Orthologous 데이터베이스의 효율적인 구축 방안

오정수* 조완영** 김태경*** 김선선*** 이종섭*** 권혜향*** 김영창***

충북대학교 "경영정보학과", "정보산업공학과", "전자계산학과", "미생물학과
misois@hanmail.net

An Efficient Methodology For The Construction Of Orthologous Database
Dept. of "Management Information System, "Information Industrial Engineering, "Computer Science, "Microbiology
Chungbuk National University

요 약
생활을 진화적으로 분석할 때, 보전적인 유전자(Conserved gene)들은 기능이 알려지지 못했던 다양한 생물학적 정보를 얻어 내는데 유용하게 쓰일 수 있다. 특히 전례히 서열이 밝혀진 지문(Conome) 데이터로부터 진화적으로 보존적인 유전자 서열의 상동성을 따른 분류를 통한 2차 데이터베이스의 구축은 생물학자들에게 다자견적인 정보를 제공할 수 있다. 이에 이러한 데이터베이스가 다양한 방법에 따라 구축 되었고 생물학자들의 연구에 활용되고 있다. 그러나 각 구축된 데이터베이스들이 생물학자들이 이용하기에 paralogs의 포함 문제점으로 인해 신뢰성이 떨어지거나 데이터베이스 생성 기간이 오래 걸리는 단점을 있다. 본 연구는 기존에 구축된 데이터베이스들의 구축 방법에 주목하고, 정보기술을 활용하여 빠르고 효과적으로 정확성을 높인 새로운 구축 방법론과 데이터베이스를 활용한 분석 시스템에 대해 제시하고자 한다.

1. 서 론

본 논문에서 실험적 단계는 유전체 분석이라고 할 수 있다. 유전체를 분석하기 위해서는 먼저 유전체 시퀀싱(Seqenescing) 작업이 실행되어야 하며 시퀀싱된 유전자를 통해 기능을 예측하는 주제(Annotation) 과정을 거치게 된다. 기술의 발달로 인해 과거에 비해 빠르게 많은 양의 시퀀싱된 데이터를 얻을 수 있게 되었고 정보기술을 활용하여 기능을 예측하고 분석하는 논문도 도시가 방법이 현재 널리 쓰이고 있다. 유전자의 기능을 예측하는데 활용되는 가장 보편적인 방법은 브라스트(Blast)과 같은 프로그램을 이용해 기존의 서열 데이터베이스와 서열의 상동성 비교를 통해 기능 예측 방법이 주로 사용된다. 그러나 사용하는 데이터베이스의 신뢰성이 떨어지면 안 되니 그 데이터베이스를 이용하여 생물학자들이 다양한 분석을 하기엔 까다로운 것이 사실이다.

본 논문은 이러한 단점을 보완하기 위해 진화적으로 보존적인 유전자들의 서열의 상동성에 따른 분류를 통한 orthologous 데이터베이스를 구축하고, 구축된 데이터베이스를 바탕으로 유전자들의 다자간적인 분석가 가능한 시스템에 대해 제시하고자 한다. 여기서 orthologous는 공동의 조상으로부터 분화(divication)되어 서로 다른 유전자에 있는 직접적으로, 관련된 유전자들의 집합이라고 정의하며 이와 반대로 paralogs는 같은 유전자 내에서 복제(duplication)에 의해 생성된 관련된 유전자들의 집합으로 정의할 수 있다. 일반적으로 같은 orthologous 관계에 있는 유전자들은 서열의 유사성을 함께 같은 기능을 갖게 된다. paralogs는

본 연구는 한국과학기구특성기초 연구사업(R01-2003-000-11723-0, R01-2001-000-00697-0)로부터 지원을 받았음

서열의 유사성을 갖고 있지만 진화적으로 기능은 완전히 달라진다. 따라서 orthologous 관계를 분류한 2차 데이터베이스의 구축은 계통발생학적(phylogenetic) 분석과 있어서, 서로 다른 종에서 나타나는 공통의 유전자 파악 및 정확한 기능 예측에 다양한 분석을 위한 기본모델이라고 말할 수 있다.

