구조적 특징에 기반한 대사 경로 드로잉 알고리즘의 설계 및 구현

이소희 송은하° 이상호 박현석
이화여자대학교
{ssoy5502°, ehsong, shlee, neo}@ewha.ac.kr

Design and Implementation of a Metabolic Pathway Drawing Algorithm based on Structural Characteristics

So-Hee Lee, Eun-Ha Song°, Sang-Ho Lee and Hyun-Seok Park
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

'생물정보학'이란 생물학적 데이터를 처리, 가공하여 정보를 얻어내는 연구 분야로 이 중 대사체학은 대사 경로 네트워크를 가시화하여 생명 활동을 이해하고자 하는 분야로, 대사 경로 내의 흐름을 한 눈에 알 수 있도록 가시화하여 보여 주는 도구가 반드시 필요하다. 따라서 본 논문에서는 새로운 '대사 경로 드로잉 알고리즘'을 제안하였다. 대사 경로 그래프의 구조로는 이분 그래프를 이용하여 가독성을 높였으며, 이 그래프가 척도 없는(scale-free) 네트워크 구조라는 것과 구조적으로 환형, 계층적, 선형 컴포넌트를 가진다는 것을 고려하여 사이즈가 큰 그래프도 적절하게 드로잉 하도록 하였다.

1. 서 론

대사체학은 생물정보학의 한 연구 분야로 대사경로 네트워크를 그래프로 가시화하여 생명 활동을 총체적으로 이해할 수 있다[1]. 대사경로는 효소에 의한 화합물 간의 변화를 보여주는 네트워크로서 이 정보를 텍스트로 보게 되면 그 본질적 복잡성 때문에 내용을 이해하기가 어렵다. 따라서 대사경로의 흐름을 한 눈에 알 수 있도록 가시화하여 보여 주는 도구가 반드시 필요하다.

기존의 '대사경로 가시화 시스템'은 크게 정적인 시스템과 동적인 시스템으로 나뉜다[1, 2, 3]. 정적인 시스템은 대사경로 다이어그램을 수작업으로 그려 이미지파일로 저장하는 것으로 KEGG와 ExPASy Molecular Biology Server가 대표적이다[2, 4]. 그리고 대사경로에 적합한 그래프 드로잉 알고리즘을 적용하여 그 결과를 보여주는 동적인 시스템의 예로는 EcoCyc이 있다[3, 5]. 정적인 시스템의 경우 대사경로가 업데이트 되었을 때 이미지를 다시 그려주어야 하고 각각에 대해 준비되어 있는 이미지 하나로만 정보를 제공한다는 단점이 있다[1, 2, 4]. 또한 EcoCyc은 대사경로 정보가 업데이트 되는 경우, 새로운 이미지를 보여주는 것이 가능하지만 규모가 큰 대사경로를 표현해주지 못 한다는 문제점이 있다[1, 3, 5].

본 논문에서는 동적인 방법에 초점을 맞추어 대사 경로 정보의 업데이트를 안정적으로 반영해주고, 여러 형태의 컴포넌트가 복합적으로 존재하는 그래프도 그려줄 수 있는 개선된 대사 경로 드로잉 알고리즘을 설계 및 구현한다.

2. 대사 경로의 특징 및 기존 연구

2.1 대사 경로의 특징

2.1.1 척도 없는(scale-free) 네트워크

척도 없는 네트워크란 주어진 네트워크에 계속하여 추가되는 새로운 노드들이 특정한 소수의 노드에 링크되려는 성질을 가지는 네트워크를 말한다[6].

WIT(What Is There) 내의 대사 경로 네트워크를 분석해보면 대부분 척도 없는 네트워크에 속한다는 것을 알 수 있다[7, 8]. 따라서 대사 경로 그래프를 레이아웃 해 줄 때, 연결성이 높은 단백질을 중심에 위치시킴으로써 가독성을 높일 수 있다[9].

2.1.2 구조적인 특징

일반적인 대사경로들을 살펴보면 대부분 환형 컴포넌트들과 계층적 컴포넌트들로 이루어져 있음을 알 수 있다[1]. 주어진 대사경로에 환형 컴포넌트가 존재하는 경우 이 컴포넌트에 '환형 레이아웃 알고리즘'을 적용하고, 이 환형 컴포넌트와 연결이 되어 있는 다른 컴포넌트들을 환형 컴포넌트의 내·외부에 적당히 레이아웃 해 줄 수 있다[1, 3]. 계층적 컴포넌트에는 '계층적 레이아웃 알고리즘'을, 대사경로 내에서 빈번히 나타나는 긴 선형 컴포넌트는 '스네이크 레이아웃 알고리즘'을 적용하는 것이 적합하다[1, 3].

