바이오패스웨이를 위한 개선된 지식표현 시스템

이민수° 박승수
이화여자대학교 컴퓨터학과
{ssue°, sspark}@ewha.ac.kr

A New Knowledge Representational System for BioPathway

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

요 약

최근 바이오인포매틱스의 발전과 함께 생물 관련 정보들이 기하급수적으로 증가하고 있다. 연구 대상도 DNA, RNA, 단백질에서 더 나아가 이들의 상호작용 및 조절 메커니즘에 의해 기능들이 어떻게 수행되는지에 관한 BioPathway까지 포함하게 되었다. BioPathway는 광대한 양의 정보를 포괄하며 구성체 사이의 유기적 관계를 나타내고 있는 것이므로 다양한 형태의 지식을 통합하며 지식의 특성에 맞게 정보를 관리하고 표현함으로써 컴퓨터 프로세싱을 용이하게 하여 정보의 부가가치를 높이는 것이 중요하다. 이러한 BioPathway를 지식표현 관점에서 체계화하고 이를 확장함으로써 궁극적으로 바이오 정보의 거대한 지식베이스를 형성할 수 있다. 본 논문에서는 다양한 종류의 BioPathway 지식을 프레임 형식에 기반하여 보다 명료하고 효율적으로 표현할 수 있는 UniPath 표기법을 제안하였다. 또한 이 표기법을 적용하여 BioPathway 지식을 그래프 형태로 편집함으로써 그 정보를 등록하며, XML 포맷으로 쉽게 변환할 수 있는 시스템을 설계하고 실제 데이터에 적용함으로써 타당성을 검증하였다.

1. 서 론

HGP(Human Genome Project), 마이크로어레이 등 대용량의 데이터를 산출해내는 HTS(High-Throughput experimental System)의 등장과 함께 생물 관련 정보들이 기하급수적으로 증가하고 있다.

BioPathway는 생체 시스템 상에서 분자들의 상호작용에 의해 기능들이 어떻게 수행되는지에 관해 연구하는 생물정보학의 한 분야이다. BioPathway 지식은 서로 유기적으로 복잡하게 얽혀있으므로 특성에 맞게 적절하게 표현하고 효과적으로 관리함으로써 지식 베이스를 구축하고 컴퓨터 프로세싱을 통해 정보의 부가가치를 높이는 것이 중요하다. 그러나 기존 시스템은 BioPathway를 이미지 맵으로 제공하여 정보를 컴퓨터상에서 처리하기 힘든 경우가 많다. 그래프 형식으로 제공하더라도 표현 영역이 특정 한 단면에만 국한되어있으며 같은 정보를 표현하여도 시스템마다 지식 표현 수준과 방식이 달라 시스템 확장 및 통합이 어렵다.

본 논문에서는 BioPathway 지식을 효과적으로 관리하며 세분화 되어있는 BioPathway 지식의 통합을 용이하도록 하기 위해 프레임(Frame) 형식으로 관리하고 이를 단일한 표기법을 사용하여 그래프 형태로 표현하기 위한 UniPath 표기법을 설계하였다. 그리고 이 표기법을 적용하여 등록된 지식을 XML 포맷으로 변환할 수 있는 시스템을 설계하고, BioPathway를 위한 그래픽 에디터를 구현하였다.

본 논문은 2장에서는 생물정보학과 BioPathway, 그리고 BioPathway 지식 표현 방법들에 대해 살펴보고, 3장에서 세분화되어 있는 BioPathway 지식을 단일 표기법으로 표현할 수 있는 UniPath 시스템을 설계하고 4장에서 UniPath 시스템을 적용함으로써 검증한다. 마지막으로 5장에서 본 논문의 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 생물정보학과 BioPathway

생물정보학(Bioinformatics)은 일반적으로 데이터베이스, 알고리즘, 기계학습 및 컴퓨터 그래픽스 등과 같은 컴퓨터 기술을 이용하여 생물학 데이터를 저장, 분석 및 해석하는 계산적 생물학(computational biology)을 의미한다.

