명료한 BioPathway 지식표현을 위한 방법에 관한 연구

이민수[0] 박승수
이화여자대학교 컴퓨터학과
(ssue, sspark)@ewha.ac.kr

A Study on Better Representation of BioPathway

Min Su Lee[0] Seung Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 바이오인포매틱스의 발전과 함께 생물 관련 정보들이 기하급수적으로 증가하고 있다. 연구 대상 자체도 DNA, RNA, 단백질에서 더 나아가 신체 조직 기관 안의 분자적 트랜젝션들과 프로세스들에 의해 기능들이 어떻게 수행되는지에 관한 BioPathway까지 포함하게 되었다. BioPathway는 광대한 양의 정보를 포괄하며, 구성체 사이의 유기적 관계를 나타내고 있는 것이므로 이를 컴퓨터로 처리하기 위해서는 보다 명료하며 직관적인 표현이 중요시 된다. 그러나 현재 사용되고 있는 시스템들은 표준화가 안된 상태로 서로 다른 표기법을 사용하고 있어서 같은 정보를 다르게 표현하게 되고, 사용하고 있는 표기법 자체도 명료하게 해석할 수 없는 경우가 많다.

본 논문에서는 기존의 BioPathway에 관해 제안된 형식적 표기법들과 실제 사용되는 시스템들의 표기법들을 비교 분석하여 BioPathway를 보다 명료하고 효과적으로 표현하기 위한 방향을 제시하고자 한다.

1. 서 론

생물 관련 정보가 무수히 쏟아지고, 유전체의 완전한 서열 정보가 밝혀짐에 따라 DNA, RNA, 단백질에 관한 연구에서 더 나아가, 분자들의 상호작용에 의해 생체 시스템상에서 기능들이 어떻게 수행되는지에 관해 활발히 연구되고 있다. 이러한 분야를 대사체학(metabolism), 또는 BioPathway라고 부른다. [1]

BioPathway를 전체적으로 파악하고, 여러 경로들이 어떻게 상호작용하는지를 보다 쉽게 이해하기 위해, 그래프 형태의 지도를 사용하여 경로를 표현한다. 현재 많은 생명공학 관련 시스템들이 BioPathway를 그래프 형태의 지도를 사용하여 서비스를 제공하고 있다. 그러나 이들 시스템들은 각각 자신들만의 표기 방법을 사용하고 있기 때문에 같은 관계(relation)를 서로 다르게 표현하고, 사용자들은 같은 표기방법을 서로 다르게 해석할 수 있다는 문제를 안고 있다. 예를 들면, 상호 관계를 의미하는 화살표가 Boeheringer Mannheim의 생화학 경로[2] 중 대사 경로에서는 화합물의 전환을 의미하는 반면, 똑같이 생긴 다른 화살표는 IUPAC-IUBMB의 신호 전달 지도에서는 조절 단계를 의미한다. 또한, 시스템들이 사용하는 표현 방법이 BioPathway에서의 모든 관계를 명료하게 표현할 수 없는 불완전한 상태이므로 표준화 된 표기법이 요구된다.

본 논문에서는 BioPathway 표기법의 표준화를 위한 기반연구로서, 기존의 BioPathway 표기법에 관한 연구와 BioPathway에 관한 그래프 형태의 지도를 제공하는 시스템들을 조사하고 표현 방법을 비교분석한다. BioPathway에 대해 제안된 형식적(formal) 표기법 들과 상용화된 바이오인포매틱스 시스템들 중 BioPathway 지도를 서비스하는 것을 주 대상으로 한다.

본 논문의 구성은 다음과 같다. 2장에서 연구의 배경이 되는 생물정보학과 대사체학, 지식 표현 방법에 대해서 알아보고, 3장에서는 BioPathway를 표현하기 위해 필요한 기본적인 특성들과 요구사항들, 그리고 BioPathway에 관한 정형화된 표기법과 그래프 형태로 생화학 경로 지도를 제공하고 있는 시스템에서의 표기법을 살펴본다. 4장에서는 3장에서 살펴본 형식적 표기법과 시스템들의 장단점을 분석한 후, 좀더 향상된 표기법을 정립하기 위한 방향을 제시한다. 마지막으로 5장에서 결론과 향후 과제를 논의한다.

