Orthologous 데이터베이스의 효율적인 구축 방안

오정수* 조완섭* 김태경** 김선신*** 이충세*** 권혜룡**** 김영창****
충북대학교 *경영정보학과, **정보산업공학과, ***전자계산학과, ****미생물학과
misojs@hanmail.net

An Efficient Methodology For The Construction Of Orthologous Database

Jeongsu Oh*, wansup Cho*, Taekyuong Kim**, Sunshin Kim***, Chungsei Rhee***,
Haeryong Kown****, Youngchang Kim****
Dept. of *Management Information System, **Information Industrial Engineering,
***Computer Science, ****Microbiology
Chungbuk National University

요 약

생물을 진화적으로 분석할 때, 보전적인 유전자(Conserved gene)들은 기능이 알려지지 못했던 다양한 생물학적 정보를 얻어 내는데 유용하게 쓰일 수 있다. 특히 완전히 서열이 밝혀진 지놈(Genome) 데이터로부터 진화적으로 보존적인 유전자 서열의 상동성에 따른 분류를 통한 2차 데이터베이스의 구축은 생물학자들에게 다차원적인 정보를 제공 할 수 있다. 이미 이러한 데이터베이스가 다양한 방법에 따라 구축 되었고 생물학자들의 연구에 활용되고 있다. 그러나 기 구축된 데이터베이스들은 생물학자들이 이용하기에 paralogs의 포함 문제점으로 인해 신뢰성이 떨어지거나 데이터베이스 생성 기간이 오래 걸린다는 단점이 있다. 본 연구는 기존에 구축된 데이터베이스들의 구축방법을 응용하고, 정보기술을 활용하여 빠르고 효과적으로 정확성을 높인 새로운 구축 방법론과 데이터베이스를 활용한 분석 시스템에 대해 제시하고자 한다.

1. 서 론

분자 생물학에서 필수적 단계는 유전체 분석이라고 할 수 있다. 유전체를 분석하기 위해서는 먼저 유전체 시퀀싱(Sequencing) 작업이 선행 되어야 하며 시퀀싱된 유전자를 통해 기능을 예측하는 주해(Annotation) 과정을 거치게 된다. 기술의 발달로 인해 과거에 비해 빠르게 많은 양의 시퀀싱된 데이터를 얻을 수 있게 되었고 정보기술을 활용하여 기능을 예측하고 분석하는 많은 도구와 방법이 현재 널리 쓰이고 있다. 유전자의 기능을 예측하는데 활용되는 가장 보편적인 방법은 블라스트(Blast)와 같은 프로그램을 이용해 기존의 서열 데이터베이스와 서열의 상동성 비교를 통한 기능 예측 방법이 주로 사용된다. 그러나 사용되는 데이터베이스의 신뢰성이 떨어 질 뿐만 아니라 그 데이터베이스를 이용하여 생물학자들이 다양한 분석을 하기엔 까다로운 것이 사실이다.

본 논문은 이러한 단점을 보완하기 위해 진화적으로 보전적인 유전자들의 서열의 상동성에 따른 분류를 통한 orthologous 데이터베이스를 구축하고, 구축된 데이터베이스를 바탕으로 유전자의 다차원적인 분석이 가능한 시스템에 대해 제안하고자 한다. 여기서 orthologous는 공통의 조상으로부터 종분화(speciation)되어 서로 다른 유전체에 있는 직접적으로 관련된 유전자들의 집합이라고 정의하며 이와 반대로 paralogs는 같은 유전체 내에서 복제(duplication)에 의해 생성되어진 관련된 유전자들의 집합으로 정의 할 수 있다. 일반적으로 같은 orthologous 관계에 있는 유전자들은 서열의 유사성과 함께 같은 기능을 갖게 되며, paralogs는 서열의 유사성을 갖고 있지만 진화적으로 기능이 완전히 틀리게 된다. 따라서 orthologous 관계를 분류한 2차 데이터베이스의 구축은 계통발생적(phylogenetic) 분석에 있어서, 서로 다른 종에서 나타나는 공통의 필수 유전자 파악 및 정확한 기능 예측과 다양한 분석을 위한 기본모델이라고 말할 수 있다.