최근의 생물정보학은 다양한 생물 종에 있어서 유전체의 염기서열을 해독하는 HGP의 결과물을 바탕으로 각 유전자의 위치와 생체내에서의 기능을 밝히며, 각 유전자 집합으로부터 시스템 전체(세포 또는 생물 개체)가 재구성될 수 있는지 여부를 조사하여 생명 작용을 시스템의 작용으로 이해하려는 연구가 이루어지고 있다. 특정 개체의 완전한 유전체 서열 정보들이 밝혀짐에 따라 생물정보학 연구자들은 정보의 효율적 통합과 검색, 새로운 분석 알고리즘의 개발, 대사 경로의 구성 및 이해, 그리고 데이터 마이닝을 통한 새로운 지식의 창출 등에 초점을 두고 연구를 진행하고 있다.

모든 생물학적 기능은 분자 상호작용의 네트워크를 통해 발현되므로 분자 상호작용에 대한 정보는 개개 분자에 대한 정보 못지않게 중요하다. 따라서 생화학 신체 조직 기관 안의 모든 형태의 분자적 상호작용들과 프로세스들을 나타내는 BioPathway에 대한 연구는 생명 현상의 신비를 해독하기 위해 필수적이라 할 수 있다[1]. BioPathway는 서로 긴밀하고 유기적으로 얽혀있지만, 기능에 따라 크게 화학물질들의 효소반응으로 일어나는 물질 수송과 에너지 변환에 관한 신진대사 경로(metabolic pathways)와 유전자 조절과 신호 전달을 포괄하는 조절 경로(regulatory pathways)로 분류할 수 있다.

2.2 BioPathway 표현 방법

최근 대부분의 지능형 시스템은 그 시스템이 적용되는 분야의 지식을 내재하고 있어서, 시스템 내에서 이러한 지식의 적절한 표현이 중요하다. 이러한 점은 특히 서로 유기적이면서 복잡하게 얽혀있는

BioPathway를 표현하고자 할 때 더욱 강조된다.

BioPathway를 보다 효과적이고 적절하게 표현하기 위해 다양한 지식 표현 방법들이 사용되고 있다. [표 1]은 상용 시스템에서 BioPathway 지식을 표현하기 위해 사용하는 대표적인 방법들을 보여준다. 각 시스템들은 한 가지 표현 방법에만 국한하지 않고, 다양한 형식으로 BioPathway 정보를 제공하고 있다.

표 1 BioPathway의 다양한 표현 방법

주요 표현 방법	시스템 예
상호작용의 순서 리스트	BIND
Peri Net	Genomic Object Net
분자 상호작용 그래프	GeneNet, KEGG, DIP
XML	CellML, SBML

3. UniPath 시스템 설계

본 장에서는 세분화 되어있는 BioPathway 지식을 단일 표기법을 사용하여 보다 명료하게 표현할 수 있는 UniPath 시스템을 제안한다.

3.1 UniPath 시스템 개요 및 특징

세포에서 일어나는 반응을 전체적으로 살펴보려면 개별적인 경로 정보를 통합할 수 있어야 하고 그것을 일반화된 표기법을 사용하여 나타내야 한다. 그러나 기존 시스템에서는 BioPathway 지식을 서로 다른 표현 방법과 레벨을 사용하여 표현하고 있으므로, 경로 정보의 확장 및 통합이 어렵다. 이러한 점은 조금씩 밝혀지고 있는 조절 경로의 단편들을 연결하고 통합하려할 때에 그 문제점이 더 커진다.

UniPath 시스템은 BioPathway 지식을 BioOntology에 기반하여 구조적 지식이나 통합적 지식의 표현에 적합한 프레임 형식으로 데이터를 관리한다. 생체 경로상의 표현 주체는 다양한 속성값을 가진 슬롯(slot)으로 구성된 프레임으로 주체사이의 관계는 프레임간의 연결관계로서 관리하며, 지식의 표현은 표현주체는 노드(node)로 주체 사이의 관계는 아크(arc)로 나타냄으로써 생체 반응을 생물학자에게 친숙한 그래프 형태로 제공하도록 하였다. UniPath 표기법은 복잡한 BioPathway 지식을 특성에 맞게 적절하게 표현하며, 신진대사, 유전자 조절, 신호 전달 등 모든 종류의 경로에 관한 정보를 생물학자들에게 친근한 표기법에 기반하여 단일 표기법으로 표현하는 것을 목표로 한다.

