C-Hunter 알고리즘을 이용한 유전자 기능 분석 구현에 관한 연구

지민근, 이강만*
강릉원주대학교 컴퓨터공학과
jmg0630@cs.gwnu.ac.kr, gangman@cs.gwnu.ac.kr

A study on the implementation for the gene functional analysis using C-Hunter algorithm

Min-Geun Ji, Gangman Yi*
Dept of Computer Science & Engineering
Gangneung-Wonju National University

요 약

최근 대량으로 발생되고 있는 유전체 정보 해석 과정 중의 하나로 유전체 내에서 유전자 기능 분석을 수행하기 위하여 본 논문에서는 C-Hunter 알고리즘을 이용하여 계통 발생 다양성을 대표하는 8종의 생물데이터를 이용하여 유전자 클러스터들을 다양한 OS에서 GUI환경으로 기능 해석을 수행 할 수 있도록 구현하였다. 본 연구를 통하여 NGS 수행 시 대량 생산되는 유전체의 유전자 기능 분석을 보다 빠르고 정확하게 다양한 환경에서 수행할 수 있을 것으로 기대한다.

1. 서론

최근 Next Genome Sequencing을 통해 분석해야 할 데이터들은 방대한 데이터로 지속 생산되고 있다. NGS를 통해 생성된 리드 데이터는 어셈블리에서 어노테이션까지 각 단계별로 다양한 분석이 요구되어진다. 특히 어노테션을 위한 기능 분석은 기존 비교 대상이 있는 레퍼런스 시퀀싱과 비교 대상이 없는 새로운 유전체의 시퀀싱을 수행하는 드노보 어셈블리로 구분되어지는데 드노보 어셈블리 같은 경우는 비교 대상이 없거나 많지 않아 기능 분석이 쉽지 않다. 또한 유전체에서 유전자 클러스터를 판별하는 대부분의 알고리즘들은 근접한 유전자들의 관계 분석을 통하여 유전자 클러스터를 판별하지만 최근 많은 문헌들을 통해 유전자 클러스터 내의 유전자의 위치는 근접 위치뿐만 아니라 원거리에도 산재할 수 있다고 보고되고 있다[1,2]. 대부분의 방법론들은 마이크로어레이 데이터나 대사경로 데이터베이스를 사용하여 염색체 내에서 공통 속성들로 연결된 유전자 그룹 찾기 방법이 사용되어진다. 하지만 종의 메커니즘에 관계없이 보다 일반화된 방법론을 이용하여 유전자 클러스터를 찾을 수 있다면 진화연구 및 유전자 클러스터 판별에 편향되지 않은 알고리즘으로 수행할 수 있을 것이다.

본 논문에서는 위와 관련된 문제들을 해결하고자 C-Hunter 알고리즘을 사용하여 원핵생물과 진핵생물에서 모두 사용될 수 있는 기능 분석 소프트웨어를 구현하였으며 보다 정확한 클러스터 판별을 위하여 그래프 기반의 어휘들로 구성된 Gene Ontology를 이용하여 최소한의 생물학적 속성 데이터를 사용하였다[3,4]. 이러한 방식은 유전자 발현이나 그 외의 다른 속성들의 제약 없이 GO를 소유한 공통 기능들만으로 분석이 가능하며 최적의 클러스터 결과물 산출을 위하여 p-value와 e-value의 통계테스트를 통해서 클러스터링을 수행한다. 본 논문에서는 다양한 크기 및 통계 수치의 클러스터의 의미 있는 다양한 분석을 위하여 윈도우즈, 리눅스, 맥 오에스에서 GUI환경으로 구현되었다.

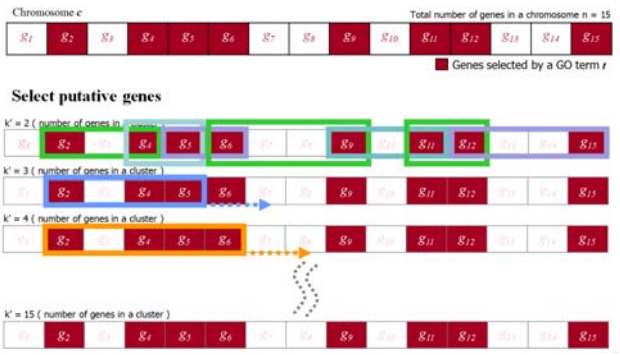
2. 데이터 구성

본 논문에서 구현한 시스템은 유전체에서 염색체 내 유전자 위치 정보인 유전자 지도와 각 유전자의 GO 어노테이션을 기본 입력 데이터로 이용한다.

기본 입력 데이터를 구성하기 위하여 계통 발생 다양성을 대표하기에 충분한 대표 종 8개 (*Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Escherichia coli, Homo sapiens, Mus musculus, Saccharomyces cerevisiae*)를 선택하였다. 각 유전체는 최저 약 25%에서 최대 약 96%의 어노테이션이 되어있다. 초기 미가공 데이터는 NCBI에서 제공되는 유전체 데이터를 사용하였다[5].

