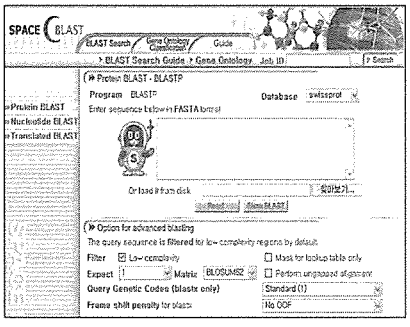
SPACE-BLAST

"리눅스 클러스터기반 대용량 유전자서열 검색솔루션"



포스데이타(대표 김광호)는 리눅스 클러스터 기반 바이오 인포매틱스 솔루션인 'SPACE-BLAST(Super Parallel Computer Engine for BLAST)'를 독자 개발하여 생명공학관련 연구 개발기간을 획기적으로 단축시킬 수 있게 되었다.

포스데이타가 개발한 'SPACE-BLAST'는 대용량 바이오 정보를 초고속으로 분석하여 유전자 서열에 대한 정보를 제공하는 솔루션이다. 특히 저비용의 리눅스 클러스터링 컴퓨터를 이용해 대량의 유전자 서열 분석은 물론 생명공학분야의 지식체계인 Gene Ontology를 이용하여 검색 결과를 요약하는 특징이 있어 연구개발 기간을 단축시킬 수 있을 것으로 기대되고 있다. 기존의 'BLAST(Basic Local Alignment Search Tool)'는 미국 생명공학연구소(NCBI)에서 제공하던 툴로서, 생물학자들이 유전자 서열의 기능을 파악하기 위해 사용하는 중요한 프로그램이지만, 고비용의 유닉스 기반에서 사용되어 쉽게 도입하기 어렵다는 단점을 가지고 있었다. 반면에 'SPACE-BLAST'는 단백질 서열 데이터베이스를 8노드 기준으로 검색할 경우, 시간당 1만 2천Sequence(유전자 서열)의 처리 능력을 갖고 있어, 유닉스 서버 대비 30%이상 향상된 성능을 갖고 있을 뿐 아니라 시스템 구축 비용 또한 3분의1 수준 이하로 저렴하다. 'SPACE-BLAST'는 국제 특허 출원 중에 있으며, 향후 신약 개발, 농업, 화학, 의료, 환경 등 생명공학 연구에 널리 적용될 전망이다. 포스데이타는 대학연구소를 비롯한 생명공학 관련 연구기관을 대상으로 'SPACE-BLAST'를 공급할 계획이며, 웹을 통한 무료 시범 서비스를 제공하고 있다(space-blast.posdata.co.kr). 바이오인포매틱스분야의 클러스터 컴퓨팅 솔루션이 개발됨에 따라 앞으로 저렴한 비용으로 장비가 보급돼 미래산업인 BT(Bio Technology)분야의 생물정보 분석 연구가 활발히 진행될 수 있을 뿐만 아니라 생명공학 관련 연구기관들이 시스템 구입에 대한 부담이 줄어들어 생명공학 발전에 크게 기여할 것으로 기대한다.

SPACE-BLAST는 향후 멀티 얼라이먼트 및 단백질 구조 분석 뿐만 아니라 서열정보관련 문헌 등을 연결시켜주는 텍스트 마이닝 기능 등을 추가하여 바이오인포매틱스분야의 토털 솔루션으로 확대해 나갈 예정이다.

문의 : 031-779-2502 담당 : 유 승식 부장

SPACE-BLAST (정보통신분장관상)

1. 작품명 : SPACE-BLAST

(Super PArallel Computer Engine for BLAST)

리눅스 클러스터기반 대용량 유전자 서열 검색 솔루션

2. 제작자 : ㈜ 포스데이타

대 표 자 : 김 광 호

개발참여자 : 유승식 박미화, 전광석, 한동훈, 김재우 외 5명

주 소 : (415-234) 경기도 성남시 분당구 서현동 276-2
포스데이타 빌딩

전 화 : (031) 779-2114

팩 스 : (031) 779-2709

E - m a I l : cluster@posdata.co.kr

3. S/W 요약 설명

SPACE-BLAST는 클러스터 수퍼 컴퓨터 구축 기술을 기반으로 대용량 유전자 서열 정보를 초고속으로 분석 가능한 병렬처리 시스템을 구축하여 대량의 서열 검색 결과를 생명공학분야 지식체계인 Gene Ontology (GO)를 이용해 서열의 기능에 대한 Global View를 제공하는 리눅스 클러스터 기반 바이오인포매틱스 솔루션이다.

