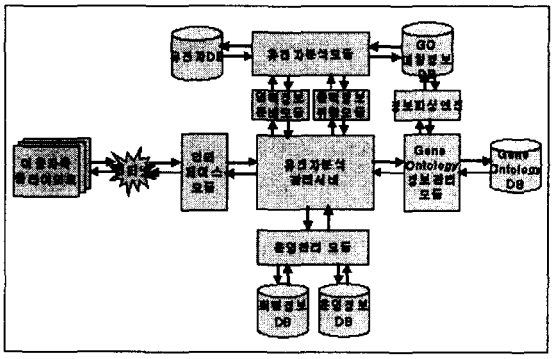
3. 유전자 서열분석 병렬처리 모형

3.1 클러스터시스템 구성도

본 논문에서 제시하는 유전자 서열분석 병렬처리 모형은 리눅스 클러스터시스템을 기반으로 한다[7]. 클러스터 시스템 구성은 [그림 1]과 같이 한대의 Management Node와 여러 대의 Computing Node로 구성되어있다. 사용자는 웹 인터페이스를 통하여 검색하고자 하는 서열을 입력 또는 업로드 시킨 후 검색을 실행한다. Management Node는 웹 서버를 통해서 전달된 입력 서열을 Computing Node 수만큼 분할한 후 각 Computing Node로 분배하여 대상 데이터베이스에 대한 서열검색 작업을 실행하도록 한다. Computing Node들은 로컬 데이터베이스를 대상으로 서열검색 작업 및 검색결과 요약을 위한 인덱스 추출 작업에 검색결과 및 인덱스 파일을 Management Node로 보낸다. Management Node에서는 검색결과를 Merge하여 사용자에게 보내주고 Gene Ontology에 의한 요약 결과 또한 Tree 구조의 GO Browser를 통하여 사용자에게 보여주어 해당 서열의 기능들을 쉽게 파악할 수 있도록 한다.



[그림 1] 유전자 서열 병렬처리 모형 클러스터 구성도



[그림 2] 유전자서열 병렬처리 및 Gene Ontology연동체계

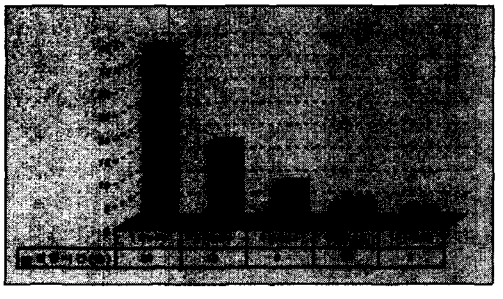
3.2 병렬처리 및 Gene Ontology 연동 알고리즘

Gene Ontology 란 Gene Ontology Consortium 에서 제공하는 생물학 관련 지식을 표현하기 위한 지식 체계로 크게 생물학 용어 controlled vocabulary 와 이를 계층적으로 표현한 Tree 구조로 되어있다[8]. [그림 2]에 나타낸 바와 같이 유전자분석 병렬모형의 구성은 크게, 자가 생산한 유전자서열의 분석을 요망하는 임의의 이용자, 예컨대, 유전자서열 생산업체, 유전자서열 생산 연구소 등에 의해 관리되는 이용자 측 클라이언트와 인터넷 망과 같은 일련의 통신망을 매개로 선택적으로 신호 연결되는 유전자서열 분석관리 서버와, 이 유전자서열 분석관리 서버에 의해 제어되는 다수의 유전자서열 분석모듈들, 정보 파싱 엔진, 그리고 Gene Ontology 매핑정보 관리모듈 등의 조합으로 이루어진다. 인터페이스 모듈은 이용자 측 클라이언트 및 유전자서열 분석관리 서버를 매개한 상태에서, 이용자 측 클라이언트로부터 출력되는 일련의 이벤트 데이터를

