웹 기반 고성능 다중서열정렬시스템 설계 및 구현

김태경°, 김훈기**, 최치환**, 정승현***, 허보경*, 조완섭****
°*한국생명공학연구원 국가생명연구자원정보센터
**충북대학교 바이오정보기술학과
***충북대학교 정보산업공학과
****충북대학교 경영정보학과
e-mail: {tkkim, bkher71}@kribb.re.kr, wscho@cbnu.ac.kr

A Web-Based High Performance Multiple Sequence Alignment System Design and Implementation

Tae-Kyung Kim°, Hun-Gi Kim**,Chi-hwan Choi**, Seung-Hyun Jung***, Bo Kyeng Hou*, and Wan-Sup Cho****
°*Korean Bioinformation Center, Korea Research Institute of Bioscience & Biotechnology
**Dept. of Bio-Information Technology, Chungbuk National University, Korea
***Dept. of Information Industrial Engineering, Chungbuk National University, Korea
****Dept. of Management Information Systems, u-BIZ BK21, Chungbuk National University, Korea

● 요 약 ●

다중서열정렬 알고리즘은 생명정보학 분야에서 서열기반의 계통분류 분석에 가장 많이 사용되며, 가장 대표적인 공개 프로그램은 ClustalW로 사용자가 로컬시스템에 설치하여 이용할 수 있다. 그러나 실제로 사용자들이 ClustalW을 설치한 후, 서열데이터의 준비, 가공, 처리 및 타 시스템과 연동 등과 같은 작업을 하는데 여러 가지 어려움이 있다. 따라서 본 논문에서는 다중서열정렬 작업을 편리하고 빠르게 수행할 수 있는 웹기반의 고성능 다중서열정렬시스템을 제안한다. 제안된 시스템의 특징은, (1) Inter-Query 라우팅 알고리즘을 통해 다수의 PC 자원을 효율적으로 활용하여 계산 성능을 극대화하였으며, (2) 사용자 편의성을 고려한 웹인터페이스의 제공을 통해 개인화된 데이터관리, 실시간 모니터링, 데이터 편집 등을 지원하여 사용자가 손쉽게 서열데이터의 수집, 관리 및 처리할 수 있도록 지원한다.

키워드: 생명정보(Bioinformatics), 클라우드 컴퓨팅(Cloud Computing), 다중서열정렬(Multiple Sequence Alignment)

I. 서론

차세대 시퀀싱 기술을 비롯한 다양한 실험도구가 출현하면서 유전체 관련 서열데이터가 급격히 증가하고 있다. [그림1]은 가장 대표적인 유전체 서열데이터베이스인 GenBank의 데이터 증가량을 보여준다. 서열데이터의 증가량은 현재 1년에 1.5배정도로 증가하고 있으며 그 속도는 점점 빨라지고 있다[4].

생명정보학은 이러한 대용량의 서열데이터를 여러 가지 IT 기술을 적용·분석하여 생명공학자들에게 생물학적 의미를 제공할 뿐만 아니라 실험 범위를 좁혀서 연구시간 및 비용의 절감 효과를 주고 있다.

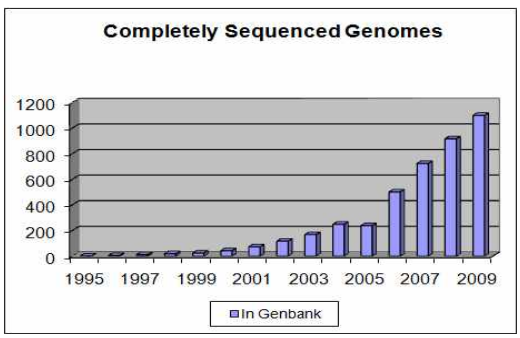

그림 1. GenBank 데이터베이스 증가추세
Fig. 1. GenBank DB Growth

현재 생명정보학 분야에서 가장 널리 사용되는 알고리즘들은 대부분 서열을 대상으로 개발되고 있다. 특히, 특정서열을 기존의 서열데이터베이스와 비교하여 그 기능(function)을 예측하기 위한 알고리즘과 서열간의 거리를 비교하여 계통분류학적 거리를 계산하는 알고리즘이 가장 많이 사용되고 있다. 서열 검색 프로그램은 NCBI BLAST[1], 서열간의 분류학적 거리 계산 프로그램은 ClustalW[2]가 가장 많이 사용되고 있다.

