세포 신호전달 경로 데이타베이스를 위한 데이타 모델링

(Data Modeling for Cell-Signaling Pathway Database)

박 지 숙 † 백 은 옥 †† 이 공 주 †††
(Jisook Park) (Eunok Paek) (Kong-Joo Lee)
이 상 혁 †††† 이 승 록 †††† 양 갑 석 ††††
(Sanghyuk Lee) (Seung-Rock Lee) (Kap-Seok Yang)

요 약 최근 유전체학과 단백질체학 분야에서 생성되는 방대한 분량의 데이타로부터 생물학적 의미를 추출해내기 위한 생물정보학적인 도구들에 대한 필요성이 크게 대두되고 있다. 본 논문에서는 세포 신호전달 경로에 관한 정보를 효율적으로 표현, 저장함은 물론 저장된 데이타로부터 생물학적 의미를 추출할 수 있도록 하기 위한 다양한 요구 조건들을 생물학자의 관점에서 분석하고, 이들 요구조건을 체계적으로 반영하여 설계한 ROSPath 데이타베이스 시스템을 제안한다.

ROSPath 데이타 모델에서는 향후의 확장성을 고려하여 불완전한 지식의 표현이 가능하도록 하며 인터넷상에서 기존의 다른 생화학 데이타베이스를 공유할 수 있는 연결성을 제공한다. 또한, 객체지향 모델을 이용하여 계층적인 구성을 제공함으로써 효율적인 검색을 지원한다. ROSPath 데이타 모델은 두 가지 주요 데이타 요소인 '바이오 개체'와 '상호작용'으로 정의된다. 바이오 개체는 세포 신호전달 경로에 관여하는 단백질과 단백질 상태 등과 같은 개개의 생화학적인 개체를 의미하고, 상호작용은 단백질 상태 전이나 화학 반응, 단백질-단백질 상호작용 등과 같은 바이오 개체들 간의 다양한 관계 및 신호전달과정을 설명한다. 제안된 ROSPath 데이타 모델을 이용하여 구성되는 복잡한 정보 네트워크는 다양한 생화학 프로세스들을 기술하고 분석하는 데에 활용할 수 있다.

키워드 : 세포 신호전달 경로, 상태 전이, 상호작용, 단백질, 데이타 모델

Abstract Recent massive data generation by genomics and proteomics requires bioinformatic tools to extract the biological meaning from the massive results. Here we introduce ROSPath, a database system to deal with information on reactive oxygen species (ROS)-mediated cell signaling pathways. It provides a structured repository for handling pathway related data and tools for querying, displaying, and analyzing pathways.

ROSPath data model provides the extensibility for representing incomplete knowledge and the accessibility for linking the existing biochemical databases via the Internet. For flexibility and efficient retrieval, hierarchically structured data model is defined by using the object-oriented model. There are two major data types in ROSPath data model: 'bio entity' and 'interaction'. Bio entity represents a single biochemical entity: a protein or protein state involved in ROS cell-signaling pathways. Interaction, characterized by a list of inputs and outputs, describes various types of relationship among bio entities. Typical interactions are protein state transitions, chemical reactions, and protein-protein interactions. A complex network can be constructed from ROSPath data model and thus provides a foundation for describing and analyzing various biochemical processes.

Key words : cell-signaling pathway, state transition, interaction, protein, data model

· 이 논문은 한국과학재단의 SRC 지원사업(이화여대 세포신호전달연구센터), IMT 2000 프로젝트인 IMT2000-C5-2(IT-BT), 2002년도 서울시립대학교 학술연구조성비, 그리고 2003학년도 서울여자대학교 교내특별연구 지원사업에 의하여 연구되었음

† 종신회원 : 서울여자대학교 정보통신공학부 조교수
jspark@swu.ac.kr
†† 종신회원 : 서울시립대학교 기계정보공학과
paek@uos.ac.kr
††† 비 회 원 : 이화여자대학교, 약학대학 제약학과/분자생명과학부
kjl@ewha.ac.kr
†††† 비 회 원 : 이화여자대학교 분자생명과학부 교수
sanghyuk@ewha.ac.kr
Leesr@ewha.ac.kr
ksyang@ewha.ac.kr

논문접수 : 2003년 6월 18일
심사완료 : 2003년 11월 4일

1. 서 론

세포 내에서 단백질 분자들은 복잡한 상호작용의 네트워크를 형성하여 세포의 성장, 분화와 사멸 등 다양한 생리적 기능을 수행하는 데 참여한다. 여러 생물학자들이 많은 노력을 통하여 특정 세포 신호전달 경로에 관여하는 것으로 밝혀낸 단백질 분자(molecule)에 관련된 정보는 그 양이 폭발적으로 증가하고 있을 뿐만 아니라 방대한 문헌에 산재되어 있어, 현재까지 밝혀진 지식 체계를 손쉽게 파악하기 어려운 것이 현실이다. 생물학 연구자로 하여금 이러한 세세한 정보를 일관성 있는 모델을 이용하여 체계적으로 종합하여 관리할 수 있도록 하는 것은 물론, 종합된 정보를 편리한 형태로 제공하여 사용하도록 하는 것이 세포 신호전달 데이타베이스 시스템 구축의 목적이다.

이상적으로는 세포 신호전달 경로에 관한 정확한 모델이 밝혀진 경우, 그 결과로써 데이타베이스를 구축하는 것이 옳겠지만, 현실적으로는 세포 내의 신호전달 경로에 관하여 완전한 정보가 알려져 있는 경우는 거의 없으며, 시간의 흐름에 따라 새로운 정보가 점진적으로 밝혀지고 있기 때문에 어느 시점에서든 완전한 모델을 기대하기 어렵다. 또한 세포 신호전달 경로 데이타베이스에는 단백질이나 화학물질(chemical)에 관한 기초적인 정보에서부터, 현재로서는 그 구체적인 메커니즘이 밝혀지지 않아 모호성을 지닌 채로 표현되어야 하는 정보들이 산재한 신호전달 경로(signaling pathway) 정보까지 다양한 정보들이 서로 다른 수준에서 다양한 형태로 표현 및 저장되어야 한다. 따라서 이와 같은 모호성 및 다양성을 지원할 수 있는 모델의 정의가 매우 중요하다. 즉, 이 모델은 생물학자들이 지니고 있는 신호전달 물질에 관한 기본적인 개념을 잘 반영함은 물론, 이를 손쉽게 정형화된 형태로 표현할 수 있도록 하여야 하고 검색을 용이하게 할 수 있는 형태로 구현이 가능하여야 한다.

본 논문에서는 세포 신호전달 경로에 관한 정보를 효율적으로 표현하고 저장하기 위하여 고려되어야 할 다양한 요구 조건들을 생물학자의 관점에서 분석하고, 이들 요구조건을 체계적으로 반영하여 설계한 ROSPath (Reactive Oxygen Species mediated cell-signaling Pathways) 데이타베이스 시스템을 위한 데이타 모델을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 신호전달 데이타베이스의 데이타 모델을 소개하며, 3장에서는 세포 신호전달 경로에 대한 정보를 효율적으로 표현하고 저장하기 위한 요구 조건들을 분석하고 이를 반영한 ROSPath 데이타 모델을 제시한다. 4장에서는 ROSPath 시스템에 세포 신호전달 경로에 관한 정보를 입력하는 과정을 설명하며, 5장에서 기존의 세포 신호전달 데이타베이스와 비교될 수 있는 ROSPath 시스템의 특징과 향후 연구 과제에 대해 설명하는 것으로 결론을 맺는다.

2. 관련 연구

세포 내의 신호전달 및 대사 경로를 체계적으로 관리하기 위하여 구축된 생물정보 데이타베이스들은 네트워크 상에서 개방적으로 사용자들에게 제공되고 있다[1-7]. 다음에서는 세포 신호전달 경로와 관련한 데이타베이스 시스템의 특징들을 살펴본다.

