바이오패스웨이를 위한 지식 표현 시스템
(UniPath: A Knowledge Representation System for Biopathways)

이 민 수 † 박 승 수 †† 강 성 희 †††
(Min Su Lee) (Seung Soo Park) (Sung Hee Kang)

요 약 최근 생물정보학의 발전과 함께 생물 관련 정보들이 기하급수적으로 증가하고 있다. 연구 대상도 DNA, RNA, 단백질에서 더 나아가 이들의 상호작용 및 조절 메커니즘에 의해 기능들이 어떻게 수행되는 지에 관한 바이오패스웨이까지 포함하게 되었다. 바이오패스웨이는 광대한 양의 정보를 포괄하며 구성체 사이의 유기적 관계를 나타내고 있는 것이므로 이를 컴퓨터로 처리하기 위해서는 보다 명료하며 직관적인 표현이 요구된다.

그러나 기존 시스템에서 사용하는 표기법들은 명료하게 해석될 수 없는 경우가 많고 표현 가능한 영역이 특정 한 단면에만 국한되어 있으며 같은 정보를 표현하여도 시스템마다 표현 레벨과 방식이 달라 시스템 확장 및 통합이 어려운 상황이다.

본 논문에서는 다양한 종류의 바이오패스웨이 지식을 체계적인 단일 표기법을 사용하여 보다 명료하고 효율적으로 표현하며 단일화되고 통일된 UniPath 표기법을 제안하였다. 또한 이 표기법을 적용하여 바이오패스웨이 지식을 그래프 형태로 편집함으로써 그 정보를 등록하며 XML 포맷으로 쉽게 변환할 수 있는 프레임 기반 지식 표현 시스템을 설계하고 실제 데이타에 적용함으로써 타당성을 검증하였다.

키워드 : Biopathways, 신진대사 경로, 조절 경로, 지식 표현, XML

Abstract Recently, the information processing of ever increasing bio-related data is becoming a very important issue. One of the main sources of these bio-data comes with the form of biopathways, which includes molecular transactions and processes that are part of biochemical systems. The information represented by biopathways includes various organic relations among its components. However, most of the current systems to represent biopathways have been initially developed without computer processing in mind, and hence suffer from inconsistencies and ambiguities.

In this paper, we propose an improved notation, called UniPath, for clear and systematic representation of biopathways. The proposed system is designed to provide a unified representation of metabolic and regulatory pathways.

We also designed and implemented a graphic editor for UniPath to draw biopathways map according to the proposed notation. The graphic editor is designed so that biopathway data can be easily transformed into XML format.

Key words : Biopathways, Metabolic Pathway, Regulatory Pathway, Knowledge Representation, XML

1. 서 론

생물학과 정보통신이 융합된 생물정보학(바이오인포매틱스) 분야에서 생물 관련 지식에 정보통신 기술 및 이론들을 효과적으로 적용하기 위해서는 적절한 형태의 지식 표현 방법을 개발하는 것이 선행되어야 한다. 특히 기하 급수적으로 증가하며 매우 복잡하고 긴밀하게 연관되어 있는 생명 현상에 관한 지식들을 체계적으로 관리하기 위한 시스템을 개발하거나 데이타마이닝, 혹은 추론 등의 인공지능 기법을 적용하고자 할때에 그 중요성이 더욱 강조된다.

바이오패스웨이(biopathways)란 생화학 신체 조직 기관 안의 모든 형태의 분자적 상호작용들과 프로세스들을 포함하는 포괄적인 의미의 용어이다[1]. 바이오패스웨이를 규명함으로써 생체 시스템 상에서 유기적으

† 종신회원 : 이화여자대학교 컴퓨터학과
ssue@ewha.ac.kr
†† 정 회 원 : 이화여자대학교 컴퓨터학과 교수
sspark@ewha.ac.kr
††† 비 회 원 : 명지대학교 교육학습개발원 교수
kangsh@mju.ac.kr
논문접수 : 2003년 3월 7일
심사완료 : 2003년 12월 17일

로 복잡하게 얽혀있는 분자들의 상호작용에 기반한 다양한 생화학 경로들에 의해 생명 현상을 유지하기 위한 기능들이 어떻게 수행되는지에 관한 연구를 수행할 수 있다.

기존에는 실험실의 실험에 의존하여 연구를 진행해 왔기 때문에 밝혀진 지식이 제한적이고 국지적이기 때문에 이들을 통합하여 바이오패스웨이에 관한 전체적인 메커니즘을 다룰 수 없었다. 따라서 바이오패스웨이를 신진대사 경로, 신호전달 경로, 단백질 상호작용 등과 같은 특정 영역으로 나누어 독립적으로 연구하여 왔다.

그러나 실제로 이러한 다양한 바이오패스웨이 작용들은 그 경계를 정확히 명시할 수 없을 만큼 하나의 전체 생체 시스템 안에서 서로 자연스러우면서 긴밀하게 연결되어 있다. 따라서 지금까지 세분화되어 연구되어왔던 바이오패스웨이 지식을 통합하여 지식 기반 시스템을 구축하고 이를 기반으로 다양한 컴퓨터·통계 알고리즘과 기법들을 적용함으로써 의학, 약학, 농학 등에 필요한 실용적 지식을 이끌어낼 수 있을 것이다.

그러나 기존의 시스템들은 대부분 생물학자들의 필요에 의해 조금씩 확장된 상향식(bottom-up) 방법으로 설계되어 연구자에 따라 지식 표현의 범위와 방법, 그리고 입도(granularity)가 다르며 지식이 명료하게 해석되기 힘든 경우가 많다. 또한 바이오패스웨이 정보는 사람이 쉽게 이해할 수 있을 뿐만 아니라 컴퓨터에 의해 쉽게 처리될 수 있도록 모델링 되어야 하는데 현재 제공되고 있는 대부분의 시스템들은 바이오패스웨이 지식을 이미지 맵(image map)형태로 제공하고 있기 때문에 기존의 지식들을 확장하거나 통합하고 데이타마이닝 기법을 적용하는 등의 컴퓨터 작업에 적용시키기가 매우 어렵다.

본 논문에서는 바이오패스웨이와 관련된 지식들을 프레임(frame) 형식으로 관리하려 할 때 각 프레임들 사이의 관계를 정립하고 이를 체계적으로 그래프 형태로 표현하기 위한 단일 표기법을 설계하고 이를 활용하는 방안을 제시하고자 한다. 이를 위해 기존에 생물학 분야에서 여러 가지 형태로 제안되었던 표현 방법에 대해 분석하였으며, 이를 바탕으로 여러 영역의 바이오패스웨이를 통합하여 표현할 수 있고 간결·명확하며 컴퓨터 처리에 적합한 UniPath 표기법을 제안하였다. 또한 이 표기법을 적용하여 바이오패스웨이 지식을 그래프 형태로 편집하며 프레임 구조로 관리하고 등록된 지식을 XML 포맷으로 변환할 수 있는 시스템을 설계하고, 이를 위한 그래픽 에디터를 구현하였다.