본 S/W는 미국 생명공학연구소(NCBI)에서 제공하는 유전자 서열 검색 툴인 BLAST(Basic Logical Alignment Tool)의 병렬처리를 통한 성능향상 뿐 만 아니라 방대한 양의 서열 검색 결과를 Gene Ontology를 이용하여 요약해 주는 것으로 신약개발 및 유전자 발굴 등의 연구기간을 획기적으로 단축 시켜 신약 개발, 농업, 화학, 의료, 환경 등 생명공학 연구에 핵심적인 역할을 할 것으로 기대한다.

SPACE-BLAST는 향후 멀티 얼라이먼트 및 단백질 구조 분석 뿐만 아니라 서열정보관련 문헌 등을 연결시켜주는 텍스트 마이닝 기능 등을 추가하여 바이오인포매틱스분야의 토털 솔루션으로 확대해 나갈 예정이다.

4. 개발 배경

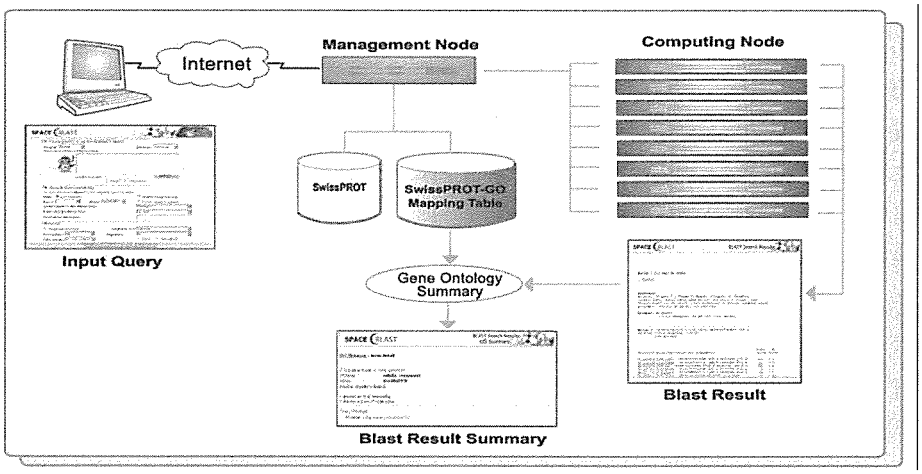
Human Genome Project과 같은 대형 Sequencing 프로젝트와 High-throughput Sequencing 기술의 발전에 의해 현재 Expressed Sequence Tag (EST)와 같은 대량의 DNA 서열들이 생산 되고 있고, 이를 효과적, 효율적으로 분석해야 할 필요성이 증대되고 있습니다. 대부분의 실험자들이 서열 분석을 위해 우선적으로 BLAST 검색을 이용하고 있는데, 대량의 서열, 검색 DB의 크기, BLAST 검색 결과의 복잡성에 의해 어려움을 겪고 있고, 이에 빠르고 정리된 결과를 보여줄 수 있는 BLAST 검색 시스템의 필요성이 증대하고 있다.

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 바이오인포매틱스 솔루션을 개발하게 되었다.

5. 시스템 개요

SPACE-BLAST의 시스템은 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 분할 한 후 각 Computing Node로 분배하여 대상 데이터베이

스에 대한 서열검색 작업을 실행하도록 한다. Computing Node들은 로컬 데이터베이스를 대상으로 서열검색 작업 및 검색결과 요약을 위한 인덱스 추출 작업 에 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 파악 할 수 있도록 한다.



[그림 1] SPACE-BLAST 시스템 구성도

6. 시스템 특징

① 저비용 고성능의 클러스터 컴퓨팅 지원
- 리눅스 클러스터 컴퓨팅 기반 가격대비 고성능 처리 능력
  : 단백질 서열데이타베이스인 Swissprot을 대상으로 검색할 경우 8 Node 기준 시간당 12,000 Sequence 처리 가능
- IT 투자비용 최소화 : 기존 바이오 클러스터 대비 1/3이하의 IT 투자비용

