클러스터링 기법을 통한 대사 네트웍의 진화적 분류

오석준[0] 정제균 장병탁
서울대학교 서울대학교 서울대학교
바이오정보기술 연구센터 생물정보학 협동과정 컴퓨터공학부
sjaugh@cbit.snu.ac.kr[0], {jgjoung, btzhang}@bi.snu.ac.kr

Evolutionary Classification of Metabolic Networks by Hierarchical Clustering

Sirk June Augh[0] Je-Gun Joung Byoung-Tak Zhang
Center for Bioinformation Technology Interdisciplinary Program in Bioinformatics School of Computer Science and Engineering
Seoul National University

요 약

현재 유전자 서열 분석이 완료된 유전체들이 점점 늘어나고 있다. 따라서 이에 대한 방대한 정보가 생성됨에 따라 다양한 생물체들에 대하여 대사 네트웍을 통한 다차원적 분석이 가능하게 되었다. 대사 네트웍은 단백질 또는 효소들의 전체적인 상호작용을 표현하기 때문에 생물학적 메카니즘에 대하여 보다 풍부한 정보를 제공해 준다. 본 논문에서는 일차원적인 유전자 서열에 의한 종의 계통 분류가 아니라 메타 수준의 생리 구조적 비교를 통하여 계통분류학에 대하여 새로운 방법의 접근을 제안하고자 한다. 제안된 방법은 기존의 상동성 비교에 의한 계통 분류와 함께 좀 더 포괄적이고 거시적인 분석을 가능하게 한다.

1. 서 론

최근 분자생물학의 눈부신 발전에 의하여 유전자 서열 분석이 가속화됨에 따라 유전 정보량은 기하급수적으로 증가하고 있다. 더불어 이러한 유전 정보에 대하여 기능을 규명하고 생물학적인 메카니즘을 밝혀 내기 위한 노력이 진행되고 있다. 또한 이들 유전 정보들은 생물학 연구의 방법론에 있어서 획기적인 전환점을 제공하였다.

일반적으로 생물학자들은 계통분류를 위한 유전자 서열 분석에 종간의 거리를 측정하기 위한 수단으로 BLOSUM, PAM[1, 2]과 같은 유사 매트릭스(similarity matrices)를 사용한다. BLOSUM, PAM을 이용한 계통분류 방법은 유전자 시퀀스를 근거로 진화트리를 구성하기 때문에 방법적인 측면에서 제한적이라고 할 수 있다. 대사 네트웍의 활용은 이러한 제한에서 탈피할 수 있는 한가지 방법일 것이다.

지금까지 대사 네트웍에 대하여 정보학 관점의 접근은 대부분 시각화와 데이터베이스 저장 및 검색 문제에 편중되어 있었다. Forst[3]는 대사 네트웍을 계통분류에 활용하는 방법론을 제시하였다. 각 네트웍 요소들간의 거리와 경로(path)에 있어서 갭(gap)를 측정함으로써 계통분류를 하였다.

본 논문에서는 대사 네트웍들을 클러스터링함으로써 계통분류를 하는 방법을 제시한다. 대사 네트웍들의 클러스터링을 위하여 사용된 알고리즘은 계층적 클러스터링 알고리즘이다[4]. 계층적 클러스터링은 생물정보학 분야에 있어서 과거에는 계통분류학에 사용된 바 있었고 최근에는 마이크로어레이 분석에 있어서 핵심적인 알고리즘으로 사용되고 있다[5]. 계층적 클러스터링은 간단한 알고리즘이면서 비교적 좋은 성능을 보이고 있으며 결과가 트리 형태로 표현되기 때문에 계통분류학 문제에 자연스럽게 응용될 수 있는 장점을 가지고 있다. 클러스터링에 있어서 클러스터간의 유사한 정도를 측정하게 되는데 이러한 근사도(approximity)의 측정 방법으로 상관계수(correlation coefficient)를 이용하였다.