GO 데이터베이스는 biological process, celluar component, molecular function 3개의 범주로 구분되어진다. 이 데이터는 “GO Database Converter”를 통해서 GO Graph로 변환된다. 데이터는 NCBI 에서 gene2accession

* 교신 저자 (corresponding author)
** 이 논문은 2013년도 강릉원주대학교 학술연구조성비 지원에 의하여 수행되었음.



(그림 1) C-Hunter 알고리즘 예

파일과 gene2go파일을 통해서 염색체 상의 각 유전자에 GO 어노테이션 정보로 사용된다.

3. 유전자 기능분석 모델

유전자 클러스터는 같은 GO 용어 또는 같은 부모 용어에 어노테이션 되어 있는 유전자들의 그룹을 의미한다. 따라서 클러스터는 크기와 길이로 구분되어지는데 크기는 염색체 상에서 클러스터 내에 어노테이션된 유전자 개수이고 길이는 염색체 상의 클러스터가 점유하고 있는 길이가 된다. C-Hunter 알고리즘은 그림1과 같이 최소 2의 가변 클러스터 크기를 사용하여 염색체 상에서 다양한 클러스터를 판별한다.

4. 시스템 설계

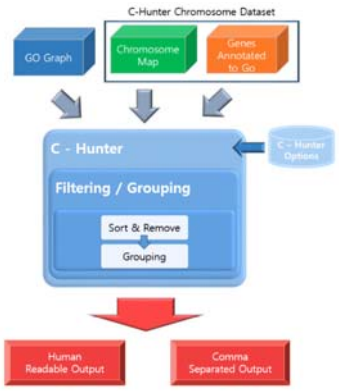
4.1 데이터베이스

본 소프트웨어에 필요한 데이터베이스는 아래와 같이 총 3개가 요구된다.

- *GO Graph* - GO DB에서 DAG 방식으로 변경된 C-Hunter 데이터베이스
- *Chromosome Map* - gene2access를 이용하여 생성된 염색체 상의 유전자 위치 정보 데이터베이스
- *Genes Annotated to GO* - gene2go를 이용하여 염색체 상의 각 유전자의 GO 어노테이션 데이터베이스 (한 유전자에 2개 이상의 어노테이션 할당이 가능하다.)

4.2 입력 파라미터

4.1에서 생성된 데이터베이스들을 기반으로 염색체 상에서 클러스터를 찾기 위해서는 최소/최대의 클러스터 크기/길이 및 통계테스트의 cutoff, 필터링/그룹핑 스텝에 대한 옵션이 필요하다[그림1]. 클러스터가 구성되기 위해서는 최소한 두 개의 어노테이트된 유전자가 존재해야하고 최대 크기는 염색체 상 최대 유전자 수가 되기 때문에 최대 유전자 수를 0으로 설정 시 최대에 대한 고려 없이 전



(그림 2) 워크플로우

체를 다 계산하기 때문에 수행시간이 늘어난다.

클러스터의 길이는 다양한 클러스터를 판별하기 위하여 각 염색체 상에서 가변적으로 수행되기 위하여 기본값 0으로 제한을 두지 않는다. 이는 인접한 유전자들로 구성된 클러스터뿐만 아니라 유전자 사이의 거리가 다양한 길이의 클러스터까지 판별 가능하기 때문이다.

4.3 통계테스트

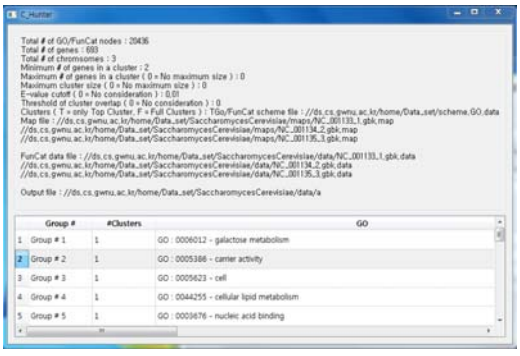
다양한 크기의 클러스터들의 다양한 염색체에서 생성되기 때문에 클러스터의 중요도 판별이 요구되어진다. 따라서 초기하분포를 이용하여 p-value와 e-value를 통해서 통계적 중요도를 판단한다.

4.4 필터링/그룹핑 단계

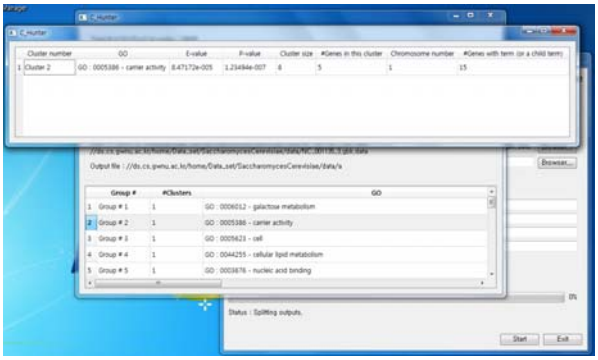
최종 결과물 생성 전에 의미 있는 결과물 산출을 위해서 필터링/그룹핑의 두 단계를 수행한다[그림2].

