분산된 대사 네트워크에 대한 경로탐색을 위한 분산 알고리즘

Distributed Algorithm to search paths in distributed metabolic pathway networks

이선아, 이건명
충북대학교 전자전기컴퓨터공학부, 첨단정보기술연구센터(AITrc)

Lee Sun-a, Lee Keon-Myoung Lee
School of Electric and Computer Engineering, Chungbuk National University and AITrc
E-mail : salee719@aicore.chungbuk.ac.kr,kmlee@aicore.chungbuk.ac.kr

요약

이 논문에서는 분산된 생물학의 대사 네트워크들이 있을 때, 이를 통합하지 않은 상태에서 경로검색을 하는 분산 알고리즘을 제안한다. 대사 네트워크는 여러 데이터베이스에 존재하며 서로 중복되는 데이터를 가지고 있다. 제안한 방법은 네트워크 사이의 중첩이 있는 부분을 하이퍼 노드로 하고, 네트워크 자체는 하이퍼 에지로 하는 추상 하이퍼 그래프를 만들어서, 이를 이용한 상위수준의 경로를 구축한다. 각 네트워크내의 중첩된 영역 간의 경로를 미리 계산해 둔 다음, 상위수준의 경로에 기반하여 분산된 대사네트워크 간에 존재하는 경로를 검색한다. 추상 하이퍼 그래프는 데이터베이스를 하이퍼 노드로 하는 것에 대한 경로탐색을 한 다음, 그 경로에 따라 데이터베이스 내에 존재하는 대사경로를 탐색한다. 이때 존재하는 대사경로가 많기 때문에 각각의 대사경로를 하이퍼 노드로 하는 추상 하이퍼 그래프를 만들어 경로를 탐색하고 나서 그 하위 노드에 대해 경로탐색을 한다. 이는 분산된 네트워크를 통합할 저장 공간 및 탐색시간을 줄일 수 있다는 장점이 있다.

1. 서론

컴퓨터의 보급과 성능 향상으로 인하여 생물학 데이터가 기하급수적으로 증가하여 그 양이 방대해졌다. 유전자 데이터베이스, 단백질 데이터베이스, 대사경로 데이터베이스 등으로 구분되어 각 연구 분야에 활발히 사용되고 있다. 이들은 분산되어 있으며 일부 데이터가 다른 데이터베이스와 중복되기도 한다. 대사경로 데이터베이스에 대해 보면, KEGG[1], EcoCyc[2], EMP[3] 등에서 대사경로 데이터베이스를 제공한다. KEGG는 전반적인 대사 네트워크 정보를 가지고 있으며 EcoCyc은 ecoli에 대한 대사 경로만을 전문적으로 다룬다. 사용자가 특정 comound가 대사경로를 통해 현재 알려진 다른 대사 네트워크에 있는 compound로 변화할 수 있는지, 또한 어떤 경로를 통해 변화하는지를 알고 싶을 때 이를 확인하기 위해서 KEGG등에서 제공하는 정보를 이용해 경로가 존재하는지를 확인하거나 다른 데이터베이스의 대사 네트워크를 확인하여야 한다. 분산된 네트워크 임의의 한 데이터베이스에서만 검색할 경우에 나온 결과와 다른 사이트에서 제공하는 데이터베이스에서 검색한 결과가 서로 다를 수도 있다는 것이다. 하나의 데이터베이스 안에서 어떤 변화를 보기 위해서는 웹서비스에서 제공하는 이미지맵을 추적하거나 자신의 컴퓨터에 해당 데이터베이스를 미러링하여 검색할 수 있는 프로그램을 이용하여 찾아야 한다. 잘 알려진 경로상의 내용이거나 이미 검색하고자 하는 compound들이 어느 대사 네트워크와 관련이 되어 있는지를 알고 있는 경우라면 경로탐색이 간단할 수도 있다. 하지만 잘 알려진 compound라 해도 그와 관련된 대사경로가

본 연구는 첨단정보기술연구센터(AITrc)를 통해서 과학재단 지원으로 수행된 것임.

