클러스터링 단백질 데이터베이스와 데이터 분산 기법을 적용한 단백질 이차구조예측 시스템 설계*

이수진*0 김재훈* 정진원** 이원태**
아주대학교 정보통신전문대학원* 연세대학교 생화학과**
{genie77⁰, jaikim}@ajou.ac.kr, {solwind,wlee}@spin.yonsei.ac.kr

Protein Secondary Structure System Design Using Clustering Protein Database and Data Distribution Scheme

Soojin Lee*0 Jai-Hoon Kim* Jin-Won Jung** Weontae Lee**
Graduate School of Information and Communication, Ajou University*
Department of Biochemistry, Yonsei University**

요 약

생물학 데이터베이스의 크기가 점점 증가함에 따라 데이터베이스를 사용하여 서열을 정렬할 경우 많은 처리시간이 필요하게 되었다. 단백질 이차구조예측 시스템에서 단백질 서열 데이터베이스를 이용해 사용자의 서열들을 정렬하는 부분에서도 많은 처리 시간을 요구한다. 본 논문에서는 단백질 데이터베이스를 비슷한 크기로 나눠 여러 노드에서 서열 정렬을 분산 처리하여 처리율을 높이고자 했다. 또한, ClustalW에서 서열들의 관계에 따라 다양한 BLOSUM을 사용하여 정렬의 정확도를 높이는 휴리스틱 전략을 적용하기 위해 기존의 데이터베이스를 클러스터링 하였다. 클러스터링된 데이터베이스의 대표서열과 사용자 서열의 거리를 비교하여 적합한 BLOSUM을 선택하여 보다 정확한 서열 정렬을 통해 단백질 이차구조예측의 정확도를 높이게 될 것이다. 본 논문에서는 대용량의 단백질 데이터베이스를 여러 노드를 사용하여 병렬 클러스터링하여 이를 이차구조예측 시스템에 적용하여 처리율과 정확도를 높이고자 하였다.

1. 서 론

BT와 IT의 결합으로 생물 정보 데이터에 대한 처리가 빨라짐에 따라 데이터가 데이터를 생성하여 생물 데이터의 정보가 점점 증가하게 되었다. 생물 정보 데이터가 점점 증가함에 따라 생물 정보학 데이터베이스를 이용하여 또 다른 정보를 구하기 위한 시스템에서는 많은 처리 시간을 요구하게 되었다. 단백질 서열 데이터베이스도 GenBank[1], PIR[1] 등과 같은 다양한 데이터베이스에서 나날이 데이터들이 증가함에 따라 이를 이용하여 정보를 구하는 시스템의 처리 효율이 점점 떨어지게 되어 보다 효율적인 처리 방법이 요구되었다. 본 논문에서는 단백질 데이터베이스를 이용하여 단백질 이차구조를 예측하는 시스템의 성능을 높이기 위해 데이터베이스의 병렬 클러스터링과 데이터 분산 기법을 적용한 이차구조예측 시스템을 설계하였다.

2. 관련연구

2.1 데이터 클러스터링 알고리즘

계층적인 클러스터링(Hierarchical Clustering)에는 하나의 데이터 아이템으로 클러스터를 시작하여 데이터의 유사도에 의해 합쳐지는 bottom-up 방식인 집적적인 클러스터링(agglomerative clustering) 알고리즘과 모든 데이터로 구성된 하나의 클러스터를 나누어 그룹을 만드는 top-down 방식인 구분적인 클러스터링(divisive clustering) 알고리즘이 있다[2].

본 논문에서는 계층적인 클러스터링 방식 중 집적적인 클러스터링 알고리즘을 적용하여 데이터베이스를 클러스터링하였다. 즉, 데이터베이스의 한 서열을 하나의 클러스터로 시작하여 나머지 데이터들과 비교하여 유사한 서열들을 골라내어 그룹으로 만들었다.

2.2 BLOSUM (Blocks Substitution Matrix)[1,3,5]

BLOSUM은 치환 행렬의 한 종류로, 아미노산 서열 중 다른 부분에 비해 보존이 잘된 부분만을 모아 만든 Block 데이터베이스로부터 만들어졌다. 본 논문에서 이용하는 PSIPRED프로그램의 경우, NCBI(National Center Biotechnology Information)[3]에서 제공하는 PSI-BLAST[1,4,5]를 적용하여 서열 정렬을 하게 되는데, PSI-BLAST프로그램의 경우 모든 데이터베이스 내의 서열에 대해서 BLOSUM62 매트릭스를 사용하여 정렬을 하게 된다. 본 논문에서는 데이터베이스내의 서열들을 병렬 클러스터링하여 여러 개의 노드에 클러스터링된 여러 데이터베이스들을 사용자 서열과의

* 본 연구는 정보통신부 정보통신선도기반기술개발사업의 지원에 의하여 이루어진 것임.