본 연구는 다음과 같이 크게 3가지 목적으로 가지고 있다. 첫째, 빠르고 신뢰성 있는 orthologous 데이터베이스를 구축한다. 둘째, 구축된 orthologous 데이터베이스를 활용하여 유전자들의 orthologous 관계 파악 및 유전자와 죽 동일한 유전자를 구별할 수 있다. 셋째, 유전체 분석이 유전자와 신뢰성이 있는 기능 예측을 가능하다. 이를 위해 본 논문에서는 컴퓨터 기술을 최대한 활용하여 기존의 구축된 orthologous 데이터베이스의 단점을 보유한 데이터베이스 구축을 위한 방법론과 orthologous 데이터베이스를 활용한 분석 시스템에 대해 개발하도록 하였다.

본 논문은 2장에서 지금까지의 orthologous 관련 데이터베이스나 방법론에 대해 소개하고, 3장에서 본 논문이 제시한 orthologous 클러스터링 방법론 및 데이터베이스 구축 방법론을 제시한다. 4장에서는 orthologous 데이터베이스를 구축한 후 분석을 하기 위한 시스템에 대해 제시하고, 마지막으로 결론과 향후 연구방향에 대해서 말한다.

2. 관련 연구

2.1 COG(Clusters of Orthologous Groups of proteins)
COG는 현재 서열이 완전히 발전한 66개의 유전자 단백질 서열의 일련의 동등성 비교를 통해 orthologous 관계를 파악하고 유사한 기능을 하는 도메인으로 나누어 그룹을 지었다. COG는 크게 다음과 같은 방법로 구축되었다[1].
1. 모든 단백질 서열의 all-against-all 비교를 수행한다.
2. 명백히 드러나는 paralog을 제거한다.
3. 비교된 서열 중 서로 best hit인 유전자들이 3개 이상이면 하나의 orthologous 그룹으로 묶어준다.
4. 기능상 유사한 것들을 도메인별로 묶어 분류한다.
이와 같은 방법으로 구축된 COG는 정보수집 및 처리기능 관련 유전자군(I,L,K,L), 세포생물 기능 관련 유전자군(D,F,G, M,L,N,P,T), 대사 기능 관련 유전자군(G,C,E,F, H,L), 기능 비슷한(R,L,S)군 등 7,469개의 Orthologous 그룹으로 구성되어 있다. COG는 위의 모든 과정을 컴퓨터 프로그래밍 등을 사용한 과정을 수행하여 구축하기 때문에 관리하기 힘든 데이터베이스를 구축하였다. 그러나 COG는 반복되는 paralogs를 완전히 제거하지 못한다. 이로 인해 기능상 유사하지 않은 도메인 그룹이 생기게 되었고 신뢰성 있는 분석을 하는데 방해 요인이 되고 있다.

2.2 KO/KEGG Orthologous

2.3 CGB(Center For Genomics and Bioinformatics)
CGB(Center For Genomics and Bioinformatics)에서는 두 종간의 자동적인 서로 비교를 통해 orthologous 그룹을 형성하는데 cut-off 방식을 적용하여 신뢰성을 높인 방법론을 제시하였다[4]. 컴퓨터 프로그래밍의 기반으로 간편하게 신뢰성이 높은 orthologous 클러스터링을 형성해 주는 장점이 있다. 그러나 두 종간만의 대상으로 해야 하는 단점이 있고 실제 데이터베이스가 구축되지 않았다.

우리는 위의 방법론의 단점을 보완하여 새로운 구축 방법론을 제시한 것이다.

3. Orthologous 데이터베이스 구축
다음은 본 논문에서 제시하는 Orthologous 그룹의 형성 방법과 데이터베이스 구축 방법에 대해 설명한다.

3.1 Orthologous 그룹의 형성
Orthologous 데이터베이스를 구축하기 위해서 먼저 orthologous 그룹을 형성하여야 한다. 그룹의 형성은 다음과 같은 단계로 거친다.
첫째, 시변형 작업이 완전히 완료된 유전자들의 단백질 서열을 가지고 클러스터를 실행시킨다. 클러스터는 단백질 서열의 비교를 수행하여 유전자들이 가지는 상호성을 분석해주는 프로그램이다. 따라서 이를 통해 우리는 상호 best-hit 된 유전자 목록을 얻을 수 있다. 이때 COG와 같이 다수의 중에 대한 all-against-all 비교가 아닌, 두 중간의 서로 비교만을 수행한다. 이는 COG에서와 같이 paralogs를 완전히 제거하지 못해 신뢰성이 떨어지는 문제점들을 해결하기 위함이다. 두 종간의 비교를 통해 나온 클러스터의 결과 중 top-hit 된 것만을 가지고 두 종간의 orthologous 그룹을 형성한다. 여기서 false positive와 false negative의 문제를 해결하기 위해 cut-off 방식을 적용한다. cut-off는 두 가지로 나누어 적용되는데 score 값에 대한 것과 두 중간의 유전자 소개도시에 관한 것이다. Score 값은 단백질 서열의 유사성이 어느 정도 일치하였나를 나타내는 통계적 값이다. 따라서 아래의 best-hit 되었다면 score 값이 낮으면 신뢰성이 떨어진다. 우리는 score의 cut-off 값을 50%로, overlap cut-off를 50%로 주었다. 이는 false positive와 false negative를 줄일 수 있는 최적의 cut-off 수치이다. 문제에 대해 최대한 신뢰성을 높이는데 데이터의 신뢰성 없이 해석할 수 있기 때문이다[4].
셋째, 두 번째 단계를 반복하여 나온 여러 개의 두 종간 orthologous 그룹을 클러스터링 한다. 그리고 그림 1에서 보듯 최소 3 종 이상에서 orthologous 그룹이 형성된 것만을 분류하여 실제 orthologous 그룹으로 인정한다.