2.2 기존 연구

2.2.1 EcoCyc에 적용된 알고리즘

EcoCyc에 적용된 알고리즘은 주어진 대사경로 그래프의 위상에 따라 '환형 레이아웃 알고리즘', '계층적 레이아웃 알고리즘', '선형 레이아웃 알고리즘'을 적절히 적용한다. 복합형인 경우 그래프 내의 가장 큰 환형 컴포넌트를 찾아 이를 제외한 나머지 부분을 연결 성분들로 나눈다. 그리고 연결 성분들 중 환형 컴포넌트와 둘 이상의 에지로 연결되어 있는 것을 찾아 그 위상에 맞게 레이아웃 한 뒤 그 주위에 가장 큰 환형 컴포넌트를 레이아웃 하고 이 연결 성분과 환형 컴포넌트를 묶어서 슈퍼노드로 간주한다. 마지막으로 나머지 연결 성분들을 각각 그 위상에 맞게 레이아웃 한 뒤 각각을 슈퍼노드로 간주하여 슈퍼노드들로 이루어진 전체 대사 경로 그래프에 '트리 레이아웃 알고리즘'을 적용한다[3].

이 알고리즘은 '트리 레이아웃 알고리즘'을 적용할 때, 슈퍼노드 내의 노드들의 실제적 배치를 고려하지 않기 때문에, 에지 크로싱을 미리 예측하여 제거하는 것이 불가능하므로 전체적인 레이아웃 결과가 항상 적절하지 않다는 단점이 있다.

2.2.2 환형·계층적인 특징에 기반을 둔 알고리즘

이 알고리즘에서는 주어진 대사 경로 그래프 내에 환형 컴포넌트가 없는 경우 '계층적 레이아웃 알고리즘'을, 그래프 자체가 환형 컴포넌트인 경우 '환형 레이아웃 알고리즘'을 적용한다. 복합적인 경우에는 그래프 내의 가장 큰 환형 컴포넌트를 찾아 이를 제외한 나머지 부분을 연결 성분들로 나눈 뒤, 환형 컴포넌트에는 '환형 레이아웃 알고리즘'을 적용하고 각각의 연결성분들을 그 위상에 맞게 환형 컴포넌트의 내·외부에 레이아웃 한다.

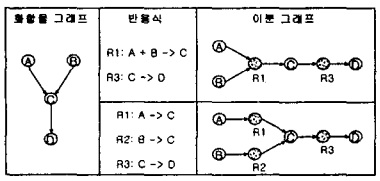
이 알고리즘은 화합물 그래프를 이용하여 대사 경로 그래프를 표현하고 있어 하이퍼에지가 빈번히 나타난다. 이러한 하이퍼에지는 레이아웃 결과에 잦은 에지 크로싱을 발생시킴으로써 가독성을 떨어뜨린다는 문제점이 있다[1].

3. 구조적 특징에 기반한 대사 경로 드로잉 알고리즘

3.1 고려 사항

대사 경로 그래프를 표현하는데 사용하는 화합물 그래프는 다음과 같은 단점들을 가지고 있다.

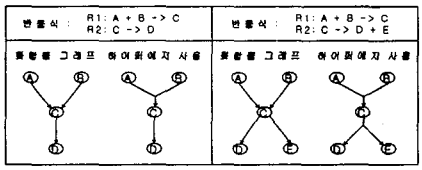
첫째, 해석이 모호한 경우가 많다[10].



[그림 1] 화합물 그래프와 이분 그래프

[그림 1]의 화합물 그래프는 두 개의 반응식으로 해석될 수 있다. 이처럼 해석의 모호성을 해결하기 위해서는 [그림 1]의 오른쪽 그림에서와 같이 하나의 노드 집합은 화합물들로, 다른 하나의 노드 집합은 반응 노드들로 이루어진 이분 그래프로 표현해 주는 것이 적합하다[10].

둘째, 하이퍼에지가 빈번하게 나타난다[1].



[그림 2] 하이퍼에지를 사용한 그래프

하이퍼에지는 화합물 그래프의 단점을 해결하기 위해 화합물 간의 다(多) 대 다(多)관계를 에지를 확장하여 표현하는 과정에서 발생된다([그림 2] 참조). 하이퍼에지를 이용한 표현은 해석의 모호성은 해결해주나 레이아웃 결과의 가독성을 떨어뜨리게 된다.

이러한 단점들을 극복하기 위하여 화합물들을 하나의 노드집합으로, 효소들을 다른 하나의 노드집합으로 하는 이분 그래프를 대사 경로 그래프 구조로 사용하였다[10].

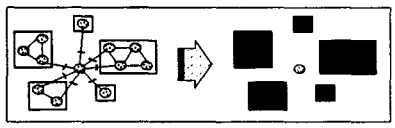
3.2 알고리즘

본 알고리즘은 입력으로 들어온 대사 경로 그래프의 구조적인 특성을 고려하여 적합한 레이아웃 모듈을 적용한다.