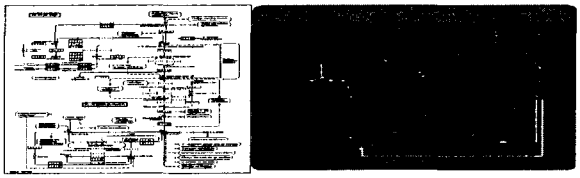
2. 관련 연구

2.1 생물정보학

생물정보학(Bioinformatics)은 생물학에서 다루는 정보의 양이 급증함에 따라 전산학, 수학, 통계학 등의 분야에서 사용되고 있는 정보처리기법을 응용하여 이를 효율적으로 생산, 관리, 활용하려는 연구분야를 총칭한다.[3] 생물정보학의 역할은 쏟아져 나오고 있는 대용량의 생물학 데이터의 저장과 처리뿐만 아니라 유전체 정보 및 생물학적 지식을 종합하여 생명의 기본 원리를 창출해내는데 있다.

2.2 BioPathway

BioPathway Consortium[1]에서는 BioPathway를 생화학 신체



[그림1] KEGG: metabolic pathway [그림2] BioCarta: signal transduction networks

* 이 논문은 BK21(Brain Korea 21)사업의 지원으로 연구되었음

조직들의 기관 안의 모든 형태의 분자적 트랜젝션들과 프로세스들을 포함하는 포괄적인 의미의 용어라 정의 한다. 이들 경로들의 일부는 선형적인 처리과정에 관련되어있지만, 대다수들은 복잡한 트리(tree)형태나, 단방향, 양방향 그래프, 또는 순환(cycle) 형태로 되어있다.

생화학적인 경로들은 크게 화학물질들의 효소반응으로 일어나는 물질 수송과 에너지 변환에 관한 신진대사 경로 (metabolic pathways), 세포 주기와 특정 유전자의 정보 발현에 관한 신호전달 경로(signal transduction networks), 유전정보 전달과 발현을 위한 처리과정에 관한 유전자 조절 경로(gene regulation networks), 약의 대사 경로(drug metabolism pathways)와 같은 클래스로 분류된다.

2.3 지식 표현(knowledge representation)

최근 대부분의 지능형 시스템은 그 시스템이 적용되는 분야의 지식을 내재하고 있어서, 시스템 내에서 이러한 지식의 적절한 표현이 중요하다. 인공지능 분야에서는 지식을 항상 참인 지식인 사실과 문제의 영역 내에서의 일련의 사건들의 전후 관계를 기술하는 절차적인 규칙, 항상 성립한다는 보장은 없으나, 대부분의 경우 성립하는 경험적인 규칙 등의 세가지 형태로 나눈다. 이러한 지식들을 시스템 내부에 표현하기 위한 방식으로 여러 표현 방식이 연구되어왔으나, 크게 분류하면 논리에 근거를 둔 형식적 표현 방식과 비형식적 표현 방식이 있다. BioPathway를 표현하기 위해서는 형식적 표현방법 보다는 비형식적 표현 방법의 frame, 의미망(semantic network) 등이 많이 사용된다.

3. BioPathway의 표현

3.1 기본적인 특성들과 요구사항

BioPathway를 그래프로 표현하면, 노드에 해당하는 것은 경로상의 주체들(entity)이고, 에지에 해당하는 것은 주체들 사이의 이벤트(event)가 된다. 또한, 경로의 종류에 따라 노드와 에지에 해당하는 것이 달라진다. 그러므로 BioOntology를 바탕으로 명료하게 주체와 이벤트를 정의해야 한다. 다음 [표1]은 많이 사용되는 대표적인 예들이다.

[표1] BioPathway 안에서의 표현 주체

경로 종류	주체	이벤트
대사 경로	반응물, 생성물, 효소	생화학 상호작용, 조절 작용(activation, inhibition)
유전자 조절 경로	세포, 단백질, 유전자, RNA, 조절물질,	생화학 상호작용, 조절 작용 (direct /indirect,, switching on/off, positive / negative effect)
신호 전달 경로	호르몬, 신경전달물질, 막 리셉터, 이온채널, 전사 인자, 금속이온 …	생화학 상호작용, 조절 작용(up/down regulation, activation/inhibition, increase/decrease)

BioPathway를 그래프로 표현하기 위해서는 기본적으로 다음과 같은 성질을 만족시켜야 한다.

