단백질 구조 정보 분석을 위한 바이오 온톨로지

남덕우[O] 예형석 진 훈 김인철
경기대학교 전자계산학과
{ndw76[O], elta, jinun, kic} @kyonggi.ac.kr

Bio-ontology for Analyzing Protein Structure Information

Deokwoo Nam[0] Hyoungsuk Ye Hoon Jin Incheol Kim
Dept. of Computer Science, Kyonggi University

요 약

생물정보학 분야에서의 온톨로지는 다양한 생물학적 의미들을 표현하는 구조로 되어 있으며, 생물학 데이터의 의미를 효과적으로 해석할 수 있는 매우 중요한 기술로 인식되고 있다. 특히 바이오 온톨로지는 생물학 데이터베이스로부터 정보에 대한 탐색과 추론 등 의미 전달 과정에서 중심적인 역할을 수행한다. 본 논문에서는 단백질 구조 예측을 지원하는 다중 에이전트시스템인 APSS내에서 각 구성원 에이전트들간에 온톨로지에 기초한 정확한 구조 정보의 전달을 통해 효과적인 단백질 구조 예측 작업을 지원하고자 한다. 이를 위하여 먼저 단백질 구조 관련 바이오 온톨로지의 설계방법을 제시하고, 이것에 기초한 실제 바이오 온톨로지의 설계에 대해 설명한다. 그리고 이렇게 구축된 단백질 구조 온톨로지를 APSS시스템 안에서 어떻게 응용하였는가에 대해서도 설명한다.

1. 서 론

생물정보학에서 관심을 갖는 데이터 집합은 데이터를 작성한 사람이나 기관, 연구 목적 등의 차이로 인해 구조적으로나 내용적, 그리고 의미적으로 볼 때 이질적인 면이 많이 존재한다[1]. 이런 문제점들을 해결하기 위해 1990년대 후반부터 생물정보학 분야에 적용된 개념이 바로 온톨로지이며 특정 분야에 관한 지식들을 온톨로지화 하였을 때 이를 이용하는 측면에서 데이터가 갖는 개념들 간의 관계 및 속성을 고려하지 않고도 쉽게 정보에 대한 해석작업을 수행할 수 있다.

APSS(Agent-based Proteomics Support System) 는 개방형 네트워크를 지향하면서도 데이터 간의 이질성을 고려하여 효과적으로 미지의 단백질에 관한 3차 구조 예측 작업을 돕기 위해 제안된 에이전트 기반의 시스템으로서 단백질의 3차 구조와 관련된 정보의 해석 및 구조 예측 작업을 위해 온톨로지를 이용하는 시스템으로 개발되고 있다[2]. APSS는 크게 자원, 분석, 관리 및 조정, 제어, 사용자 인터페이스 에이전트들로 구성되며 본 논문에서는 자원 에이전트인 PDB 에이전트와 관리 및 조정 에이전트인 CODY 에이전트 사이에서 동작하는 온톨로지에 관하여 설계하고 응용에 대하여 기술하고자 한다. 2장에서는 생물학 정보를 다루는 온톨로지 시스템들에 관하여 소개하고 3장에서는 바이오 온톨로지 및 APSS 내에서 PDB 에이전트가 수행하는 데이터를 온톨로지로 구축하기 위한 설계과정, 마지막으로 4장에서는 이와 같이 구축된 바이오 온톨로지 응용시스템으로서의 APSS에 관해 기술하고자 한다.

2. 관련연구

온톨로지는 전산학 분야에서 사용된 초기부터 생물정보학 분야에도 적용되어 많은 온톨로지들이 개발되었는데 이를 소개하면[4] Gene Ontology(GO)는 알려진 진핵세포 (eukaryote)의 잘 알려진 생물학적 기능을 중심으로 분류한 유전자 서열정보에 관한 온톨로지로 DNA 대사작용, 분자 기능, 다른 세포에도 적용될 수 있는 진핵(성숙)세포 온톨로지를 포함한다. Molecular Biology Ontology(MBO)는 개념과 관계를 분자생물학을 중심으로 포함하고 있는 온톨로지이다. Tambis Ontology(TaO)는 단백질 , 제한효소, 2차-3차 구조 등의 다양한 정보 검색과 분석을 위한 온톨로지로서, Tambis 시스템은 사용자의 단일 질의에 대해 다중 검색을 지원하며 단일한 결과를 제공하는 특징을 갖는다.