(1) DIP/LiveDIP (http://dip.doe-mbi.ucla.edu/)

DIP(the Database of Interacting Proteins) 데이타베이스가 단백질의 물리적 상호작용에 초점을 맞춘 것이라면 LiveDIP 데이타베이스는 단백질 상태 및 상태 전이에 기초한 생물학적인 상호작용에 주안점을 둔 시스템이라고 설명할 수 있다. DIP에서는 물리적인 상호작용에 관여하는 단백질의 아미노산 서열 정보, 도메인의 구조나 범위 등과 같은 데이타를 저장하는데 반해, LiveDIP에서는 각각의 단백질이 서로 다른 단백질 상태로서 존재하고 이들이 모여 하나의 단백질 상태 공간을 구성하도록 모델링하였다[1].

(2) Transpath (http://transpath.gbf.de/index.html/)

Transpath 시스템은 유전자 조절 경로 (gene-regulatory pathways)에 대한 정보를 다루는 데이타베이스 시스템이며, 인간, 쥐 등의 서로 다른 종(species)에서의 전사 인자(transcription factors)의 조절에 관여하는 경로에 주안점을 두고 있다. 이는 신호전달 경로에 관여하는 요소들 즉, 호르몬, 수용체(receptors), 효소(enzymes), 전사 인자 및 그들 간의 상호작용 등을 저장한 객체 지향 데이타베이스 시스템으로, 현재 POET Software의 상용 객체지향 데이타베이스 시스템을 사용하여 구축되었고, 신호전달 경로를 표현할 수 있는 그래픽 툴을 제공하고 있다[2,3].

(3) KEGG (http://kegg.genome.ad.jp/kegg/)

KEGG(Kyoto Encyclopedia of Genes and Genomes) 시스템은 유전자나 분자 사이의 상호작용으로 구성되는 정보 전달 경로를 이용해 분자와 세포의 생물학 지식을 전산화하였다. KEGG 시스템에서는 한 쌍의 유전자나 분자들의 상호작용을 이진관계(binary relation)로 모델링 하는데, 이미 저장된 관계로부터 새로운 관계를 추론할 수 있고 이들의 계층구조를 구성함으로써 유전자나 분자간의 기능적/구조적/진화적인 관계들을 표현할 수 있다[4].

(4) STKE (http://stke.sciencemag.org/)

STKE(Science's signal Transduction Knowledge Environment)는 세포 신호전달경로 검색을 지원하는 여러 종류의 툴과 접근 방식들을 혼합하여 정의한 시스템이다. 이 시스템에서는 신호전달 경로를 구성하는 기본 데이타가 구성 요소(component)로 정의되며 이들 간의 관계는 연결도(connections map)에 의해 도식화될 수 있다[5].

(5) AfCS (http://www.cellularsignaling.org/)

AfCS(Alliance for Cellular Signaling)에서는 신호전달 경로에 관여하는 분자들에 대한 기존 연구 결과들을 효율적으로 관리하기 위하여 AIMS(Alliance Information Management System)을 개발하였는데, 신호전달 경로에 대한 구체적인 데이타 모델은 아직까지 개발되지 못한 상태이다[6].

(6) AMAZE (http://www.ebi.ac.uk/research/pfbp/)

AMAZE 시스템은 단백질의 분자적 기능 및 조직 내 생화학적인 단계 즉 전달 경로에 대한 정보를 저장한다. 또한, 대사 경로, 유전자 조절, 신호전달 등의 대사적(metabolic) 전달 경로를 포함하고, 불완전한 정보의 점진적인 명세 등의 기술적 이슈에 대한 연구가 진행 중이다. AMAZE 시스템에서는 객체 지향 개념을 도입하여 생화학 개체(Biochemical Entity)와 상호작용이라는 주요 객체들을 모델링 하였다[7].

3. ROSPath 데이타베이스의 데이타 모델

본 장에서는 세포 신호전달 경로 데이타베이스의 구축을 위한 데이타 모델의 설계를 위한 요구사항들을 생화학자들의 관점에서 분석한다. 우선 전반적인 설계 목표에 대해서 살펴 본 이후에 세포 신호전달 경로 데이타베이스에서 표현해야 하는 데이타의 유형 및 모델의 핵심적인 부분에 관해 서술한다.

3.1 요구조건 분석

3.1.1 계층적 구성

본 연구에서 제안하는 세포 신호전달 경로를 위한 모델의 가장 핵심적인 특징은 계층적 구성이다. 세포 신호전달 경로를 밝히기 위해서는 신호전달에 관여하는 다양한 단백질에 대한 표현은 물론, 단백질 간의 상호작용의 표현, 이들 단백질에서 일어날 수 있는 다양한 단백질 합성 후 변형(post-translational modification)의 표현, 단백질 간의 산화-환원 반응에 대한 표현, 또한 이들 반응이 어떻게 조절되는지를 나타내는 신호전달에 대한 표현 등 여러 계층의 다양한 정보가 복합적으로 표현되어야 한다. 계층적으로 구성된 정보는 높은 수준의 정보는 알려져 있으나 구체적인 정보는 알려져 있지 않은 경우에도 입력을 허용하므로 세포 신호전달 경로와 같이 끊임없이 새로운 사실이 밝혀져 데이타베이스에 추가되어야 하는 상황에 적합하다. 즉, 현재에는 구체적인 실험결과를 얻을 수 있는 방법론이 개발되지 않아 모호한 수준의 정보만이 알려져 있으나 향후 실험방법과 도구의 발달에 따라 구체적인 정보가 밝혀질 경우, 이전에 입력된 모호한 수준의 데이타와 새로 입력되는 구체적인 수준의 데이타가 서로 그 내용적으로는 모순된 것이 없이 다만 표현 수준의 차이만 존재한다면 계층적인 모델은 이들 여러 수준의 표현을 수용하기에 바람직한 형태가 된다. 입력의 측면에서 뿐 아니라 데이타베이스의 내용을 사용자에게 제공할 때에도 계층적인 정보의 구성을 이용하면 사용자가 원하는 수준에서 정보를 검색하고 제공하는 것이 가능하게 된다는 장점이 있다.

3.1.2 불완전 정보의 표현

본 논문에서는 세포 신호전달 경로와 관련된 정보를 가능한 한 상세히 서술할 수 있도록 함과 동시에, 현재까지는 실험적으로 그 구체적인 내용이 밝혀지지 않았지만 생물학자들이 제한적으로나마 알고 있는 지식을 입력할 수 있도록 하는 모델의 설계가 요구된다.

ROSPath 데이타베이스를 위한 모델링 과정에서 이와 같이 제한적인 지식을 표현해야 할 필요성은 많은 곳에서 발견되었다. 예를 들어, 단백질의 구체적인 상태 변화에 대해서는 알지 못하지만 특정한 단백질이나 화학물질에 의해 다른 단백질의 상태가 변화됨으로써 신호가 전달된다고 믿는 경우가 있을 수 있는데 이런 경우 단백질의 구체적인 상태를 명시하지 않고도 신호전달 관계를 표현할 수 있어야 한다. 또 다른 예로, 세포 신호전달의 단계 중 알려지지 않은 여러 단백질이 중간에서 매개하여 단백질 A에서 단백질 B로 신호가 전달된다고 믿을 만한 근거가 있는 경우를 들 수 있다. 이 때, 중간에서 매개체로 작용하는 여러 단백질의 정체를 구체적으로 표현하지 않고도 단백질 A에서 단백질 B로 신호전달이 이루어진다는 정보를 표현할 수 있어야 한다. ROSPath 데이타 모델에서는 이와 같이 완전하게 밝혀지지 않은 부분도 현재의 지식수준에서 데이타베이스화할 수 있고, 향후 이 부분에 대하여 새로운 내용이 밝혀질 경우 이를 쉽게 수정할 수 있도록 하였다.

