분산 생물정보 DB에 대한 GO 기반의 통합 시맨틱 질의 기법

(Integrated Semantic Querying on Distributed Bioinformatics Databases Based on GO)

박 형 우 † 정 준 원 † 김 형 주 ††
(Hyoung-Woo Park) (Jun-Won Jung) (Hyoung-Joo Kim)

요 약 최근 여러 생물학 연구 집단들은 연구의 효율 향상을 위해 그들의 연구 결과를 서로 공유하기 위한 노력을 하고 있다. 뿐만 아니라, 공통의 어휘를 이용하여 유전자의 기능을 기술하기 위해 통제된 어휘들로 이루어진 Gene Ontology(GO)라는 온톨로지를 구축하였다. 하지만 현재까지도 각 연구 집단들의 데이타는 분산되어 있고, 기존의 시스템들은 이처럼 분산된 데이타들에 대한 통합 질의를 지원하지 않고 있을 뿐 아니라, 각 연구 집단의 독자적인 어휘들과 GO와의 대응 관계에 대한 의미가 명확하게 기술되어 있지 않아 통합 시맨틱 질의가 근본적으로 불가능한 상태이다.

본 논문에서는 대응 관계의 의미를 결정하는 기법과, 통합 시맨틱 질의를 지원하는 인터페이스를 제안하였다.

먼저, 문자열 규칙과 다중도 분석 등을 통해 이러한 대응 관계의 의미를 반자동으로 결정해 주고 이렇게 결정된 대응 관계의 의미를 GO와 통합하여 통합 온톨로지를 생성해 주는 AutoGOA 시스템을 제안하였다.

또한, 대표적인 메타데이타 기술 모델인 RDF 모델을 이용하여 여러 데이타들을 통합하고 이렇게 생성된 통합 온톨로지를 이용하여 통합 시맨틱 질의를 지원하는 인터페이스인 GOGuide II를 제안하였다.

키워드 : Gene Ontology, RDF, 온톨로지, 시맨틱 웹

Abstract Many biomedical research groups have been trying to share their outputs to increase the efficiency of research. As part of their efforts, a common ontology named Gene Ontology(GO), which comprises controlled vocabulary for the functions of genes, was built. However, data from many research groups are distributed and most systems don't support integrated semantic queries on them. Furthermore, the semantics of the associations between concepts from external classification systems and GO are still not clarified, which makes integrated semantic query infeasible.

In this paper we present an ontology matching and integration system, called AutoGOA, which first resolves the semantics of the associations between concepts semi-automatically, and then constructs integrated ontology containing concepts from GO and external classification systems. Also we describe a web-based application, named GOGuide II, which allows the user to browse, query and visualize integrated data.

Key words : Gene Ontology, RDF, Ontology, Semantic Web

1. 서 론

생물정보학의 발달과 더불어, 여러 종들의 유전자 기능을 밝혀내는 작업이 여러 연구 집단들에 의해 이루어지게 되었다. 이후 연구의 효율을 높이기 위해 공통의 어휘를 사용할 필요성이 제기되었고, 이러한 필요에 따라 여러 연구 집단들은 Gene Ontology 컨소시엄을 결성하고 Gene Ontology(GO)[1]라는 공통의 온톨로지를 구축하였다. GO에는 생물학적 과정을 나타내는 개념들

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성지원사업(IITA-2005-C1090-0502-0016)과 BK21의 지원을 받아 수행되었음

† 학생회원 : 서울대학교 전기.컴퓨터공학부
hwpark@idb.snu.ac.kr
jwjung@idb.snu.ac.kr
†† 종신회원 : 서울대학교 전기.컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2006년 1월 26일
심사완료 : 2006년 5월 9일

과 세포의 각 부분을 나타내는 개념들, 그리고 분자의 기능을 지칭하는 개념들이 포함되어 있을 뿐 아니라, 이들 간의 의미적 관계(동의어, 상위어, 하위어 관계 등)들이 정의되어 있다. 또한, 여러 연구 집단들은 그들이 유전자의 기능을 기술하는 데 사용하는 어휘들이 GO 상의 어떤 개념들과 밀접한 관련이 있는지 명시함으로써 다른 연구 집단들이 그들의 데이타를 잘 이해할 수 있도록 돕고 있다.

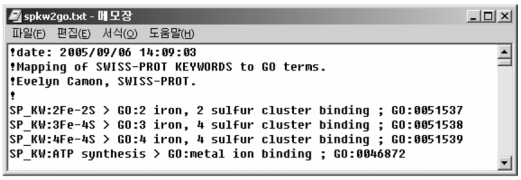

그림 1 UniProt의 어휘들과 GO와의 대응 관계

하지만 그림 1에서 볼 수 있듯이 현재 이러한 대응 관계는 그 의미가 명확하게 기술되어 있지 않다. 각 연구 집단들은 자신들의 웹 사이트에 그들의 데이타와 함께 이러한 대응 관계를 단순 하이퍼링크의 형태로 제공하고 있을 뿐이다. 따라서 연구자들은 이러한 링크를 통해 개별 사이트를 탐색해야 한다.

예를 들어, 다음과 같은 질의를 생각해 보자.

질의 1. GO 상에서 UniProt[2] 키워드인 'Transcription'과 동의어이거나 'Transcription'의 하위어인 개념들을 찾고, 이러한 개념들과 동일한 의미를 갖는 작용들을 Reactome[3]에서 찾아라. 뿐만 아니라 이렇게 찾은 작용들의 선행 작용들과 후행 작용들, 그리고 올소로구스(otrhologous) 작용들도 찾아라.

현재와 같이 생물 정보 DB가 분산되어 있고 통합 시맨틱 질의가 지원되지 않는 상황에서 연구자들은 다음과 같은 과정을 거쳐야 원하는 결과를 얻을 수 있다.