관계에 따라 적합한 BLOSUM들을 선택, 적용하여 서열 정렬의 정확성을 높이고자 하였다.

2.3 단백질 이차구조예측

단백질 이차구조예측은 Chou-Fasman[1] and GOR [1]과 같은 통계학적인 방법이 예전에 많이 사용되었고, 최근에는 기계학습에 의한 알고리즘들이 개발, 적용되어 80%에 이르는 정확도를 보이고 있다.

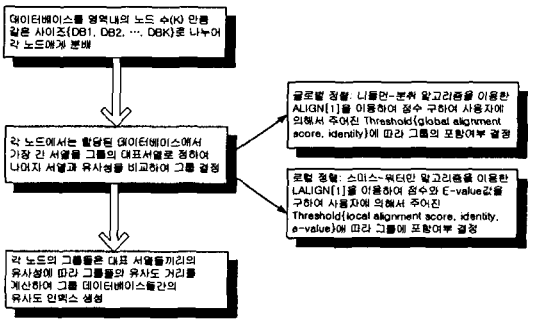
본 논문에서 높은 정확도를 보이는 PSIPRED[1,5]를 애플리케이션으로 수정하여 사용하였다. 이 프로그램은 단일 서열의 이차구조를 예측하는 프로그램으로, 전처리 과정으로 서열을 정렬하는 PSI-BLAST 프로그램과 PSSM(Position Specific Scoring Matrix)를 PSIPRED 프로그램의 입력 값으로 사용하기 위해 NCBI Toolkit[3,5]이 있다. 본 논문에서는 사용자 서열과 데이터베이스내의 서열과의 관계에 대한 고려 없이 BLOSUM62만을 사용하여 서열을 정렬하는 PSI-BLAST 프로그램을 클러스터링 데이터베이스의 대표 서열과 사용자의 서열과의 관계를 구해 적합한 BLOSUM을 선택하여 서열을 정렬할 수 있게 하였다. 이것은 ClustalW[1] 프로그램의 휴리스틱 전략가운데 각 정렬을 수행할 때 예상 진화 거리에 기반한 다른 측정 행렬을 사용하여 근연관계에 있는 서열일 경우 근연관계에 최적화된 측정 행렬을 사용하고, 원연관계에 있는 서열일 경우 원연관계에 최적화된 측정 행렬을 사용하는 전략을 이차구조예측 시스템에 적용하여 좀 더 정확한 서열의 정렬을 통해 PSSM을 만들어 이차구조의 예측을 높여 보고자 한다.

3. P-dbclustr: 병렬 단백질 데이터베이스 클러스터링 프로그램 구현

Genbank의 경우 14개월 단위로 데이터베이스의 사이즈가 2배씩 증가하므로, 데이터베이스를 사용하여 서열 정렬하는 PSI- BLAST 프로그램의 효율은 계속 떨어질 수 밖에 없다. 왜냐하면, PSI-BLAST 프로그램의 경우 데이터베이스를 메모리에 올려놓고 비교하기 때문에, 메모리보다 데이터베이스의 사이즈가 커질 경우 데이터베이스를 메모리에 맞게 여러 번 올려놓고 비교해야 하므로 효율이 떨어지고, 비교해야 할 대상이 많아지기 때문에 비교 시간과 처리 시간이 길어지게 된다. 이를 효율적으로 처리기 위해 데이터베이스를 일정한 크기로 나누어 클러스터를 이용하여 여러 노드에서 병렬로 처리할 수 있다. 하지만, 의미 없이 일정한 크기로 데이터베이스를 나누어 처리하는 것보다 기존의 데이터베이스를 클러스터링 하여 유사한 서열들로 이루어진 여러 개의 데이터베이스로 만들어 사용자의 서열에 따라 각각의 데이터베이스에 대해 적합한 BLOSUM을 사용하여 서열 비교 및 정렬하게 된다면, 처리 효율의 증가뿐만 아니라 좀 더 정확한 서열 정렬을 얻을 수 있을 것이다. 기존의 생물학 데이터베이스를 하나의 컴퓨터에서 클러스터링 하려면 수십 시 간의 처리시간이 필요 한다. 이러한 처리 시간을 단축하기 위해 여러 노드를 사용하여 병렬로 데이터베이스 클러스터링하여 처리 효율을 높이고자 한다.