3.1.3 확장성

위에서 서술한 바와 같이 ROSPath의 데이타 모델에서는 제한된 지식을 현재 알려진 지식수준에서 표현하고 향후 새로운 정보를 점진적으로 추가하면서 더 구체화할 수 있도록 하는 것을 지식의 계층적인 구성에 의해 제공하고자 한다. 그러나 한 발 더 나아가면, 현재로서는 어떤 체계를 가지고 지식을 계층적으로 구성할 만한 정보를 가지고 있지 못하지만 미래에 이와 같은 체계를 새로이 추가하는 일을 손쉽게 할 수 있도록 모델

을 구성하는 일 또한 매우 중요한 측면이다. 이를 위해서 데이타 모델의 논리적인 구성을 객체-관계 형태로 유지하여 데이타 모델을 물리적인 데이타베이스 스키마로 구현하는 것과 독립되게 하였다.

3.1.4 외부 데이타베이스 연결

위에서 언급한 다양한 수준의 정보 중에서는 기존의 다른 데이타베이스에서 각 수준에 맞는 전문적인 정보는 물론 이를 활용하는 데에 필요한 소프트웨어 툴을 제공하는 경우가 있다. 이와 같은 외부 데이타베이스 정보는 세포 신호전달 경로 데이타베이스의 활용에 필요하다고 인정되는 수준에서 ROSPath의 고유한 정보와 함께 제공하는 것이 바람직하다.

그러나 이 정보를 모두 ROSPath 데이타베이스에 중복하여 유지하는 것은 정보의 최신성을 유지하기 어려워 바람직하지 않다. 이와 같은 정보는 외부 데이타베이스에 대한 하이퍼링크로서 제공하도록 하여 지속적으로 수정과 보완이 이루어지는 외부 데이타베이스의 최신 정보를 사용자가 손쉽게 이용할 수 있도록 한다. 이들 외부 데이타베이스의 정보 중 단백질의 서열 정보와 같이 변경 가능성이 매우 적고 세포 신호전달 경로 데이타베이스에서 빈번히 접근할 것으로 예상되는 데이타에 한해서는 ROSPath 데이타베이스 내에서 중복적으로 유지함으로써 정보에 대한 빠른 접근을 허용함과 동시에 네트워크 자원을 효율적으로 사용하도록 한다.

3.2 데이타 모델의 특징

ROSPath 데이타 모델에서는 위에서 언급한 요구사항들을 만족시키기 위하여 객체지향 데이타 모델을 이용하였다. 객체지향 데이타 모델이 제공하는 클래스 계층구조는 데이타 모델의 계층적 구성을 지원하고 데이타 모델의 확장을 용이하게 한다.

3.2.1 클래스 계층 구조

그림 1은 ROSPath 데이타베이스의 클래스 계층구조를 보여 준다. ROSPath 데이타 모델의 최상위 레벨에서는 세포 신호전달 경로를 표현하는데, 신호전달 경로는 신호전달 스텝(signaling step)이라 부르는 조절/신호전달 관계의 집합으로 표현된다. 신호전달 스텝은 바이오 개체(Bio Entity)라 불리는 단백질 또는 화학물질 사이에 존재하는 다양한 종류의 상호작용으로 정의되는데, 여기서의 상호작용은 단백질 사이의 결합에 의해 발생할 수도 있고, 하나의 단백질이 효소로 작용하여 다른 단백질의 상태를 변경시켜 발생할 수도 있으며, 단백질의 상태변화 과정에 화학물질이 촉매로 작용하는 경우 등 매우 다양한 조절관계나 신호전달을 나타낸다. 세포 신호전달 스텝을 표현하는 데에 사용되는 단백질은 그것이 신호전달에 관여하는 양상에 따라 각각 특정한 단백질 상태에 있는 것으로 정의된다. 하나의 단백질은 단백질 합성 후 변형에 따라 서로 다른 상태로 정의될 수 있고, 세포 내 위치에 따라, 단백질의 분해 또는 다른 단백질과의 결합에 따라, 리간드(ligand)의 종류에 따라 서로 다른 상태로 정의가 가능하며, 이들 단백질의 상태 간의 변화는 상태 전이로 정의할 수 있다.

3.2.2 불완전 정보의 표현

ROSPath 데이타 모델의 특징 중 하나는 세포 신호전달에 관여하는 단백질을 세분화하여 각각의 단백질 상태를 표현하도록 한 것이다. 단백질이 신호전달에 참여할 때에는 동적으로 변화하며 특정한 상태로 존재하는 경우가 많은데, 각 단백질이 서로 다른 상태라고 정의할 수 있는 요인을 단백질 합성 후 변형, 세포 내에서의 위치, 올리고머 상태(oligomeric state), 리간드, 단백질 분해효소에 의한 절단 형태(proteolytic form) 등으

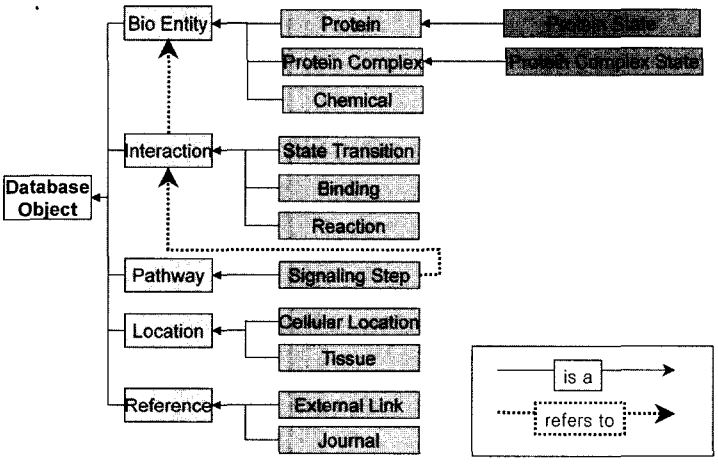

그림 1 ROSPath 데이타 모델의 클래스 계층구조

로 분류하고 각 경우에 더욱 자세한 분류는 필요에 따라 상세 분류로서 제공하였다. 예를 들어, 단백질 합성 후 변형에는 인산화(phosphorylation), 시스테인 산화(cysteine oxidation), 당화(glycosylation), 메틸화(methylation), 아세틸화(acetylation) 등이 포함되고, 인산화는 다시 어느 아미노산에 대한 것이냐에 따라 세린(serine), 트레오닌(threonine), 타이로신(tyrosine)으로 구성되어 있다[8]. 또한 각 경우 변형 부위(modification site)에 관한 정보가 알려져 있는 경우, 이를 포함시킬 수 있다. 이렇게 함으로써 단백질의 상태를 표현함에 있어서, 생물학적으로 밝혀진 정보를 최대한 상세하게 그리고 체계적으로 나타낼 수 있도록 하였다.

위와 같은 체계적인 구성을 이용하게 되면, 사용자가 자신의 지식을 매우 상세한 수준까지 표현하여 입력할 수 있게 됨은 물론, 경우에 따라서는 매우 제한적인 지식이라도 표현이 가능하게 된다. 위의 예를 이용하면, 해당 단백질 상태를 서술함에 있어서 구체적인 단백질 합성 후 변형의 위치는 물론 그 구체적인 종류 즉, 세린, 트레오닌, 타이로신 중 어느 것인지를 모르는 경우에도 데이타 모델에서 제공하는 체계 내의 적당한 수준에서 인산화를 서술하는 것이 가능하다.

3.2.3 확장성

분자생물학의 발전에 의해 새로이 밝혀질 내용들을 최대한 수용할 수 있도록, 가능하면 입력형태에 제한을 적게 하였다. 예를 들어 단백질 합성 후 변형의 경우, 변형의 종류와 변형 부위를 동시에 여러 개 정의할 수 있는데, 이는 신호전달 경로에 따라서는 변형되는 종류와 그 조합이 다양할 수 있으며, 동시에 상호작용을 하는 단백질에 따라서도 변형이 변할 수 있기 때문이다.