1) UniProt에서 키워드 'Transcription'을 검색해 GO 상의 개념들로의 하이퍼링크들을 찾는다.
2) GO 상에서 1)에서 찾은 개념들 중 'Transcription'과 동의어이거나 'Transcription'의 하위어인 개념들을 찾고, 이렇게 찾은 개념들의 하위 개념들도 모두 찾는다.
3) 2)에서 찾은 개념들 각각에 대해 Reactome 상의 작용에 대한 하이퍼링크들을 찾고 이들 중 2)에서 찾은 개념과 동일한 의미를 갖는 Reactome 작용들 및 그들의 선행 작용들과 후행 작용들, 그리고 올소로구스 작용들도 찾는다.

연구자들이 이러한 과정을 거쳐 원하는 자료를 찾기 위해서는 수 시간에서 수 일이 소요될 수도 있다. 따라서, 여러 사이트들의 데이타 간의 복잡한 의미적 관계를 효율적으로 검색하기 위해서는 무엇보다도 먼저 대응 관계들의 의미가 결정되어야 한다. 그림 2는 대응 관계의 의미 결정을 통해 앞의 질의1에 대한 결과가 명확해지는 과정을 보여준다.

또한, 오른쪽 그림은 대응 관계의 의미가 결정된 후의 통합 RDF(Resource Description Framework) 모델을 보여주고 있다. RDF는 데이타 간의 연관관계와 부가정보를 기술하는 표준으로서, 그래프 상에서 술어(predicate)는 화살표로 표현되고, 주어(subject)와 목적어(object)는 자원(resource)의 경우 타원으로, 상수 값(literal)의 경우 직사각형으로 표현된다. 이처럼 대응 관계의 의미가 결정되고 데이타가 통합된 상태에서 통합 시맨틱 질의를 통해 원하는 결과를 얻을 수 있다는 것을 알 수 있다.

본 논문에서는 먼저 여러 연구 집단들의 분류 체계와 GO 사이의 대응 관계를 분석하여 대응 관계의 의미를 결정하고 RDF 모델을 이용하여 여러 연구 집단들의 데이타들을 통합하는 반자동 기법을 제시한다. 이와 더불어, 본 논문에서는 이렇게 통합된 데이타들에 대해 통합 시맨틱 질의를 손쉽게 수행할 수 있도록 도와주는 인터페이스를 제시한다. 본 논문에서 제안하는 시스템은 연구자들이 원하는 결과를 효율적으로 질의할 수 있도록 도와주며, 시맨틱 질의를 통해 복잡한 관계를 질의할 수 있게 해 준다. 뿐만 아니라 현재의 분산된 환경에서는 발견하기 힘든 정보들을 통합 질의를 통해 손쉽게 발견할 수 있도록 해 준다.

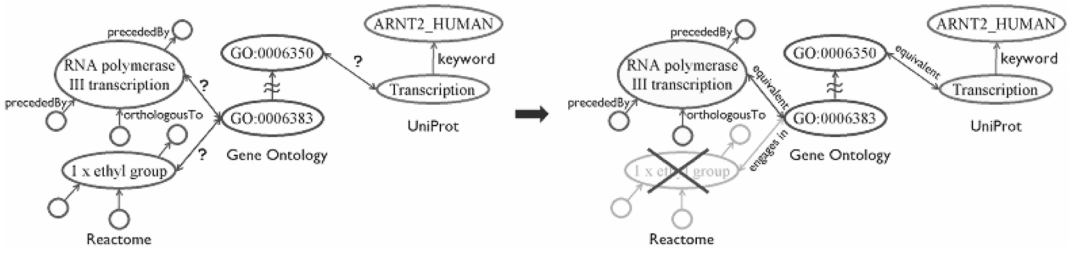

그림 2 대응 관계의 의미 결정

본 논문의 구성은 다음과 같다. 먼저 2장에서는 생물정보학 분야에서 주로 사용되는 다른 온톨로지 브라우징 시스템들에 대해 살펴본다. 그리고 3장에서는 본 논문에서 사용되는 데이타 모델과 질의 언어에 대해 설명하고 기존의 온톨로지 통합 기법에 대해 살펴본다. 4장에서는 본 논문에서 구현한 통합 시맨틱 질의 시스템의 구조와 각 구성 요소들에 대해 설명한다. 5장에서는 본 논문에서 구현한 시스템의 구현 결과를 살펴보고, 데이타 통합 결과가 얼마나 실제 연구자들의 판단 결과와 일치하는지 검증한다. 마지막으로 6장에서는 본 연구의 의미를 정리하고 향후 연구를 기술한다.

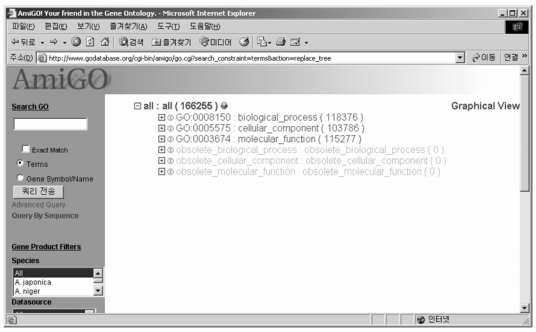

그림 3 AmiGO

2. 관련 연구

현재 생물학 온톨로지와 관련해 유전체들에 대한 검색을 지원하는 시스템들 중 가장 널리 사용되는 AmiGO [4], 특정 데이타베이스에 대해 제한적이지만 논리적 복합 질의가 가능한 GOView & GOGet[5], 그리고 시맨틱 질의를 지원하는 GOGuide[6]에 대해 살펴보자.