본 연구는 다음과 같이 크게 3가지 목적을 가지고 있다. 첫째, 빠르고 신뢰성 있는 orthologous 데이터베이스를 구축한다. 둘째, 구축된 orthologous 데이터베이스를 활용하여 유전자들의 orthologous 관계 파악 및 필수 유전자와 종 특이적 유전자를 구별 할 수 있다. 셋째, 유전체 분석시 유전자의 신뢰성 있는 기능 예측이 가능하다. 이를 위해 본 논문에서는 컴퓨터 기술을 최대한 활용하여 기존의 구축된 orthologous 데이터베이스의 단점을 보완한 데이터베이스 구축을 위한 방법론과 orthologous 데이터베이스를 활용한 분석 시스템에 대해 설명하도록 하겠다.

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 향후연구방안에 대해서 알아본다.

2. 관련 연구

현재 orthologous 관련 데이터베이스는 미국 NCBI의 COG[4]와 일본 KEGG의 KO[5]를 들 수 있다. 각각의 데이터베이스는 재료와 방법에 있어 차이를 보인다. 여기서는 각각의 데이터베이스 및 방법론에 대해 소개한다.

본 연구는 학국과학재단 특정기초 연구사업(R01-2003-000-11723-0, R01-2001-000-00097-0)으로부터 지원을 받았음

2.1 COG(Clusters of Orthologous Groups of proteins)
COG는 현재 서열이 완전히 밝혀진 66개의 유전체의 유전자 단백질 서열의 일대일 상동성 비교를 통해 orthologous 관계를 파악하고 유사한 기능을 하는 도메인으로 나누어 그룹을 지었다. COG는 크게 다음과 같은 방법으로 구축되었다[1].
1. 모든 단백질 서열의 all-against-all 비교를 수행한다.
2. 명백히 드러나는 paralog를 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 정보축적 및 처리기능 관련 유전자군([J],[k],[L]), 세포생리 기능 관련 유전자군([D],[O],[M],[N],[P],[T]), 대사 기능 관련 유전자군([G],[C],[E],[F],[H],[I]), 기능 미확인 ([R],[S])군 총 74059개의 Orthologous 그룹으로 구성되어 있다. COG는 위의 모든 과정을 컴퓨터 프로그래밍 등의 정보기술을 활용하여 구축하였기 때문에 편리하게 데이터베이스를 구축하였다. 그러나 COG는 방법론상 paralogs를 완전히 제거하지 못하였다. 이로 인해 기능상 명확하지 않은 도메인 그룹이 생기게 되었고 신뢰성 있는 분석을 하는데 방해 요인이 되고 있다.

2.2 KO(KEGG Orthologous)
KO는 COG와는 달리 KEGG가 가지고 있는 pathway 데이터베이스로부터 수작업(manually)을 통해 구축되어 기능상 정확히 분류가 되었다[2]. 현재 계통발생학적 profile을 활용하여 매트릭스를 사용한 트리를 만들어 구축하는 방법에[3] 대해 연구되어 지고 있지만 기간이 오래 걸린다는 점과 새로운 종의 추가 시 일일이 추가해야 하는 단점이 있다. 또한 웹 기반의 HTML 형식의 테이블 기반에 분석도구도 단순해 생물학자들이 다양한 분석에 활용하는데 어려움이 있다.