또한 입력 과정에서 현재의 데이타 모델에서는 체계화 할 수 없는 지식을 수용하기 위하여 생물학자가 수시로 주석을 달 수 있도록 하였다. 예를 들면, 각 입력 페이지에는 자연어로 임의의 설명을 추가할 수 있는 필드를 제공하고 있으며, 현재의 입력 환경에서 제공하는 다양한 목록 즉, 세포내 위치(cellular location), 단백질 합성 후 변형, 리간드의 종류 등에 대한 추가를 시스템 관리자를 통하여 할 수 있도록 하였다.

마지막으로 데이타 모델 자체는 생물학자들의 지식을 체계화할 수 있는 계층구조로 구성하였으나, 실제로 이를 입력하는 과정에서는 이 데이타 모델이 입력 환경에 그대로 드러나지 않는 경우도 많다. 생물학자들은 많은 정보를 단백질을 중심으로 이해하고 구성하는 경향이 있다. 생물학자들이 단백질을 중심으로 이해하고 있는 지식을 ROSPath 데이타 모델에서 신호전달을 위한 기본적인 단위가 되는 신호전달 경로로 재구성하는 일은 신호전달 데이타베이스를 구축함에 있어서 필수적인 부분이지만 이를 가능한 한 입력 환경에서 담당하도록 하여, 실제로 데이타베이스의 내용을 입력할 분자생물학자의 지식 구조가 쉽게 반영할 수 있도록 하였다.

3.3 데이타 모델

다음에서는 ROSPath에서 표현하고자 하는 데이타를 그 유형에 따라 서술하고 각 데이타 타입이 어떻게 모델링되었는지를 설명한다. 그림 2는 PDGF(Platelet-Derived Growth Factor) 신호전달 경로를 ROSPath 데이타 모델을 사용하여 모델링 한 결과를 보여주며, 이 예를 사용하여 ROSPath 데이타 타입을 설명한다[9].

3.3.1 단백질

ROS 세포 신호전달 경로에 관여하는 것으로 알려진 각 단백질은 하나의 데이타 개체로서 모델링 된다. 세포 신호전달 경로에 관한 생물학적인 지식은 주로 단백질을 중심으로 구성되어 있는데, 예를 들면 외부의 자극에 의해 촉발된 '신호'가 단백질 간의 상호작용에 의해 다

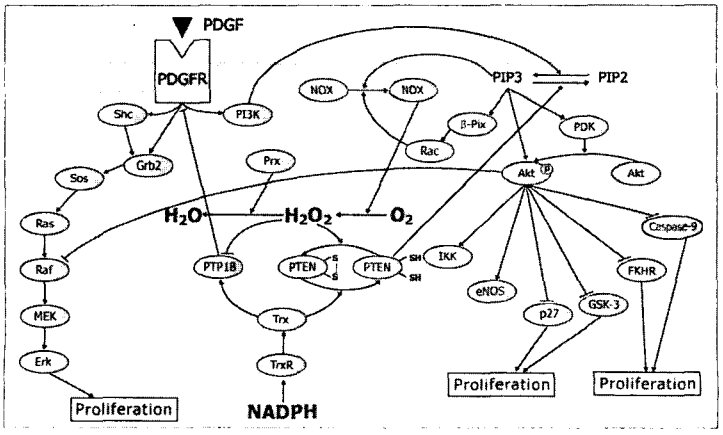

그림 2 PDGF 신호전달 경로

른 단백질로 '전달'되거나 세포 내에서 단백질이 다른 위치로 이동함으로써 신호가 전달된다고 여기는 등 연관된 단백질의 활동을 통하여 서술되고 있다. 따라서 특정 신호전달경로에 관여하는 것으로 알려진 단백질에 관한 관련 정보들을 종합적으로 표현하고 제시하여야 한다.

각 단백질에 대해서는 기존의 데이타베이스에서 다양한 정보를 제공하므로 이들 외부 데이타베이스의 단백질 정보 중에서 세포 신호전달 경로에 관한 연구에 유용하다고 판단되는 주석(annotation)은 ROSPath 데이타베이스와 연결하여 ROSPath 사용자들이 쉽게 관련 정보를 접근할 수 있도록 하고, 그 외에 신호전달 경로 연구에 도움이 된다고 판단되는 데이타는 데이타베이스 구축 과정에서 직접 입력할 수 있도록 한다.

현재 단백질 관련 정보는 단백질을 고유하게 식별하기 위한 접근 번호(accession number)와 버전 번호, 단백질의 이름 및 동의어, 단백질의 생물종(source organism), 핵산 및 아미노산 서열(nucleic acid/amino acid sequence), 염색체내에서의 위치, 3차원 구조, 신호전달 도메인(signaling domain), 세포 내 위치 및 티슈(tissue) 정보, 효소 반응(enzymatic reaction), 세포내의 활성(cellular activity), 문헌정보로 구성되어 있다.

위의 데이타 중, 이름 및 동의어와 서열 정보는 NCBI의 GenBank[10] 데이타베이스와 Swiss PROT [11] 데이타베이스의 내용을 참조하여 ROSPath 데이타베이스에서 중복되게 제공하며 동의어는 사용자가 임의로 추가할 수 있도록 한다. 염색체 상의 위치 (chromosome localization), 3차원 구조, 신호전달 도메인, 문헌정보는 각각 GenBank, PDB[12] , InterPRO[13], PUBMED[14]를 참조할 수 있는 링크(link)를 제공한다. 세포내의 활성은 Gene Ontology[15] 데이타베이스의 분류 체계를 이용하여 Swiss PROT의 단백질에 대하여 작성해 놓은 주석을 기본으로 하고 사용자가 추가적으로 정의할 수 있도록 한다. 생물종, 세포 내 위치, 티슈 정보, 효소 반응 정보는 사용자가 새로 입력할 수 있도록 하고, 이외에도 입력자가 추가하고자 하는 정형화되지 않은 정보가 있는 경우 임의의 텍스트를 사용하여 주석을 추가할 수 있도록 하였다. 그림 2에서 타원형으로 표시된 노드, 즉, Shc, Grb2, Sos, Ras, Raf 등은 각각 하나의 단백질을 표현하고 있다.

3.3.2 단백질 상태(Protein state)

세포 내에서 신호전달에 관여하는 단백질은 그 역할에 따라 다양한 양상으로 존재한다. 그림 2의 PDGF 신호전달 경로에서 예를 들면, PTEN은 산화된 형태인 PTEN-S-S와 환원된 형태인 PTEN-SH-SH 두 가지 상태로 존재하고 있으며, 이 중에서 PTEN-SH-SH가 PIP3를 PIP2로 변화시키는 역할을 한다. 이와 같이 하나의 단백질이 그 동적인 상태에 따라 기능적인 역할이 달라질 수 있으므로 단백질의 상태를 하나의 데이타 개체로 모델링 하여 표현하였다. ROSPath 세포 신호전달에 관여하는 단백질은 적어도 하나 이상의 단백질 상태를 갖는 것으로 가정하였는데, 구체적인 상태 정보가 밝혀지지 않은 경우라고 하더라도 하나의 기본 상태가 자동으로 생성되어 사용되도록 하였다. 하나의 단백질을 서로 다른 상태에 있는 것으로 정의하는 기준은 여러 가지가 있을 수 있는데, ROSPath 데이타 모델은 다음 다섯 가지 측면에서 단백질 상태를 구별한다.

첫째, 단백질 합성 후 변형의 종류에 따라 구분한다. 단백질 합성 후 변형으로 표현할 수 있는 종류는 이미 40여 가지로 제공되고 있고, 이들은 두 단계로 분류되어 있다. 하나의 단백질 상태에 대하여 여러 종류의 변형을 명시할 수 있으며, 변형의 종류를 적당한 수준에서 표현할 때 변형 부위에 관한 구체적인 위치 정보도 함께 표현할 수 있다.

둘째, 어떤 리간드를 갖느냐에 따라 구분한다. 리간드로 사용할 수 있는 물질은 현재 20여 가지의 화학물질이나 단분자 (small molecule)의 목록에서 다수의 선택을 할 수 있으며 이 목록에 새로운 리간드를 추가할 수도 있다.