2.1 AmiGO

GO 컨소시엄에서 구현한AmiGO는 웹 기반 시스템으로서, GO 상의 개념들에 대한 설명을 제공하고, 개념들 간의 관계와 타 데이타베이스와의 연결 관계를 보여준다. AmiGO의 전체적인 기능은 다음과 같다.

- GO 브라우징
- GO 상의 개념들과 그에 연관된 유전체들 검색
- 특정 조건에 맞는 유전체들 필터링
- 타 데이타베이스/사이트로의 하이퍼링크

이처럼 AmiGO의 기능은 주로 GO의 브라우징 및 검색에 초점이 맞추어져 있다. 따라서 타 데이타베이스/사이트와의 연관 관계를 알기 위해서 연구자는 하이퍼링

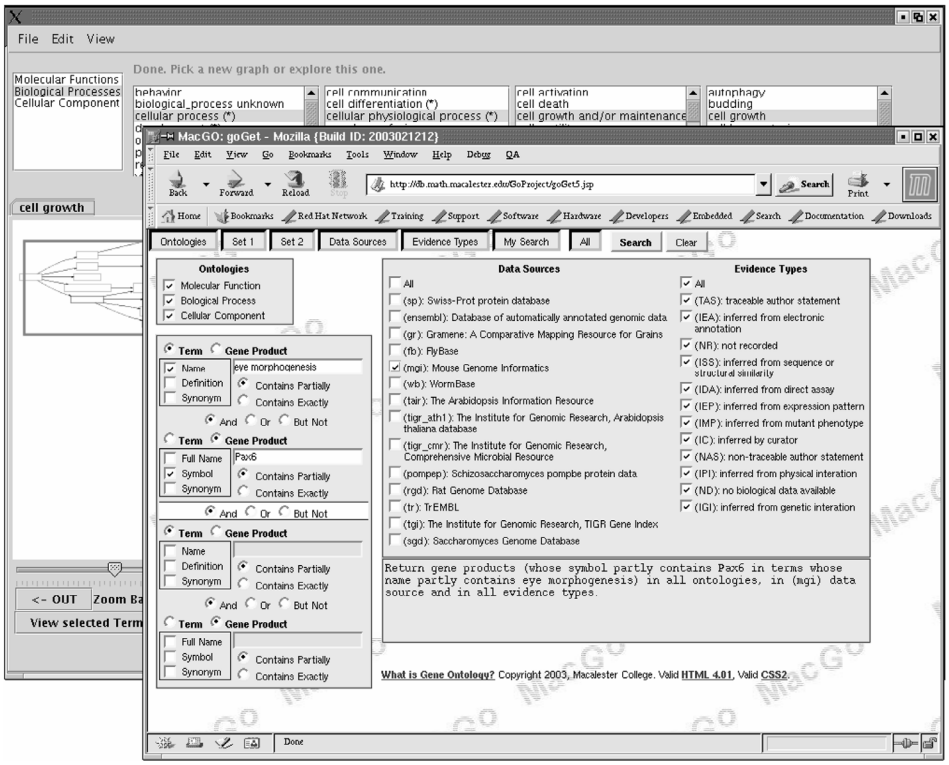

그림 4 GOView & GOGet

크를 이용해 직접 해당 데이타베이스/사이트를 방문하여야 한다. 뿐만 아니라, 현재 AmiGO는 GO 상의 특정 개념을 검색하거나 특정 유전체를 검색하는 식의 단순한 요구 사항에 대해서는 그 결과를 제공해 줄 수 있지만, 특정 개념의 하위 개념이나 특정 개념들의 공통 상위 개념을 찾는 식의 복잡한 시맨틱 질의에 대해서는 단순 브라우징 외에 별다른 방법을 제공하고 있지 못하다.

2.2 GOView & GOGet

GOView는 GO 상의 개념들 간의 관계를 DAG 형태로 보여주며, 사용자가 원하는 항목들을 선택하면 GOGet을 띄워 선택된 항목들에 관한 상세 정보를 보여준다. GOGet은 GOView에서 선택된 항목들을 보여주는 기능 외에도, 사용자에 의해 선택된 특정 저장소들로부터 특정 조건을 만족하는 데이타를 가져와 보여주는 기능을 갖고 있다. 사용자는 특정 저장소에서 특정 증거형(evidence type)을 갖는 개념이나 유전체를 검색할 수 있고, 여러 검색 결과들에 대한 논리 연산(AND, OR 등)의 결과를 볼 수도 있다. 그러나 이러한 논리 연산으로는 검색 결과들을 합치거나 걸러낼 수 있을 뿐이고, 온톨로지 상의 개념들 간의 관계를 명시하는 것은 불가능하다. 즉, 논리적 포함과 배제는 가능하지만 시맨틱 질의와 같은 고급 질의는 불가능하다.

GOGet은 여러 저장소의 데이타들을 통합해서 보여준다는 점에서 단순한 GO 브라우저에 비해 훨씬 편리하다. 그러나 GOGet은 대상이 되는 저장소들이 동질적인 구조를 갖는다는 전제를 필요로 하며, 개념들 간의 관계보다는 유전체에 초점을 맞추고 있고, 관계형 데이타 모델에 따라 데이타가 저장되어 있어 통합 시맨틱 질의가 구조적으로 불가능하다는 한계를 지닌다.

2.3 GOGuide

본 연구의 전 단계로서 구축된 생물학 온톨로지 검색 시스템인 GOGuide는 현재 기본적으로 다른 온톨로지 브라우징 시스템들이 제공하는 브라우징과 검색 기능을 통합적으로 제공할 뿐 아니라, GO 상의 개념들 간의 복잡한 의미적 관계를 활용한 질의 기능을 제공하고 있다. 연구자들은 시맨틱 질의를 통해 단순 키워드 검색으로는 찾기 힘든 복잡한 관계들을 손쉽게 찾을 수 있다. 하지만 GOGuide 역시 타 데이타베이스나 타 사이트로의 하이퍼링크를 제공할 뿐, 다른 연구 집단들의 데이타에 대한 통합적인 질의는 수행하지 못한다.