이와 같이 통합된 바이오패스웨이 지식을 이용하여 세포에서 서로 긴밀하게 연관된 경로 정보들을 전체적으로 보다 명료하게 표현할 수 있고, 결과적으로 생체 시스템 상에서 다양한 생화학 경로들에 의해 생명 현상을 유지하기 위한 기능들이 어떻게 수행되는지에 관해 다양한 알고리즘 및 기술들을 적용하여 보다 깊이있는 연구를 진행할 수 있다. 또한 그래픽 에디터 형태의 시스템을 통해 새로운 지식을 빠르게 업데이트할 수 있으며, 프레임 구조로 지식을 관리함으로써 컴퓨터 프로세싱이 용이해질 수 있다.

본 논문은 2장에서는 생물정보학과 바이오패스웨이, 그리고 바이오패스웨이 지식 표현 방법들에 대해 살펴보고, 3장에서 세분화되어 있는 바이오패스웨이 지식을 단일 표기법으로 표현할 수 있는 UniPath 시스템을 설계하고 4장에서 UniPath 시스템을 적용함으로써 검증한다. 5장에서는 UniPath 표기법과 시스템에 관해 논하고, 마지막으로 6장에서 본 논문의 결론과 향후 연구 방향을 제시한다.

2. 관련연구

2.1 생물정보학과 바이오패스웨이

최근의 생물정보학은 다양한 생물 종에 있어서 유전체의 염기서열을 해독하는 인간 게놈 프로젝트의 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. 이를 위하여 서열을 해독하는 유전체학(genomics), 유전자의 발현을 연구하는 전사체학(transcriptomics), 단백질에 대한 연구인 단백체학(proteomics), 그리고 그들 서로의 상호작용을 총체적으로 이해하고자하는 systems biology에 대한 활발한 연구가 진행되고 있다.

모든 생물학적 기능은 분자 상호작용의 네트워크를 통해 발현되며, 분자 상호작용에 대한 정보는 추상화된 상위 수준에서의 생물학적 기능 분석을 가능하게 하므로 개개 분자에 대한 정보 못지않게 중요하다. 따라서 생화학 신체 조직 기관 안의 모든 형태의 분자적 상호작용들과 프로세스들을 나타내는 바이오패스웨이에 대한 연구는 생명 현상의 신비를 해독하기 위해 필수적이라 할 수 있다[1].

2.2 바이오패스웨이 지식 표현 방법

바이오패스웨이관련 지식을 보다 효과적이고 적절하게 표현하기 위해 다양한 표현 방법들이 사용되고 있다. 표 1은 상용 시스템에서 바이오패스웨이 지식을 표현하기 위해 사용하는 대표적인 방법들을 보여준다. 각 표현 방법마다 장점과 그에 따른 용도가 다르므로 대부분의 시스템들은 한 가지 표현 방법에만 국한하지 않고, 다양한 형식으로 바이오패스웨이 정보를 제공하고 있다.

표 1 바이오패스웨이의 다양한 표현 방법

주요 표현 방법	시스템 예
상호작용의 순서 리스트	BIND
Petri Net	Genomic Object Net
Markup Languages	XML(CellML, SBML), ASN.1(NCBI)
분자 상호작용 그래프	GeneNet, KEGG, PathDB

2.2.1 상호작용의 순서 리스트

바이오패스웨이는 특정 세포 기능을 조정하는 상호작용들의 네트워크이므로, BIND(The Biomolecular Interaction Network Database)에서는 특정 경로를 그것을 이루는 상호관계의 ID를 순서대로 나열함으로써 표현한다[2]. 바이오패스웨이를 상호관계의 순서 리스트로 표현하는 것은 그것을 이루는 상호관계에 관한 정보를 얻기에는 편리하지만, 상호관계 ID의 나열로는 분기되거나 수렴되는 경로 흐름 등을 표현할 수 없어 전체적인 경로의 흐름을 파악하기 힘들다는 단점이 있다.

2.2.2 Petri Net

시간과 장소 정보에 따른 바이오패스웨이의 흐름을 시뮬레이션하기 위한 세부사항을 쉽게 관리하기 위해 Genomic Object Net은 Petri Net의 확장 형태인 HFPN(Hybrid Functional Petri Net)을 사용한다[3]. 바이오패스웨이를 표현하기 위해 Petri Net을 사용하는 것은 각 상호작용에 생화학적 지식에 근거한 변이 함수를 적용시킬 수 있고 각 객체의 상태를 명료하게 구분할 수 있으나, 생화학자가 이해하기에는 표현 방법이 익숙하지 않다는 단점을 안고 있다.

2.2.3 Markup Languages

바이오패스웨이 지식 표현을 위해 사용되는 Markup Language로는 ANS.1과 XML 포맷이 있다. Markup Language를 사용하는 것은 바이오패스웨이 지식을 온톨로지나 분류 체계에 기반하여 표현할 수 있으며, 플랫폼 독립적으로 데이타를 전송하고 처리하기에 매우 용이하므로 시스템의 확장 및 연계를 위해 필수적이라 할 수 있다. 최근에는 바이오패스웨이를 위한 표준적인 XML 포맷에 관한 연구가 활발히 이루어지고 있으며, 대표적인 것으로는 SBML(System Biology Markup Language)과 CellML 등이 있다[4,5].

2.2.4 분자 상호작용 그래프

바이오패스웨이를 분자 상호작용 그래프로 표현하는 것은 시각적으로 경로 지식을 표현하므로 유기적이며 복잡한 바이오패스웨이를 쉽게 이해할 수 있는 좋은 방법이다. 또한, 최종 사용자인 생화학자에게 익숙하다는 장점을 가지고 있어 KEGG, GeneNet, PathDB 등 거의 대부분의 바이오패스웨이 시스템에서 사용된다[6-8].

바이오패스웨이를 그래프 형태로 나타내기 위한 표기법은 경로의 흐름을 명확하게 이해할 수 있도록 직관적(intuitive)이어야 하고, 객체간의 관계의 복잡도에 따라 적합(fitting)하게 디자인되어야 하며, 한가지 표기법으로 모든 종류의 네트워크를 나타낼 수 있도록 표현력(expressive)이 강해야 하고, 표기법과 그 해석 사이의 관계가 정형적(formal)으로 정의되어야 한다. 또한, 연구 분야 확대에 의한 바이오패스웨이 지식 범위의 확장을 고려하여 확장성있게(extensible) 디자인해야 한다.