현재 생명공학자들이 다중서열정렬 작업을 수행하기 위한 절차는, (1)분석하고자 하는 종들을 정하고, 각 종들 간에 비교할 공통 유전자를 선정한 후, (2) 공통유전자 서열을 수집하고, (3)종간의 공통 유전자 서열에 대하여 그루핑을 수행한다. 마지막으로, (4)그루핑된 서열 그룹에 대하여 다중서열정렬을 수행한다. 이러한 일련의 작업에 있어서 서열데이터 수집 및 관리의 어려움이 있을 뿐만 아니라, 처리해야할 작업이 많아 처리 시간이 오래 걸린다.

본 논문에서는 생명공학자들이 ClustalW 알고리즘을 웹 기반에서 편리하게 활용할 수 있는 시스템을 제시한다. 본 시스템의 장점은 다음과 같다. 첫째, 하부 자원으로 PC를 연결한 클러스터를 활용하여 성능을 극대화한다[3]. Inter-Query 라우팅 알고리즘을 고안하여 효율을 개선하였다. 둘째, 웹 기반에서 개인화 서비스를 지원한다. 셋째, 웹 환경에서 생물학자들이 실험을 설계하고 데이터를 수집할 수 있다. 넷째, 처리한 결과들을 다음 분석 도구로 연계할 수 있다.

논문의 구성은 다음과 같다. 2장에서는 다중서열정렬과 관련 도구와 서비스들에 대해 살펴보고, 3장에서는 제안된 시스템의 구조와 작업 분배 라우팅 알고리즘을 제시한다. 4장에서는 사용자 인터페이스와 알고리즘에 대한 성능평가 결과를 제시하고, 5장에서 결론을 다룬다.

II. 관련 연구

생명정보학 분야에서 다중서열 정렬알고리즘은 서열을 기반으로 계통분류학적 거리를 예측하는데 핵심적으로 사용되고 있다. 이러한 필요에 따라 여러 가지 프로그램들이 출현하였다. 가장 대표적인 프로그램은 ClustalW 이며, 그 이외에 MAFFT[5], MUSCLE[6], DCA[7] 등이 있다.

각 프로그램은 서열분석을 위한 유사한 기능을 제공하고 있으나 약간의 성능 차이가 존재한다. 그러나 실제적으로 이러한 프로그램들은 최종 사용자인 생명공학자들이 적극적으로 활용하는데 여러 가지 한계가 있다. 첫째, 대부분의 경우 소스코드 또는 바이너리를 제공하여 사용자가 전산시스템에 설치하여 활용할 수 있도록 하고 있다. 생명공학자들은 전산시스템에 세팅하고 활용하는 것이 쉽지 않으며, 분석 프로그램을 활용하는데 상당한 시간이 걸리는 실정이다. 둘째, 웹 기반으로 개인화된 서비스를 지원하지 않고 있다. EBI에서 제공하는 MUSCLE의 경우, 웹을 통해 다중서열정렬을 수행할 수 있는 환경을 제공하고 있으나[6], 하나의 서열 그룹에 대해서만 처리가 가능하고 개인화된 서비스도 제공하지 않는다. 셋째, 단일 시스템에서 활용하기 위한 목적으로 개발되어, 고성능 처리 서비스 및 분산처리 환경을 지원하지 않는다.

III. 시스템 아키텍처 및 라우팅 알고리즘

3.1 시스템 아키텍처

제안된 시스템은 크게 다음 [그림 2]와 같이 사용자 인터페이스, 웹 서버, 클러스터 미들웨어, 시스템 데이터베이스로 4 계층구조로 구성되어져 있다.

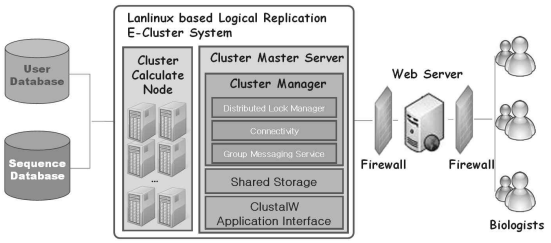

그림 2. 시스템 아키텍처
Fig. 2. System Architecture

사용자 인터페이스 계층은 웹으로 구현되었으며, 개인화된 서비스를 제공한다. 특히 AJAX로 구현되어 하나의 페이지에서 다중서열정렬 작업에 필요한 모든 기능을 활용할 수 있다. 또한, 활용하는 컴퓨팅 자원과 처리 중인 작업들에 대한 실시간 모니터링 기능을 제공한다.