3.2.1 노드의 연결성을 고려한 레이아웃

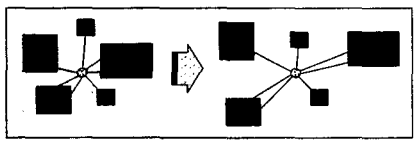
노드의 연결성을 고려한 레이아웃 알고리즘은 대사 경로 그래프 상에 존재하는 연결성이 높은 노드들을 적절하게 위치시키자는 아이디어를 기반으로 제안되었다.

첫 단계로는 [그림 3]과 같이 연결성이 높은 노드들에 인접한 에지들을 제거한 후, 남은 부분들을 연결 성분들로 나누어 각각의 연결 성분을 각각 하나의 슈퍼노드로 간주한다.



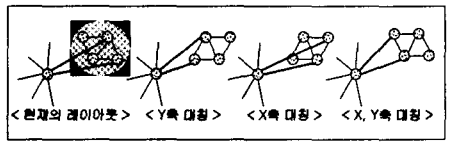
[그림 3] 연결 성분을 슈퍼노드로 그룹핑

각각의 슈퍼노드에 해당하는 연결 성분들을 레이아웃 해 준 후 각 연결 성분의 가로, 세로의 값을 이용하여 해당하는 슈퍼노드의 가로, 세로의 길이로 지정해주고 [그림 4]와 같이 선택된 노드, 슈퍼노드 그리고 제거했던 에지들로 이루어진 변형된 입력 그래프에 '스프링 임베딩 알고리즘'을 적용한다.



[그림 4] 슈퍼노드를 이용한 그래프의 변환

마지막 단계로 에지 크로싱을 최소화하기 위해 슈퍼노드 내의 레이아웃 되어 있는 연결 성분을 적절하게 대칭 시켜준다([그림 5] 참조). 이 경우 슈퍼노드 내의 노드와 선택된 노드 간을 연결하는 에지들의 길이의 합이 가장 작은 것을 골라 그 경우를 연결 성분의 배치 상태로 결정한다.



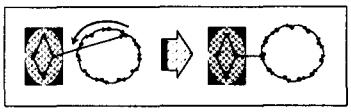
[그림 5] 슈퍼노드 내의 배치 결정

3.2.2 환형·계층적인 특징을 고려한 레이아웃

환형·계층적인 특징을 고려한 레이아웃 알고리즘은 대사 경로 그래프 상에 존재하는 환형 컴포넌트를 적절하게 배치하기 위하여 고안된 알고리즘이다.

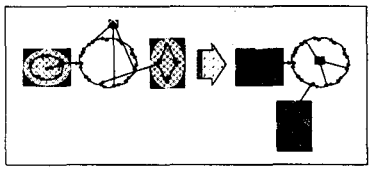
첫 단계로는 선택된 환형 컴포넌트에 인접한 에지들을 제거한 후, 그래프의 남은 부분들을 연결 성분으로 나눈다. 그리고 환형 컴포넌트를 제외한 각각의 연결 성분을 각각 하나의 슈퍼노드로 간주한다. 다음에는 환형 컴포넌트에 '환형 레이아웃 알고리즘'을 적용한 후, 이 컴포넌트를 슈퍼노드의 무게 중심에 따라 회전시킨다. 이 과정에서, 노드의 수가 3개 이하이면서 환형 컴포넌트에 연결된 에지의 수가 가장 많은 하나의 슈퍼노드를 내부 컴포넌트로 체크해 둔다. 그 다음 가장 큰 슈퍼노드를 선택하여 이 슈퍼노드의 무게 중심 쪽에 환형 컴포넌트를 위치시킨다. 그리고 이 슈퍼노드에 대한 환형 컴포넌트의 무게 중심을 구한 후, 이 무게 중심이 슈퍼노드의 무게 중심에 최대한 가까워지도록 환형 컴포넌

트를 회전한다([그림 6] 참조)



[그림 6] 환형 컴포넌트의 회전

레이아웃 된 환형 컴포넌트는 고정한 채로 '스프링 임베딩 알고리즘'을 적용한다. ([그림 7] 참조).



[그림 7] 스프링 임베딩 알고리즘 적용

마지막 단계로 슈퍼노드 내의 연결 성분을 적절하게 대칭 시켜주어 슈퍼노드내의 배치를 결정한다.