신진대사 경로는 조절 경로에 비해 그 메커니즘이 간단하므로 그래프로 표현하기가 상대적으로 쉽다. 그러나 조절 경로는 널리 허용되는 분류법(taxonomy)이나 대표적인 경로 지도가 없고 조절 경로를 전체적으로 이해할 수 있는 메커니즘이 확립되지 않은 상태이다[9]. 따라서 시스템마다 고유 목적에 기반을 둔 표기법을 사용하여 바이오패스웨이 지식을 제공하므로 시스템마다 표현 방법과 레벨이 다르다. 또한 표현 범위도 개별적인 바이오패스웨이 영역으로 제한함으로 인해 시스템 확장 및 통합이 어렵다.

바이오패스웨이 정보를 단일한 표기법으로 표현하기 위해 제안된 대표적인 형식적 표기법(formal notation)으로 Eberhard Voit의 신진대사 경로에 관한 표기법[10]과 Kurt Kohn의 단백질 상호작용에 관한 표기법[11], 그리고 Isabelle Pirson의 조절작용에 관한 표기법[12]이 있다. 기존에 제안된 바이오패스웨이 지식 표현을 위한 표기법은 그 범위가 제한되어 있으므로 전체적으로 생체 내에서 서로 긴밀하게 연결되어 있는 경로 정보를 함께 표현하고자 할 때에는 한계에 부딪힐 수 밖에 없다.

2.3 그래프 기반 바이오패스웨이 지식 표현 시스템

바이오패스웨이 지식을 제공하는 기존 시스템은 경로를 그림 지도(hand-drawn image map) 형식으로 제공하거나 데이타베이스안의 정보에 그래프 레이아웃 알고리즘을 적용하여 자동으로 생성된 지도(auto-generated map)를 제공한다.

수작업으로 만든 그림 지도는 각 경로의 의미를 이해하고 있는 사람이 작성하므로, 직관적으로 이해하기 쉽게 경로 정보를 높은 밀도로 잘 정리하여 제공한다는 장점이 있다. 그러나 이 방법은 고정적이며 기존에 정의된 경로만을 보여주므로, 새롭게 밝혀지는 지식들을 업데이트하기 힘들고, 경로 정보를 검색하거나 새로운 가치있는 정보를 얻기 위해 다양한 데이타마이닝 기법을 적용시키기 힘들다는 단점을 안고 있다. 또한 경로 지도를 만드는 정형적인 방법이 없으므로 표현상의 불일치

성이 발생할 수 있다.

그래프 레이아웃 알고리즘을 사용하여 자동으로 경로 지도를 생성하는 경우는 데이타베이스에 쿼리를 적용하여 즉각적으로 경로 지도를 생성할 수 있으며 전문가의 수작업이 필요하지 않다는 장점이 있다. 그러나 경로 지도가 자동 생성됨으로 인해 경로 지도가 직관적이지 않을 수 있다는 단점이 있다.

3. UniPath 시스템 설계 및 구현

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

본 장에서는 세분화 되어있는 바이오패스웨이 지식을 프레임 형식으로 통합하여 관리하려 할 때 각 프레임들 사이의 관계를 체계적으로 그래프 형태로 표현하기 위한 UniPath 표기법을 제안한다. 그리고 UniPath 표기법을 적용한 그래픽 에디터 형태의 지식 표현 시스템을 통해 입력받은 경로 정보를 프레임 구조로 관리하고 XML 포맷으로 호환하는 방법에 대해 설명한다.

3.1 UniPath 표기법 개요

UniPath 표기법은 복잡한 바이오패스웨이 지식을 특성에 맞게 적절하게 표현하며, 필수요소만 적용시킨 단일 표기법으로 신진대사, 유전자 조절, 신호 전달 등 모든 종류의 경로에 관한 정보를 생물학자들에게 친근한 표기법에 기반하여 간결하게 표현함으로써, 바이오패스웨이 지식을 컴퓨터상에서 처리하기 용이하도록 하는 것을 목표로 하였다.

UniPath 표기법은 상호작용이 일어나는 장소를 나타낼 수 있도록 세포 내 위치 정보를 표현할 수 있다. 또한, 표 2와 같은 신진대사와 조절 경로의 주체의 형(type: DNA, RNA, Protein, ...)과 다양한 주체 사이의 관계를 표현할 수 있다.

주체의 형이나 세포 내 위치에 의해 직관적으로 구분 가능한 주체 사이의 관계는 세분화하지 않고, 기본 표기법만으로 다양한 경로를 간결하게 표현할 수 있도록 디자인하였다. 또한 생략된 부분이 알려져 있는 부분인지, 아직 밝혀지지 않은 부분인지 구분할 수 있으며, 실험에 의해 명백히 밝혀진 관계가 아닌 컴퓨터의 계산에 의해 추정되는 부분들도 구분되게 명시적으로 표현함으로써 보다 지식 표현에 있어서의 불확실성과 모호성을 완화시켰다.

3.2 UniPath 표기법 구성 및 설계

UniPath 표기법은 바이오패스웨이 지식을 작용이 일어나는 세포 내 위치 정보를 표현해주는 '세포 내 위치

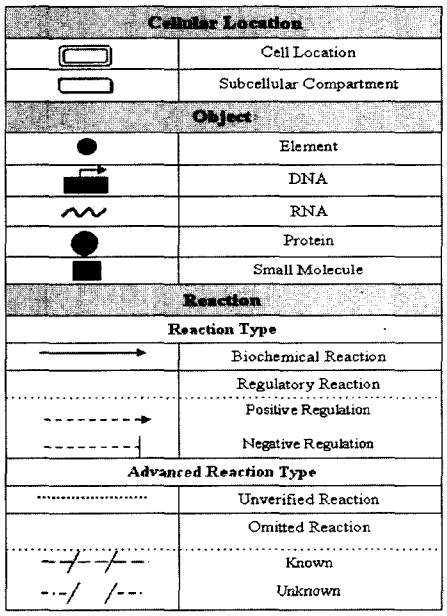


그림 1 UniPath 표기법

표 2 신진대사 · 조절 경로의 표현 주체와 작용

		신진대사 경로	조절 경로	
주체		반응물(기질, 생성물), 효소	원소, DNA, RNA, 단백질, 작은 분자	
작용	생화학 작용	화학적 변환	연결(bind), 연결 해제(release), 인산화(phosphorylate), 비인산화(dephosphorylate), 상태의 변화(state change), 장소의 변화(translocate)	
	조절 작용	효소의 촉매 작용	양적 조절 작용	활성화(activation), 증가(increase), 스위치 켜기(switching on), ...
			음적 조절 작용	억제(inhibition), 감소(decrease), 스위치 끄기(switching off), ...