기존의 시스템들과 본 논문에서 제안한 시스템의 기능상 차이점은 표 1과 같다.

3. 통합 대상 및 기법

이 장에서는 본 연구에서 통합 대상으로 삼은 UniProt과 Reactome에 대해 간략히 서술하고 본 연구에

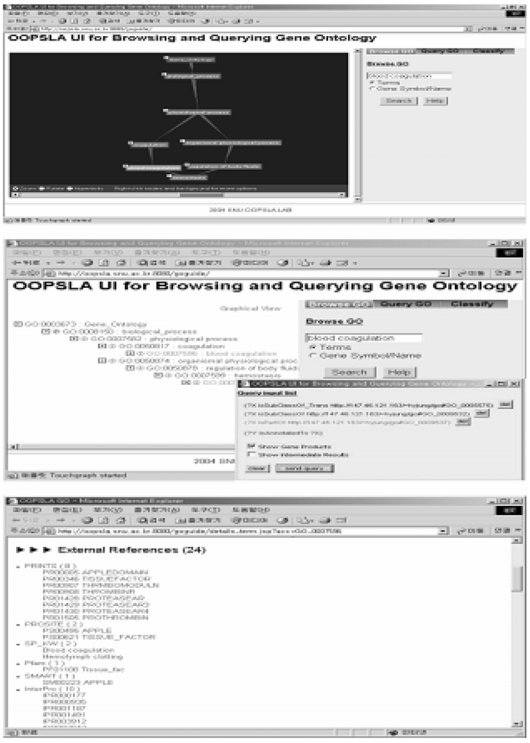

그림 5 GOGuide

표 1 AmiGO, GOView & GOGet, GOGuide, GOGuide II 기능 비교

	데이타 통합 방식	질의 방식
AmiGO	하이퍼링크	키워드 검색
GOView & GOGet	하이퍼링크	논리식
GOGuide	하이퍼링크	시맨틱 질의
GOGuide II	통합 온톨로지	시맨틱 질의

사용된 온톨로지 대응 및 통합 기법을 소개한다.

3.1 생물학 관련 저장소들

3.1.1 UniProt

UniProt은 미국 국립 인간 유전체 연구소(NHGRI)와 국립보건원(NIH) 산하 연구 기관들에 의해 구축된 세계 최대의 연합 단백질 데이타베이스(United Protein Database)로서, Swiss-Prot, TrEMBL, PIR 등의 단백질 데이타베이스들을 통합한 것이다. UniProt에는 단백질의 이름과 위치, 참고 문헌, 키워드, 타 저장소 내의 데이타들과의 관계 등이 기술되어 있다.

3.1.2 Reactome

Reactome은 생물학적 반응 경로와 관련된 데이타가 집약되어 있는 대표적인 저장소로서, Cold Spring Harbor 연구소, 유럽 생물정보학 연구소(EBI) 등의 기관들에 의해 구축되었다. Reactome에는 생물학적 과정(path-

way)의 명칭과 그에 대한 설명 및 선행 작용들, 후행 작용들, 올소로구스 작용들 등의 정보가 담겨 있다.

3.2 온톨로지 대응 및 통합 기법들

여러 기업체나 조직들은 개별적으로 자신의 관심 영역에 대한 온톨로지를 가지고 있으며, 이러한 온톨로지들은 상당부분 겹치기도 하고 서로 밀접하게 연관된 개념들을 포함하기도 한다. 따라서, 온톨로지를 통합하기 위해서는 먼저 온톨로지 상의 개념들 간의 상호 대응 관계의 의미를 명확히 해야 한다. 본 논문에서는 대응 관계의 의미를 결정하기 위해 언어적(linguistic) 방법과 다중도(multiplicity)를 이용한 방법, 그리고 관리자에 의한 수동적 방법을 복합적으로 사용하였다. 본 연구에서 사용한 알고리즘들은 다음과 같다.

- Porter stemming algorithm[7]: 단어에서 모든 파생 접사와 굴절접사를 제거하여 어근을 찾는 알고리즘이다. 예를 들어, “connected”, “connecting”, “connection”, “connections”를 어근화하면 모두 “connect”로 통일된다. 본 연구에서는 서로 같은 어근을 갖는 단어들이 서로 다른 언어로 인식되는 것을 막기 위해 전처리 단계에서 이 알고리즘을 적용하였다.
- 불용어 제거: 관사나 접속사와 같이 별로 중요하지 않은 단어들을 전처리 단계에서 제거함으로써 문자열 비교를 용이하게 하였다.
- Levenshtein distance(edit distance)[8]: 서로 다른 두 문자열 간에 얼마나 차이가 나는가를 나타내는 지표로서, 하나의 문자열에 최소 몇 회의 삽입/수정/삭제 연산을 가해야 다른 문자열과 같아지는가를 나타낸다. 여기서 삽입 연산은 문자열의 처음이나 중간 또는 끝부분에 하나의 문자를 삽입하는 것을, 수정 연산은 문자열에서 하나의 문자를 다른 문자로 변경하는 것을, 삭제 연산은 문자열에서 하나의 문자를 삭제하는 것을 뜻한다. 예를 들어, “GUMBO”에서 “U”를 “A”로 바꾸고 끝에 “L”을 붙이면 “GAMBOL”이 되므로, “GUMBO”와 “GAMBOL” 사이의 Levenshtein distance는 2가 된다.

4. 시스템 구현

앞에서 살펴본 바와 같이 현재 UniProt, Reactome 등 여러 사이트들이 각자의 방식으로 데이타를 저장하고 있다. 뿐만 아니라 AmiGO와 GOView & GOGet 등 현존하는 시스템들은 이처럼 분산된 데이타들에 대한 통합 질의를 지원하지 않고 있다.