② 확장성
- 사용자의 처리성능 요구에 따라 네트워크 연결만으로 Computing노드 확장 가능

③ 고가용성
- 특정 노드의 장애발생시 해당 노드를 제외한 지속적인 시스템 서비

스 제공 가

④ 사용자편리성

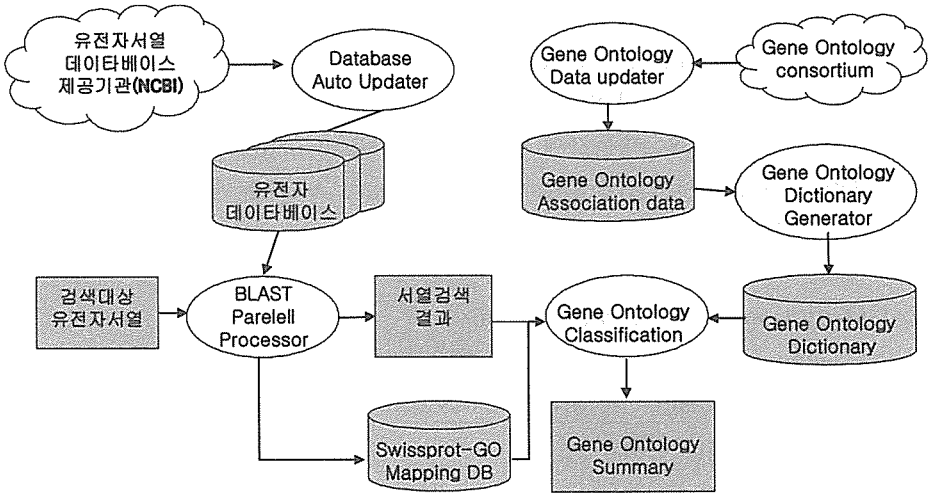
- 웹기반(Web based)의 인터페이스(Interface)로 시용자 편의성 제공

⑤ 시스템관리 용이성

- 통합시스템 제어 및 모니터링, 소프트웨어 설치 및 분배 자동화

⑥ BLAST관련 Data Base 자동 Update 기능 제공

7. 주요 기능



[그림 2] SPACE-BLAST 프로그램 구성도

① BLAST Parallel Process Module

- Sequence Split Module

  사용자가 입력한 서열을 정해진 node의 개수에 따라 나누게 되고, 나누어진 서열들을 computing node의 해당 디렉토리에 File로 저장함.

- Parallel BLAST Distribute Module

  Management node는 각각의 computing Node들의 상태를 점검한 후 적절하게 Job을 분배하여 실행하도록 한다.

- Job Status Monitering Module
  Management Node는 Computing Node의 현재 Job실행 상황을 모니터링하여 사용자에게 작업진행 상태 및 예상 종료시간을 알려주고 Computing Node에서의 작업이 완료된 것을 감지한 순간 각 Computing Node의 검색결과 및 인덱싱 파일을 management Node로 가져옴.
- GO index extract Module
  서열검색결과 파일을 파싱하여 Gene Ontology 용 인덱스를 추출하는 모듈로서 Make Shelve Module에서 만들어진 Gene Ontology Dictionary를 이용한다.
- Result Merge Module
  각각의 Computing Node이 가지고 있는 서열검색결과 및 Gene Ontology 인덱싱 파일을 Management Node로 가지고 와서 서열검색결과는 Merge후 사용자에게 보여주고 Gene Ontology 인덱싱 파일은 DB에 저장하여 Gene Ontology Summary Module에서 사용할 수 있도록 한다.

② GO dictionary Generate Module

- Gene Ontology Consortium에서 제공하는 SWISSPROT-GO Mapping Table을 파싱하여 Swissprot의 ID와 Gene Ontology의 ID를 DB화 하는 모듈.

③ Gene Ontology Summary Module

- Web Browser 상에서 GO 계층 트리를 이용하여 사용자가 입력한 서열이 어느 Gene Ontology Category에 해당하는 지를 보여주는 모듈로서 GO Browser에서 해당 Category를 눌었을 때, 거기에 해당하는 내용을 GO-Seqence DB를 이용하여 해당되는 Sequence를 뿌려준다.

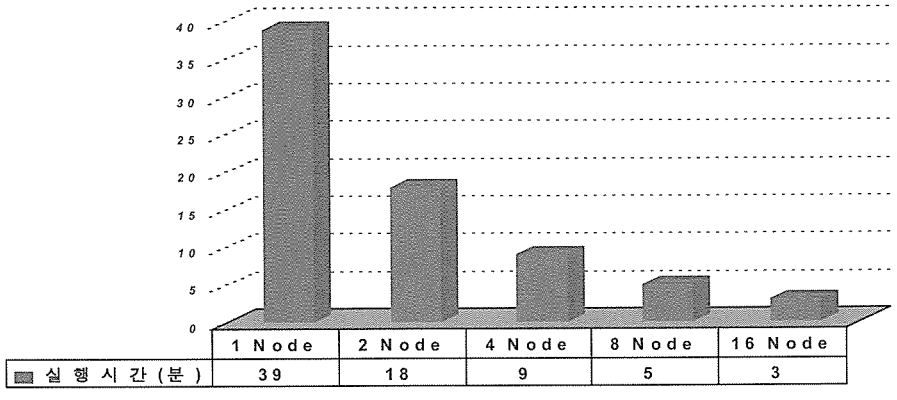
④ Database Auto Update Module & Gene Ontology data Update Module