정보', 그래프의 노드에 해당하는 '주체의 형(object type)', 그리고 에지에 해당하는 '주체 사이의 작용(reaction between objects)'으로 나누어 그림 1과 같이 정의된다. 이와 같은 세가지 요소를 사용하여 간결하게 표현함으로써, 각 작용이 일어나는 세포 내 위치와 주체의 형, 그리고 작용의 형태에 기반하여 문맥에 맞게 주체 사이의 작용과 각 주체의 역할을 문맥에 맞게 해석할 수 있다.

3.2.1 세포 내 위치 정보

바이오패스웨이 상에서 각각의 상호작용이 세포의 어느 위치에서 일어나는지에 관한 정보는 생물학자에게 상당히 중요하다. 이를 위해서, 세포 간 영역을 표현해 주는 이중막 구조인 '세포막'과 핵이나 미토콘드리아와 같은 세포 내 영역을 표현하는 단일막 구조인 '세포 내 세부 영역'을 정의함으로써 상호작용의 위치 정보를 시각적으로 표현할 수 있도록 하였다.

UniPath 시스템을 사용하여 표현한 인슐린의 조절작용-그림 2를 살펴보면, 세포 외부에 있는 단백질인 인슐린(insulin)이 세포 내부에 인슐린 수치가 높아졌다는 신호를 보내주는 호르몬이라는 것, 세포막 위의 단백질들은 외부 신호를 받아들여 세포 안으로 전달하는 리셉터(receptor) 역할을 한다는 것, 그리고 핵 안의 단백질은 전사 요소(transcription factor)로서 핵 안의 일련의 DNA를 활성화 시킨다는 것 등을 알 수 있다. 이와 같이, 상호작용이 일어나는 장소인 세포 내 위치 정보에 대한 표현은 가독성과 직관성을 향상시켜 사용자의 이해를 도우며 위치 정보에 의해 명백하게 구분 가능한 표현 주체나 작용의 세분화를 막아준다는 장점이 있다.

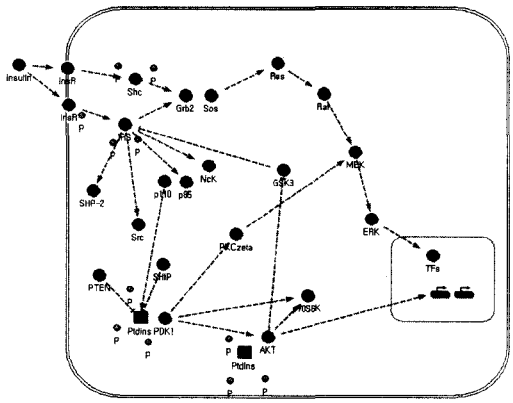


그림 2 UniPath로 표현한 인슐린 조절작용

3.2.2 주체의 형

표현 주체를 정의함에 있어, 주체의 형만 고려하고 주체의 경로 안에서의 역할은 배제하였다. 그림 2에서 볼 수 있듯이, 세포 내 위치 정보와 주체의 형에 의해 기본적인 주체의 역할을 구분할 수 있다. 주체의 기능에 대한 구분이 필요한 경우에는, 필요에 따라 특정 색상지도(color map)를 적용하여 해당 노드의 색을 바꿔주는 등의 시스템 차원의 부가 서비스를 제공함으로써 주체의 기능을 표현하는 것이 가독성과 직관성을 향상시킨다. 주체의 형은 EcoCyc의 온톨로지에 기반하여 선별하고 [13], GeneNet에서 사용하는 주체 표기법을 참고하여 디자인하였다[7].

생물연구자들은 특정 유전자와 그것이 발현된 RNA와 그것이 번역된 단백질은 -대소문자 구별을 원칙으로 하지만- 같은 이름으로 표현한다. 따라서 한 가지 이름이 여러 상태로 해석할 수 있다는 모호성을 내제하고 있다. UniPath 표기법은 주체의 형을 구분함으로써, 같은 이름일지라도 그것이 어떤 상태인지 직관적으로 알 수 있게 해준다. 그림 3은 UniPath를 사용하여 세포주기 중 DNA가 복제되는 S기의 일부를 표현한 것이다. 핵 안의 단백질들이 특정 유전자를 발현시키기 위해 전사 요소로서 어떻게 작용하며 그에 따라 어떤 유전자들이 RNA로 전사되는지를 명료하게 파악할 수 있다.

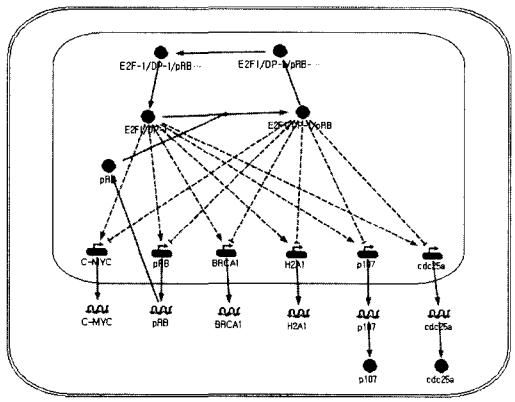


그림 3 UniPath의 세포주기 표현

3.2.3 주체 사이의 작용

주체들 사이의 작용은 주체의 형과 세포 내 위치 정보에 의해 구분 가능한 에지들의 불필요한 세분화를 막기 위해 Voit[10]와 Pirson[12]이 제안한 표기법들 중 필수 요소들만을 선별하여 정의하였다. 작용의 종류를 신진대사 경로와 조절 경로 모두에 적용될 수 있는 '생화학 작용'과 생화학 작용에 영향을 주는 '조절 작용'으로 양분하였다. 조절작용은 '양적 조절 작용'과 '음적 조절 작용'으로 나누어 양적 조절 작용은 녹색 점선 화살표로, 음적 조절 작용은 적색 점선 막대 선으로 구별하여 직관성을 높였다. 그림 3에서 핵 안의 단백질들이

DNA를 조절하는 작용 중 양적 조절작용은 특정 DNA의 발현 촉진(switch on) 역할을 하며, 음적 조절작용은 발현 억제(switch off)역할을 한다는 것을 문맥상 알 수 있듯이, 각 작용은 주체의 형과 세포 내 위치에 따라 세분화하여 해석될 수 있다.