복잡하거나 잘 모르는 대상 일수도 있다. 또한 검색하고자 하는 compound들에 대해 몇 가지 데이터베이스에서 검색하고자 할 때, 각 사이트에서 제공하는 이미지맵을 추적하는 데에 걸리는 시간이 오래 걸린다. 마지막으로 데이터베이스를 미러링한 경우이다. 검색대상이 되는 데이터베이스들을 모두 미러링하여 자신의 컴퓨터에 저장하였을 경우, 저장 공간적인 문제와 방대한 데이터에 대한 탐색시간 문제가 발생한다. 이 논문에서는 여러 데이터베이스들을 통합하지 않고 임의의 compound가 어떤 작용을 통해 다른 compound로 변화하는지를 확인할 수 있는 알고리즘을 제안하고자 한다. 이 논문은 2장에서 제안하는 알고리즘을 설명하고 3장에서 결론 및 앞으로의 작업에 대해 말하고자 한다.

2. DBGET/LinkDB

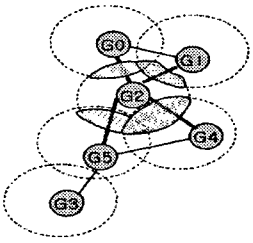
KEGG의 데이터베이스는 대사경로 데이터베이스, 실험을 통해 얻어진 유전자와 단백질 정보, 그리고 세포의 처리과정과 관련된 화학적인 compound와 반응에 대한 정보를 통합하여 제공한다[4]. 이중 DBGET/LinkDB[4]는 통합된 데이터베이스로부터 정보를 추출하는 시스템이다. DBGET/LinkDB는 검색을 위해 플랫파일 형태의 데이터를 이용한다. 데이터는 데이터베이스 이름과 엔트리 이름으로 구분된다. LinkDB는 많은 분자 생물학 데이터베이스들을 포함하는 데이터베이스인 GenomeNet에 직접적인 연결정보들로 이루어져 있다. 사용자가 검색창에서 데이터베이스 이름과 엔트리의 이름을 선택한 후 검색을 시작하면 LinkDB는 연결정보를 이용해 그와 관련된 데이터베이스와 엔트리 이름이 출력된다. 결과를 보고 정보를 더 보고 싶을 경우 출력된 결과에 해당하는 맵을 볼 수 있다. 우리는 DBGET/LinkDB와 같은 1차적인 연결정보뿐만 아니라 1차 이상의 경로를 거쳐 도달할 수 있는 엔트리에 대해서도 검사할 수 있다.

3. 제안하는 알고리즘

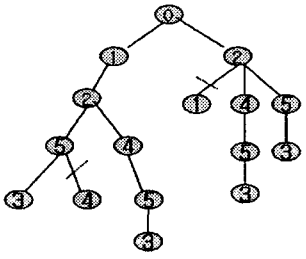
3.1 문제 정의 및 알고리즘

특정 compound(A)가 다른 compound(B)로 변화하는 과정을 알고 싶을 경우, compound(A)가 포함된 대사 네트워크(G0)로부터 변화된 compound(B)가 포함된 대사 네트워크(G3)까지의 경로를 찾아야 한다. G0와 G3 서로 다른 데이터베이스에 존재한다고 가정한다.

분산된 데이터베이스들 각각을 하나의 노드로 가정하고 이를 하이퍼 노드라 한다. 하이퍼 노드들이 서로 연결이 된 경우 즉, 임의의 두 대사경로 데이터베이스에 대해 서로 공통된 요소를 포함할 경우에 이들은 서로 연결되었다고 하며 이들을 연결하는 선을 하이퍼 에지라고 한다. 또한 하이퍼 노드와 하이퍼 간선으로 표현한 그래프를 추상하이퍼 그래프(Abstract Hyper Graph : AHG)라 한다. 또한 각 에지의 weight는 1로 가정한다.