BioPathway 지식을 프레임 형식으로 관리하고 그래프 형태로 표현함으로써 BioPathway의 효과적인 검색 및 추론 과정을 가능하도록 하였으며 복잡한 지식을 구조적으로 관리하며 세분화된 BioPathway 지식을 통합하는 것을 보다 용이하게 하였다.

3.2 UniPath 표기법 구성 및 설계

UniPath 표기법은 BioPathway 지식을 작용이 일어나는 세포 내 위치 정보를 표현해주는 '세포 내 위치 정보', 그래프의 노드에 해당하는 '주체의 형(object type)', 그리고 아크에 해당하는 '주체 사이의 작용(reaction between objects)'으로 나누어 [그림 1]과 같이 정의된다.

주체의 형이나 세포 내 위치에 의해 직관적으로 구분 가능한 주체 사이의 관계는 세분화하지 않고, 기본 표기법만으로 다양한 경로를 간결(compact)하게 표현할 수 있도록 디자인하였다.

생물학자에게 상당히 중요한 정보인 상호작용이 일어나는 장소인 세포 영역을 명확하게 표현해주는 것은 가독성과 직관성을 향상시켜 사용자의 이해를 도우며 위치 정보에 의해 명백하게 구분 가능한 표현 주체나 작용의 세분화를 막아준다는 장점이 있다

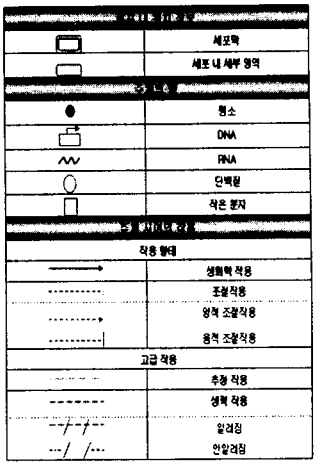

그림 1 UniPath 표기법

구분되게 표현할 주체의 형은 EcoCyc의 바이오 온톨로지에 기반하여 선별하고[2], GeneNet에서 사용하는 주체 표기법을 참고하여 디자인하였다[3]. 그리고 주체들 사이의 작용은 주체의 형과 세포 내 위치 정보에 의해 구분 가능한 에지들의 불필요한 세분화를 막기 위해 Voit[4]와 Pirson[5]이 제안한 표기법들 중 필수 요소들만을 선별하여 정의하였다.

예측 알고리즘이나 데이터 마이닝 등의 기법을 통해 얻은 데이터는 가치 있지만 실험적으로 검증되지 않았기 때문에 오류를 포함하고 있을 수 있다. UniPath 표기법에서는 이러한 추정된 작용을 명시할 수 있으며, 경로 지도상에 생략된 부분이 있다면 그 부분이 현재 밝혀져 있는 부분인지의 여부를 구분할 수 있도록 하였다.

3.3 UniPath 표기법의 XML 포맷으로의 변환

경로에 관한 정보는 경로를 구성하는 각 주체와 주체사이의 작용들에 관한 세부작용들을 명세한 상호작용의 리스트로서 표현될 수 있다. 구축된 BioPathway 지식은 사실상의 산업계 표준(De facto Standard)인 SBML(System Biology Markup Language)[6]로 쉽게 변환 시킬 수 있다. 예를 들면, UniPath로 표현한 LAT와 Grb2의 상호작용은 [그림2]과 같이 XML 포맷으로 변환시킬 수 있다.