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

우리는 위의 방법론의 단점을 보완하여 새로운 구축 방법론을 제시할 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 위해서는 먼저 orthologous 그룹을 형성하여야 한다. 그룹의 형성은 다음과 같은 단계는 거친다.
첫째, 시퀀싱 작업이 완전히 완료된 유전체들의 단백질 서열을 가지고 블라스트를 실행시킨다. 블라스트는 단백질 서열의 비교를 수행하여 유전자들끼리의 서열의 상동성을 분석해주는 프로그램 이다. 따라서 이를 통해 우리는 상호 best-hit 된 유전자 목록을 알 수 있다. 이때COG와 같이 다수의 종에 대한 all-against-all 비교가 아닌, 두 종간의 상호 비교만을 수행한다. 이는 COG에서와 같이 paralogs를 완전히 제거하지 못해 신뢰성이 떨어지는 문제점을 해결하기 위함이다.
둘째, 두 종간의 비교를 통해 나온 블라스트의 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 종간의 유전자의 overlap 정도에 관한 것이다. Score 값은 단백질 서열의 유사성이 어느 정도 일치하였나를 나타내는 통계적 값이다. 따라서 아무리 best-hit 되었더라도 score 값이 낮으면 신뢰성이 떨어진다. 우리는 score의 cut-off 값을 50bit로, overlap cut-off를 50%로 주었다. 이는 false positive와 false negative를 낮출 수 있는 최적의 cut-off 수치이다. 문제에 대해 최대한 신뢰성을 높이면서 데이터의 손실 없이 해결할 수 있기 때문이다[4].
셋째, 두 번째 단계를 반복하여 나온 여러 개의 두 종간 orthologous 그룹을 클러스터링 한다. 그리고 그림 1.에서 보듯 최소 3 종 이상에서 orthologous 그룹이 형성된 것만을 분류하여 실제 orthologous 그룹으로 인정한다.

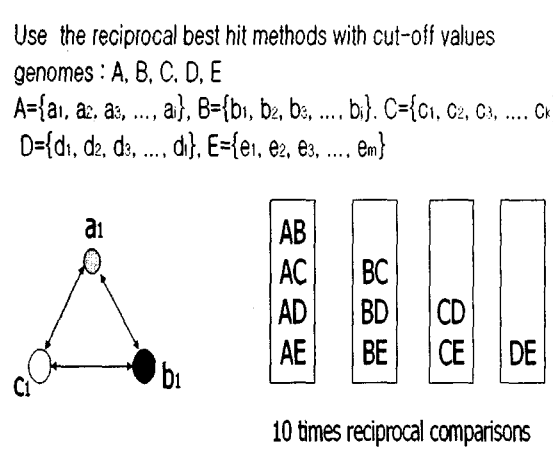

그림 1. Orthologous 그룹의 형성

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

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

3.2 데이터베이스 구축

최종적으로 나온 클러스터링 된 orthologous 그룹을 생물학자들이 다양한 분석을 하기 위해선 데이터베이스로 구축해야 한다. 데이터베이스에는 클러스터링 결과에 대한 정보뿐만 아니라 서열, 기능 및 기타 등등 다른 여러 가지 정보들을 포함하고 있어야 한다. 이를 위해 데이터베이스 구축을 위한 여러 프로그램의 개발이 요구된다. 또한 무엇보다 다차원적인 분석을 염두에 둔 데이터베이스의 스키마의 설계가 무엇보다 중요하다.

4. 분석 시스템

그림 2.는 우리가 개발할 전체 시스템의 개요이다.