- 직관성(Intuitive): BioPathway 지도를 통해 상호작용의 의미를 명확하게 이해할 수 있어야 한다.
- 적절성(Fitting): 간단한 관계들은 간결한 표기 방법으로, 복잡한 관계들은 복잡한 표기 방법으로만 표현할 수 있다.
- 표현력(Expressive): 한가지 표기 방법으로 모든 종류의 BioPathway를 나타낼 수 있어야 한다.
- 정형성(Formal): 주어진 관계와 그것을 표현한 그래프 상에 1대1 대응관계가 만족되어야 한다. 즉, 주어진 관계를 그래프로 구성할 수 있어야 하고, 그래프를 통해 상호 관계를 재구성할 수 있어야 한다.

[표2] BioPathway의 명료한 표현을 위한 요구사항

이슈	요구사항
불완전한 정보에 대한 명료한 표현	주체나 관계들의 특성에 기반한 제약 조건의 명시 미결정 주체의 명료한 표현 미결정 관계의 명료한 표현
명확히 구분되어져야 하는 정보에 대한 표현	생략된 부분의 표현 계산에 의해 추정되는 결과로 나온 상호 관계의 구별된 표현

위의 기본적 성질 외에, BioPathway 그래프를 명료하게 표현하기 위해 다음과 같은 사항들에 관한 구분이 요구된다. [표2]

3.2 BioPathway에 관한 형식적 표기법

대표적인 BioPathway에 관한 대표적인 세 가지의 형식적 표기법에 대해 살펴본다.

· Eberhard Voit[4]의 대사 경로에 관한 표기법은 간략한 표기법만을 사용하여 대사 경로를 표현하고 있다. 기본적인 표기법 사이의 1:1 대응관계가 만족되지 않는다.

· Kurt Kohn[5]은 가능한 단백질 상호작용에 관한 표기를 할 수 있는 기본 표기법을 제안하였다. 매우 상세하지만, 그 표현 가능한 영역이 단백질 상호 작용에만 제한되어 있어서, 전반적인 생화학 메커니즘을 표현하기에는 부족하다는 단점을 안고 있다. 또한, 잘 알려져 있지 않거나 추정되는 것들에 대해서는 명료한 표현이나 간략한 표현 자체가 허용되지 않는다. 그리고, 표현 방법 자체가 기존에 쓰이던 것들과는 완전히 다르므로, 이 표기법을 사용하려면 이 표기법에 대한 전반적인 학습이 요구된다.

· Isabelle Pirson[6]은 생화학 메커니즘 전반에 걸쳐 적용될 수 있는 표기방법을 생물학자들이 이해하기 쉽도록 기존에 많이 쓰이던 표기법들에 기반하여 제안하였다. 화살표 위에 인덱스와 색깔을 사용하여 신호전달 경로에서의 작용 시간도 표현하면서 상호작용의 직관성을 높였다. 화학물질의 전환이나 수송과 조절 정보의 흐름의 구분을 명확히 하고, 단백질의 활동에 있어서의 즉각적/지연 효과(direct/delayed effect)들을 구분해서 표현한다는 장점도 있다. 또한 어떤 부분이 시스템상에서 학습된 것인지, 어떤 부분은 다른 시스템에서 가져온 지식인지, 어떤 것들은 서열상에서 알려지지 않은 것인지 구분할 수 있다. 그러나 불완전한 정보를 구별하여 표현할 수 있는 방안을 제시하고 있지 못하다.

3.3 Graphical BioPathway를 제공하는 시스템들

Bioinformatics에 관련된 시스템들 중에서 BioPathway에 관한 정보를 제공하는 시스템들이 많이 있다. 그들의 대부분은 BioPathway에 관한 정보를 그래프 형태에 기반하여 제공한다.

3.3.1 대사 경로

ExPASy(Expert Protein Analysis System)[2]는 Boehringer Mannheim의 생화학 경로의 대사 경로 지도를 제공한다. 반응이 관측 되는 유기체(organism)의 종류와 조절작용, 반응에서의 역할, 효소 활동의 중감과 속도, 이화, 동화 관계를 표현할 수 있다. 그러나 생화학적 단계를 의미하는 기본 화살표는 화학적 변환이나 조절 등 여러 뜻으로 해석될 수 있으므로 모호함이 따르게 된다. 또한 대사 화합물들을 화학식으로 표현하여, 작은 노드로 각 주체들을 표현하는 것에 비해 전체적인 경로의 흐름을 이해하기가 힘들다.