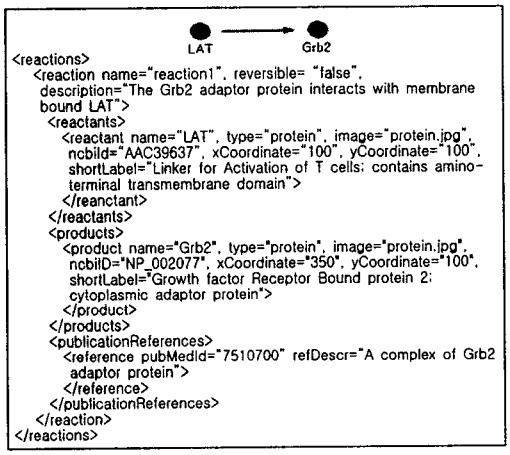

그림 2 UniPath 표기법의 XML 포맷으로의 변환

4. UniPath 시스템 구현 및 평가

4.1 시스템 구현

BioPathway 지식을 제공하는 기존 시스템은 경로를 그림 지도

(image map) 형식으로 제공한다. 이러한 방법은 새롭게 알려지는 지식들을 업데이트하기 힘들고, 경로 정보를 검색하거나 새로운 가치있는 정보를 얻기 위해 다양한 마이닝 기법을 적용시키기 힘들다는 단점을 안고 있다.

UniPath 시스템에서는 UniPath 표기법을 적용시킨 경로 지도 편집기를 사용하여 BioPathway를 편집하고 세부정보를 입력함으로써 지식을 등록하도록 하여, 경로 정보를 기본적인 상호작용의 네트워크로서 인식할 수 있으며 새로운 지식을 즉각적으로 업데이트할 수 있도록 하였다. 그럼으로써, 복잡한 BioPathway 정보를 컴퓨터가 처리하기가 용이한 프레임 형식으로 관리하며, 단편적인 정보를 서로 연결하여 효과적으로 처리할 할 수 있도록 하였다. UniPath 시스템은 Windows 2000 운영체제에서 Visual C++ 6.0을 사용하여 개발하였다.

4.2 UniPath 시스템 적용

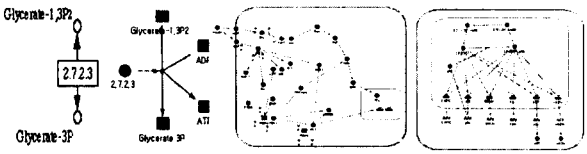

그림 3 KEGG와 UniPath 표기법 비교

그림 4 UniPath로 표현한 인슐린 조절작용

그림 5 UniPath의 세포주기 표현

[그림 3]는 KEGG[7]와 UniPath 표기법으로 해당과정의 일부를 표현한 것이다. KEGG의 표기법은 에지 위에 해당작용을 조절해주는 효소의 EC번호를 써줌으로써 조절작용을 표현한다. 이러한 방법은 대사과정의 부산물로 나오는 에너지와 물질 대사에 관해 표현할 수 없고, 표현 범위를 조절 경로까지 확장시킬 수가 없다. 반면, Unipath 표기법은 생물학자들에게 친근한 형태를 띄고 있으며, 해당작용의 주체의 형과 부수적으로 생성되는 에너지나 대사산물들을 명확히 표현할 수 있다. 또한 반응의 촉매 역할을 하는 효소의 조절작용을 구분되게 나타낼 수 있어, 조절 경로에도 이 표기법을 적용시킬 수 있다.

또한 UniPath 표기법은 세포 내 위치 정보를 통해 주체의 형만으로도 기본적인 주체의 역할을 구분할 수 있다[그림 4]. 주체의 형에 대한 명시를 통해 특정 유전자와 그것이 발현된 RNA와 그것이 번역된 단백질은 -대소문자 구별을 원칙으로 하지만- 같은 이름으로 표현함으로써 초래되는 모호성을 사전에 방지하여, 같은 이름일지라도 서로 다른 형의 주체를 명료하게 구분할 수 있다[그림 5].

4.3 UniPath 시스템과 기존 시스템의 표기법 비교

기존 시스템에서 사용하는 표기법을 신진대사 경로와 조절경로 나누어 UniPath 표기법과 비교하였다.