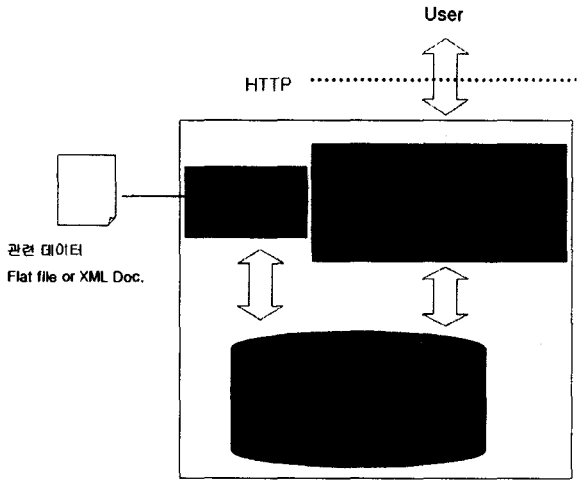

그림 2. 분석 시스템의 개요

구축된 데이터베이스만으로는 정보기술에 친숙하지 않은 일선의 생물학자들이 다양한 분석을 하기 힘들다. 따라서 생물학자들이 완성된 Orthologous 데이터베이스를 가지고 능동적으로 자신이 연구하는 유전자나 생물 종에 대해서 다양한 분석을 하기 위해서는 적합한 도구 들이 하나의 시스템 안에 포함되어야 한다. 기존의 COG의 경우 웹상에서 HTML로 데이터베이스에 있는 정형화된 분석 결과만을 보여주기 때문에 실제 생물학자들이 연구하기엔 한계가 있다. KO와 같은 경우도 정확하기는 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 들이다. orthologous clustering tool은 추가 적인 시퀀싱이 끝난 종에 대해 기존의 데이터베이스와 비교하여 새롭게 형성된 orthologous 그룹을 추가시켜주는 기능을 한다. 분석도구는 웹으로 접근하며 사용자들의 편의를 위해 GUI기반으로 구성한다. 분석도구에는 유전자 예측 컴퍼넌트, orthologous viewer 컴퍼넌트, 마이닝 컴퍼넌트를 포함하고 있다. 유전자 예측 컴퍼넌트는 사용자가 자신이 연구중인 유전자에 대해 기능을 예측하고자 할때 orthologous 데이터베이스와의 비교를 통해 기능을 예측해주고 자신의 유전자의 진화적 위치를 바로 확인 할 수 있다. Ortologous viewer 컴퍼넌트는 사용자가 보기엔 편하고 알기 쉽도록 ortologous 관계 및 기타 분석 결과를 비주얼하게 보여준다. 마이닝 컴퍼넌트는 데이터 마이닝 기법을 적용하여 예측하지 못하는 생물학적 관련성을 분석할 수 있도록 도와준다. 이와 같은 분석도구 들이 완벽히 기능을 수행하기 위해서는 앞에서 지적 했듯 데이터베이스의 구축작업이 매우 중요하다.

5. 결론 및 향후연구

과거에 비해 급격히 생성되는 염기서열로부터 최대한 다양한 정보를 추출하기 위해서 계통분류학적으로 보존적인 유전자들의 서열의 상동관계를 분석해야 한다. 본 논문은 이를 위해 기존에 구축되어진 orthologous 데이터베이스의 단점을 보완한 구축방법론을 제시하고 구축된 데이터베이스를 바탕으로 한 분석시스템에 대해서 제시 하였다. 현재 데이터베이스 구축작업에 있으며 앞으로 genome project가 완료된 170종 이상의 prokaryote에 대해 데이터베이스를 구축 할 것이다. 서열비교 작업의 특성상 시간이 굉장히 많이 걸리는 작업이기 때문에 우선 몇 종에 대한 데이터베이스 구축을 하고 차차 추가하는 방법을 쓸 것이다. 동시에 분석시스템을 개발하고 최종적으로 pathway 데이터베이스 시스템과의 통합을 통해 다양한 생명현상에 대한 연구에 많은 기여를 할 것이다.

[참고문헌]

[1] Roman L. Tatusov, Michael Y. Galperin, Darren A. Natale and Eugene V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution" , *Nucleic Acids Res.* 28, 33-36, 1999

[2] Yamanishi, Y., Akiyasu C. Yoshizawa, Itoh, M., Katayama, T.,'Kanehisa, M., "Extraction of Organism Groups from Whole Genome Comparisons" , *Genome Informatics* 14, 438-439, 2003

[3] Bono, H., Goto, S., Fujibuchi, W., Ogata, H., Kanehisa, M., "Systematic Prediction of Orthologous Units of Genes in the Complete Genomes" , *Genome Informatics* 9, 32-40, 1998

[4] Maido Remm, Chrisian E.V. Storm, Erik L.L. Sonnhammer, " Automatic Clustering of Orthologs and in-paralogs from Pairwise Species Comparisons" , *J. Mol. Biol.* 314, 1041-1052, 2001

[5] NCBI(National center for Biotechnology Information) available : http://www.ncbi.nlm.nih.gov/COG

[6] KEGG(Kyoto Encyclopedia Genes and Genomes) available : http://www.genome.ad.jp/kegg/kegg2.html