[그림 1] 하이퍼 그래프 AHG

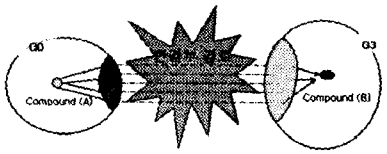


[그림 2] 하이퍼 노드 G0~G3까지의 모든 경로

[그림 1]은 6개의 하이퍼 노드를 가지는 추상 하이퍼 그래프(AHG)이다. 추상 하이퍼 그래프의 G0에서 G3까지 도달하려면 [그림 2]와 같은 경로를 지나가야 한다. [그림 1]에서 보면 빨강색 점선 동그라미는 각각의 데이터베이스의 대사 네트워크 영역을 표시하며, 색칠한 영역은 네트워크와 네트워크 사이에 공통으로 가지고 있는 부분이다. 이 영역에 해당하는 노드가 연결점이 되어 다른 네트워크에서의 검색을 시작할 수 있도록 한다. [그림 1]에서 색칠한 부분은 G2와 연결된 네트워크들을 연결해주는 노드들의 집합을 표현한 것이다. [그림 2]를 통해 G0에서 G3까지의 경로는 모두 4가지임을 알 수 있다. 모든 하이퍼 노드에 대하여 다른 모든 노드들에 대한 경로를 미리 계산한다. 이때 가능한 모든 경로에는 중복된 노드도 포함시켜야 한다. 추상 하이퍼 그래프 자체가 추상화되었기 때문에 하위의 어떤 노드가 선택될지 알 수 없기 때문에 중복된 노드를 제거하지 않는다.

실제 네트워크에서 경로를 탐색하기 위해서는 [그림 2]에 나타난 모든 경로에 대하여 검색을 하여야 한다. [그림 2]를 보면 같은 노드를 통과하는 횟수가 많다는 것을 알 수 있다. 이점을 이용하여 차수가 높은 노드 중 일부에 대하여 그 노드를 경유하는 연결부들 사이의 경로를 미리 계산하여 보다 빠르게 검색할 수 있도록 한다. 차수가 큰 하이퍼 노드들로 이루어진 그래프일 경우에는 그 효율이 더 커진다.

compound(A)에서 compound(B)까지의 대사경로를 탐색하기 위해서는 먼저 미리 계산된 A와 B를 포함하는 네트워크, 즉 하이퍼 노드들 간의 경로(P)를 참고하여 G0에서 다음 연결 하이퍼 노드인 G1과 G2의 연결부로의 경로를 탐색한다. G0의 compound(A)에서 G1이나 G2와의 연결부까지 탐색을 한 다음, 같은 방법으로 다음 하이퍼 노드와의 연결부위까지의 경로를 탐색하여 최종 노드인 compound(B)를 포함하는 네트워크의 연결부까지의 경로를 검색한다. 이때 중간에 연결부위가 많은 하이퍼 노드를 경유할 경우에는 미리 계산된 경로를 이용할 수 있다. 그런 다음, 연결부로부터 compound(B)까지의 가능한 경로를 탐색하면 된다. [그림 3]은 위에서 설명한 검색방법을 간단히 그림으로 보인 것이다.

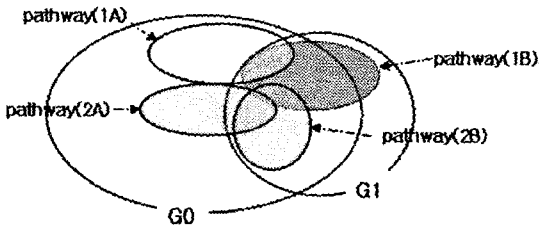


[그림 3] 탐색과정

경로탐색은 최단경로를 선택하는 것이 아니라 모든 경로를 검색한 다음에 평가하도록 해야 한다. 이 문제의 의도가 최단 거리가 아닌 네트워크 사이에 가능한 모든 경로를 통해 사람들이 경로를 확인하고자 하는 것이 목표이다.

