Ortholog 데이터베이스를 이용한 생물 경로 재구축 시스템

정태성 [1], 오정수 [2], 조완섭 [2]
{정보산업공학과 [1], 경영정보학과 [2]} 충북대학교
{mispro, ofang, wscho}@cbnu.ac.kr

Pathway Reconstruction System using Orthlogs Database

Taesung Jung[1], Jeongsu Oh[2], Wansup Cho[2]
Information Industrial Engineering[1], Management Information Systems[2]
Chungbuk National University

요 약

현재 국내외 적으로 많은 대사경로 재구축을 위한 소프트웨어들이 개발 보급되고 있다. 그러나 기존의 소프트웨어들은 유전자 서열의 주해 작업이 끝난 게놈에 대해서만 가능하다. 따라서 대사경로를 예측하고자 할 경우는 주해 작업이 선행되어야 하는 어려움이 있었다. 본 논문에서는 주해 작업이 완료되지 않은 유전자 서열로부터 유전자의 기능 예측뿐만 아니라 대사경로를 예측할 수 있는 시스템을 제안한다. 제안된 시스템은 Orthologous 데이터베이스를 활용하여 새롭게 밝혀진 유전자 서열을 대상으로 비교적 정확성이 높은 대사경로를 예측하는 기능을 제공한다. 이 방법을 통해 주해 작업이 완료되지 않은 유전자 서열을 이용하여 서열 내에 포함된 유전자의 기능을 예측할 뿐만 아니라 예측된 유전자 정보를 이용하여 대사 경로를 예측할 수 있다.

1. 서 론

유전자의 생물학적 기능을 밝히고 세포 내 상호작용을 이해하는 것은 post-genome era 의 가장 중요한 작업 중 하나이다. 이런 목적을 위한 고전적인 방법은 먼저 어떤 형질의 원인이 되는 유전자를 발견하고 그 유전자의 구조를 밝히는 것이었다. 그러나 염기서열 해독의 자동화, 고속화, 대용량화가 급진전되고 컴퓨터를 사용한 정보처리 기술이 발전함에 따라 먼저 유전체의 서열분석을 통해 기능화 되고 있다[1]. 생물학자들은 새롭게 밝혀진 유전자의 단편적인 기능적, 진화적 분석뿐만 아니라 생명체 안에서 유기적으로 이루어지는 다양한 생명 현상에 대해 알고 싶어 한다.

특히 관심이 있거나 연구하고 있는 게놈이나 유전자에 대해 대사경로(Metabolic Pathway) 상에서 경로 예측 또는 유전자 예측 등의 분석을 수행하곤 한다. 대사경로 재구축 시스템은은 이러한 연구를 수행하는데 있어 매우 유용하게 쓰일 수 있다. 현재 국내외 적으로 많은 대사경로 재구축 소프트웨어들이 개발 보급되고 있다[2][3][4][5][6][7][8]. 그러나 기존의 소프트웨어는 유전자 서열의 주해(Annotation) 작업이 완료되어 효소 번호(EC-Number)나 유전자 이름(Gene Name)이 밝혀진 것을 대상으로 대사경로를 재구축하는 것이 대부분이다. 또한 서열을 가지고 있다고 하더라도 알려지지 않은 유전자(Unknown Gene)가 나오는 것이 빈번하여 정확성이 떨어지는 문제점이 있었다. 따라서 기존의 시스템들은 주해 작업이 완료되지 않은 유전자 서열에 대해서는 경로 재구축을 수행할 수 없다.

지구상에 존재하는 생물들이 생물학적 진화기간 동안 공통 조상의 유전자(Ancestral Gene)의 종분화와 복사에 의하여 각 생물의 유전체에 분포하게 되었고, 생물체의 공통 생명현상과 각 분류단계에 특이한 생명현상을 띄게 되었다. 이때 공통조상의 한 유전자로부터 종분화 되어 서로 다른 종에 속하는 유전자들 Orthologs 라 한다. 같은 Orthologs 내에 구성원들은 서열의 유사성과 같은 기능을 갖게 된다. 따라서 계통발생학적 계보 (Phylogenic Lineage)에 있어 Orthologs 는 서열이 불완전하게 밝혀진 게놈에서 많은 기능 부위를 자동으로 예견할 수 있는 기본이 되며 생물체들의 다양한 생명현상과 공통적으로 나타내는 필수기능(housekeeping function)을 연구, 분석하는데 유용하다. 이러한 Orthologs 정보를 이용하면 주해 작업이 완료되지 않은 서열 정보에서 유전자의 기능을 예측할 수 있으며, 이렇게 예측된 유전자 정보를 이용하여 경로 재구축을 수행할 수 있다.

본 논문에서는 이러한 Orthologs 데이터베이스를 구축하고 이를 이용하여 주해(Annotation) 작업이 완료되지 않은 서열에 대해서 경로 재구축을 하는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 관련연구에 대해서 알아보고, 3장에서 제안된 시스템 구조와 경로 재구축 과정에 대해서 설명한다. 4장에서는 결론 및 향후 연구로 맺는다.