그림 4는 UniPath를 사용하여 다양한 조절작용을 표현한 예이다. 여기에서 겹세로줄은 세포막을 의미하고 세로줄은 핵막을 의미한다. UniPath 표기법의 기본 작용 형태, 즉 생화학 작용과 양적·음적 조절 작용의 세 가지 작용만으로는 다양한 생체 내의 작용들을 표현하기에 무리가 있다. 그러나 UniPath 표기법은 이 세가지 작용과 함께 '세포 내 위치 정보'와 '주체의 형'을 함께 표현함으로써 주체의 형과 세포내 위치 정보에 기반하여 문맥에 맞게 다양한 작용들을 간결하면서 직관적으로 표현할 수 있다.

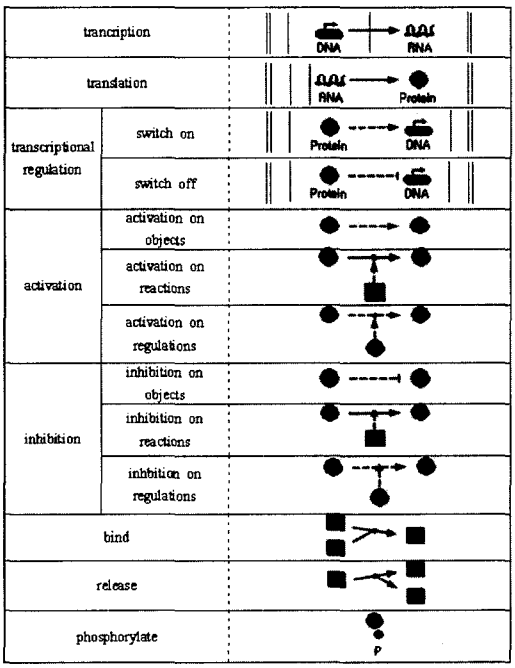


그림 4 UniPath로 표현한 조절작용의 표현 예

예측 알고리즘이나 데이타 마이닝 등의 기법을 통해 얻은 데이타는 가치 있지만 실험적으로 검증되지 않았기 때문에 오류를 포함하고 있을 수 있다. 또한 경로 지도상에 생략된 부분이나 아직 명확하게 밝혀지지 않은 부분들을 구분되게 표현하지 않으면 표현 레벨의 불규칙함(granularity)과 경로 지도의 잘못된 해석을 초래할 수 있다. 그러나 아직까지 이러한 부분들을 고려하며 바이오패스웨이 지식을 제공하는 시스템이 없다. UniPath 표기법에서는 이러한 추정된 부분을 명시할 수 있으며, 경로 지도상에 생략된 부분이 있다면 그 부분이 현재 밝혀져 있는 부분인지 아직 밝혀지지 않은 부분인지를 구분할 수 있도록 하였다. 추정 작용과 생략 작용은 생화학 작용과 조절 작용에 해당하는 화살표의 점선의 종류를 바꿔줌으로써 적용시킬 수 있다. 생략된 부분에 대한 구분된 표현과 추정된 작용을 구분하기 위한 표기법의 사용 예는 그림 5와 같다.

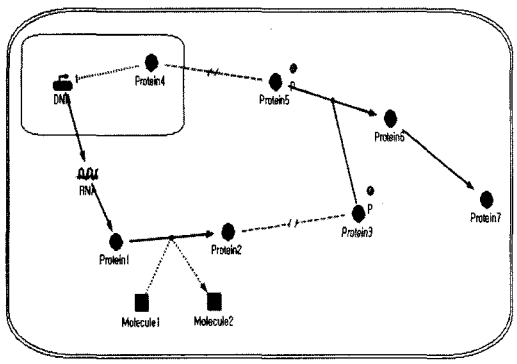
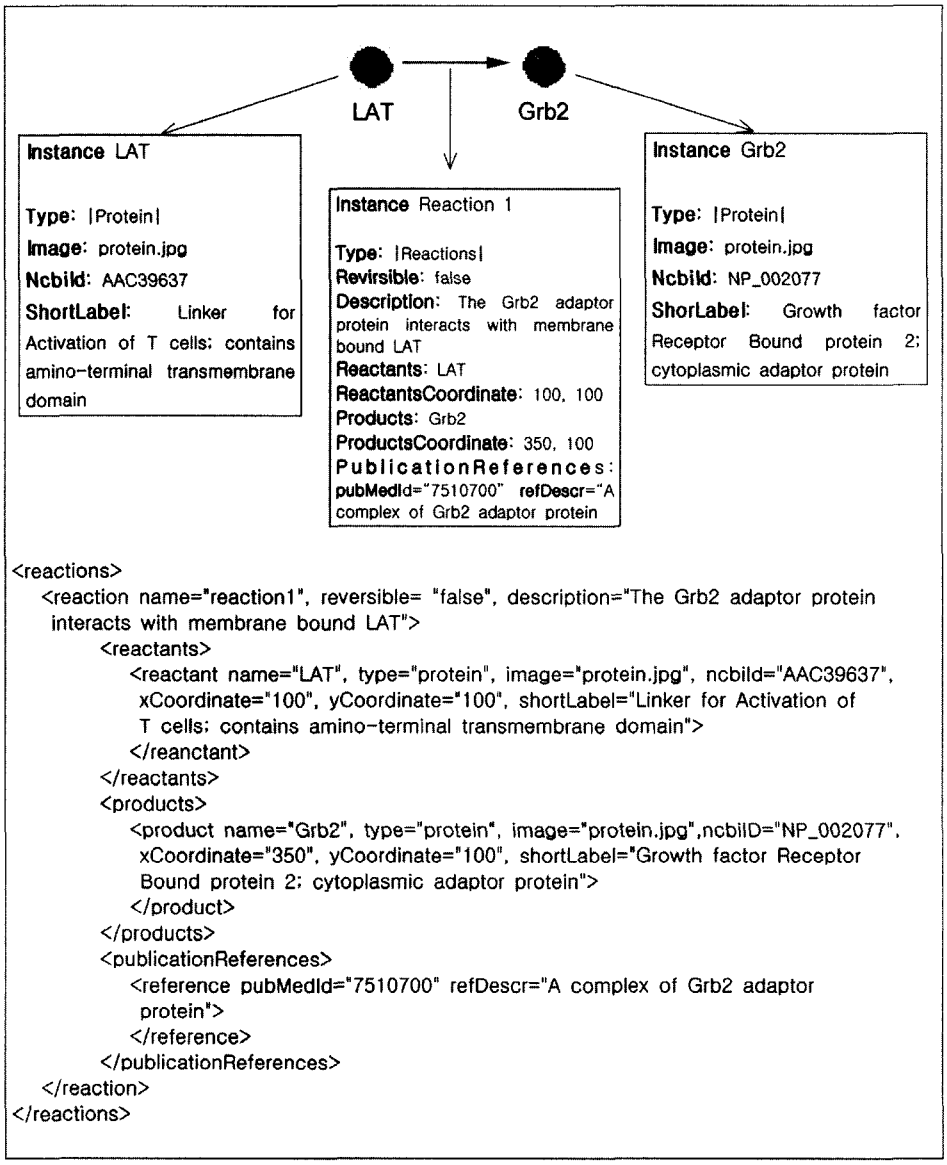