교토 대학의 KEGG(Kyoto Encyclopedia of Genes and Genomes)[7]는 대사 경로 표기를 위해 비교적 간단한 표기법을 사용하고 있다. 기본적으로 대사 화합물을 노드로 하고 효소 작용을 에지로 하는 그래프 형태이다. 대사 화합물과 효소로 이루어져있는 KEGG의 대사 경로는 효소들의 순서화된 나열 이라는 것을 명료하게 보여준다. 그러나 KEGG도 상호 작용의 종류가 상세히 구분되어있지 않으므로, 해석 시 혼돈을 초래할 수 있다.

WIT(What Is There?)[8]는 대사경로 모델을 지원하는 시스템

으로 작은 범위의 경로 지도를 제공하며 Voit의 형식적 표기법과 유사한 형태를 띄고 있다. 간략하고 생물학자에게 친근한 표현을 사용하여 해석이 편리하나, 명료한 작용 종류를 나타내지 못한다.

3.3.2 유전자 조절 경로

유전자 조절 경로에 관한 그래픽 서비스를 하는 GeneNet[9] 데이터베이스는 객체 지향적 접근법으로 계층적인 유전자 조절 네트워크를 표현하고 있다. 기본 구조는 유전자와 RNA, 단백질 등과 그들의 상호작용과 조절 작용과 같은 사건들로 구성된다. 유전자 네트워크는 유기체 수준과 세포 수준, 그리고 유전자 수준으로 표현을 하고, 표현 주체의 상태에 따라 명료한 표현이 가능하도록 되어있다. ExPASy[2]에서는 세포와 분자 처리 과정(Cellular and molecular processes) 지도는 대사경로 표기법을 기본으로 확장한 표기법으로 세포 내에서의 유전자 발현 조절을 나타내고 있다. 세포와 분자 처리 과정 지도상에서는 효소와 그 활성화 여부를 표현할 수 있다. 그리고 유전자와 그의 산물을 읽기 쉽게 표현 하고 있다.

3.3.3 신호전달 경로

신호전달 경로는 새로운 정보들이 밝혀짐에 따라 그 복잡도가 날이 갈수록 커져가기 때문에 모든 것을 명료하게 표현하기 힘든 경로이다.

KEGG[7]에서는 신호 변환 경로에서 선 위에 인덱스를 붙임으로써 활성화(activation)와 억제(inhibition)를 구별하고, 유전자 조절과 신호 전달 정보를 수록한 데이터베이스인 BioCarta [10]는 다양한 표기 방법을 사용하여, 각 주체와 상호작용 관계를 보다 직관적으로 이해할 수 있도록 표현한다.

4. 분석 및 BioPathway 표기법 발전 방향

4.1 분석

통일된 BioOntology가 정립되지 않은 상황에서 현재 사용되고 있는 BioPathway에 관한 표현법을 3절에서 살펴보았다. 자신들만의 BioOntology를 구축하고 이를 바탕으로 표기법을 정의한 시스템도 존재하기는 하나, 대부분의 시스템들의 대사경로에 관한 명료한 표기법은 정립되지 않은 상태이다. 표기법을 정의했다 하더라도 넓은 범위로 확장시킬 경우, 해석 시 의미가 모호해지는 경우가 많고, 불완전한 정보나 명확히 구분되어져야 하는 정보들에 대해 명료한 표현을 제공하지 못하고 있다.

4.2 BioPathway 표기법의 발전 방향

사용자에게 보다 직관적이고 명료한 BioPathway 정보를 제공하기 위해서는 대사과정 안의 각 주체들의 명확한 유형을 구분해야 하고, 각 상태를 명료히 나타내야 한다. 각 구역, 세포 영역, 상태(state), 유전자, RNA, 단백질, 그리고 기타 대사 물질들(substance)을 명료하게 구분해야 한다. 그리고 단백질의 상태도 활성화 상태와 비활성화 상태, monomer, homodimer, heterodimer, multimeric complex를 구분할 수 있어야 한다.