Use the reciprocal best hit methods with cut-off values.

![ClustalO diagram]

그림 1. Orthologous 그룹의 형성

벽계, 결과로 나온 orthologous 그룹 중 몇 개를 KO와 비교를 통해 통계적 해석을 찾는다. 이 과정은 신뢰성이 높은 KO와의 비교를 통해 새로운 구축한 결과의 신뢰성을 확인하고, 발견된 통계적 검정을 바탕으로 피드백을 통해 세 번째 단계에 적용한다. 따라서 더욱 정교하게 orthologous 그룹을 형성하는 최적의 조건을 부여하여 가장 신뢰성을 높은 orthologous 그룹을 형성할 수 있다.

이상의 단계를 거쳐 형성된 orthologous 그룹을 다시 기능성 도메인으로 클러스터링 하는 작업을 한다. 이 작업은 생물학자들의 수작업을 통해 수행된다.

3.2 데이터베이스 구축
최종적으로 나온 클러스터링된 orthologous 그룹을 생물학자들이 다양한 분석을 하기 위한 데이터베이스로 구축해야 한다. 데이터베이스에는 클러스터링 결과에 대한 정보뿐만 아니라 서열, 기능 및 기타 동등한 여러 가지 정보들을 포함하 고 있어야 한다. 이를 위해 데이터베이스 구축을 위한 여러 프로그램의 개발이 요구된다. 또한 무엇보다 다차원적 분석을 엽두에 둔 데이터베이스의 스키마의 설계가 무엇보다 중요하다.

4. 분석 시스템

그림 2는 우리가 개발한 전체 시스템의 개요이다.

![User](image)

관련 데이터
Flat file or XML Doc.

그림 2. 분석 시스템의 개요.

구축된 데이터베이스만으로는 정보기술에 침숙하지 않은 일반 시민의 생물학자들이 다양한 분석을 하기 힘들다. 따라서 생물학자들이 원활한 Orthologous 데이터베이스를 가지고 동적으로 자신이 연구하는 유전자나 생물 종에 대해서 다양한 분석을 하기 위해서 적합한 도구들이 하나의 시스템 안에 포함되어야 한다. 기존의 COG의 경우 웹에서 HTML로 데이터베이스에 있는 정보를 분석 결과만을 보여주기 때문에 실제 생물학자 들이 연구하기에 한계가 있다. KO와 같은 경우도 정보를 제공하기는 하지만 같은 문제가 있다. 따라서 분석시스템이 생물학자들에게 요구되고 있다.

우리가 향후 개발할 시스템은 크게 3가지로 구성되어 있다. 데이터베이스, orthologous clustering tool, 그리고 GUI 기반의 분석도구 등이다. orthologous clustering tool은 추가적인 키워드와 라인어 통합된 orthologous 그룹을 추가시켜주는 기능을 한다. 분석 도구는 높은 정확도 및 사용성을 위해 GUI기반으로 구성한다. 분석도구에는 유전자 예측 컴퓨터코드, orthologous viewer 컴퓨터코드, 마이닝 컴퓨터코드를 포함하고 있다. 유전자 예측 컴퓨터코드는 사용자가 자신이 연구중인 유전자에 대해 기능을 예측하고자 할때 orthologous 데이터베이스의 비교를 통해 기능을 예측하므로 자신의 유전자들의 전화적 위치를 바로 확인할 수 있다. Ortologous viewer 컴퓨터코드는 사용자가 보기