유전자서열 분석관리 서버가 처리하기 용이한 형식으로 변환한 후, 변환 완료된 이벤트 데이터를 유전자서열 분석관리 서버로 전달하는 역할을 수행함과 아울러, 유전자서열 분석관리 서버로부터 출력되는 일련의 정보들을 이용자 측 클라이언트로 전송하는 역할을 수행한다. BLAST 검색 대상이 되는 유전자정보 데이터베이스는 NCBI에서 제공되는 Public 데이터베이스로 주기적인 업데이트를 실시하여 최신의 유전자정보를 가질 수 있도록 한다. BLAST 검색결과의 요약정보를 제공하기 위한 Gene Ontology 데이터베이스는 Gene Ontology Consortium 에서 제공하는 데이터베이스로 해당 데이터베이스에서 유전자정보관련 아이디를 추출하여 매칭 정보 데이터베이스를 생성하여 이용자가 요청한 해당 유전자서열에 대응되는 간략화된 형태의 필수 유전자사열 요약정보를 임시 생성하고, 이용자의 본격적인 요약정보 요청 시, 이 유전자서열 요약정보를 탄력적으로 제공함으로써, 이용자 측에서 자 기관이 원하는 유전자서열 분석결과를 별다른 어려움 없이 손쉽게 판독할 수 있도록 유도할 수 있다. 데이터베이스에 저장되어 있는 유전자서열 정보를 병렬 처리할 수 있는 일련의 모듈환경을 구축하고, 이를 통해, 데이타베이스에 저장되어 있던 일련의 유전자서열 검색결과가 최적의 속도로 이용자에게 온라인 제공될 수 있도록 유도함으로써, 해당 이용자의 분석결과 대기시간을 최소화하였다.

4. 성능검증

본 논문에서 제시한 유전자분석 병렬처리 모형의 성능평가를 위해서 [그림 1]과 같이 시스템을 구성하였다. Management Node는 인텔 Xeon 1GHz 4 CPU와 2GB RAM으로 구성되어 있고, Computing Node는 인텔 펜티엄III 2 CPU와 1G RAM으로 구성되어 있다. Node간 Interconnection을 위한 네트워크는 Gigabit Ethernet을 사용하였다[9]. 검색대상 데이터베이스는 생명공학분야에서 가장 널리 쓰이는 단백질 데이터베이스인 Swissprot을 이용하였다. Swissprot 데이터베이스는 약 13만건의 단백질서열로 이루어져있다[10].



[그림 3] 노드 증가에 따른 성능 향상율

입력 Sequence는 1000 Query를 기준으로 하였고 각 노드 증가에 따른 성능 향상율은 [그림 3]과 같다. BLAST는 노드증가에 따라 성능이 선형적으로 향상됨으로써 시스템의 확장성이 뛰어남을 알 수 있다.

5. 향후 과제 및 결론

본 논문에서는 이미 알려진 유전자서열 정보를 생명공학 연구자들이 최대한 효율적으로 활용 할 수 있도록 유전자서열 분석을 위한 병렬처리 모형을 개발하였다. 유사성 분석을 위한 BLAST의 처리 성능을 극대화하기 위하여 클러스터 컴퓨팅의 장점을 최대한 활용하여 데이타베이스에 저장되어 있는 유전자서열 정보를 병렬로 군집 처리 할 수 있는 일련의 모듈환경을 구축하였다. 이를 통해 데이타베이스에 저장되어 있던 일련의 유전자서열 정보가 최적의 속도로 이용자에게 온라인 제공될 수 있도록 유도함으로써, 해당 이용자의 분석결과 대기시간을 최소화 시킬 수 있는 유전자분석의 병렬처리모형을 제시하였다. 뿐만 아니라 생명공학 분야의 지식체계라 할 수 있는 Gene Ontology와 연동하여 이용자가 요청한 해당 유전자서열에 대응되는 간략화된 형태의 필수 유전자서열 요약정보를 임시 생성하고, 이용자의 본격적인 요약정보 요청시, 이 유전자서열 요약정보를 탄력적으로 제공함으로써, 이용자 측에서 자 기관이 원하는 유전자서열 분석결과를 별다른 어려움 없이 손쉽게 판독할 수 있도록 하였다. 향후 유전자분석 결과의 정합성을 향상시키기 위하여 멀티얼라이먼트 및 단백질 구조 분석 뿐 아니라 서열정보관련 문헌 등을 연결시켜주는 텍스트마이닝 기능 등을 연동한 통합 모델에 관한 지속적인 연구가 요구된다.

참고문헌

[1] Lee Rowen, Gregory Mahairas and Leroy Hood, "Sequencing the Human Genome", SCIENCE, 278, pp605-607.

[2] David W. Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press, 2001.

[3] Lincoln Stein, "Genome annotation: From sequence to biology", Nature Genetics Review , Vol2, pp493, July 2001.

[4] http://www.ncbi.nlm.nih.gov/BLAST

[5] http://www.isb-sib.ch/

[6] http://www.ebi.ac.uk/services/index.html

[7] International Technical Support Organization, "Linux HPC Cluster Installation", IBM, June 2001.

[8] http://genome-www.stanford.edu/GO

[9] http://www.linuxdoc.org/HOWTO/Beowulf-HOWTO-4.html - Beowulf cluster system design how-to document Web site

[10] http://kr.expasy.org/sprot