3.2 제안한 알고리즘의 특징

2.1의 가정에서 정의한 것처럼 추상 하이퍼 그래프의 하이퍼 노드는 하나의 데이터베이스를 의미한다. 이 노드 안에는 여러 대사 네트워크가 서로 연결되어 있다. 그러므로 각 하이퍼 노드가 또 다른 하위 하이퍼 그래프가 된다. [그림 4]의 내용을 보면, 빨간선의 동그라미인 G0와 G1은 앞서 2.1에서 보인 추상 하이퍼 그래프의 하이퍼 노드이다. 각각의 하이퍼 노드는 그 내부에 여러 대사경로를 가질 수 있다. 예로 G0는 pathway(1A)와 pathway(2A)를 가지며 G1은 pathway(1B)와 pathway(2B)를 가진다. pathway(1A)와 pathway(1B)는 서로 같은 대사경로로 G0, G1 데이터베이스가 공통으로 가지고 있는 대사경로이다. 하지만 각 데이터베이스의 일부만 공통으로 가지며 일부는 각 데이터베이스에만 가지는 대사경로이다. pathway(2A)와 pathway(2B) 또한 대사경로이다. 하지만 이는 공통으로 가지고 있는 부분이 pathway(2B) 자체이므로 G0에서의 경로만 검색을 해도 되는 경우이다. pathway(1A)와 pathway(1B)와 같은 경우라면 연결되었다고 하며 pathway(2A)와 pathway(2B)인 경우는 연결되지 않았다고 한다.



[그림 4] 추상 하이퍼 그래프 안의 하위 추상 하이퍼 그래프

이들 대사경로들은 [그림 4]에서처럼 같은 대사 네트워크가 아니더라도 서로 중복될 수 있다. 제안한 알고리즘은 이러한 중복된 노드들의 정보를 가지고 있으며 차수가 높은 하이퍼 노드에 대해서는 노드와 그에 연결된 연결부와의 경로를 미리 계산한다.

4. 결론

이 논문은 분산된 네트워크 사이의 관계를 이용하여 데이터베이스 단계에서의 추상 하이퍼 그래프를 만들어 하이퍼 노드에 대한 경로를 탐색한 다음, 그 경로에 따라 각 데이터베이스에 존재하는 대사 네트워크를 하이퍼 노드로 하여 또 하나의 추상 하이퍼 그래프를 만든다. 두 번째로 만든 추상 하이퍼 그래프에 대한 탐색경로를 기반으로 하여 실제 노드에 대한 경로를 탐색한다. 추상 하이퍼 그래프의 경로탐색과 그

하위 노드에서의 경로탐색은 모두 최단 거리이기보다는 모든 경로를 탐색하여야 한다.

분산된 대사경로 데이터베이스를 통합하지 않고 검색이 가능하도록 하기 위해 대사경로 데이터베이스의 네트워크에 대해 경로탐색을 할 수 있도록 서로 중복된 데이터 정보를 이용한다. 이를 통해 모든 데이터베이스를 저장하지 않아도 검색이 가능하며 중복된 연결부와 높은 차수를 가지는 노드와의 경로를 저장함으로써 탐색 시간을 줄일 수 있다.

5. 참고문헌

[1] KEGG(Kyoto Encyclopedia of Genes and Genomes) - http://www.genome.jp/kegg

[2] EcoCyc(E.coli Genes and Metabolism) - http://www.ecocyc.org

[3] EMP(Enzymes and Metabolic Pathway) - http://www.empproject.com

[4]http://www.genome.jp/about_genomement/service

[5]Alexander Goesmann, Martin Haubrock, Folker Meyer, Jörn Kalinowski and Robert Giegerich, "pathfinder : reconstruction and dynamic visualization of metabolic pathways", vol. 18, no.1, pages 124-129, 2002