병렬 컴퓨팅 기반의 유전체 분석 워크벤치

선충현[1], 이관수[1], 박학수[2]
[1] 한국정보통신대학교, [2] 한국과학기술정보연구원
{chsun, gsyi}@icu.ac.kr, {hspark}@kisit.re.kr

Genomic Analysis Workbench Based on Parallelized Computing

Choong Hyun Sun, Gwan Su Yi[1], Hark-Soo Park[2]

요 약

최근 바이오 데이터 분석에는 데이터 양의 급격한 증가와 이에 따른 문제의 복잡성도 함께 증가하고 있다. 이 결과 다양한 분석 툴들의 유연한 조합과 고성능, 고처리 컴퓨팅이 가능한 분석 시스템이 절실히 요구되고 있다. 본 논문에서는 병렬 컴퓨팅 환경을 이용하고 워크플로우 기반에서 다양한 생물정보 분석 툴들을 자유롭게 조합하여 작업을 수행할 수 있는 바이오워크벤치를 소개한다. 바이오워크벤치 내에는 컴퓨팅 자원 및 작업 정보에 대한 모니터링 툴, 각 툴 들과 데이터를 손쉽게 가공할 수 있도록 고안된 인터페이싱 툴, 워크플로우 디자인 툴을 포함 하고 있다. 이 기능모듈을 활용함으로써 다양한 생물정보 분석 툴을 이용하는 과정에서 효율적인 분석을 수행을 지원하는 바이오 워크벤치의 기능 및 아키텍쳐을 제시한다.

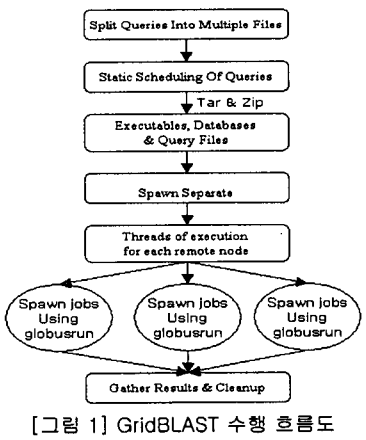
1. 서 론

고도화된 생물공학 기술과 유전체 서열규명연구의 진척에 힘입어 분석될 서열데이터의 양은 급격하게 증가하고 있다. 또한 생물학 데이터 양의 증가와 함께 풀어내고자 하는 문제의 복잡성 또한 점점 커져가는 상황에서 이 방대한 서열정보를 분석하는 연구는 고용량, 고처리 컴퓨팅자원이 절실하게 되었다. 이러한 요구에 힘입어 현재 크게 두 가지 측면에서 컴퓨팅 자원을 이용한 연구가 진행되고 있다. 그리드 기술[1]이라고 하는 범지역적, 가상 공동체적인 연합을 통해서 분산된 컴퓨팅 자원을 확보하는 방안이 그 하나이고 여러 노드로 구성된 클러스터를 효율적으로 통합 및 관리, 작동시켜 우수한 컴퓨팅 능력을 활용하는 방법이 있다. 최근 몇몇 그룹에서는 그리드 기술 및 클러스터를 이용한 생물정보학 연구 분석의 예를 보고한 바가 있다[2,3]. 그러나 대부분의 생물정보학 애플리케이션들이 단순 서열정열 비교에 그치고 있어서 종합적인 데이터 분석을 기대하기는 어려운 상황이다. 그러나 실질적으로 생물정보 분석에서는 분석 툴간의 유연한 조합을 이용한 통합적인 분석이 진행되어야만 양질의 결과를 얻어낼 수 있다. 이를 위해서 Condor를 이용한 고처리 컴퓨팅 능력을 확보하고 분석 툴간의 워크플로우 디자인을 가능케 하여 생물정보 분석의 두 가지 큰 문제에 대한 해법을 제시하며 또한 사용자의 애플리케이션에 대한 친화도와 참여도를 높이기 위한 기능을 제공한다.