3. 바이오 온톨로지와 개념분석

온톨로지는 인공지능 분야에서 다양한 영역을 서술하기 위해 개발되었고, 영역의 지식과 공유 기반으로서의 어플리케이션을 제공하는 수단으로 제안되었다. 그 후 온톨로지의 중요성은 점차적으로 생물정보학 분야에도 적용되었고 이를 연구하고 구축하려는 연구가 시작되었다[3]. 생물정보학 분야에 적용되어 구축된 온톨로지를 바이오 온톨로지라 하며 이를 구축해야 하는 필요성에 대하여 예를 들어 설명하면 다음과 같다.

의미의 이질성(semantic heterogeneity)은 두 가지 형태로 나타나게 되는데, 하나의 용어가 다른 의미를 갖거나,

두 개의 용어가 하나의 의미를 갖는 경우이다. 그러나 생물정보학에서의 이질적인 형태 또는 의미를 갖는 데이터베이스에서의 의미 해석은 다양한 형태로 나타날 수 있다. 예를 들어 고분자 집합체로서 촉매 역할을 하는 peroxidase라는 효소에 대한 표현은 PDB 데이타베이스에서 제공하는 mmCIF 형식과 PROSITE 데이터베이스의 포맷에서 다음과 같이 다르게 표현되고 있다.

MOLECULE: CY1TOCHROME C PEROXIDASE; mmCIF DE Cytochrome c peroxidase; PROSITE

이런 형태의 데이터 간의 이질성 문제를 해결하기 위해서는 두 가지의 다른 의미에 대한 의미를 정의하여 매칭하는 기능이 요구된다. 이런 기능은 직접적으로 기능하는 프로그램을 사용하여 두 가지 표현결과를 매칭하는 방법 또는 온톨로지의 사용을 통해 추론 단계를 거치면서 의미를 연결하는 형태로 동작할 수 있다. 바이오 온톨로지는 이러한 이질적인 생물정보 데이터들 간의 개념을 통합하고 의미를 연결하여 기능을 밝히는데 이용된다.

여러 데이터 자원들로부터 온톨로지 모델을 통하여 개념을 추출하는 과정은 일반적인 형태의 수식으로 표현할 수 있다. 전체를 포괄하는 온톨로지 개념은 여러 데이터베이스에서 사용하는 용어의 합으로 표현 된다. 이것은 이질적 데이터들 개념을 C_n라고 할 때 ($C_1 \cdots C_n$)에 대한 개념의 합집합에서 π_p 속성과 σ_q의 조건으로 개념을 추출할 때 다음과 같은 형태로 나타난다.

$$Concept \Leftrightarrow \pi_p \sigma_q (C_1 \cup C_2 \cup ... \cup C_n)$$

이렇게 얻어지는 개념은 아래와 같이 다른 개념에 대한 재귀적 개념으로 형성될 수 있게 된다

$$(Concept1\,(\,Concept2\,(\,Concept3...)))$$

이 식은 로컬 데이터베이스에 대해 질의를 통해 정보를 얻어 내거나, 여러 자원을 통합적으로 운용하는 시스템의 경우에도 동일하게 적용된다.

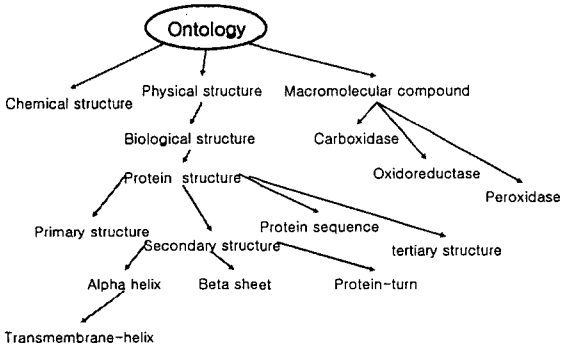

그림 1 온톨로지의 계층적 구조

[그림 1]은 데이터베이스로부터 얻어진 개념에 대해 바이오 온톨로지의 개념 모델을 통하여 구하고자 하는 개념의 위치를 탐색하는 과정을 나타낸다. 데이터베이스에서 얻어낸 정보에 대하여 온톨로지 모델에서의 의미를 찾을 수 있다는 것은 해당 정보에 대해 의미적인 통합을 시도할 수 있음을 의미한다. 다만 온톨로지의 의미에 대해서는 도메인 지식에 대하여 알고 있는 전문가의 도움을 거친 후에 비로소 정확한 의미를 갖는다는 제약을 갖고 있다. 이것은 아무리 확실하다고 판단되는 개념을 가지고 작성된 온톨로지 모델일지라도 생물학적 영역지식을 고려하여 구축되어야 한다는 것을 의미한다. 바이오 온톨로지는 생물학적 정보에 대해 생물학적 분류에 잘 맞아야 하고, 다른 사용자 관점을 고려하여 이를 포함해야 하며, 생물학적 개념과 그 관계에 대하여는 연산이 가능하여야 한다[1][4].