셋째, 단백질의 4차 구조인 올리고머 상태에 따라 구분한다. 이량체(Dimer)인지, 삼량체(trimer)인지, 이종복합체(heteromer)인지를 구별하고, 이종 복합체인 경우 각각의 구성요소가 되는 단백질의 상태를 명시하는데, 이는 아래의 복합 상태에서 자세히 설명한다.

넷째, 세포 내의 위치에 따라 구분한다. 현재는 세포내의 위치가 10여 개로 구분되어 목록으로 제공되는데 이 목록에서 다수를 선택할 수 있으며 목록에 새로운 항목을 추가할 수 있다.

마지막으로, 단백질 분해효소에 의한 절단 형태에 따라 구분하는데, 이 때 절단 부위(cleavage site)를 아는 경우에는 이를 표현한다.

그림 2에서는 Akt라는 단백질의 상태로서 Akt-p가 표현되어 있다. 이는 Akt에 인산화라는 단백질 합성 후 변형이 적용되어 정의된 상태이다.

3.3.3 복합단백질(Protein Complex)

단백질 중에는 구성요소가 되는 여러 단백질이 모여서 하나의 복합체로서 특정한 기능을 수행하는 복합단백질이 있는데, 이러한 복합단백질도 하나의 데이타 개체로서 모델링 한다. 복합단백질은 이미 정의된 단백질 리스트를 참조하여 그 구성요소를 정의하는데, 복합체 구성 후의 안정성 여부에 따라 안정한 복합체(stable complex)와 일시적인 복합체(transient complex)로 구

분한다. 일시적인 복합체는 단백질간의 상호작용에 의한 일시적인 결합에 의한 것으로 정의하는데, 이와 같은 정보는 BIND[16]나 DIP과 같은 단백질 상호작용 데이타베이스에 저장된 정보를 활용할 수 있다.

3.3.4 복합단백질 상태(Protein Complex State)

복합단백질 역시 그 역할에 따라 다양한 상태로 존재할 수 있으며, 복합단백질 상태를 각각 하나의 데이타 개체로서 모델링 하였다. 복합단백질의 경우에도 언제나 하나 이상의 복합단백질 상태를 갖는 것으로 가정하였는데, 여러 개의 구성 단백질 상태(component protein state)로 구성됨에 따라 복합단백질 상태를 구분하는 기준을 정의할 때 다음의 특성들을 고려하여야 한다.

첫째, 단백질 합성 후 변형의 종류를 정의할 때 어떤 구성 단백질에 변형이 발생한 것인지를 명시할 수 있어야 하며, 2개 이상의 구성 상태 사이에서 변형을 정의할 수 있도록 한다. 예를 들어, 펩타이드간 디설피드 결합(inter-chain disulfide bond)의 경우 복합단백질 상태의 각 구성요소가 되는 단백질 상태 사이에서 디설피드 결합(disulfide bond)을 정의해야 하기 때문이다.

둘째, 복합단백질 상태의 활성(activity)을 명시할 때 어떤 구성 단백질 상태가 활성인지를 명시할 수 있어야 한다.

셋째, 복합단백질의 상태를 정의할 때 명시하는 세포 내 위치정보와 복합단백질을 정의할 때 명시하는 세포 내 위치정보 사이에 불일치가 없도록 하여야 한다.

3.3.5 결합(Binding)

앞서 설명한 복합단백질이 구성 단백질의 결합에 의해 복합체를 구성하는 과정은 결합으로 모델링 된다. 예를 들어, 그림 2에서 PDGFR에서 PI3K로 '신호'가 전달되는 과정은 PDGFR가 인산화되어 PDGFR-p로 상태 전이된 후에 PDGFR-p에서 복합단백질[PDGFR-p, PI3K]이 생성되는 결합으로 설명할 수 있다. 결합에 의해 생성되는 일시적인 복합단백질은 항상 두개의 결합 구성요소로 구성된다고 가정하였으며 결합 방법과 결합 부위가 복수로 정의될 수 있다. 또한, 결합을 정의할 경우에는 그 결과가 되는 일시적인 복합단백질을 자동적으로 생성하여 저장한다.

3.3.6 단백질 상태 전이(Protein State Transition)

특정한 상태로 존재하는 단백질은 다른 상태로의 전이를 통해 신호전달에 참여하는 경우가 많다. LiveDIP 데이타베이스의 경우에는 이와 같은 상황을 상태전이의 종료 상태로의 신호전달로 모델링 하였다. 단백질 상태 전이의 종료 상태만을 신호전달에 포함시켜 모델링 하게 되면 하나의 단백질에 대해서 여러 가지의 상태 전이가 일련의 순서를 가지고 일어난다고 하더라도 이 순서 내의 어느 구체적인 상태 전이에 의해 신호전달이 이루어졌는지를 판단할 수 없게 된다. 본 논문에서는 단백질 상태 사이의 전이를 하나의 독립적인 데이타 개체로 정의한다. 상태 전이에 관련된 단백질이 하나의 상태에서 다른 상태로 전이되는 내용에 따라 변형의 종류, 올리고머 상태의 변화, 리간드의 변화, 세포 내 위치의 변화, 단백질 분해효소에 의해 절단된 형태의 변화 등으로 정의할 수 있는데, 여기에 부가적으로 입체구조 변화(conformational change) 여부를 주석으로 달 수 있도록 하였다. 상태 전이를 정의함에 있어서 시작 상태와 종료 상태에 대한 정의가 명백하게 잘 되어 있는 경우 이들 두 상태 간의 차이로부터 상태 전이에 대한 정보를 유추할 수 있다.

그림 2에서 PTEN은 두 가지 상태를 가진 것으로 표현되어 있는데, 각각은 PTENreduced(PTEN(SH)2)와 PTENoxidized(PTEN-S-S)이고 이 둘 사이에 상태 전이가 양방향으로 정의되어 있으며 각각 산화와 시스테인 환원에 의한 상태 전이로 정의된다[17].

3.3.7 반응(Reaction)

결합과 단백질 상태 전이를 제외한 모든 신호전달 과정은 반응으로 모델링한다. 여러 단백질이 작용하여 새로운 하나 또는 그 이상의 물질로 변화되는 연합/분리(association/dissociation) 과정이나 화학 반응을 그 예로 들 수 있다. 따라서 연합/분리 각 단계의 구성 단백질과 화학반응에 참여하는 화학물질을 하나의 독립적인 데이타 개체로 표현하고 이들 사이의 관계로서 반응을 정의한다.

그림 2의 PIP2와 PIP3는 화학물질이며 이들 사이에 가능한 양방향의 화학 반응을 반응으로 정의하였다.

3.3.8 신호전달 스텝

그간 생물학자들은 세포 내의 신호전달 경로를 일종의 방향 그래프로서 표현하여 왔다. 이때 그래프의 각 노드는 흔히 단백질 분자 또는 화학물질 등을 나타내고, 그래프의 각 에지는 하나의 분자가 다른 분자에게 '신호'를 '전달'한다는 개념적인 관계를 나타낸다. 여기서 말하는 신호전달이란 생물학자가 해석하는 '논리적'인 관계로서 일정한 화학적 또는 생물학적 작용을 표현하기 위한 '물리적' 관계를 나타낸다고 볼 수 없다. 예를 들어 어떤 경우에는 A라는 분자가 인산화 되어 그 활성이 없어짐으로써 A가 다른 분자 B와 결합할 수 없게 되기 때문에 A를 인산화시키는 효소 X가 B로 신호를 전달한다고 해석하는 것이 가능한 반면, 하나의 단백질이 세포 내에서 이동하여 세포질에서 핵 안으로 위치가 바뀌는 것도 하나의 신호전달로 해석하기도 한다. 그림 2에서는 H_2O_2가 PTP1B를 산화시킴으로써 PTP1B에게 신호를 전달한다고 해석하였다. 이와 같이 화학적, 생물학적으로는 전혀 다른 종류의 작용이라 할지라도 동일

하게 '신호를 전달'하는 것으로 해석함으로써 세포가 외부로부터 받은 자극이 일련의 '신호전달' 과정을 통하여 세포의 생성과 사멸 등에 영향을 미치는지를 이해하고자 하는 것이다.