Management Node는 유전자 서열 데이터 베이스 및 Gene Ontology Data를 제공하는 해당사이트로부터 추가된 부분 또는 필요한 경우 전체 데이터베이스를 주기적으로 다운로드 한 후 Computing Node의 데이터베이스를 일괄적으로 업데이트 시킨다.

8. 성능 평가

SPACE-BLAST의 성능평가를 위해서 [그림 3]과 같이 시스템을 구성하였다. Management Node는 인텔 Xeon 1GHz 4 CPU와 2GB RAM으로 구성되어 있고 Computing Node는 인텔 팬티엄III 2 CPU와 1G RAM으로 구성되어있다. Node간 Interconnection을 위한 네트워크는 Gigabit Ethernet을 사용하였다.
검색대상 데이터베이스는 생명공학분야에서 가장 널리 쓰이는 단백질 데이터베이스인 Swissprot을 이용하였다. Swissprot 데이터베이스는 약 13만건의 단백질서열로 이루어져있다. 입력 Sequence는 1000 Query를 기준으로 하였고 각 노드 증가에 따른 성능 향상율은 [그림 4]과 같다. SAPCE-BLAST는 노드증가에 따라 성능이 선형적으로 향상됨으로써 시스템의 확장성이 뛰어남을 알 수 있다.

[그림 3] 노드증가에 따른 검색 성능



9. 적용분야 및 향후 발전방향

신약 개발, 농업, 화학, 의료, 환경 등의 생명공학 연구를 수행함에 있어 가장 기본적이고 필수적인 툴로서 클러스터 컴퓨팅을 통한 고성능 분석 지원 및 대량 서열 검색 결과 요약 서비스를 통하여 연구 생산성 향상 및 편리

성을 제공 할 뿐만 아니라 저렴한 비용으로 장비가 보급돼 생명공학 관련 연구기관들이 시스템 구입에 대한 부담이 줄어들어 생명공학 발전에 크게 기여할 수 있을 것으로 기대한다.
향후 멀티얼라이먼트 및 단백질 구조 분석 뿐만 아니라 서열정보관련 문헌 등을 연결시켜주는 텍스트마이닝 기능 등을 추가하여 바이오인포매틱스분야의 토털 솔루션으로 확대해 나갈 예정이다.

10. 개발기간 및 투입인원수

개발단계	개발기간	투입인원	공수	비 고
분석/설계	'02. 4. 1 ~ 4. 30	3	3	- 바이오인포매틱스분야의 클러스터 컴퓨팅적용 타당성 조사 - 클러스터 시스템 인프라 최적화 설계 및 장비 도입 - 시스템 기본설계 및 상세설계
시스템 개발	'02. 5. 1 ~ 7. 31	10	30	- 클러스터 시스템 구축 및 프로그램 개발 - 단위모듈 테스트
통합 테스트	'02. 8. 1 ~ 8. 31	10	10	- 대용량 서열에 대한 시스템 부하 테스트 및 성능 검증
상품화	'02. 9. 1 ~ 9. 30	5	5	- 웹 인터페이스 디자인 및 매뉴얼 제작
계			48(MM)	

11. 개발언어, TOOL

구 분		갯 수	사 양
Server	Management Node	1	Intel Xeon 4CPU 2GB of Main memory One Host Bus Adapter One 10/100 Ethernet NIC One Gigabit NIC
	Computing Node	16	Intel Pantium III 1GB of Main memory One 10/100 Ethernet NIC One Gigabit NIC
S/W	OS		Redhat7.1 kernel 2.4.9
	clusterware		Job scheduler
	BLAST		NCBI BLAST2.0
	C,java,Python		개발 Tool
Network	Gigabit Switch	1식	
	Ethernet Switch	1식	
Storage	San Storage	1	FAStT200

12. 사용시스템 (개발 및 테스트)

구분	프로그램명	비고
Web Program	Javascript	
System program	Python,C++	