본 논문에서 제안하는 GOGuide II는 다음과 같은 기능들을 지원함으로써 기존 시스템들의 단점을 보완하였다.

- 통합 시맨틱 질의
- 통합 온톨로지 브라우징
- 통합 온톨로지 그래프 탐색

또한, 본 논문에서는 GOGuide II에서 필요로 하는 통합 온톨로지 생성을 도와 주는 도구인 AutoGOA를 제시하였다.

4.1 시스템 구조

본 시스템의 전체적인 구조는 그림 6과 같다.

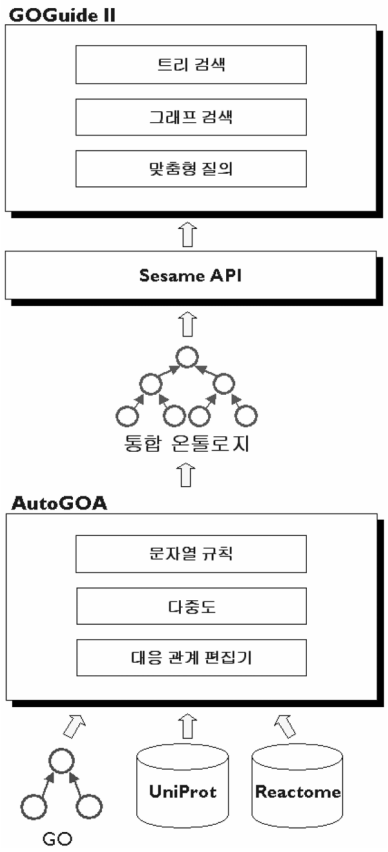

그림 6 전체 시스템 구조

AutoGOA는 여러 연구 집단들의 개념 분류 체계와 GO의 대응 관계를 읽어들인 후, 문자열의 유사성, 다중도 등을 이용해 자동으로 의미를 결정한다. 또한, 이 결과를 관리자가 수정, 보완할 수 있도록 도와주는 인터페이스를 제공한다. 관리자의 검토 후 최종 결과는 통합 온톨로지에 저장된다.

GOGuide II는 GOGuide를 통합 시맨틱 질의가 가능하도록 확장한 것으로서, AutoGOA에 의해 통합된 온톨로지를 이용해 통합 시맨틱 질의를 수행한다.

4.2 대응 관계의 의미 결정

AutoGOA는 가장 먼저 대응 관계의 의미를 분석한다. 분석은 크게 다음과 같은 네 단계로 이루어진다.

- 전처리 단계: 대상 항목들에 대해 어근화, 불용어 제거 등을 적용해 문자열 비교에 적합한 형태로 만든다.
- 문자열 규칙 적용 단계: 정규 표현식으로 기술된 문자열 규칙들을 적용해 대응 관계의 의미를 자동으로 결정한다.
- 다중도에 의한 판별 단계: 대응 관계의 다중도에 따라 의미를 예측할 수 있다.
- 수정 및 보완 단계: 앞의 세 단계를 거친 결과를 관리자가 검토하여 대응 관계를 수정 및 보완할 수 있는 인터페이스를 제공한다.

4.2.1 전처리 단계

일반적으로 동사의 경우 하나의 어근으로부터 여러 형태의 단어들이 파생될 수 있다. 또한, 연자 부호나 쉼표 같은 문장 부호 및 전치사 등의 위치와 종류에 따라 한 단어가 여러 형태로 변형될 수 있다. 따라서, 문자열 규칙 적용 이전에 문장 부호를 제거하고 전치사 등의 불용어(stop word)를 제거하고 단어들을 어근화(stemming)하는 등의 전처리 과정을 거쳐야 한다. 그림 7은 “activation of pro-apoptotic gene products”에 이러한 일련의 전처리를 수행하는 과정을 보여준다.

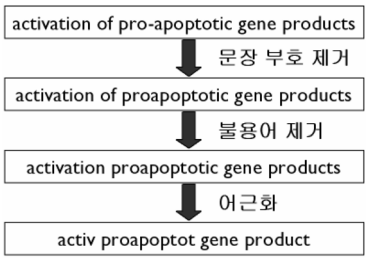

그림 7 전처리 수행 과정

4.2.2 문자열 규칙 적용 단계

전처리 단계를 거친 후에는 문자열 비교를 통해 대응 관계의 의미를 결정한다. 먼저 패턴 규칙들을 읽어들인 후 대응 관계 각각에 대해 패턴 규칙들을 적용한다. 패턴 규칙은 정규 표현식을 이용하여 기술된다. 이는 정규 표현식을 이용하면 복잡한 패턴 규칙들을 잘 나타낼 수 있을 뿐만 아니라 생물정보학 용어들의 문법적 특성을 패턴 규칙에 잘 반영할 수 있기 때문이다. 그림 8은 Alginate biosynthesis와 Alginic acid biosynthesis에 문자열 규칙을 적용하여 동치 관계로 결정되는 것을 보여주고 있다.

뿐만 아니라, 패턴 규칙 외에도 두 문자열 간의 Levenshtein distance를 유사성 판단의 근거로 삼는다. 예를 들어, 두 문자열 간의 Levenshtein distance가 매우 작다면 두 문자열은 동의어 관계에 있다고 볼 수 있다.