표 2 신진대사 경로 표현을 위한 표기법 비교

	UniPath	ExPASy	KEGG	WIT	PathDB	UM-BBD
사용자에게 친근함	○	○	△	○	△	△
생화학 작용/조절 작용	○	○	○	○	○	△
에너지 대사/효소 작용	○	○	△	○	○	×
세포 내 위치 표현	○	×	△	○	×	×
추정 작용 구분	○	×	×	×	×	×
생략 작용 구분	○	×	△	×	×	×
표현 범위의 확장성	○	△	△	△	×	×

표 3 조절경로 표현을 위한 표기법 비교

	UniPath	GeneNet	ExPASy	TransPath	KEGG
사용자에게 친근함	○	○	○	○	△
생화학 작용/조절 작용	○	○	○	×	△
활성화 여부 표현	○	○	○	○	×
다양한 조절 작용의 구분	○	○	△	○	△
세포 내 위치 표현	○	○	×	○	△
추정 작용 구분	○	×	×	×	×
생략 작용 구분	○	×	×	×	×
표현 범위의 확장성	○	△	△	△	×

위의 평가 기준으로 표기법들을 비교한 결과, UniPath 표기법은 다양한 종류의 경로를 각 경로의 요구사항에 맞게 표현 할 수 있으며, 생물학자에게 중요한 정보인 세포 내 위치 정보와 아직 상용 시스템에서 적용되지 않은 추정작용과 생략작용을 표현함으로써 보다 BioPathway 지식을 명료하게 표현할 수 있음을 확인하였다.

5. 결론 및 향후 연구

본 논문은 세분화 되어있던 BioPathway 지식표현을 위한 UniPath 시스템을 제안하였다. 본 논문의 의의는 크게 세 가지로 다음과 같다.

첫째, UniPath 표기법은 간결하면서도 강력한 표현력을 가지며, 신진대사 경로와 조절 경로를 하나의 통합된 방식으로 표현함으로써, 세분화되어 있는 BioPathway 지식을 통합할 수 있다.

둘째, UniPath 표기법은 플랫폼 독립적으로 데이터를 통합하며 처리하기에 적합한 XML 포맷으로도 손쉽게 변환할 수 있다. XML 포맷은 복잡한 BioPathway 지식을 바이오 온톨로지를 바탕으로 표현하고 공유하기에 적합하다.

셋째, UniPath 표기법을 적용하여 BioPathway 지식을 그래프 형태로 편집하여 세분화되어있는 BioPathway 지식을 등록하고 프레임 형식으로 관리할 수 있는 그래픽 에디터를 설계하고 구현함으로써, BioPathway 지식을 쉽게 저장·수정하며 확장이 용이한 시스템의 프로토타입을 제시하였다.

결과적으로 UniPath 시스템은 복잡하고 체계화가 필요한 BioPathway 지식을 통합하고 보다 명료하게 표현함으로써 전산처리가 용이하도록 하였으며, 이를 통하여 자료 검색이나 데이터 마이닝 등의 정보 부가가치를 높일 수 있도록 하였다.

참고문헌

[1] BioPathway Org. http://www.biopathways.org
[2] EcoCyc. http://www.ecocyc.org
[3] GeneNet. http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet
[4] Voit, Eberhard, *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, pp.25-28, Cambridge press. 2000
[5] Pirson, Isabelle *et al.*, "The Visual Display of Regulatory Information and Networks", Trends in Cell Biology, vol.10, pp.404-408, Oct. 2000
[6] SBML. http://www.cds.caltech.edu/erato/sbml/
[7] KEGG. http://www.genome.ad.jp/kegg/
[8] Stevens, R., C. Goble & S. Bechhofer, "Ontology-based Knowledge Representation for Bioinformatics", Bioinformatics, vol.1, no.4, pp.398-414, 2000
[9] Fukuda, K., T. Takagi, "Knowledge Representation of Signal Transduction Pathway", Bioinformatics, vol.17, no.9, pp.829-837, 2001