ROSPath 모델에서는 세포 신호전달 경로를 표현함에 있어서 의미적으로 가장 기본적인 단위의 구성요소가 되는 것을 신호전달 스텝이라 부른다. 위에서 서술한 바와 같이 각 신호전달 스텝은 실제로는 다양한 화학적 또는 생물학적 작용을 나타내므로 ROSPath 모델에서도 다양한 형태로서 존재하게 된다. 예를 들어 하나의 신호전달 스텝은 하나의 단백질 상태나 화학물질에 의해 조절되는 다른 단백질의 상태 전이, 혹은 단백질 상태에 의해 조절되는 화학반응으로 표현될 수 있다.

그림 2에서 H_2O_2는 PTP1B-SH에서 PTP1B-SOH로의 상태 전이를 활성화함으로써 신호를 전달한다. 즉, H_2O_2와, PTP1B-SH에서 PTP1B-SOH로의 상태 전이 사이에 신호전달 스텝을 정의한 것이다. 이처럼 신호전달 경로가 화학물질을 촉매로 하는 상태 전이로 모델링 될 수 있다. 이 경우 만약 H_2O_2에 의해 활성화되는 단백질의 상태 변화에 관해 밝혀진 바가 없다면 이를 단순히 H_2O_2와 단백질 사이의 신호전달 스텝으로만 모델링 할 필요가 있다. 그림 2에서 PIP3와 PDK 사이에 정의된 신호전달 스텝은 이런 경우가 된다. 이와 유사하게, PI3K와 PIP2 → PIP3 반응 사이에서와 같이 단백질과 화학 반응에 대해 신호전달 스텝을 정의할 수도 있고, 단백질과 단백질 상태 전이 사이에서 신호전달 스텝을 정의하는 것도 가능하다. 이 보다 불확실한 경우에는 단순히 두 개의 단백질 사이에 신호전달을 정의하는 것도 필요하다. 그림 2에서 Ras와 Raf 사이의 신호전달 스텝이 이런 경우이다. 즉, 신호전달 스텝의 시작과 끝에는 각각 단백질, 단백질 상태, 복합단백질, 복합단백질 상태, 화학물질 등이 정의될 수 있고, 결합, 단백질 상태 전이, 또는 반응으로 양단간의 관계를 명시할 수 있어 다양한 조합으로 신호전달을 정의할 수 있다.

위와 같이 신호전달 경로가 완전히 밝혀지지 않아 불확실한 신호전달 과정을 모델링해야 하는 경우를 좀더 확장하여 생각한다면, 직접적으로는 신호전달 스텝이 알려진 바가 없다고 해도 알려지지 않은 어떤 경로로 신호전달이 된다고 믿는 경우가 있을 수도 있고 이를 표현해야 할 필요가 있을 수 있다. 그림 2에는 이와 같은 예가 없으나 만약 Ras와 MEK사이의 신호전달 물질인 Raf가 밝혀지지 않은 경우, Raf가 관여한다는 것은 모르지만 Ras에서 MEK로 신호가 전달된다는 것을 표현할 필요가 있다면, 직접적으로는 신호전달을 정의할 수 없으므로 "간접 신호전달 (indirect reaction)"이라고 하는 형태의 특별한 신호전달 스텝을 사용하도록 한다.

위에서 정의하는 각 신호전달 스텝은 그 종류에 따라 활성화(activation), 억제(inhibition), 그리고 간접 여부로 규정할 수 있다. 그림 2에서는 보통의 화살표로 (→) 표시된 신호전달이 활성화, 화살표의 끝이 막힌 것으로 (⊣) 표시된 신호전달이 억제, 중간이 끊어진 화살표 (⇢)를 사용하여 표시된 신호전달은 간접 형태를 나타낸다.

3.3.9 신호전달 경로

ROSPath에서 하나의 경로는 신호전달의 각 단계를 구성하는 신호전달 스텝의 집합으로 모델링 된다. 앞에서 설명한 바와 같이 신호전달 스텝은 다시 단백질, 단백질 상태, 단백질 상태 전이, 화학물질, 결합, 반응 등을 이용하여 표현된다.

4. ROSPath 데이타의 입력 시스템

본 논문에서는 세포 신호전달 경로를 위한 데이타 모델의 설계와 이를 이용한 데이타의 입력에 주안점을 두고 있다. 현재 ROSPath 시스템[18,19]은 Linux (Redhat 7.2) 환경에서 Oracle9i (9.2.0)를 이용하여 구현되어 있으며 웹 환경에서 개방적으로 사용될 수 있도록 Apache 1.3.27 웹 서버와 Resin 2.1.6 웹 응용 서버에 기반을 두어 구현하였다. 다음은 그림 2의 PDGF 신호전달 경로에 포함된 PTEN의 입력 단계를 설명함으로써 실제로 생물학자가 어떤 데이타 모델링 과정을 거쳐서 ROSPath 데이타베이스에 입력하게 되는지를 보인다.

특정 단백질의 정보는 ROSPath 홈페이지에서 인증절차(Author Login)를 거친 후 상단 'Edit' 메뉴의 'Protein' 버튼을 누를 때 나타나는 그림 3의 화면에서 입력될 수 있다. 이 화면에서 사용자는 입력하고자 하는 단백질의 이름이나 접근 번호를 우선 입력한 뒤 'External Info' 버튼을 눌러 현재 정의하려는 단백질에 관한 기본적인 정보를 외부 데이타베이스로부터 자동적으로 입력한다. 일부 정보는 하이퍼 링크 형식으로 제공되어 외부의 웹 데이타를 직접 접근할 수 있으며, 세포내 위치 이동 정보 즉, 세포내의 위치, 티슈 정보, 효소 반응의 정보는 사용자가 직접 입력을 하도록 하였고, 세포내의 활성의 경우 유전자 온톨로지 브라우저(gene ontology browser)인 QuickGO에서 제공하는 내용을 확인하여 사용할 수 있도록 하였다. 세포내의 활성에 QuickGO에서 제공하고 있는 것 외에 새로운 정보를 추가하려면 직접 유전자 온톨로지 분류 체계를 검색하면서 적당한 용어를 추가할 수 있도록 하였다. 참고문헌의 입력은 현재 입력하고자 하는 단백질의 이름 및 동의어를 키워드로 사용한 PUBMED 검색 결과를 웹 브라우저에서 볼 수 있도록 하였다. 입력자는 이 검색 결과에서 원하는 문헌만을 선택하여 ROSPath로 가져올 수 있다. 그림 3은 이와 같이 입력된 PTEN의 예를 보여준다.