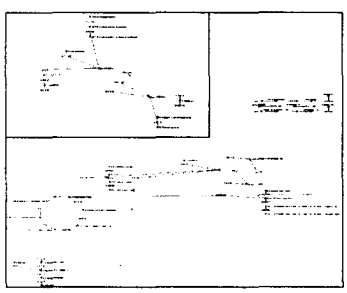
3.2.3 계층적 구성요소의 레이아웃

계층적 구성요소의 레이아웃 알고리즘은 위의 두 조건을 만족시키지 못한 그래프에 대해 적용되는 알고리즘으로, 선형 구조는 그 길이에 따라 '직선형 레이아웃', '스네이크 레이아웃'을 적용한다. 트리 구조일 경우에는 '트리 레이아웃'을, 아무 조건에도 속하지 않는 구조일 경우에는 '계층적 레이아웃'을 적용한다.

4. 실험 및 실험 결과

본 실험은 마이크로소프트 윈도우 98 환경에서 수행하였으며, Java 프로그래밍 언어와 Java 기반 그래프 라이브러리인 YFiles를 사용하여 구현하였다[11].

제안한 알고리즘을 적용한 결과 불필요한 에지 크로싱이 줄었고 하이퍼에지를 제거하여 가독성을 높임으로써 기존 시스템의 단점이 보완되었음을 알 수 있다. [그림 10]의 예를 보면 노드의 연결성을 고려한 레이아웃이 적절하게 적용되고, 복합적인 경우에도 환형·계층적인 특징을 고려한 레이아웃, 계층적 구성요소의 레이아웃이 적절하게 적용되어 있음을 볼 수 있다.



[그림 8] 레이아웃 알고리즘의 적용

또한 노드 사이즈에 따른 수행시간을 측정해 본 결과 노드 사이즈가 약 200인 경우 약 10초의 시간이 걸리는 것을 알 수 있다. KEGG 내의 대사 경로의 노드 사이즈는 대부분 100 내외 이므로, 드로잉 결과를 얻기 위해 그다지 많은 시간이 소요되지는 않는다.

5. 결론 및 향후 연구과제

본 논문에서는 '대사 경로 가시화 시스템' 구축에 있어서 필요한 그래프 레이아웃 알고리즘을 제안 및 구현하였다. 이 알고리즘은 기존 시스템들의 단점들을 극복하기 위하여 이분 그래프 구조의 대사 경로 그래프를 입력으로 사용하였으며, 각각의 대사 경로 그래프의 특징에 적합하게 레이아웃을 해 주었다. 이러한 결과는 비록 완벽하지 않다 하더라도 최소의 수정을 통하여 가장 적합한 드로잉 결과를 얻을 수 있는 초기 드로잉으로써 유용하게 사용될 수 있다.

향후 과제로는 생물체 내의 생명 활동을 총체적으로 이해하기 위해서 대사 경로뿐만 아니라, 조절 경로와 신호 전달 경로를 같이 레이아웃 해 줄 수 있는 그래프의 구조 및 레이아웃 알고리즘의 연구가 필요하다.

6. 참고 문헌

[1] M. Y. Becker and I. Rojas, "A Graph Layout Algorithm for Drawing Metabolic Pathways," BIOINFORMATICS, Vol. 17, No. 5, pp.461-467, 2001.

[2] M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, "The KEGG Databases at GenomeNet," Nucleic Acids Research, Vol. 30, No. 1, pp.42-46, 2002.

[3] P. D. Karp and S. Paley, "Automated Drawing of Metabolic Pathways," Third International Conference on Bioinformatics and Genome Research, pp.225-238, 1994.

[4] R. Appel, A. Bairoch and D. Hochstrasser, "A New Generation of Information Retrieval Tools for Biologists: The Example of the ExPASy WWW Server," Trends in Biochemical Sciences, Vol. 19, pp.258-260, 1994.

[5] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides and S. Gama-Castro, "The EcoCyc Database," Nucleic Acids Research, Vol. 30, No. 1, pp.56-58, 2002.

[6] K. -I. Goh, B. Kahng and D. Kim, "Universal Behavior of Load Distribution in Scale-Free Networks," Physical Review Letters, Vol. 87, No. 27, 2001.

[7] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. -L. Barabasi, "The Large-scale Organization of Metabolic Networks," NATURE, Vol. 407, pp.651-654, 2000.

[8] 정하웅, 강병남, "복잡계의 이해-네트워크의 구조적 성질 및 그 응용," 물리학과 첨단기술 10권 23, 2001.

[9] P. Holme, M. Huss and H. Jeong, "Subnetwork Hierarchies of Biochemical Pathways," BIOINFORMATICS, Vol. 19, No. 4, pp.532-538, 2003.

[10] Y. Deville, D. Gilbert, J. Helden and S. Wodak, "An Overview of Data Models for the Analysis of Biochemical Pathways," Proceedings of International Workshop on Computational Methods in System Biology, p.174, 2003.

[11] R. Wiese, M. Eiglsperger and P. Schabert, "The Y-files Graph Library: Documentation and Code available at http://www-pr.informatik.uni-tuebingen.de /yfiles," 2000.