위의 표현 주체들은 관련된 2차 데이터들과 깊은 연관 관계를 가지고 있다. 따라서 이러한 연관 관계를 명료하면서도 효율성 있게 표현하여야 한다. [표3]

또한, 생체 안의 모든 대사과정들은 세포 안의 단백질들과 단백질 화합물, 그리고 핵 안의 유전자들 사이의 작용으로 표현할 수 있다. 기본적으로 이들의 상호작용을 명확하게 구분지어 표현할 수 있도록 표기법을 정의하여야 한다. [표4]

한 화면에 모든 정보를 보여주려고 하면 전체적으로 직관성이 떨어지게 되고, 오히려 중요 정보들이 누락될 수도 있다. 보여주고자 하는 정보를 다양한 분석 레벨로 나누어 경로를 표현하고, 서로 다른 유기체나 조직 상에서 생화학 경로를 비교하도록 하며, 주석의 상세 정도를 단계별로 제공한다면 이러한 점들을 보완할 수 있다.

[표3] 표현 주체에 따른 평가 기준과 관련 데이터

데이터	평가 기준	2차 데이터	2차 데이터 평가 기준
생물학적 객체	표현력	장소, 표현법, 서열, 구조	존재 유무, 표현 방법, 문맥상의 독립성 (2차적 데이터에 의해 생물학적 문맥이 구축됨)
생화학적 작용	표현력 효율성	장소, 빈도, 지연, 작용 형태	
개념적 관계	표현력 효율성	유전자 데이터	

[표4] 명료하게 구분해야 할 주체들 사이의 관계

주체	관계 종류
유전자 - 유전자	동형(homologue)
유전자 - 단백질	유전자가 단백질로 번역(translation)
	단백질이 유전자의 발현을 조절
단백질 - 단백질	한 단백질이 다른 단백질을 인산화 (활성화)시킴
	단백질이 결합
	리셉터로서의 단백질이 리간드와 결합
	리셉터가 다른 단백질들을 활성화 시킴

또한, 경로 검색 결과를 정보의 레벨이나 다양한 기준에 따라 경로를 강조하거나 사소한 경로들은 생략하는 기능도 필요하다.

기본적으로 대사 경로에 관한 형식적 표기법을 정립한 후, 좀더 넓은 범위인 유전자 조절 경로와 신호전달 경로까지 표현할 수 있도록 하여 한가지 표기 방법으로 모든 종류의 경로들을 나타낼 수 있어야 한다.

5. 결론 및 향후 연구

본 논문에서는 BioPathway를 표현하기 위해 요구되는 사항들을 정리하고, 현존하는 형식적 표기법과 시스템 상의 표기법들에 대한 비교분석을 바탕으로 향상된 BioPathway 표기법의 방향을 제시했다.

앞으로는 통일된 BioOntology에 기반하고 중요한 요소인 시간과 인과관계가 반영된, 체계적이면서도 확장 가능한 단일한 경로 표기법이 만들어져야 한다. 이미지 형태의 그래픽 표현에서 지양하여, 계속 업데이트되는 데이터들을 즉각 반영하여 전체적으로 재구성될 수 있도록 그래프 알고리즘에 기반하여 계속 발전되어야 한다.

6. 참고 문헌

[1] BioPathway Org. http://www.biopathways.org/
[2] ExPASy. http://kr.expasy.org/
[3] Bertone, P. & Mark Gerstein, " Integrative Data Mining: The New Direction in Bioinformatics" IEEE Engi-neering in medicine and biology vol. 20, pp.33-40 July/Aug. 2001
[4] Voit, Eberhard, *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists,* Cambridge press. 2000
[5] Kohn, Kurt, " Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems" Molecular Biology of the Cell Vol.10 pp.2703-2734, Aug. 1999
[6] Pirson, Isabelle et al. " The Visual Display of Regulatory Information and Networks" Trends in Cell Biology Vol.10 pp.404-408, Oct. 2000
[7] KEGG. http://www.genome.ad.jp/kegg/
[8] WIT. http://wit.mcs.anl.gov/WIT2/
[9] GeneNet. http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet
[10]BioCarta. http://www.biocarta.com/