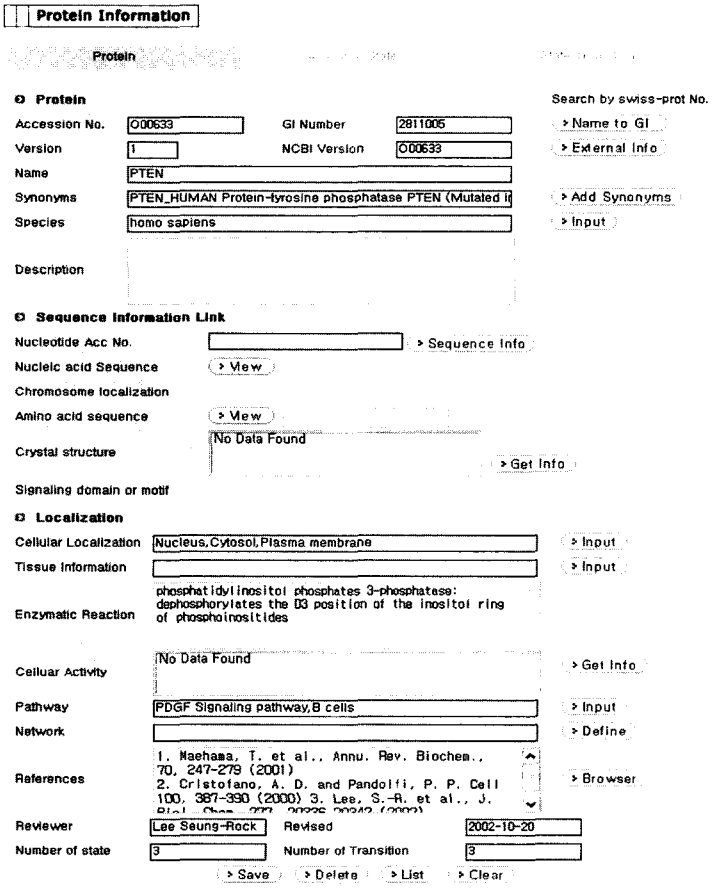

그림 3 단백질 정보 입력 화면

단백질을 정의하게 되면 하나의 디폴트 상태가 자동으로 생성된다. 디폴트 상태 이외에 추가적으로 새로운 상태를 정의하려면 그림 4에 나와 있는 단백질의 상태 정보 입력 화면을 이용한다. 여기서 단백질 합성 후 변형 정보나, 리간드, 세포주(cell line), 복합체 형성(complex formation), 복합체 구성성분(complex component), 세포내의 위치, 절단된 부분 형태(truncated form) 등과 같은 정보들은 우측의 'Select' 버튼을 눌렀을 때 나타나는 팝업 메뉴를 이용해 입력할 수 있다. PTEN의 경우 단백질 합성 후 변형으로서 산화되어 있는 상태와 시스테인 환원이 된 상태를 각각 정의할 수 있다.

일단 단백질의 상태 정보가 정의되면 서로 다른 두 상태 간의 전이 정보를 정의할 수 있다. 그림 5의 상태 전이의 입력 페이지에서는 시작 상태와 종료 상태를 이미 정의된 단백질 상태의 리스트에서 선택할 수 있다. 상태 전이를 정의하는 두 상태에 대해서 앞서 입력한 내용이 충분한 경우, 시작과 종료 상태 사이의 차이를 구하여 현재 정의하고자 하는 상태 전이에 대한 정보를 자동으로 생성하여 제공한다. 입력자는 이를 검토한 후, 자동으로 생성된 상태 전이 내용을 수정할 수도 있다. PTEN의 경우 시스테인 환원 상태를 시작으로 하고 산화 상태를 종료 상태로 하는 상태 전이를 정의할 수 있는데, 이 경우 두 상태 사이의 전이가 단백질 합성 후 변형의 차이에 의해 표현된다. 이때 상태 전이를 촉진하는 물질인 H_2O_2를 상태 전이 정의의 일부로 포함할 수 있다.

이와 같이 정의된 단백질의 상태 전이를 근거로 신호전달 스텝을 정의하고자 할 경우에는 *신호*를 전달하는 분자와 전달 받는 분자를 우선 명시하고, 이 둘 사이의 관계가 구체적으로 어떤 상태 전이 또는 단백질 사이의 결합에 근거한 것인지를 정의해 주어야 한다. PTEN의 경우, H_2O_2를 '신호'의 전달자로, PTEN을 '신호'의 피전달자로 정의한 다음, 이와 같은 개념적이고 논리적인 신

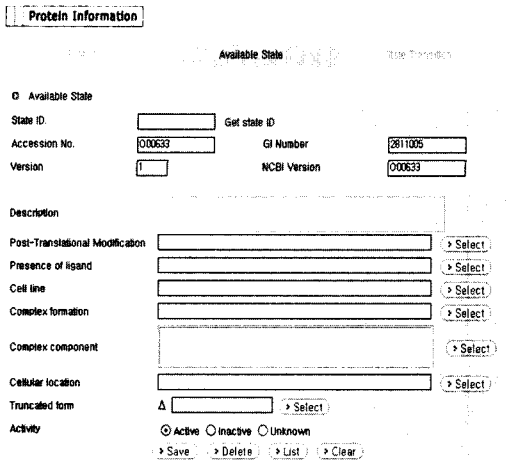

그림 4 단백질 상태 입력 화면

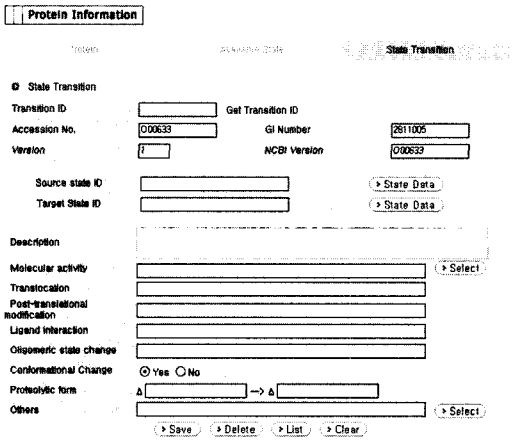

그림 5 단백질 상태 전이 입력 화면

호전달 스텝이 생화학적으로는 PTEN이 산화되는 상태 전이에 H2O2가 촉매로 작용하는 것에 근거한 것임을 표현한다.

5. 결 론

본 장에서는 ROSPath 데이타베이스 시스템의 특징을 기존 데이타베이스 시스템과 비교하여 기술하고 향후 연구방향에 대해 설명한다. 표 1은 ROSPath 데이타베이스 시스템을 신호전달 경로에 관한 주요 데이타베이스들과 모델링 대상, 데이타 모델, 주요 특징의 측면에서 비교한 것이다. 이들 시스템은 신호전달 경로에 참여하는 분자 또는 단백질을 주요 모델링 대상으로 하였으며, 관계 데이타 모델 또는 객체 지향 데이타 모델을 기반으로 한다. ROSPath 데이타베이스 시스템의 특징을 기존의 세포 신호전달 데이타베이스들과 데이타 모델의 관점에서 비교할 때 다음과 같이 몇 가지 특징을 언급할 수 있다. 첫째로 가장 기본적인 분자 구성을 단백질 상태로 모델링 하였다는 점을 들 수 있다. 본 연구에서는 단백질의 상태 정보를 체계적이고 포괄적으로 도입함과 동시에 상태에 관한 불완전한 지식의 표현을 허용하였다. 둘째, 신호전달경로를 독립적인 데이타 개체로 표현하였다. 신호전달 경로를 명시적인 데이타 객체로 표현할 것인가의 선택은 단순하게 개개의 신호전달경로를 표현할 때에는 차이가 없을 수도 있으나 서로 간의 연관성이 있는 여러 신호전달경로의 네트워크를 표현하여 궁극적으로 시스템 생물학을 지향하는 현재의 추세

표 1 주요 신호전달 경로 데이타베이스의 비교

데이타베이스	모델링 대상	데이타 모델	주요 특징
LiveDIP	Biological protein interactions	relational data model	• 단백질의 상태를 중심으로 단백질 상태 사이의 상호작용을 모델링
Transpath	Gene-regulatory pathways	object-oriented model	• 대사 작용에 관여하는 분자와 단백질을 중심으로 경로 모델링 • 경로를 독립적인 객체로 모델링
KEGG	Metabolic pathway	deductive and object-oriented model	• 유전자나 분자 사이의 상호작용에 대한 정보전달 경로를 모델링
AFCS	Cellular Signaling molecules	object-relational model	• 신호전달에 대한 여러 실험기관의 연구결과들을 종합, 관리하기 위한 시스템을 제공하고 있으나, 신호전달 경로를 표현하기 위한 데이타 모델이 확정되지 못한 상태임
AMAZE	Biological process or pathways	object-oriented model	• entity-relationship에 근거한 구체적인 데이타 모델 존재 • 단백질 상태에 관한 표현 미흡
ROSPath	cell-signaling pathways	object-oriented model	• 단백질 상태 및 복합단백질 상태의 전이와 결합, 반응 등을 모델링하며, 신호전달 경로를 독립적으로 모델링 • 불완전 정보의 명세