2. 관련 연구

2.1 경로 재구축(Pathway Reconstruction)

일반적으로 기존의 경로 구축은 다음과 과정을 거치게 된다. 첫번째, 유전자 정보를 입력한다. 입력 데이터는 유전자 정보를 포함하고 있는 게놈 서열 또는 효소 번호 또는 유전자 목록 등을 이용한다. 두번째, 기존의 경로 데이터베이스에서 입력된 정보와 일치하는 경로 데이터를 검색한다. 세번째, 검색된 경로 데이터에 대해서 필터링(관련성이 높은 경로 데이터만 추출)을 하여 경로 재구축을 완료한다. 즉, 기존의 시스템들은 주해작업이 완료된 서열에 대해서만 경로 재구축이 가능하다.

2.2 경로 재구축 시스템

KEGG 는 대표적인 대사경로 데이터베이스이다. KEGG 를 이용하면 다양하고 유연성이 있는 경로 재구축을 수행할 수 있다. 그러나, 주해 작업이 완료되지 않은 게놈 서열에 대해서는 경로 재구축 수행이 불가능하다. EcoCyc 는 E-Coli 에 대해서 잘 구축된 대사경로 데이터베이스이다. Ecocyc 에서도 서열 정보를 이용하여 경로 재구축을 지원해주고 있으나, 이 또한 주해 작업이 완료된 서열만을 대상으로 하고 있다.

2.3 Orthologs 데이터베이스

COG[10]는 현재 서열이 완전히 밝혀진 66 개의 유전체의 유전자 단백질 서열의 일대일 상동성 비교를 통해 orthologous 관계를 파악하고 유사한 기능을 하는 도메인으로 나누어 그룹을 지었다.

KO[9]는 대사경로 (metabolic pathways)와 조절경로 (regulatory pathways) 에서 추출한 orthologs genes 로 구성된 데이터베이스이다. KO 는 기능이 확실히 밝혀진 pathways 로부터 수작업으로 작성되었기 때문에 기능별로 거의 정확히 분류되었다.

이러한 데이터베이스들을 이용하면 주해작업이 완료되지 않은 새로운 서열에 대해서 유전자의 기능을 예측할 수 있으며 생명현상을 이해하는데 큰 도움이 될 수 있다.

3. 시스템 구성

3.1 orthologs 데이터베이스 구축

제안된 시스템은 크게 3 단계를 거쳐 Orthologs Group 을 클러스터링 한다. 먼저, paralogs 를 줄이기 위해 클러스터링 하고자 하는 genomes 중 각각의 두 genome 에 관해 BLAST 를 사용 reciprocal best hits 을 찾는다. 이때 score cut-off 와 overlap cut-off 를 적용하여 정확성을 높인다. 이 작업은 많은 시간이 걸리는 작업으로 우리는 그리드 컴퓨팅을 통해 시간을 단축한다.

두 번째로, reciprocal best hits의 결과 가운데 임의의 3 종을 뽑아 clustering을 한다. 이는 3종 이상에서 Orthologs 한 것만을 하나의 group으로 묶기 위함이다. 마지막으로, 3종간 group에 나머지 종들을 더해 클러스터링을 하여 Orthologous 그룹을 구축한다. 우리는 효과적인 클러스터링을 위해 데이터베이스를 활용한 클러스터링 알고리즘을 개발하였다. 이를 통해 모든 과정을 수작업 없이 자동적으로 Ortologous group을 클러스터링 한다. 그림 1은 Orthologs 데이터베이스 구축 과정을 보여준다.

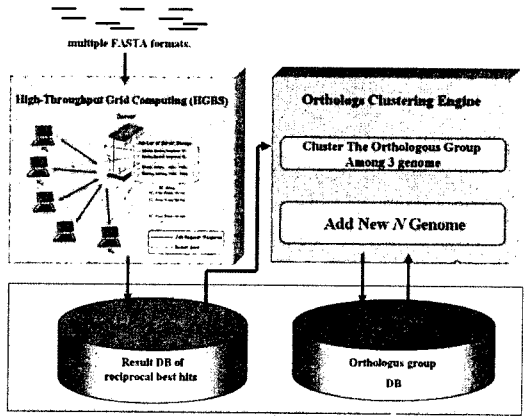

그림 1 Orthologs 데이터베이스 구축 시스템 구조

그림 1 에서 보듯이 제안된 시스템은 그리드 시스템을 기반으로 자동화된 엔진을 사용하여 Orthologs 데이터베이스를 구축하였다.

3.2 Orthologs 데이터베이스를 이용한 경로 재구축

제안된 시스템은 기존의 시스템들과 마찬가지로 주해 작업이 완료된 서열 정보 또는 유전자 정보 등을 이용하여 경로 재구축을 지원한다.

또한, 주해 작업이 완료되지 않은 서열에 대해서는 Orthologs 데이터베이스를 이용하여 경로 재구축을 지원한다. Orthologs 데이터베이스를 이용한 경로 재구축 과정은 그림 2 와 같다.