3. 단백질 구조 온톨로지의 설계

단백질의 3차 구조 예측을 지원하는 시스템인 APSS 에서 사용되어야 할 온톨로지는 다음의 4가지 사항들을 고려하여 개발되어야 한다[2]. 첫째, 다양한 형태로 이미 개발되었거나 개발 중인 단백질 자원들의 특성을 고려해야 한다. 현재 많은 사람들에게 공개되어 이용되는 단백질 자원들(예를 들어, PDB, SWISSPROT, PIR, BLAST, CLUSTALW 등)은 저마다의 목적을 가지고 서로 다른 사람 또는 기관에서 개발되었기 때문에 사용하는 데이터의 형식이나 저장방식 등이 다르다. 둘째, APSS 시스템은 위와 같은 다양한 자원들을 개방형 네트워크 시스템인 에이전트시티 네트워크 상에서 상호 쉽게 운용할 수 있도록 설계됨으로 인해 온톨로지 모델에 대한 접근이 자유로워야 하며 수정 및 추가 작업이 용이해야 한다. 셋째, 기존에 개발된 바이오 온톨로지들과 호환되면서도 이를 재사용할 수 있어야 한다. 넷째, 기존의 온톨로지들이 생물학 분야에서 발생하는 다양한 데이터를 모두 고려하여 개발됨으로 인해 폭넓은 개념들을 포함하는 대신 특정 세부 분야에 관해서 비전문화된 형태로 개발된 것에 반해 본 온톨로지는 단백질의 3차구조정보 예측과 관련되어 특화된 형태의 정보를 포함해야 한다. 이와 같은 내용들을 고려하여 단백질의 3차 구조정보 예측에 이용될 온톨로지를 설계하는 방법은 다음과 같다.

① 자원 에이전트 영역의 결정 : 고유의 데이터베이스로부터 정보를 추출하는데 있어 자원 에이전트의 정보를 얼마나 폭넓게 혹은 얼마나 깊은 의미를 추출할 것인가를 고려해야 한다.

② 기존의 바이오 온톨로지를 재사용 및 특화된 정보를 추가 : 용어의 생성에 있어서 원칙적으로 기존의 온톨로지에 존재하는 용어들을 이용하여 온톨로지 모델을 생성하지만 3차구조 예측작업과 관련된 내용들을 추가해야 한다.

③ 중요한 단백질 구조 관련 용어를 열거 : 자주 사용되는 용어에 대해서 의미의 중복을 피할 수 있도록 대표적 용어들을 상위계층으로 두어야 한다.

④ 클래스 및 클래스 계층 정의 : 클래스의 정의는 상/하위 구조의 계층을 고려해야 하며 클래스간의 관계는 속성들로 표현하고 정의하되 해당 자원의 정보 저장 및 표현 방식을 고려해야 한다.

⑤ 인스턴스 생성 : 클래스 수준에서 해당하는 값을 입력하여 인스턴스를 생성한다.

APSS 온톨로지의 계층적 모델은 [그림 2]와 같다.

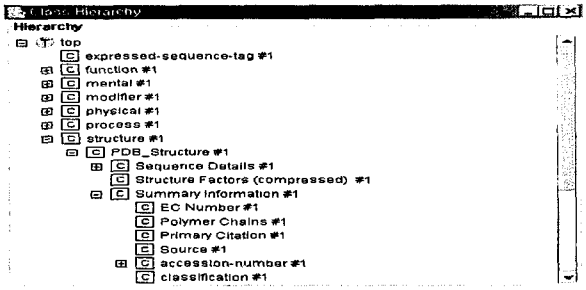

그림 2 APSS 온톨로지 모델

[그림2]는 PDB 데이터베이스에서 제공하는 정보를 대상으로 온톨로지 모델을 구축한 것이다. 그리고 단백질의 기능, 프로세스, 화학구조 등에 대한 용어들도 포함함으로써 넓은 의미영역에서도 이용될 수 있는 온톨로지로서의 구조를 가지도록 했으며, 이 외의 클래스들은 재사용성 및 확장성을 고려하여 포함된 것이다.