2. 관련 연구

2.1. Condor

Condor[4]는 대표적인 그리드 미들웨어 중이 하나이지만 완벽한 그리드 환경을 구축하기 위해서 원초적인 기반이 되는 기능만을 제공하는 globus나 legion와는 다른 특성을 보인다. 즉 Condor는 기초 기반 기능보다는 고처리 컴퓨팅을 요하는 애플리케이션의 안정적 동작 및 관리를 위한 많은 기능을 지원하고 있다. 그 대표적인 기능은 수행중인 분석작업의 안정적 완료를 위해서 checkpoint와 migration를 지원하고 있다. 그리고 remote system call를 사용하여 분석 작업에 앞서 데이터 이동을 하지 않고서도 로컬 컴퓨터에서 작업을 수행하는 기능을 제공한다. 작업을 수행하기 앞서 사용자는 자기가 원하는 자원의 사양을 미리 선택하여 실행을 할 수가 있는 class advertisement mechanism을 사용한다. DAGMAN이라고 하는 수행 작업들간의 우위성 및 의존관계를 바탕으로 워크플로우 디자인을 지원하는 근간을 제공한다. 이런 다양한 기능은 Condor를 로컬 클러스터 시스템을 효율적으로 관리 및 운용할 수 있게 해준다. 아울러 로컬 클러스터가 공용 아이피로 네트워크 설정이 된다면 본래 기능인 그리드 컴퓨팅 자원으로서 언제든지 기능할 수가 있는 장점이 있다.



[그림 1] GridBLAST 수행 흐름도

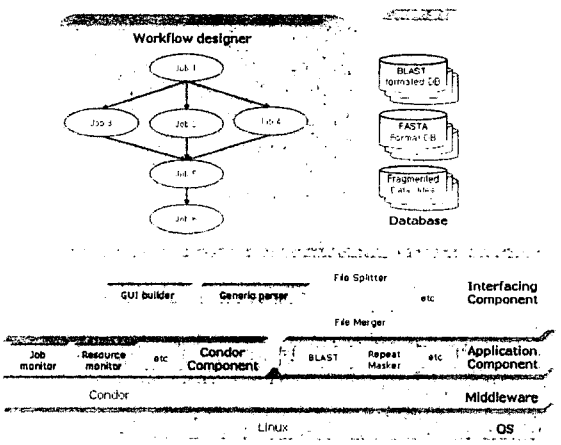
2.2. GridBLAST

BLAST(Basic Local Alignment Search Tool)[2]은 local alignment를 생성하여 유전자와 단백질 서열 유사성을 계산하는 툴로서 생물정보 분석에 있어서 가장 널리 사용되고 있는 기초 분석 툴이다. 그림 1은 query file이 분석되어 가는 과정을 보여주고 있다. BLAST는 이 구동환경에서 그 자체로만 구동하고 있을 뿐이고 다른 생물정보 분석 툴과의 연계성이 요구된다.

2.3. Tarverna

Tarverna는 웹 서비스 기술을 기반으로 현재 제공되고 있는 여러 생물학 분석 시스템을 서비스로 활용하여 워크플로우를 설계할 수 있는 생물정보학 툴이다[3]. 이것은 생물정보 분석을 위한 각지에 흩어져 있는 원격 컴퓨팅 툴과 정보 저장소를 제공하고 있다. 대표적인 특징으로 수행 중인 작업의 오류 극복 기능, nested workflows 지원, 프로세스 보고, 원격지 데이터 검색이 가능하다. 워크플로우는 scufl(simple conceptual unified flow language)로 구현되어 있다. 워크플로우 작성시 입력된 정보는 XML 파일에 의해서 모든 정보가 저장되어 관리된다.