필터링 단계는 서브셋 클러스터들을 제거하는 단계이다. 큰 e-value의 클러스터들이 작은 e-value 클러스터의 서브셋으로 포함되는 경우, 작은 e-value의 큰 클러스터가 더 의미 있기 때문에 서브셋 클러스터는 제거한다.

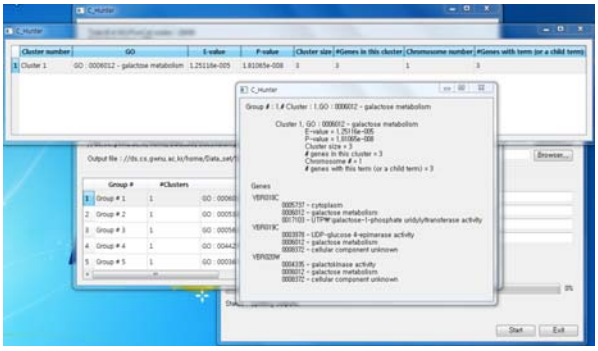
그룹핑 단계는 대표 클러스터를 설정하고 하위 클러스터들을 그룹의 멤버 클러스터로 설정하는 단계이다. 클러스터 e-value 정렬을 통해 의미 있는 대표 클러스터를 선별로 분석이 보다 용이하다.



(그림 3) Top cluster



(그림 4) 대표 클러스터 내에 클러스터 목록



(그림 5) 클러스터 내에 유전자와 각 유전자의 GO 정보

4.5 최종 결과

결과물은 총 3단계로 나뉘어서 분석이 가능하다. 첫 단계는 사용된 옵션들과 그룹 되어 있는 대표 클러스터 리스트로 결과물이 구성된다[그림3]. 각 대표 클러스터들은 e-value로 정렬되었으며 각 그룹 내 클러스터의 개수 및 연결된 GO 리스트로 구성된다.

각 대표 클러스터 내에 최소 1개 이상의 클러스터들은 각 클러스터에 어노테이트된 GO와 e-value, p-value, 클러스터 크기, 길이, 염색체 번호 및 GO 그래프에서 하위 용어에 어노테이트된 유전자의 수를 포함한 전체 유전자의 수의 정보를 표현한다. 리스트의 정렬은 e-value를 기준으로 한다[그림4].

마지막 단계에서는 각 클러스터의 세부 정보를 제공한다. 각 클러스터는 소속된 클러스터 그룹 및 GO 등의 기본 클러스터 정보뿐만 아니라 각 유전자에 DAG방식의 GO 그래프에서 자식 노드 GO까지 고려한 GO 정보를 각 유전자에 표시함으로 대표 어노테이션과 각 유전자의 어노테이션 정보를 같이 제공하여 기능 분석에 용이하다[그림5].

5. 개발 환경 및 구현

본 논문에서 수행된 연구의 구현은 Qt-4.8.5 library와 Qt Creator SDK를 이용하여 C++로 개발되었다[6]. 컴파일 환경은 윈도우, 리눅스, 맥 오에스에서 컴파일 되어 각 다른 환경의 OS에서 동일한 GUI를 제공한다.

6. 결론

본 논문은 원핵생물과 진핵생물에서 기능적으로 연관된 유전자 클러스터를 찾기 위해 C-Hunter 알고리즘을 사용하여 기본 클러스터 정보를 다양한 OS의 GUI환경에서 판별이 가능하도록 하였다. 통계테스트를 이용하여 대표 클러스터 단위 분석이 가능하도록 구현이 되었다. 앞으로 NGS를 통해 생성되는 대용량 유전자 데이터를 처리 및 분석하기 위해서는 다양한 환경의 분석 시스템이 요구되어지고 있다. 현시점에서 본 연구는 이러한 이슈들 해결하기 위한 방법 중의 하나가 될 것이다.

참고문헌

[1] J. M. Lee and E. L. L. Sonnhammer, “Genomic gene clustering analysis of pathways in eukaryotes” Genome Res., vol. 5, pp. 875-882, 2003.

[2] L. D. Hurst, C. P¶al and M. J. Lercher, "The evolutionary dynamics of eukaryotic gene order" Nat Rev Genet., vol. 5, pp. 299-310, 2004.

[3] Yi G, Sze SH, Thon MR. "Identifying clusters of functionally related genes in genomes" Bioinformatics. 2007 May 1;23(9):1053-60

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene ontology: tool for the uni¯cation of biology" Nat Genet., vol. 25, pp. 25-29, 2000.

[5] http://www.ncbi.nlm.nih.gov/

[6] http://qt-project.org/