로 볼 때, 신호전달경로를 독립적인 데이타 객체로서 표현하는 것이 필수적이다. 셋째, 여러 단백질의 상호작용으로 새로운 복합체를 구성하는 결합 과정과 결합의 결과물인 복합단백질을 독립적으로 모델링하였다. 복합단백질과 복합단백질 상태도 독립적으로 신호전달에 참여할 수 있게 하였으며, 화학 반응이나 연합/분해 작용도 반응의 범주로 간주하여 신호전달 과정으로 정의하였다. 넷째, ROSPath 데이타 모델에서는 아직 완벽하게 밝혀지지는 않았으나 불완전한 형태로라도 생물학자들이 가지고 있는 지식을 쉽게 표현할 수 있도록 하기 위해 가능한 한 많은 부분을 계층적인 구성을 이용하여 표현하였다. 사용자는 전체 계층구조상에서 자신의 지식을 표현할 수 있는 적당한 수준을 선택하여 그 수준에서만 서술하면 된다. 마지막으로 확장성에 대한 고려를 들 수 있다. 세포 신호전달에 관련된 단백질에 관한 정보는 끊임없이 생산되고 있으며 새로운 정보가 밝혀짐에 따라 기존에 가지고 있던 데이타 모델에 대한 수정 보완 또한 필연적일 것으로 예상된다. 특히 세포 신호전달에 관련한 데이타베이스는 최근에 그 구축이 시작되어 많은 지식이 축적되어 있지 않고, 새로이 생성되는 데이타를 수용하기 위한 모델 수정이 빈번할 수 있으므로 객체지향 모델링을 하는 것이 바람직하다고 판단하였다. 그러나 기존의 생물학 데이타베이스들과의 연결 또는 통합 검색 측면과 시스템 안정성 측면에서 고려할 때 관계 데이타베이스 시스템을 사용하는 것이 바람직하므로, 본 연구에서는 객체지향 방법론을 활용하여 논리적인 데이타 객체를 구성하되, 이의 구현은 관계 데이타베이스 시스템을 사용하였다.

현재 ROSPath 데이타베이스 시스템은 저장된 정보의 브라우징이나 키워드 검색을 제공하며, 임의의 두 단백질 상태 사이에 가능한 신호전달 경로에 대한 검색 결과를 경로 뷰어 (pathway viewer)를 통해 시각적으로 제공한다. 또한, 생화학 분야의 전문가들이 신호전달 정보를 그림으로 표현하고 입력할 수 있게 하는 그래픽 에디터에 대한 연구가 진행 중이다.

참 고 문 헌

[1] Duan, X., Xenarios, I., and Eisenberg, D., *"Describing Biological Protein Interactions in Terms of Protein States and State Transitions : THE LiveDIP DATABASE," Mol. Cell. Proteomics*, 1:104-116, 2002.

[2] Heinemeyer, T., Chen, X., Karas, H., Kel, A., Kel, O.V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F. and Wingender, E., *"Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms," Nucleic Acids Res*, 1;27(1):318-322, 1999.

[3] Krull, M., Voss, N., Choi, C., Pistor S., Potapov, A., Wingender, E., *"Transpath: an integrated database on signal transduction and a tool for array analysis," Nucleic Acids Research*, 31:97-100, 2003.

[4] Kanehisa, M. and Goto, S., *"KEGG: kyoto encyclopedia of genes and genomes," Nucleic Acids Res*, 1;28(1):27-30, 2000.

[5] Williams, B.R.G., *"Signal Integration via PKR," Science's STKE*, Vol.2001, Issue.89, re2, 2001.

[6] Li, J., Ning, Y., Hedley, W., Saunders B., Chen, Y., Tindill, N., Hannay, T., Subramaniam, S., *"Molecular Pages database,"* Nature, 420:716-717, 2002.

[7] Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilber, D. and Wodak, S.J., *"Representing an analysing molecular and cellular function in the computer," Biol Chem*, 381(9-10):921-935, 2000.

[8] Kim, H.J., Song, E.J., and Lee, K.J., *"Proteomic analysis of protein phosphorylations in heat shock response and thermotolerance", J Biol Chem.*, 277 (26):23193-23207. 2002.

[9] Tallquist, M, Sorjano, P, and Klinghoffer, R., *"Growth factor signaling pathways in vascular development," Oncogene*, 18:7917-7932, 1999.

[10] http://www.ncbi.nlm.nih.gov/Genbank/index.html

[11] http://www.ebi.ac.uk/swissprot/

[12] http://www.rcsb.org/pdb/

[13] http://www.ebi.ac.uk/interpro/

[14] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[15] Lee, S.R., Yang, K.S., Kwon, J., Lee, C., Jeong, W., and Rhee, S.G., *"Reversible inactivation of the tumor suppressor PTEN by H_2O_2," J Biol Chem.*, 277(23):20336-20342, 2002.

[16] http://www.geneontology.org/

[17] http://www.bind.org/

[18] Park, J.S., Park, J.K., Lee, S.H., Lee, S.R., Lee, K.J. and Paek, E., *"ROSPath: a database of Reactive Oxygen Species mediated cell-signaling Pathways", Keystone Symposia 2003 Abstract Book, Proteomics: Technologies and Applications*, pp.70, 2003.

[19] http://rospath.ewha.ac.kr

박 지 숙

1990년 한국과학기술원 전산학과(공학사) 1992년 서울대학교 컴퓨터공학과(공학석사). 1998년 서울대학교 컴퓨터공학과(공학박사). 1999년~2000년 한국전자거래진흥원 EC진흥부. 2000년~2002년 삼성 SDS Biz. Modeling팀. 2002년~현재 서울여자대학교 정보통신공학부 조교수. 관심분야는 데이타베이스, 멀티미디어 시스템, 생물정보학

백 은 옥

1985년 서울대학교 전자계산기공학과(공학사). 1991년 Stanford University *Computer Science Dept.*(전산학 박사) 1992년~1995년 서울대학교 컴퓨터신기술공동연구소. 1995년~2000년 LG전자기술원. 2001년~현재 서울시립대학교 기계정보공학과 조교수. 관심분야는 인공지능, 지식표현 및 추론, 생물정보학

이 공 주

1977년 이화여대 약대 제약학과(약학사) 1979년 한국과학기술원 생명공학과(이학사). 1986년 Stanford University, Dept. of Chemistry(Ph.D). 1986년~1988년 Stanford Medical School (Post-doctoral fellow). 1989년~1994년 한국표준과학연구원(선임연구원). 1994년~현재 이화여자대학교 약학대학 제약학과/분자생명과학부 교수. 관심분야는 stress와 혈관신생 관련 신호전달관련 연구, 프로테오믹스를 이용한 system biology, 단백질 변형연구

이 상 혁

1985년 서울대학교 화학과(이학사). 1987년 서울대학교 화학과(이학석사). 1994년 Cornell University 화학과(이학박사) 1994년~1995년 Princeton University (박사후연구원). 1995년~현재 이화여자대학교 화학과(교수). 2000년~2001년 미국 국립보건원 암센터(방문연구원). 2002년~현재 이화여자대학교 분자생명과학부 부교수. 관심분야는 생물정보학, 유전체학, 데이타마이닝

이 승 록

1987년 서울대학교 미생물학과 이학사 1990년 서울대학교 미생물학과 이학석사 1995년 서울대학교 미생물학과 이학박사 1996년~2001년 National Institutes of Health, USA, Visiting fellow. 2001년~현재, 이화여자대학교 분자생명과학부 조교수. 관심분야는 활성산소에 의한 신호전달

양 갑 석

1987년 서울대학교 미생물학과(이학사) 1989년 서울대학교 미생물학과(이학석사) 1997년 서울대학교 미생물학과(이학박사) 1997년~2002년 미국 국립보건원 Post-doctoral fellow. 2002년~현재 이화여자대학교 분자생명과학부 연구교수. 관심분야는 활성산소종에 의한 세포신호전달, Bioinformatics, Proteomics