3. 시스템 아키텍쳐

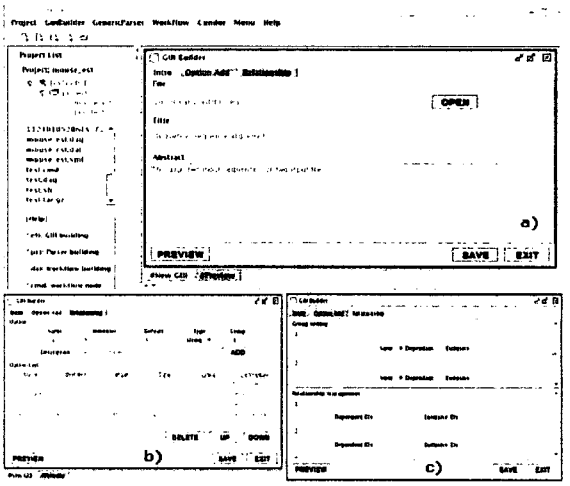


[그림 2] 바이오 워크벤치의 시스템 아키텍쳐

그림 2는 바이오 워크벤치의 시스템 아키텍쳐로서 유연한 생물정보 분석 툴간의 결합과 병렬 컴퓨팅 자원을 제공하여 효율적인 생물정보 분석을 가능하도록 필요한 기능을 정의하였으며 4가지 주요한 기능 모듈을 보여주고 있다(Workflow designer, GUI builder, Generic parser 와 Condor tool). 바이오 워크벤치의 OS는 Linux이며 미들웨어로 Condor 6.2를 설치하였다. Java, PERL, GCC 실행환경을 지원하고 있다. 현재 널리 사용되고 있는 생물정보 분석 툴들로 RepeatMasker, FASTA, BLAST, Smith-Waterman 알고리즘, TGICL 이 사용 가능하도록 내장되어 있고 BLAST 포맷팅된 서열데이타 파일과 FASTA format의 다양한 유전체 데이터 파일이 존재한다. 이러한 생물학 데이터와 생물정보 분석 툴들, 4가지 주요기능을 기반으로 사용자는 원하는 분석 작업수행의 편리하게 할 수 있다. 그리고 수행된 분석작업은 Condor tool에서 제공하는 모니터링 기능을 바탕으로 안정적인 작업완료를 제공하고 있다.

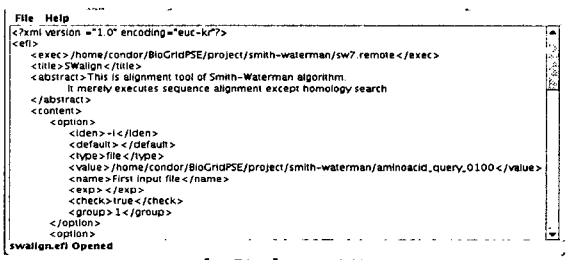
4. 구현

4.1 GUI Builder



[그림 3] GUI Builder

GUI Builder는 그림에서와 같이 실행파일 설정 부분(a), 옵션 설정부분(b), 옵션간의 의존관계 설정부분(c)으로 구성되어 있으며 사용자 인터페이스가 없는 실행파일에 대해서 사용자가 직접 인터페이스를 설계하도록 해준다. 입력된 정보는 efi라는 XML파일에 저장이 되며 수정이 가능하게 되어 있어 동적이 GUI를 지원할 수가 있다. 아래 그림은 efi 파일에 대한 내용을 보여주고 있다.

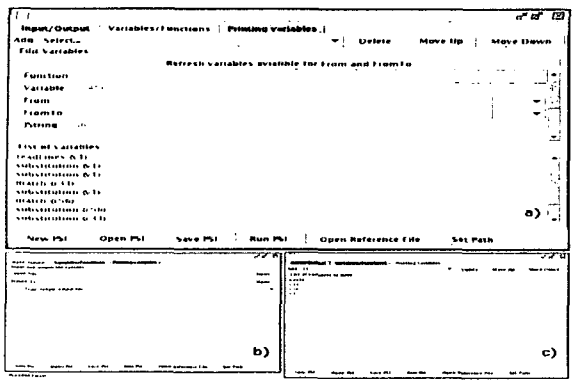


[그림 4] EFI 파일

Efi 파일은 GUI Viewer를 통해서 사용자가 지정한 인터페이스로 변환이 된다. 또한 병렬처리를 위해 필요한 스크립트 파일(Condor CMD file)을 자동 생성하도록 도와준다.

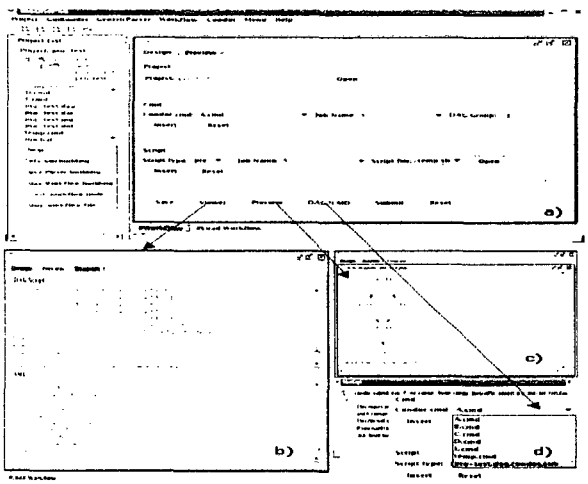
4.2. Generic Parser

Generic parser[그림 5]는 PERL의 정규 표현식과 XML를 이용하여 특정 실행파일을 통해서 얻어진 결과에서 사용자가 원하는 정보만을 파싱하도록 하는 기능을 지원하고 있다. 생물정보 분석 툴의 경우 결과 포맷이 다양하여 툴간의 정보전달이 원활치 못할 때가 많이 발생하게 된다. 이 경우 데이터 파싱이 필요한 결과 파일을 generic parser를 이용해서 처리할 수가 있다. 이 모듈 또한 XML파일에 사용자가 입력한 입력, 출력 파일 정보, 파싱 타입, 파싱 대상 데이터, 출력 타입 등이 모두 기록되어 수정 관리를 할 수 있다.