그림 5 생략·추정 작용 표현의 예

3.3 Frmae과 XML 포맷으로의 바이오패스웨이 지식 표현

UniPath 시스템에서는 그림 지도를 사용한 방법과 경로지도를 자동 생성하는 방법의 장단점을 고려하여, UniPath 표기법을 적용시킨 경로 지도 편집기를 구현하였다. 그래픽 에디터를 사용하여 바이오패스웨이를 편집하고 세부정보를 입력함으로써 지식을 온톨로지에 기반하여 프레임 형식으로 데이타를 관리하도록 하여, 경로 정보를 기본적인 상호작용의 네트워크로서 인식할 수 있으며 새로운 지식을 즉각적으로 업데이트할 수 있도록 하였다. 그럼으로써, 복잡한 바이오패스웨이 정보를 컴퓨터에서 처리되기 용이한 형태로 관리하며, 단편적인 바이오패스웨이 정보를 확장시키고 서로 연결하여 효과적으로 처리할 수 있다. 또한 바이오패스웨이 지식을 프레임 형식으로 관리하고 그래프 형태로 표현함으로서 바이오패스웨이의 효과적인 검색 및 추론 과정을 가능하게 하였으며, 복잡한 지식을 구조적으로 관리하며 세분화된 바이오패스웨이 지식을 통합하는 것을 보다 용이하게 하였다. UniPath 시스템은 Windows 2000 운영체제에서 Visual C++ 6.0와 MS SQL Server 2000을 사용하여 개발하였다.

에디터를 통해 입력받은 바이오패스웨이와 입력받은 주체와 관계의 세부 정보는 단백질 상호작용에 관해 명세한 BIND 데이타 모델에 기반하여 관리된다[2]. 즉,




```
<reactions>
   <reaction name="reaction1", reversible= "false", description="The Grb2 adaptor protein
    interacts with membrane bound LAT">
        <reactants>
           <reactant name="LAT", type="protein", image="protein.jpg", ncbild="AAC39637",
            xCoordinate="100", yCoordinate="100", shortLabel="Linker for Activation of
            T cells; contains amino-terminal transmembrane domain">
           </reanctant>
        </reactants>
        <products>
           <product name="Grb2", type="protein", image="protein.jpg",ncbilD="NP_002077",
            xCoordinate="350", yCoordinate="100", shortLabel="Growth factor Receptor
            Bound protein 2; cytoplasmic adaptor protein">
           </product>
        </products>
        <publicationReferences>
           <reference pubMedId="7510700" refDescr="A complex of Grb2 adaptor
            protein">
           </reference>
        </publicationReferences>
   </reaction>
</reactions>
```

그림 6 프레임과 XML 포맷으로 표현한 상호작용의 예

경로에 관한 정보는 경로를 구성하는 각 주체와 주체사이의 작용들에 관한 세부작용들을 명세한 상호작용 프레임의 리스트로서 표현될 수 있다.

BIND 데이타 모델을 사용하여 프레임 형식으로 구축한 바이오패스웨이 지식은 작용에 관한 설명은 속성(attribute)으로, 관련되어있는 주체들은 요소(element)로 표현함으로써 쉽게 XML 포맷으로 변환시킬 수 있다. 한 예로서, UniPath로 표현한 LAT와 Grb2의 상호작용은 그림 6과 같은 프레임과 XML 포맷으로 변환시킬 수 있다.

4. UniPath 시스템 적용 및 평가

4.1 UniPath 시스템 적용

UniPath 표기법을 검증하기 위해 신진대사 경로와 조절 경로에 해당하는 실제 바이오패스웨이 경로 데이타를 구현한 UniPath 시스템을 사용하여 편집하고 상용 시스템의 표기법과 비교하였다.

- **신진대사 경로: 해당과정(Glycolysis)**

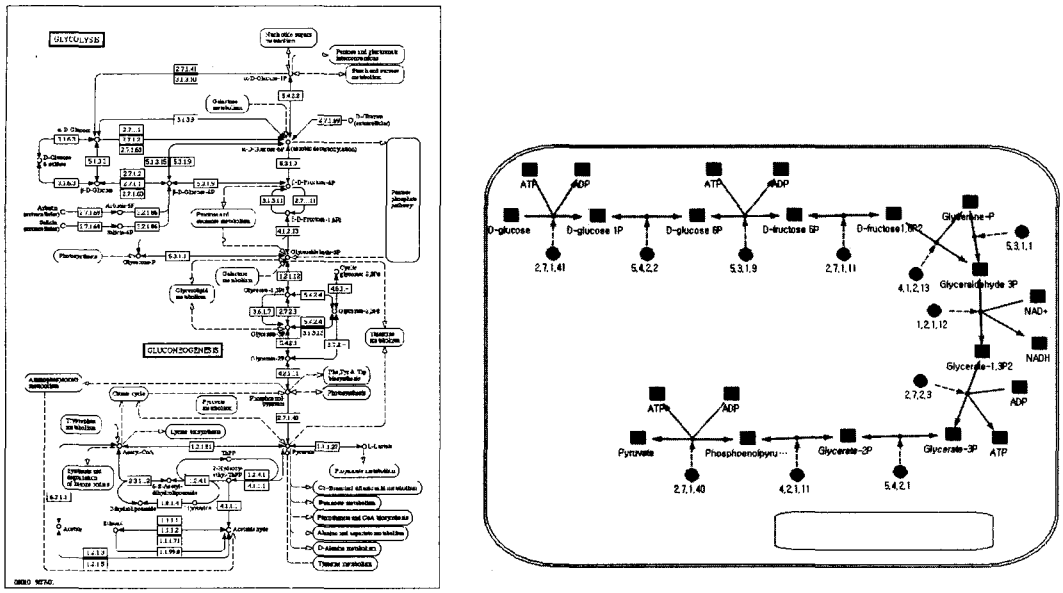


그림 7 KEGG와 UniPath의 해당과정 경로 지도

그림 7은 KEGG에서 제공하는 해당과정 그림 지도와 UniPath로 편집한 해당과정 경로 지도이다. KEGG는 경로의 초점을 촉매 역할을 하는 효소의 흐름에 맞추어서 해당과정의 핵심 흐름인 반응물과 생성물은 상대적으로 작게 표현하며, 해당과정의 부산물은 표현하지 않는다. UniPath는 생물학자가 쉽게 접하는 경로 그림의 형태를 띄고 있으며, 해당작용에 있어서의 주체의 형과 부산물들을 명확히 구분할 수 있다. 또한 반응의 촉매 역할을 하는 효소의 작용을 구분되게 표현할 수 있고, 해당작용이 일어나는 장소가 세포질이라는 것을 직관적으로 알 수 있다.