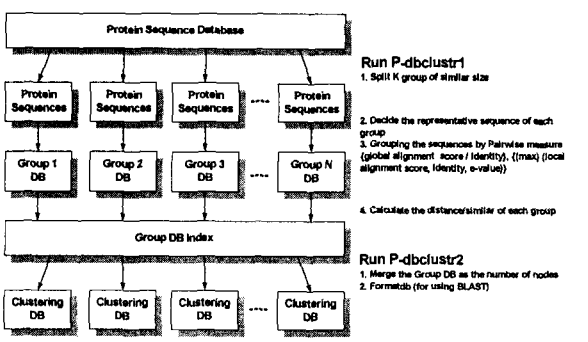
본 논문에서는 병렬 클러스터링 방법을 P-dbclustr1과 P-dbclustr2로 두 가지 단계로 나누었다. 첫 번째 단

첫 번째 단계에서는 데이터베이스를 사용 가능한 클러스터 노드 수만큼 나누어 각 노드에서 그림 1의 알고리즘을 실행한다.



[그림 1] P-dbclustr1 알고리즘

두 번째 단계에서는 사용자의 요청이 들어오면 첫 번째 단계에서 생성한 그룹 데이터베이스의 거리 인덱스를 사용하여 사용 가능한 클러스터 노드 수만큼 유사한 크기로 그룹 데이터베이스들을 합하고 각각의 클러스터링된 데이터베이스에서 가장 긴 서열을 대표서열로 선택한다. 이렇게 생성된 클러스터링 데이터베이스를 이차구조예측시스템에 적용하여 사용자의 서열에 따라 클러스터링 데이터베이스의 대표서열과의 비교를 통해 적합한 BLOSUM을 사용하여 서열 정렬을 하게 된다. 그림 2는 P-dbclustr에서의 병렬 클러스터링 방법을 보여준다.



[그림 2] P-dbclustr 에서의 병렬 클러스터링 방법

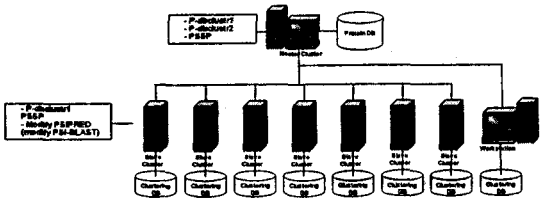
4. 단백질 이차구조예측 시스템 설계

앞의 장에서 생성된 그룹 데이터베이스들과 그룹 데이터베이스 인덱스를 사용하여 P-dbclustr2로 각각의 생성된 그룹 데이터베이스들을 사용할 수 있는 노드 수만큼 클러스터링 데이터베이스를 만든다. 본 장에서는 이차구조예측 시스템의 테스트베드의 설명과 본 논문에서 제안하는 단백질 이차구조예측 시스템에 대해 설명하고자 한다.

4.1 테스트베드

본 논문에서는 8개의 노드로 이루어진 클러스터와 워크스테이션 1개로 시스템을 구축하였다. 클러스터는 1개의 마스터 클러스터와 7개의 슬레이브 클러스터로 구성되 어 있다. 마스터 클러스터에는 단백질 데이터베이스의 병렬 클러스터링을 위한 P-dbclustr1과 P-dbclustr2

프로그램을 설치하고 사용자의 서열을 받아 각 노드에게 나누어주고 결과를 받아 처리하는 PSSPS(Protein Secondary Structure Prediction System)가 설치된다. 프로그램을 설치하고 사용자의 서열을 받아 각 노드에게 나누어주고 결과를 받아 처리하는 PSSPS(Protein Secondary Structure Prediction System)가 설치된다. 마스터 클러스터는 입력값과 결과값을 주고 받고, 결과값들을 취합하므로 컴퓨팅 노드에서는 제외하였다.