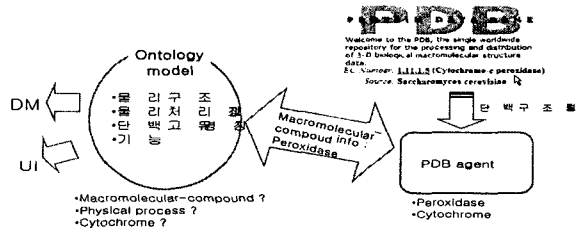

그림 3 온톨로지의 매핑

[그림 2]의 온톨로지 모델을 이용하여 데이터 분석 및 해석과정을 나타낸 것이 [그림 3]이다. 우측의 PDB 에이전트는 Paroxidase라는 날(raw) 데이터로서의 단어 정보만을 가지며 어떠한 관계나 의미에 대한 분석도 이루어지지 않은 상태이다. 이것은 에이전트가 데이터베이스로부터 가져온 내용을 단순 텍스트 정보로만 보유하고 있음을 의미한다. 온톨로지 모델에 매칭된 후 데이터는 에이전트 간 메시지 전달 과정에서 공통적 의미로 해석되어 에이전트들 사이에 상호 교환되고 공유될 수 있는 기반을 제공한다.

4. 단백질 구조 온톨로지의 응용

지금까지 서술한 온톨로지를 사용하기 위해서는 온톨로지를 생성하고 탐색 및 정보를 전달하기 위한 환경이 필요하다. 이를 위하여 CODY 에이전트에서는 온톨로지 생성을 위한 도구로서 PROT□G□를 사용하고 이의 해석 및 추론을 위해 JESS 엔진을 사용한다. JESS 엔진은 매우 효율적으로 데이터에 규칙을 적용할 수 있다는 것과, Java 언어로 작성된 모든 클래스들과 라이브러리를 사용할 수 있다는 장점을 갖는다. PROT□G□는 다양한 온톨로지 변환 기능을 통해 자유로운 변환작업이 가능하며 온톨로지 모델을 생성하는 역할을 담당한다. 또한 PROT□G□는 생성된 온톨로지에 대해 JESS 엔진의 추론결과를 바로 적용할 수 있다는 장점이 있다.

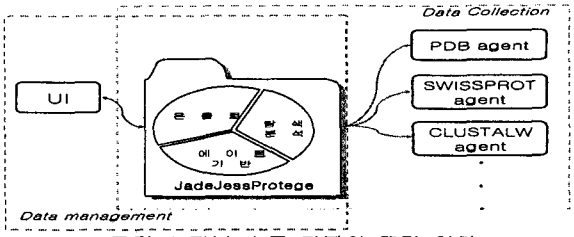

그림 4 정보 습득 과정의 중간 역할

[그림 4]는 JESS엔진과 PROT□G□를 이용한 환경을 도식화한 것이다. 우측은 자원 에이전트들로부터 정보를 추출하는 부분이며, 추출된 정보와 CODY의 온톨로지 매칭을 통해 정보를 습득한다. 좌측은 UI에이전트에서 질의 요청 시 CODY가 자원 에이전트로의 메시지 전달을 결정할 때 도움을 주게 된다. 이 때의 온톨로지는 JESS 엔진을 통해 질의에 대한 지식 기반 검색을 할 수 있도록 한다.

5. 결론 및 향후 연구

본 논문에서는 단백질 3차 구조 예측을 위한 바이오 온톨로지 설계의 중요한 고려사항과 설계방법을 제시하였다, 또한 이것의 응용으로서 단백질 구조 예측을 지원하는 다중 에이전트시스템인 APSS에서 PDB 에이전트와 CODY 에이전트 사이에서 이용될 온톨로지 모델을 구축하였다. 바이오 온톨로지에 기초한 단백질 구조 데이터의 보다 의미적인 해석으로 인해 CODY 에이전트는 능률적인 작업을 수행할 수 있다. APSS에 사용된 바이오 온톨로지 모델은 추후 확장을 통해 다양한 자원들로부터의 단백질 정보를 해석하고 추론하여 단백질의 3차 구조 예측을 위해 이용될 수 있다.

참고문헌

[1] A. Silvescu, J. Reinoso-Castillo, C. Andorf, V. Honavar, and D. Dobbs. Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

[2] 김현식 외. (2003). " 단백질 구조예측을 위한 에이전트 시티 네트워크 기반의 다중 에이전트 시스템" , 춘계정보과학회 Proceeding. Vol. 30, No. 1, pp.461-463

[3] P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A. Brass. An Ontology for Bioinformatics Applications. Bioinformatics, 15(6):510-520, 1999.

[4] R. Stevens, C.A. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. Briefings in Bioinformatics, 1(4):398-416, November 2000.