기존 계통분류와의 가장 큰 차이점은 유전자 서열만을 가지고 상동성을 분석했을 경우에는 생물학적 시스템의 전체 문맥을 보지 못하고 단백질 각각을 독립적으로 분석하여 비교하는 효과를 가지지만 대사체(metabolome) 정보를 이용하여 단백질의 상호작용의 구조를 고려하게 된다면 반대로 시스템 전체를 고려하여 분석하는 결과를 얻게 된다.

2장에서 분석을 위한 준비 과정으로 대사 네트웍의 구축에 대해서 살펴본다. 3장에서는 계층적 클러스터링 기법을 통하여 종들간의 대사 네트웍들을 비교함으로써 계통분류를 하는 방법과 이 방법을 통한 분석 결과를 기술한다. 마지막으로 4장에서 결론을 제시하고 향후 계획에 대해서 언급한다.

2. 대사 네트웍의 구축

대사 네트웍은 효소(enzyme)들에 의해 촉매화되는 대다수의 화학적 반응을 구성하고 있다. 대사 네트웍을 연구하기 위해 가장 많이 활용되고 있는 데이터베이스는 KEGG(Kyoto Encyclopedia of Genes and Genomes)이다[6]. KEGG는 2001년 12월 현재 약 4000 대사 반응(metabolic reaction)들을 포함하고 있다.

표 1은 해당작용(glycolysis)에 대하여 대사 경로가 알려진 종들을 나열한 것이다. 여기서 도메인이 같다면

진화적 거리가 가깝다고 말할 수 있지만 절대적인 것은 아니다. 예를 들어, archaea에만 유일하게 존재하는 특정 경로가 없을 수도 있다. 즉 서로 경로들을 공유하는 경우가 많다고 해석할 수 있다.

표1 분석을 위한 주요 종 목록

종	도메인
Aquifex aeolicus	bacteria
Bacullus subtilis	bacteria
Deinococcus radiodurans	bacteria
Escherichia coli	bacteria
Yersinia pestis	bacteria
Haemophilus influenzae	bacteria
Rhodobacter capsulatus	bacteria
Helicobacter pylori	bacteria
Clostridium acetoburylicum	bacteria
Mycobacterium tuberculosis	bacteria
Mycoplasma genitalium	bacteria
Mycoplasma pneumoniae	bacteria
Streptococcus pneumoniae	bacteria
Streptococcus pyogenes	bacteria
Arabidopsis thaliana	eukaryota
Saccharomyces cerevisiae	eukaryota
Schizosaccharomyces pombe	eukaryota
Caenorhabditis elegans	eukaryota
Mus musculus	eukaryota
Drosophila melanogaster	eukaryota

그림 1은 해당작용에 대하여 지금까지 밝혀진 종들의 대사 네트웍들을 통합하여 유추한 결과이며 KEGG 데이터 베이스에 저장되어 있다. 사실 각각의 종에 대하여 KEGG 데이터 베이스에서 보여지는 해당작용에 대한 정보는 이미지화된 정적인 것으로서 컴퓨팅에 바로 사용하기 어렵다. 따라서 반응에 대한 EC번호를 추출하여 대사 네트웍을 자동으로 재구성할 필요가 있다. 만약 KEGG에 특정 종에 대한 네트웍이 없다고 할지라도 시퀀싱이 되어 있다면 그 종에 대한 ORF(open reading frame)들을 분석하여 네트웍을 구성할 수도 있다.