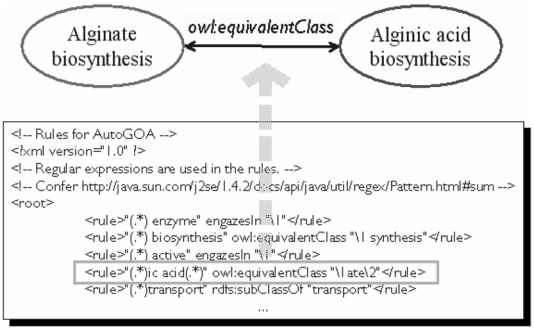

그림 8 문자열 규칙 적용의 예

4.2.3 다중도에 의한 판별 단계

대응 관계의 다중도란 대응 관계에 관여하는 대상의 수를 의미한다. 예를 들어, 한 개의 UniProt 키워드가 두 개의 GO 상의 개념들과 대응 관계를 가진다면 다중도는 2가 된다. 대응 관계는 그 다중도에 따라 크게 일대일 관계, 일대다 관계, 다대일 관계, 다대다 관계로 나눌 수 있다. 이러한 다중도를 분석하면 대응 관계의 의미에 대한 힌트를 얻을 수 있다.

먼저, 그림 9와 같은 일대다 대응관계를 살펴보자.

생물학적 과정인 ‘xenobiotic metabolism’과 ‘response to toxin’이 둘 다 ‘Detoxification’과 대응 관계를 가지므로, ‘Detoxification’이 두 개념들의 상위 개념이라고 간주할 수 있고, 따라서 그림 10과 같이 의미가 결정된다.

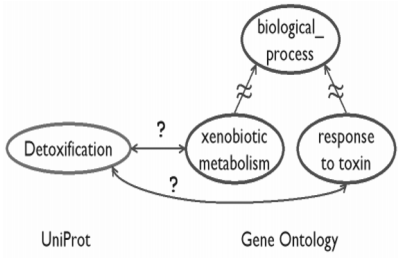

그림 9 첫 번째 예 – 의미 결정 전

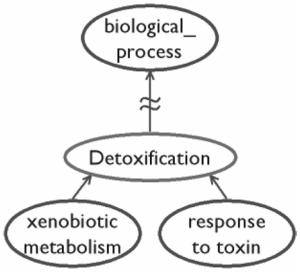

그림 10 첫 번째 예 – 의미 결정 후

모든 일대다 관계의 의미가 앞의 예와 같이 결정되는 것은 아니다. 그림 11과 같은 경우를 생각해 보자.

'Dynein'과 'dynein complex'가 전처리 단계와 문자열 규칙 적용 단계를 거쳐 이미 동의어 관계로 판명났다고 가정하자. 이 경우 'dynein complex'는 세포의 구성 요소이고 'motor activity'는 분자의 기능을 나타내는 개념이므로 서로 동의어 관계에 놓일 수 없다. 따라서, 세포의 구성 요소인 'Dynein'이 'motor activity'라는 기능에 관여한다고 보는 것이 합리적이다. 그러므로 그림 12와 같이 의미를 결정할 수 있다.

다대일 관계와 다대다 관계에 대해서도 이러한 방식으로 의미를 결정할 수 있다.

4.2.4 수정 및 보완 단계

구문론적(syntactic) 방법과 다중도를 이용한 방법만으로 대응 관계의 의미를 완벽히 결정하는 것은 불가능하다. 따라서 AutoGOA는 자동으로 결정된 의미를 전문적 지식을 가진 관리자가 수정, 보완할 수 있도록 도와주는 인터페이스를 제공한다. 관리자는 기존의 대응 관계를 수정하거나 삭제할 수 있고, 새로운 대응 관계를 추가할 수도 있다.

대응 관계들의 의미가 결정되면 AutoGOA는 여러 연구 집단들의 분류 체계와 GO를 통합하여 통합 온톨로지에 저장한다.

4.3 통합 시맨틱 질의 지원

온톨로지 및 데이타 통합의 궁극적 목표는 통합 시맨틱 질의라 할 수 있는데, 이를 위해서는 통합 시맨틱 질의를 지원하는 인터페이스가 필요하다. 따라서 본 논문에서는 현재 시맨틱 질의는 지원하고 있으나 GO에 대한 질의만 허용하는 GOGuide를 확장하여 통합 시맨틱 질의가 가능한 GOGuide II를 구현하였다.

GOGuide II는 통합된 데이타에 대한 트리 탐색, 그래프 탐색, 키워드 검색, 시맨틱 질의 등의 서비스를 제공한다. 사용자는 트리 탐색과 그래프 탐색을 통해 GO 상의 개념들 뿐 아니라 다른 연구 집단들의 분류 체계도 살펴볼 수 있다. 또한, 특정 단어를 검색하면 통합된 온톨로지 상에서 관련된 모든 대상들을 볼 수 있다. 뿐