- 조절 경로: 콜레스테롤(cholesterol) 조절 경로

그림 8은 UniPath 시스템으로 편집한 콜레스테롤 조절 메커니즘에 관한 경로 지도이다. 간단한 주체의 형과

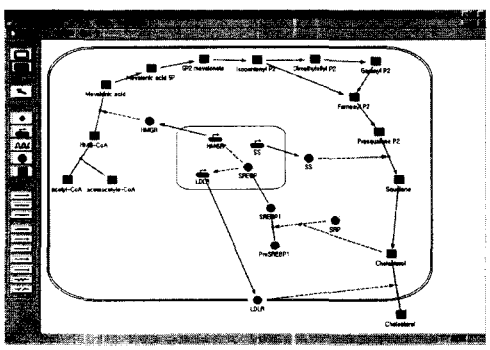


그림 8 UniPath로 표현한 콜레스테롤 조절 경로

작용의 종류만으로 각 상호작용의 의미를 파악할 수 있으며, 세포 내 위치 정보 구분이 명확하여 콜레스테롤 조절 사이클의 전반적인 흐름이 직관적이며 명료하게 표현된다.

4.2 UniPath 시스템과 기존 시스템의 표기법 비교

기존 시스템에서 사용하는 표기법을 UniPath 표기법과 비교하였다. 각 시스템마다 지식 표현 레벨과 제공하는 경로 정보의 영역이 다르므로 절대적인 비교는 사실상 불가능하다. 또한 대부분의 조절 경로는 노드와 에지를 가진 그래프 형태가 아닌 그림 형태가 대부분이다. 따라서, 경로 지도 중 그래프 형태인 것만을 대상으로 하고 비교 기준을 크게 신진대사 경로와 조절 경로로 나누어 다음과 같은 기준에 의해 각 표기법의 표현력을 비교해 보았다.

신진대사 경로를 표현하기 위해서는 우선 실질적 경로 정보 사용자인 생화학자에게 친근한 표기법을 사용해야 하며, 생화학 작용과 생화학 작용에 영향을 미치는 조절 작용을 명료하게 구분하여 표현해야 한다. 또한 물질 대사 과정과 맞물려 일어나는 에너지 대사와 조효소 작용을 함께 표현해야 한다. 그리고 작용이 일어나는 세포 내 위치 정보를 표현해주어야 하며, 추정 작용과 생략 작용을 구분해 주어야 한다. 또한 경로 정보의 확장 및 연계를 위해 표기법의 표현 범위를 고려하였다.

조절 경로 표기법의 평가 기준은 신진 대사 경로에 적용되는 기준 외에, 다양한 형태의 조절 작용의 표현력 여부와 단백질의 활성화 표현 여부를 더 고려하여 주어야 한다.

표 3 신진대사 경로 표현을 위한 표기법 비교

	UniPath	ExPASy[14]	KEGG[6]	WIT[15]	PathDB[8]	UM-BBD[16]
사용자에게 친근함	○	○	△	○	△	△
생화학 작용/조절 작용	○	○	○	○	○	△
에너지 대사/효소 작용	○	○	△	○	○	×
세포 내 위치 표현	○	×	△	○	×	×
추정 작용 구분	○	×	×	×	×	×
생략 작용 구분	○	×	△	×	×	×
표현 범위의 확장성	○	△	△	△	×	×

표 4 조절경로 표현을 위한 표기법 비교

	UniPath	GeneNet[7]	ExPASy[14]	TransPath[17]	KEGG[6]
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	○	△	△	△	×

5. 논 의

UniPath 표기법은 바이오패스웨이 정보의 실제적 사용자인 생물학자들에게 보다 간결하고 직관적인 경로 정보 서비스를 제공하며, 다양한 경로를 단일 표기법을 사용하여 표현할 수 있도록 설계되었다. 동시에 그 지식을 컴퓨터에서 효과적으로 활용할 수 있도록 조직적으로 설계되었다.

바이오패스웨이 분야에 대한 통일된 온톨로지가 아직 완벽하게 정립되지 않은 상황이고, 각 시스템에서 사용하는 표기법들이 서로 다르고 경로 정보의 표현 레벨도 다르다. 따라서, UniPath 표기법이 다른 모든 종류의 표기법들 전체를 모두 포괄하는 초집합(superset)이라는 것이나 다양한 모든 종류의 바이오패스웨이 지식을 표현할 수 있다는 것을 명확하게 증명할 수는 없어, 우리는 사용자에게 제공되는 서비스 측면에서 표 3, 표 4와 같은 표기법 비교를 수행하였다. 그 결과 UniPath 표기법은 신진대사 경로와 조절 경로 모두 각 경로의 요구사항에 맞게 표현 할 수 있으며, 생물학자에게 중요한 정보인 세포 내 위치 정보와 아직 상용 시스템에서 적용되지 않은 추정작용과 생략작용을 표현해 줌으로써 보다 바이오패스웨이 지식을 명료하게 표현할 수 있음을 확인하였다.

UniPath 표기법을 적용한 그래픽 에디터 형태의 UniPath 시스템을 사용하여 편집한 정보는 EcoCyc 온톨로지에 따라 각 생물학적 주체들을 나타내는 프레임들과 그들 사이의 의미적 관계를 표현하는 프레임들로써 표현된다. 그래픽 에디터를 통해 지식 베이스 관리를 함으로써 새로운 지식이 들어왔을 때 즉각적으로 미리 정의된 규칙에 의해 간단한 형태의 오류 체크를 수행할 수 있다. 또한 프레임 기반 지식 표현 방법을 채택함으로써, 주체의 온톨로지에 기반하여 각 주체에 관련된 데이타들을 보다 체계적으로 관리할 수 있다. 또한 프레임들 사이의 링크 정보와 주체의 기능적 온톨로지를 이용하여 기능 기반 질의 기능을 제공하여 특정 경로 정보만 추출하거나 특정 특징을 만족하는 경로를 예측할 수 있다.