먼저 주해 작업이 안된 서열 정보를 입력한다. 두번째로 비교하고자 하는 종(Specie) 목록을 선택한다. 상동성 비교 분석은 상당한 처리 시간을 요하는 작업이다. 모든 종에 대해서 상동성 비교를 수행할 경우 결과를 기대하기는 힘들다. 따라서 빠른 결과를 도출하기 위해서 비교할 종을 한정한다. 세번째로는 선택한 종의 유전자 그룹과 BLAST 를 이용하여 상동성 비교를 수행한다. 네번째로는 상동성 비교 결과 중에 유사성이 높은 유전자 그룹을 선택한다. 마지막으로, 선택된 유전자 그룹의 정보를 이용하여 기존의 경로 재구축 방식을 수행한다.

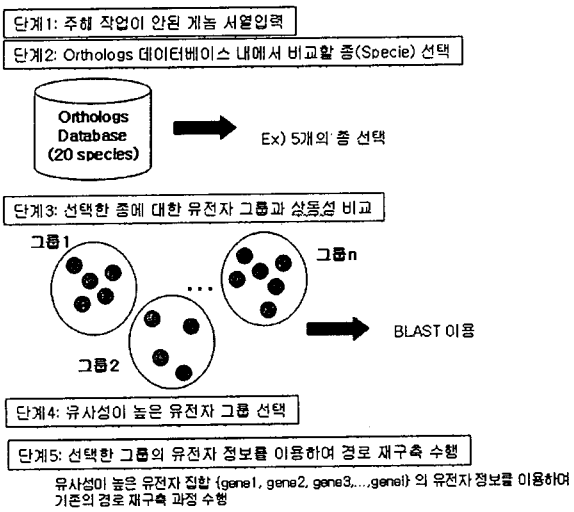

그림 2 경로 재구축 과정

3.3 평가

본 논문에서는 경로 재구축 평가를 다음과 같이 수행한다. 입력한 서열 정보를 주해 작업이 완료된 서열을 입력한다. Orthologs 데이터베이스를 이용하여 경로 재구축을 수행한 결과와 기존의 방식대로 경로 재구축을 수행한 결과를 비교한다.

4. 결론

경로 재구축은 생물학자들에게 유전자의 기능 예측 및 생물 현상을 분석하는 매우 유용한 방식이다. 그러나 기존의 시스템들은 주해 작업이 완료된 서열의 효소 정보나 유전자 정보를 이용하여 경로 재구축 기능을 제공하고 있다.

본 논문에서는 기존의 경로 재구축 시스템과 더불어 주해 작업이 완료되지 않은 게놈 서열 정보를 이용하여 유전자 정보를 예측하고 경로를 재구축할 수 있는 시스템을 제안하였다. 제안된 시스템을 이용하면 주해 작업이 완료 되지 않은 게놈 서열을 이용하여 유전자 기능을 예측할 수 있을 뿐만 아니라 경로 재구축도 수행할 수 있다.

향후 연구로 제안된 시스템을 완성하고, 이를 이용하여 주해 작업이 완료 되지 않은 게놈 서열의 유전자 기능을 예측하고 예측된 결과의 정확도와 경로 재구축의 신뢰성 정도를 평가하고자 한다.

[참고문헌]

[1] Andrew Finney, Micael Hucka Systems Biology Markup Language(SBML) Level 2 : Structrues and Facilities for Model Definitions 2003. 6

[2] Goesmann,A., Haubrock,M., Meyer,F., Kalinowski,J. and Giegerich,R. " PathFinder: reconstruction and dynamic visualization of metabolic pathways" , Bioinformatics, 18, pp.124-129, 2002

[3]Kanehisa,M. and Goto,S. " KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res., 28, pp. 27-30, 2000

[4] Karp,P.D., Paley,S. and Romero,P. " The Pathway Tools software," Bioinformatics, 18: pp. S225-S232, 2002

[5] Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. " EcoCyc: electronic encyclopedia of E. coli genes and metabolism." Nucleic Acids, Res., 27, pp. 55-58, 2002

[6] Krishnamurthy,L., Nadeau,J., Ozsoyoglu,G., Ozsoyoglu,M., Schaeffer,G., Tasan,M. and Xu,W. " Pathways database system: an integrated system for biological pathways," Bioinformatics, 19, pp. 930-937, 2003

[7] Kuffner,R.M., Gonzales,M., Steadman,P., Woldek,D.K., Jankowitz,R.J., Boinoff,J.R., Montoya,L., Peterson,T.F., Bulmore,D.L. and Blanchad,J.B. " PathDB: in The Molecular Biology Database Collection: 2004 update," Nucleic Acids Res., 32(Database issue), D3-D22, 2004

[8] Overbeek,R., Larsen,N., Pusch,G.D., D' Souza,M., Selkov,Jr.,E., Kyrpides,N., Fonstein, M., Maltsev,N. and Selkov,E. " WIT: integrated system for highthroughput genome sequence analysis and metabolic reconstruction," Nucleic Acids Res., 28, pp. 123-125, 2000

[9] COGs official homepage. http://www.ncbi.nlm.nih.gov/COG/

[10] KO official homepage. http://www.genome.jp/kegg/ko.html