[그림 5] Generic parser

4.3. Workflow Designer

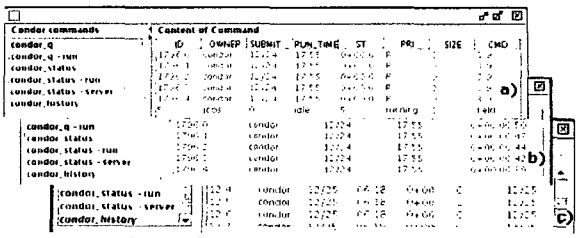


[그림 6] Workflow designer

Workflow designer는 4.1과 4.2에서 사용자가 생성한 CMD 파일들과 데이터 파싱 스크립트, 그리고 전처리/후처리를 담당하는 스크립트 등을 조합하여 수행작업 간의 워크플로우를 디자인 할 수 있도록 한다. 워크플로우 엔진으로는 Condor에서 지원하는 DAGMAN(Directed Acyclic Graph)을 활용하여 구현하였다[4]. 워크플로우 디자인에 사용될 모든 요소가 준비된 a) 부분을 통해서 워크플로우를 정의하는 DAG script와 XML파일이 생성되는 부분 b), GraphViz[5]를 활용하여 워크플로우 다이아그램을 보여주는 c) 부분으로 구성이 되어 있다. 또한 생성된 워크플로우(DAG script)는 다른 워크플로우 디자인시 inner workflow로 재활용될 수 있다.

4.4. Job & Resource Monitoring

Condor는 Central manager를 중심으로 컴퓨팅 노드들이 Condor pool이라는 Computing cluster를 구성하게 된다. 그리고 pool에 있는 모든 자원과 제출된 수행작업의 모니터링 기능을 제공하고 있으며 이를 위한 모니터링 viewer를 구현하였다. 구성은 5가지 종류의 모니터링 기능과 1개의 수행작업 로그기록 기능으로 이루어져 있다.



[그림 7] Job & Resource 모니터링

- Condor_q: 제출된 모든 수행작업에 대한 리스트
- Condor_q -run: 실행 중인 작업에 대한 리스트
- Condor_status: Condor pool내 모든 노드에 대한 상태 정보
- Condor_status -run: 작업 실행 중인 노드에 대한 상태 정보
- Condor_status - server: 모든 노드에 대한 물리적 성능 정보
- Condor_history: 제출된 모든 수행작업들에 대한 기록정보

5. 결론

본 논문에서는 병렬 컴퓨팅 자원과 유연한 생물정보 분석 툴간의 조합을 가능케 하여 효과적인 분석수행을 지원하는 바이오 워크벤치에 대한 모델을 제시하였고 이를 구현하였다. 그리고 사용자로 하여금 GUI builder, Generic parser를 이용해서 사용자 친화적인 인터페이스를 직접 설계하고 원하는 결과를 추출하도록 하였으며 워크플로우는 재사용이 가능하다. 생물정보 분석에 있어서 한 애플리케이션 내에서 복합적인 정보 분석처리를 한다는 것은 매우 유용하고 효율적인 것임을 알 수 있다. 향후 연구에서는 본 논문에서 구현된 바이오 워크벤치에 다양한 생물정보학 관련 툴들을 적용하여 기능을 보완해야 하며 생물학자들의 사용자 피드백을 통해서 개선점을 반영해야 할 것이다.

6. 참고문헌

[1] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. "Condor - a distributed job scheduler." In Thomas Sterling, editor, Beowulf Cluster Computing with Linux. MIT Press, Oct 2001

[2] A. Krishnan. Gridblast: "High throughput blast on the grid". In 2nd International Conference on Natural Products, Singapore, Jul 2002.

[3] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, "Taverna: a tool for the composition and enactment of bioinformatics workflows", Bioinformatics, Vol. 20 no.17, pp. 3045-3054, 2004.

[4] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. "Condor - a distributed job scheduler." In Thomas Sterling, editor, Beowulf Cluster Computing with Linux. MIT Press, Oct 2001.

[5] Gansner,E.R. and North,S.C. "An open graph visualization system and its applications to software engineering." Softw. Pract. Exper., 00(S1), 1-5, 1999.