UniPath 시스템은 그림 지도 형식의 직관적이고 이해하기 쉽다는 장점과 그래프 레이아웃 알고리즘을 이용하여 자동 생성된 지도의 장점인 업데이트가 쉽다는 장점을 가지고 있다. 그러나 모든 경로 정보를 하나하나 편집함으로써 등록해야한다는 단점이 있다. 따라서 대량의 경로 정보를 보다 손쉽게 처리하기 위하여, 데이타베이스의 바이오패스웨이에 관한 정보를 그래프 레이아웃 알고리즘을 사용하여 경로 지도를 자동으로 생성한 후, 해당 경로의 의미를 이해하는 사람이 보다 직관적으로 경로 지도를 편집한 후 저장하도록 UniPath 그래픽 에디터의 기능을 확장하는 것이 바람직하다. 더 나아가서 Fukuda가 제안한 신호전달 경로의 지식 표현 방법을 참고하여[9], UniPath 표기법을 슈퍼노드(supernode)를 사용하는 복합 그래프 구조(compound graph structure)로 확장하여 바이오패스웨이 정보를 표현하면 경로 정보의 다양한 입도와 계층적 구조까지 컴퓨터상에서 잘 다룰 수 있을 것이다.

6. 결 론

본 논문은 세분화 되어있는 바이오패스웨이 지식을 단일 표기법으로 표현할 수 있는 UniPath 표기법을 제안하고, 이 표기법을 적용한 프레임 기반 바이오패스웨이 지식표현 시스템을 구현하였다. 본 논문의 의의는 크게 세 가지로 다음과 같다.

첫째, 신진대사 경로와 조절 경로를 하나의 통합된 방식으로 표현할 수 있는 간결하면서도 강력한 표현력을 가진 UniPath 표기법을 제안하였다. UniPath 표기법은 복잡한 바이오패스웨이의 상호관계를 세포 내 위치와 주체의 형, 그리고 필수적인 작용의 종류만을 사용하여 구분할 수 있도록 하며, 기존 시스템에서 적용되지 않았던 추정 작용이나 생략 작용을 구분하여 명시할 수 있도록 디자인되었다. UniPath 표기법을 사용함으로써 세분화되어 있는 바이오패스웨이 지식을 통합하여 표현할 수 있다. 그럼으로써 생체 시스템상에서 분자들의 상호작용에 기반하여 다양한 생화학 경로들에 의해 기능이 어떻게 수행되는지에 관해 전체적으로 연구할 수 있다.

둘째, UniPath 표기법을 적용하여 바이오패스웨이 지식을 그래프 형태로 편집하고 이를 프레임 구조로 관리하는 그래픽 에디터를 설계하고 구현함으로써, 바이오패스웨이 지식을 쉽게 저장·수정하며 확장이 용이한 지식 표현 시스템의 프로토타입을 제시하였다.

셋째, UniPath 시스템 안에서 표현된 지식은 플랫폼 독립적으로 데이타를 통합하며 처리하기에 적합한 XML 포맷으로도 손쉽게 변환할 수 있다. XML 포맷은 복잡한 바이오패스웨이 지식을 바이오 온톨로지를 바탕으로 데이타를 표현하고 공유하기에 적합하다.

결과적으로 UniPath 시스템은 복잡하고 체계화가 필요한 바이오패스웨이 지식을 통합하고 보다 명료하게 표현함으로써 세분화되어있는 경로 정보를 전산처리가 용이한 구조로 관리하도록 하였으며, 이를 통하여 자료검색이나 데이타 마이닝 등의 정보 부가가치를 높일 수 있도록 하였다.

참 고 문 헌

[1] BioPathways Consortium. http://www.biopathways.org
[2] Bader, G., C. Hogue, "BIND - A Data Specification for Storing and Describing Biomolecular Interactions, Molecular Complexes and Pathways," Bioinformatics, vol.16, no.5, pp. 465-477, 2000.
[3] Matsuno, H., A. Doi, Y. Hirata, S. Miyani, "XML Documentation of Biopathways and Their Simulations in Genomic Object Net," Genome Informatics vol.12, pp. 54-62, 2001.
[4] SBML. http://www.cds.caltech.edu/erato/sbml
[5] CellML. http://cellml.org
[6] KEGG. http://www.genome.ad.jp/kegg/kegg2.html
[7] GeneNet. http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet
[8] PathDB. http://www.ncgr.org/pathdb
[9] Fukuda, K., and T. Toshihisa, "Knowledge representation of singal transduction pathways," Bioinformatics, vol.17, no.9, pp. 829-837, 2001.
[10] Voit, Eberhard, Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists, pp. 25-28, Cambridge press. 2000.
[11] 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.
[12] Pirson, Isabelle et al., "The Visual Display of Regulatory Information and Networks," Trends in Cell Biology, vol. 10, pp. 404-408, Oct. 2000.
[13] Karp, P. D., M. Riley, et al., "The EcoCyc Database," Nucleic Acids Research, vol. 30, no. 1, pp. 56-58, 2002.
[14] ExPASy. http://www.expasy.org
[15] WIT. http://wit.mcs.anl.gov/WIT2
[16] UM-BBD. http://umbbd.ahc.umn.edu
[17] TransPath. http://www.biobase.de/pages/products/transpath.html

이 민 수

2001년 이화여자대학교 수학과 학사. 2003년 이화여자대학교 컴퓨터학과 석사. 2003년~현재 이화여자대학교 컴퓨터학과 박사과정. 관심분야는 바이오인포매틱스, 지식표현, 데이타마이닝

박 승 수

1974년 서울대학교 수학과 학사. 1976년 한국과학기술원 전산학 석사. 1988년 미국 텍사스 대학 전산학 박사. 1988년~1991년 미국 켄사스대학 컴퓨터학과 조교수. 1991년~현재 이화여자대학교 컴퓨터학과 교수. 관심분야는 인공지능, 데이타마이닝, 바이오인포매틱스

강 성 희

1991년 이화여자대학교 전자계산학과 학사. 1995년 이화여자대학교 전자계산학과 석사. 2001년 이화여자대학교 컴퓨터학과 박사. 2001년~현재 명지대학교 교육학습개발원 교수. 관심분야는 인공지능, 에이전트, 데이타마이닝, 바이오인포매틱스