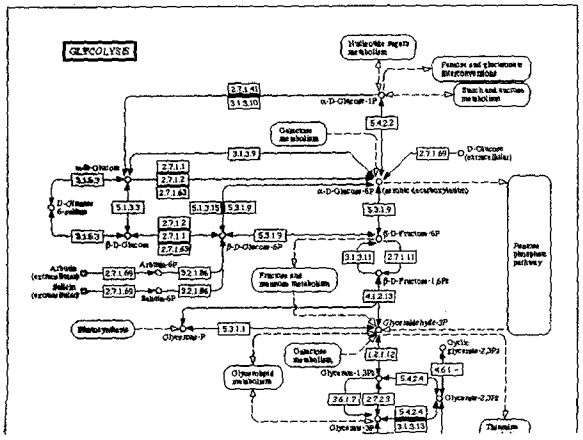

그림 1 해당작용(glycolysis)에 대한 참고 대사 네트웍

네트웍을 재구성해야 하는 궁극적인 목적은 효소에 의한 각 반응 간의 거리를 계산하기 위한 것이다. 생체 내의 대사 네트웍을 고려해 볼 때 특정 신호가 주어진다면 유전자들이 발현하게 되고 특정 반응에 관여하게 된다. 따라서 각각의 유전자들은 특정 효소들과 매핑이 될 수 있다. 각 반응에 참여하는 유전자 간의 거리는 어떤 진화적인 절차를 거쳤는지에 대한 중요한 실마리를 제공해 줄 수 있다. 이러한 정보는 각 종들간의 차이를 판별해 주는 중요한 척도로 작용할 수 있다.

3. 대사 네트웍의 진화적 분류

3.1 데이터의 표현

우선 각각의 대사 네트웍을 노드 집합 V와 링크 집합 E를 가진 무방향성의 그래프 $G = (V, E)$로 간략화 하자. 여기서 노드 집합 V는 효소의 집합을 나타내고 링크 집합 E는 노드들이 연결되는지의 여부를 나타낸다. 정확하게는 대사 네트웍은 방향성을 가져야 하지만 문제의 간략화를 위하여 생략하기로 한다.

각각의 대사 네트웍은 클러스터링을 위한 입력데이터 $D_i = \{d_1, d_2, \ldots, d_n\}$으로 전처리과정을 거치게 된다. 입력데이터 d_i는 노드 집합 V 내의 모든 노드의 쌍 v_j와 v_k에 대하여 거리 $\Delta v_{jk} = |v_j - v_k|$으로 계산된다. 각 링크의 가중치 값은 균일하게 1로 부여했다. 따라서 만약 노드의 쌍 v_j와 v_k가 네트웍 상에 있어서 서로 직접 연결되어 있으면 거리 Δv_{jk}는 1의 값을 가지게 된다. 사실 각 링크의 가중치를 차등화해야 좀더 정확한 클러스터링이 될 수 있다. 그러기 위해서는 좀 더 많은 생물학적 고려와 검증이 필요하게 된다.

3.2 계층적 클러스터링 알고리즘

클러스터링 알고리즘은 객체들(objects)의 쌍들 사이의 근사도 측정을 기반으로 하여 객체들 또는 항목들(items)을 그룹화한다. 여기서는 모든 대사 네트웍들이 객체들에 해당된다. 계층적 클러스터링 알고리즘은 이러한 근사도 정보를 활용하여 트리(tree)를 형성한다. 본 논문에서는 두 객체들의 근사도를 간단히 상관 계수로 측정하였다.

표 2 계층적 클러스터링 알고리즘

Agglomerative Hierarchical Algorithm:

Given:
A set O of objects $\{o_1,\ldots,o_n\}$
A distance function $DF(c1,c2)$
1. **for** $i = 1$ to n
$c_i = \{x_i\}$
end for
2. $C = \{c_1,\ldots,c_b\}$
3. $l = n+1$
4. **while** C.size > 1 **do**
a) (c_{min1}, c_{min2}) = minimum $DF(c_i,c_j)$ for all c_i,c_j in C
b) Remove c_{min1} and c_{min2} from C
c) Add $\{c_{min1},c_{min2}\}$ to C
d) $l = l + 1$
end while

표 2는 기본적인 계층적 클러스터링 알고리즘을 기술하고 있다. 거리함수(distance function)는 단일연결(single link)과 그룹-평균(group-average)을 포함한 여러 가지 방법에 의해서 클러스터 간의 유사도를 계산할 수 있다. 단일연결 방법은 두 클러스터들 내의 어느 두 객체들 사이에서 가장 짧은 거리를 가지고 두 클러스터의 거리를 계산한다. 그룹-평균 방법은 클러스터 내의 모든 객체들에 대하여 평균을 계산하여 두 클러스터 간의 거리를 계산한다.

3.3 실험 및 분석 결과

대사 네트웍들 내의 모든 노드들에 대해서 노드쌍들을 설정하고 각각의 네트웍 내의 노드쌍들의 거리값를 계산할 때 서로 연결이 되지 않는 노드쌍에 대해서는 최대 거리값을 부여하였다.