미들웨어에서는 클러스터 자원들을 관리하기 위한 노드 관리자, 작업들에 대한 정보와 노드 정보를 관리하고 있는 메타데이터, 그리고 작업을 노드에 분산하기 위한 라우터로 구성되어 있다. 다중서열정렬 작업의 성능을 높이기 위해서 다수 노드에 작업을 분배하는 것이 미들웨어에서의 핵심 기능이다. 3.2 절에서 작업 라우팅을 위한 핵심 알고리즘을 제시한다.

데이터베이스 영역에는 사용자가 처리하기 원하는 서열정보를 비롯한 사용자 개인정보를 관리한다.

3.2 작업 라우팅 알고리즘

다중서열정렬 작업은 유전자 그룹에 대한 처리를 수행하는 Inter-Query 타입의 알고리즘이 적용된다. 즉 독립적인 작업들이 다수 존재하는 것으로 하나의 작업을 하나의 컴퓨팅 노드에 분배하는 방식으로 문제를 해결할 수 있다. 하지만 노드에 작업을 분배하는 방식에 따라 성능과 사용자 편의성이 크게 달라질 수 있다. [그림 3]은 다중서열 정렬을 위해 제안된 Inter-Query 라우팅 알고리즘이다.

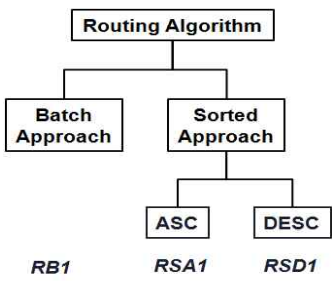

그림 3. Inter-Query 라우팅 알고리즘
Fig. 3. Inter-Query Routing Algorithm

RB는 First-in-First-out 방식이며, RSA방식은 작은 문제부터 처리하는 방식이며, RSD는 큰 문제부터 처리하는 방식이다. RB는 구현이 간단하며, RSA는 중간결과의 양을 극대화할 수 있으며, RSD는 최종결과의 처리시간을 줄일 수 있는 장점이 있다. 제안된 시스템에서는 사용자가 선택적으로 작업 분배 알고리즘을 선택할 수 있다.

IV. 다중서열정렬 서비스 구현 및 평가

4.1 웹 기반 서비스 구현

웹 인터페이스는 크게 모니터링, 프로젝트 생성, 최종결과 확인의 3가지의 영역이 있다. [그림 4]는 메인 웹 인터페이스로서 로그인 영역, 프로젝트 리스트 영역, 작업노드 모니터링 영역을 보여준다. [그림 5]는 다중서열정렬 프로젝트 생성화면으로 파일 업로드 방식과 설계방식을 지원한다. [그림 6]은 프로젝트내의 작업 처리 결과를 보여준다. 각 유전자 그룹마다 하나의 결과가 있어 사용자가 선택적으로 다운로드 받을 수 있다.

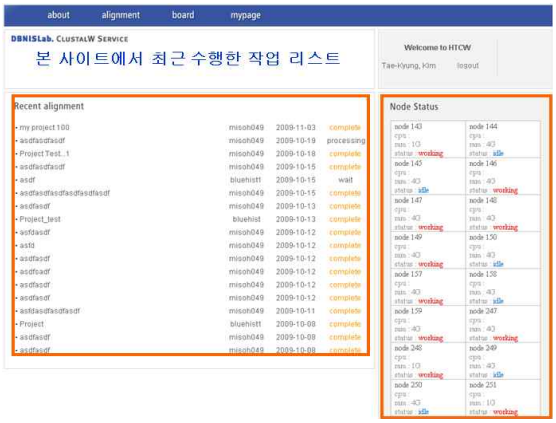

그림 4. 서비스 초기화면
Fig. 4. Initial Interface

그림 5. 프로젝트 생성화면
Fig. 5. Project Creation Interface

그림 6. 프로젝트 내의 작업처리 결과
Fig. 6. Project Result List

4.2 작업 라우팅 알고리즘 평가

여러 개의 다중서열정렬 작업에 대한 분산 작업의 효율을 높이기 위해 3.2절에서 라우팅 알고리즘을 제안하였다. 성능 평가는 16대의 PC를 활용하여 189개의 유전자 그룹에 대해 수행하였다. 성능은 응답시간 (Response Time), 성능향상(Speedup), 효율(Efficiency)의 3가지 측정하였다.