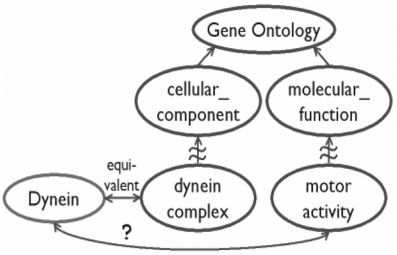

그림 11 두 번째 예 – 의미 결정 전

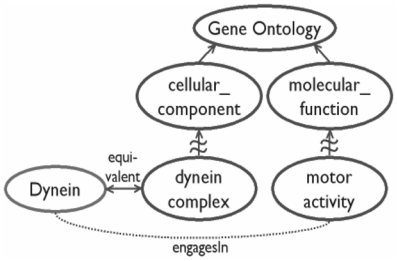

그림 12 두 번째 예 – 의미 결정 후

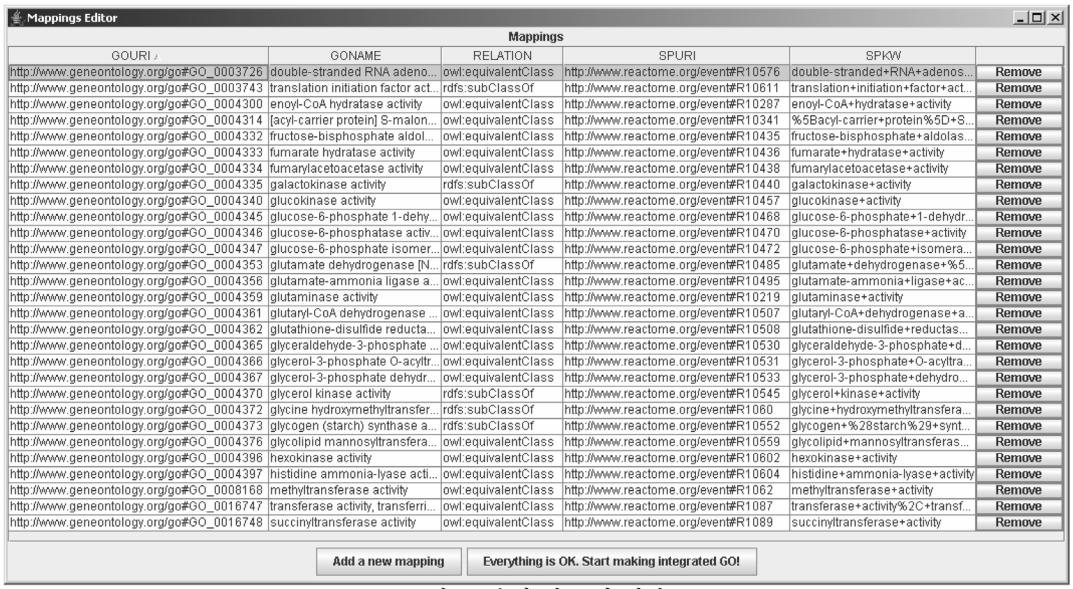

그림 13 수정 및 보완 단계

만 아니라 시맨틱 질의 인터페이스를 통해 이러한 검색 결과들과 특정 관계를 갖는 데이타들을 찾을 수 있다. 시맨틱 질의는 그림 14와 같이 내부적으로 Sesame API에서 사용하는 RDF 질의 언어인 SeRQL(Sesame RDF Query Language)로 변환되어 처리된다. 그림 15와 같이 질의 결과에는 GO 와 타 사이트 상의 관련 정보들이 함께 표시된다. 즉, UniProt의 관련 단백질들과 Reactome의 관련 정보들(선행 작용, 후행 작용, 올소로구스 작용 등)이 일목요연하게 정리되어 한 페이지 상에 출력된다.

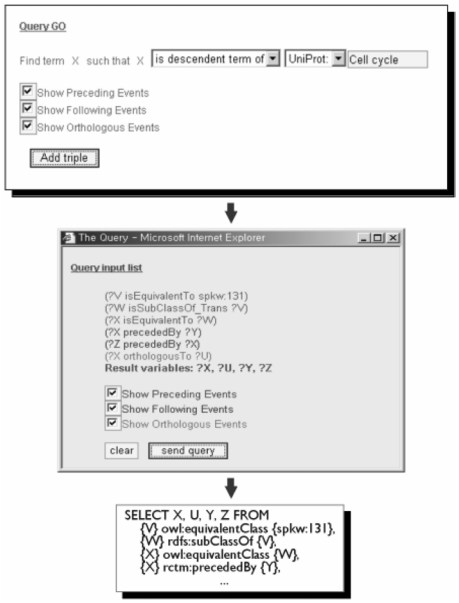

그림 14 질의 변환 과정

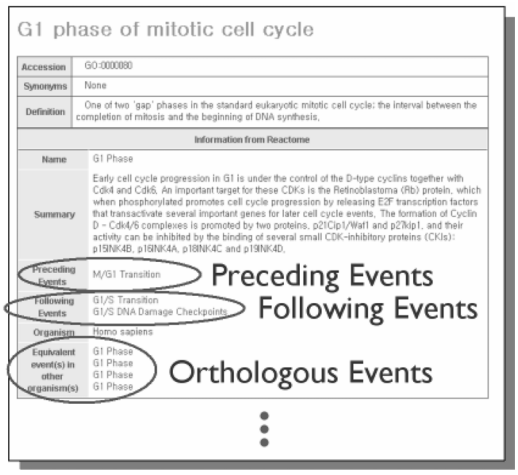

그림 15 질의 수행 결과

5. 실험 및 분석

5.1 실험 환경

본 시스템에 사용된 자바의 버전은 J2SE 1.4이고, 실험에 사용된 하드웨어의 사양은 펜티엄4 1.7GHz CPU, 512MB RAM이며, 운영 체제는 윈도우2000이다. 본 시스템의 온톨로지 통합의 정확성을 검사하기 위해, 2004년 10월에 공개된 대응 관계들을 학습 집합(training set)으로, 2005년 10월에 공개된 대응 관계들을 검증 집합(test set)으로 이용했다. 즉, 2004년에 공개된 대응 관계들에 대해 AutoGOA의 자동 의미 결정 기법들(관리자에 의한 수정 및 보완은 제외)을 적용한 결과와 2005년에 공개된 대응 관계들을 비교하였다. 여기서 학습 집합과 검증 집합 모두 대응 관계의 의미가 명시되어 있지 않으므로, 대응 관계의 존재 유무를 정확성의 기준으로 삼았다. 즉, 검증 집합에서 대응 관계가 명시되어 있으면 동의어 관계가 성립한다고 보고, 대응 관계가 검증 집합에 존재하지 않으면 동의어 관계가 성립하지 않는다고 보았다.

검증 집합에 대응 관계의 의미가 명시되어 있지 않으므로, 동의어 관계 이외의 다른 관계들에 대한 의미 결정의 정확성에 관해서는 정량적 분석이 불가능하였다.