[그림 3] 테스트베드 구성도

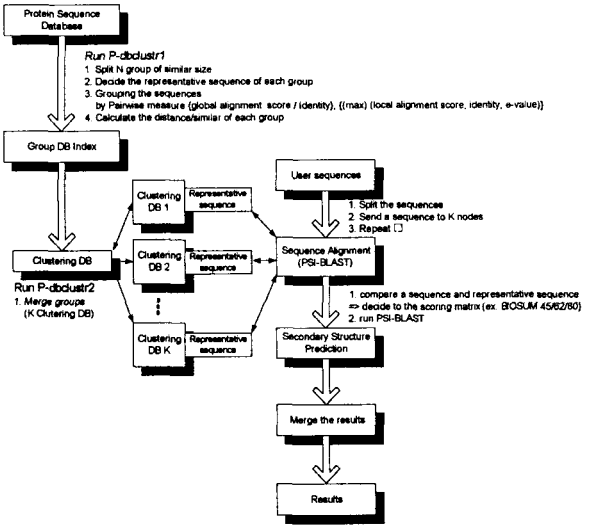
그림 3은 본 논문에서 구성한 테스트베드의 구성도를 보여준다. 컴퓨팅 노드로 슬레이브 클러스터 7대와 워크스테이션 1대에는 병렬 클러스터링시에 마스터 클러스터에서 전송 받은 부분 데이터베이스에 대해 그룹 데이터베이스를 만들어 마스터 클러스터에 전송하는 부분으로 되었다. 표1는 테스트베드의 시스템 사양이다.

[표 1] 시스템 사양

시스템 구성	대수	OS	H/W	MEM-ORY
마스터 클러스터	1	Linux Kernel 2.4.2	Intel Pentium 3	512M
클러스터 슬레이브	7	Linux Kernel 2.4.2	Intel Pentium 4	512M
워크스테이션	1	Linux Kernel 2.4.13	Intel Pentium 4	512M

4.2 단백질 이차구조예측 시스템 구조 설계

사용자가 단백질 서열(들)의 이차구조예측을 요청하면, 그림 3과 같이 단백질 이차구조예측 시스템 3장의 P-dbclustr1에서 생성한 그룹 데이터베이스 인덱스를 가지고 P-dbclustr2로 사용할 노드 수에 맞게 클러스터링 데이터베이스를 생성하여 각 노드에 전송한다. 사용자의 단백질 서열들은 단일 서열들로 나누어지어 단일 서열 하나씩 처리하게 된다. 사용자의 단일 서열은 각 노드에서 클러스터링 데이터베이스의 대표 서열과 거리를 구해 45%미만의 동일성을 가지면 BLOSUM45, 45%~80%의 동일성을 가지면 BLOSUM62, 80%이상의 동일성을 가지면 BLOSUM80을 선택하여 서열 정렬을 하여 이차구조예측을 위한 입력 값으로 PSSM(Position Specific Scoring Matrix)를 생성하게 된다. 이 매트리스를 사용하여 각 노드에서는 이차구조예측 결과값이 나오고 이 결과 값은 마스터 클러스터에 전송된다. 마스터 클러스터에서는 최종 결과를 생성하여 사용자에게 전송하게 된다. 그림 4은 병렬 단백질 데이터베이스 클러스터링 프로그램과 이를 통해 생성되는 클러스터링 데이터베이스와 데이터 분산 기법이 적용된 단백질 이차구조예측 시스템을 보여준다.



[그림 4] 제안된 단백질 이차구조예측 시스템 구성도

5. 결론 및 향후 과제

본 논문에서는 단백질 데이터베이스의 병렬 클러스터링 방법과 데이터 분산 기법을 적용하여 단백질 이차구조예측 시스템의 성능을 높이기 위한 시스템을 설계하였다. 본 논문에서 제안된 P-dbclustr와 단백질 이차구조예측 시스템은 대용량의 데이터베이스에 대한 클러스터링을 위해 병렬 처리를 하여 보다 빠르게 클러스터링된 데이터베이스를 생성하고 생성된 클러스터링 데이터베이스들의 대표 서열들과 사용자의 서열(들)과의 정렬을 통해 적합한 BLOSUM을 사용하여 처리 성능과 정확도를 높일 것이다. 향후 과제에서는 그리드를 이용하여 시스템을 확장할 예정이다.

참고문헌

[1] Cyntbia Gibas and Per Jambeck, " Developing Bioinformatics: Computer Skills," O' REILLY, 2001.

[2] O. Trelles, M.A. Andrade, A. Avlenicia, E.L. Zapata and J.M. Carazo, " Computational space reduction and parallelization of a new clustering approach for large groups of sequences," *Bioinformatics, Vol. 14, No. 5*, pp. 439-451, 1998.

[3] NCBI: http://www.ncbi.nlm.nih.gov

[4] Stephen F. Altschul, Thomas L. Madden, Alejandro A.Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller and David J. Lipman, " Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, Vol. 25, No.17, pp. 3389-3402, 1997.

[5] David T. Jones, " Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices," J. Mol. Biol. 292: pp. 195-202, 1999.