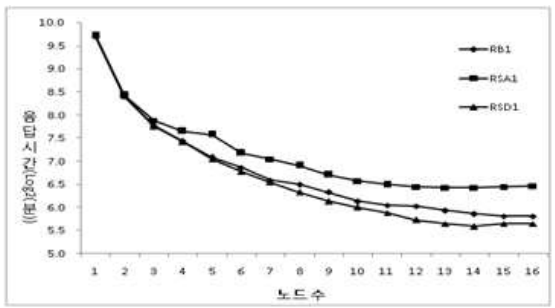

그림 7. 응답시간 측정
Fig. 7. Response Time

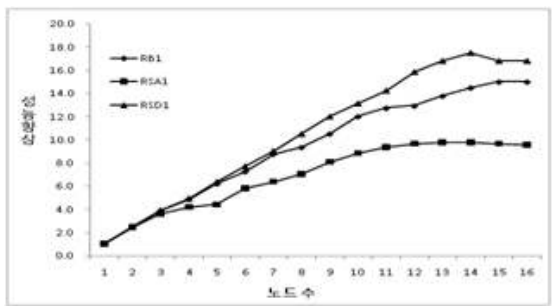

그림 8. 성능향상 측정
Fig. 8. Speedup

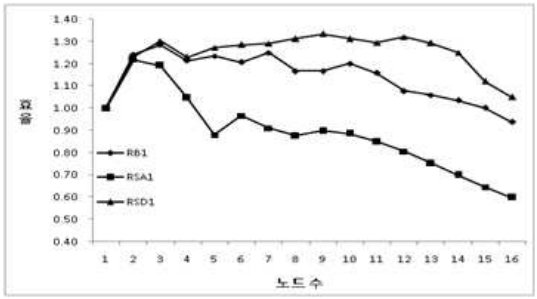

그림 9. 효율측정
Fig. 9. Efficiency

[그림 7], [그림 8], [그림 9]를 통해 노드 증가에 따라 응답시간이 비례하여 줄어들고 있음을 확인할 수 있다. 특히 응답시간은 RSD 알고리즘이 가장 우수함을 알 수 있다. 또한 성능향상 정도는 노드의 증가에 따라 비례하여 향상되고 있으며, 효율도 높음을 알 수 있다. 응답속도 관점에서는 RSD가 가장 뛰어나고, 초기 처리결과에 대한 접근은 RSA가 좋다.

Ⅴ. 결론

본 논문에서는 다중서열정렬작업의 편의성과 성능을 높이기 위한 웹기반 시스템을 제안하였다. 다중서열정렬 작업은 유전체를 연구하는 생명공학자들이 가장 많이 사용하는 도구로 기존에는 설치와 데이터관리 및 처리, 그리고 성능의 한계가 있었다. 그러나, 제안된 시스템은 편리한 사용자 인터페이스의 제공, 고성능 라우팅기법의 도입, 라우팅 알고리즘의 선택가능 등의 기능을 제공하므로 사용자 편이성과 분석성능이 크게 향상되었다. 향후 연구계획은 구축된 시스템과 다른 분석시스템간의 연동을 통해 서열분석 결과의 활용성을 극대화할 예정이다.

참고문헌

[1] S. F. Altschul, W. Gish, W. Miller., “Basic Local Alignment Search Tool,” Journal of Molecular Biology, 215:403-410, 1990.

[2] D. Higgins and P. Sharp, “CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer,” Gene, 73, 237-244, 1988.

[3] Tae-Kyung Kim and Wan-Sup Cho. “CCGRID construction based on Etherboot technology and its Utilization to Sequence analysis,” *Journal of the Korean Data & Information Science Society*, Volume 10, Number 6, December 2005, pages 569-580.

[4] GOLD database, http://www.genomeonline.org

[5] MAFFT, http://mafft.cbrc.jp/alignment/server/

[6] MUSCLE, http://www.ebi.ac.uk/Tools/muscle/index.html

[7] DCA, http://bibiserv.techfak.uni-bielefeld.de/dca/