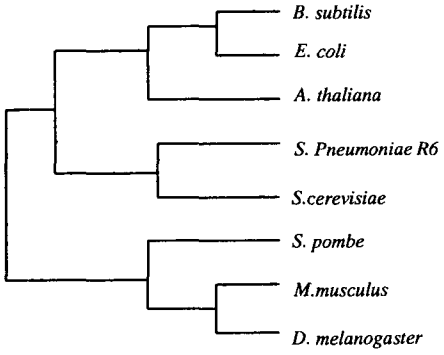

그림 2 해당작용 대사 네트웍들의 클러스터링 결과

그림 2는 대사 네트웍의 클러스터링에 의해 진화트리 구성에 대한 결과를 보여주고 있다. 대사 네트웍에 의해서 얻어진 진화트리의 가지들은 기존의 진화트리 구성방법에서 보여주고 있는 트리 가지들이 내포하는 진화적 거리와는 다를 수 있다. 기존의 시퀀스 기반의 계통 분류는 종들 간의 특정 유전자 시퀀스가 다른 정도를 기반으로 한다. 이 방법에서 간과하고 있는 점은 특정 유전자가 여러 가지 기능을 하는 경우들을 놓칠 수 있다는 것이다. 이러한 경우는 생물학적인 대사흐름의 문맥을 파악하지 못한다면 발견하기 어려운 부분이다.

4. 결론 및 향후계획

지금까지 대사 네트웍을 이용하여 계통분류를 하는 새로운 접근 방법에 대해서 살펴보았다. 계통분류를 위하여 본 논문에는 대사 네트웍들을 클러스터링하는 방법을 선택하였다. 클러스터링 방법으로는 계층적 클러스터링 알고리즘 사용했으며 이 알고리즘을 사용함으로써 자연스럽게 계통 분류 트리를 형성하였다. 클러스터링을 위한 근사도는 상관계수를 이용하였다. 이러한 분석을 위하여 KEGG 데이터로부터 네트웍을 재구성하고 효소간의 거리를 측정하는 전처리 과정이 요구되었다. 해당작용 대사 네트웍에 대한 분석 결과는 생물학적으로 상당히 의미있는 결과를 도출하였다. 앞으로 이러한 방법은 기존의 방법에 대한 가이드 역할을 할 것으로 기대하며 지속적인 개선과 확장이 요구된다.

향후 다양한 클러스터링 기법을 적용하여 비교 분석함에 따라 성능을 개선할 계획이다. 현재는 입력 데이터가 단지 연결 정보만을 사용했지만 보다 세부적인 분석을 위해서 부가적인 정보를 추가하는 것도 성능 개선의 한가지 방법일 것이다. 또한 한가지 문제가 되는 것은 만약 종에 대하여 전체 대사 네트웍을 클러스터링하기 위해서는 대용량의 메모리가 요구되며 계산의 복잡성 또한 증가할 것이다. 따라서 이를 해결하는 새로운 클러스터링 방법이 고안되어야 할 것이다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업과 IMT-2000 과제에 의하여 지원되었음.

참고문헌

[1] Dayhoff, M., Schwartz, R., and Orcutt, B., A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Vol. 5, pp. 345-352, 1978.

[2] Henikoff, S. and Henikoff, J., Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci*, USA 89, 10915-10919, 1992.

[3] Christian V. Forst and Klaus Schulten, Evolution of metabolisms: a new method for the comparison of metabolic pathways, *Journal of Computational Biology*, Vol. 6, No. 3/4, pp. 343-360, 1999.

[4] Johnson, S. C., Hierarchical clustering schemes, *Psychometrika*, Vol. 2, pp. 241-254, 1967.

[5] Eisen, M. B, Spellman, P. T., Brown, P. O. and Botstein, D., Cluster analysis and display of genome-wide expression patterns., *Proc. Natl Acad. Sci.* USA, Vol. 95, pp. 14864-14868, 1998.

[6] Ogata, H., Goto, S., K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, Vol. 27, pp. 29-34, 1999.