5.2 실험 결과

표 2는 UniProt과 GO의 대응 관계에 대한 오분류표이고, 표 3은 Reactome과 GO의 대응 관계에 대한 오분류표이다. 먼저, 표 2를 살펴보면, 학습 집합에 존재하는 총673개의 대응 관계들 중 360개가 의미 결정 기법에 의해 동의어 관계로 결정되었는데, 이 중 325개가 검증 집합에도 존재하는 것으로 드러났다. 또한, 동의어 관계가 아니라고 결정된 216개의 관계들 중 97개만이 검증 집합에 존재하는 것으로 드러났다. 따라서, 정분류율은 (216 + 325) / (216 + 35 + 97 + 325) ≒ 0.80이 된다. 다음으로, 표 3을 살펴보면, 학습 집합에 존재하는 총1964개의 대응 관계들 중 813개가 의미 결정 기법에 의해 동의어 관계로 결정되었는데, 이 중 741개가 검증 집합에도 존재하는 것으로 드러났다. 또한, 동의어 관계가 아니라고 결정된 1151개의 관계들 중 341개만이 검증 집합에 존재하는 것으로 드러났다. 따라서, 정분류율은 (810 + 741) / (810 + 72 + 341 + 741) ≒ 0.79가 된다. 2004년 데이타에 존재하던 개념들이 2005년 데이타에서 사라지거나, 2005년 데이타에 없던 개념들이 2005년 데이타에 추가되기도 했다는 점을 감안하면, 의미 결정 기법이 동의어 관계 여부를 비교적 정확하게 결정한다고 볼 수 있다. 실제로 변경 사항을 자세히 살펴본 결과 대부분의 오분류는 사라지거나 새로이 추가되는 개념들 때문에 발생하는 것을 확인할 수 있었다.

표 2 오분류표(UniProt-GO)

		분류범주	
		X(동의어 아님)	O(동의어)
실제범주	X	216	35
	O	97	325

표 3 오분류표(Reactome-GO)

		분류범주	
		X(동의어 아님)	O(동의어)
실제범주	X	810	72
	O	341	741

동의어 관계 외의 다른 관계들에 대해서는 정량적인 분석이 불가능하지만, 실험 결과 학습 집합에서 상위어/하위어 관계로 결정된 것들 중 대부분이 검증 집합의 대응 관계에 반영된 것을 확인할 수 있었다. 예를 들어, 학습 집합에 의미 결정 기법들을 적용한 결과 UniProt의 'Bacteriocin transport'가 GO의 'transport'의 하위어로 결정되었는데, 실제로 2005년 GO에는 'bacteriocin transport'라는 개념이 'transport'의 하위어로 새로 추가되었고, 검증 집합에는 GO의 'bacteriocin transport'와 UniProt의 'Bacteriocin transport' 사이의 대응 관계가 존재한다.

6. 결론 및 향후 연구

본 논문에서는 여러 연구 집단의 분류 체계와 GO 사이의 대응 관계의 의미를 반자동으로 결정해 주는 시스템인 AutoGOA와 통합 시맨틱 질의를 지원하는 브라우저인 GOGuide II를 제안하였다.

AutoGOA는 전처리 단계, 문자열 규칙 적용 단계, 다중도에 의한 판별 단계, 그리고 수정 및 보완 단계를 거쳐 대응 관계의 의미를 결정한다. 전처리 단계에서는 문자열에서 문장 부호와 불용어를 제거하고 문자열을 어근화한다. 문자열 규칙 적용 단계에서는 대응 관계들에 대해 패턴 규칙들을 적용하여 대응 관계의 의미를 결정한다. 다중도에 의한 판별 단계에서는 대응 관계의 다중도를 분석하여 대응 관계의 의미를 결정하나. 끝으로 수정 및 보완 단계에서는 자동으로 결정된 의미를 관리자가 수정, 보완할 수 있도록 도와주는 인터페이스를 제공한다.

GOGuide II는 통합된 데이타와 온톨로지에 대한 트리 탐색과 그래프 탐색, 키워드 검색, 통합 시맨틱 질의 등을 지원한다. GOGuide II의 통합 시맨틱 질의 인터페이스를 통해 연구자들은 분산된 데이타들에 대해 각각 검색을 하여 원하는 결과를 찾던 기존의 비효율적인 방식에서 벗어나, 여러 사이트들의 데이타 간의 복잡한 의미적 관계를 포함한 질의를 통해 원하는 결과를 편리하고 신속하게 얻을 수 있다.

본 논문의 자동 시맨틱 결정 기법의 정확도를 검증해 본 결과, 동의어 관계에 대해서는 전문가들의 판단과 자동 의미 결정 기법의 결과가 대부분 일치하는 것을 확인할 수 있었다. 뿐만 아니라, 동의어 관계 이외의 다른 관계들에 대해서도 비교적 정확하게 대응 관계의 의미를 결정하는 것을 확인할 수 있었다.

본 논문에서는 생물학 온톨로지와 단백질 데이타, 그리고 생물학적 과정에 대한 데이타를 주로 다루었다. 향후 PPI, subcellular localization, Blast, PubMed 등과 같은 전반적인 Omics(일정한 수준의 생물학적 분자들과 정보의 집합체) 데이타의 통합과 분석에 대한 연구가 요구된다. 따라서 현재 Omics 데이타에 대한 통합 분석 시스템인 OASIS(Omics Analysis for microbial organisms)를 구축 중이며, 여기에 본 논문의 자동 의미 결정 기법을 적용할 계획이다.

참 고 문 헌

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. *Nature Genet* 25:25-29, 2000.

[2] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, Lai-Su L. Yeh. The Universal Protein Resource(UniProt). *Nucleic Acids Research* 33:D154-D159, 2005.

[3] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* 33:D428-432, 2005.

[4] AmiGO. http://www.godatabase.org/cgi-bin/amigo/go.cgi *Gene Ontology Consortium*, 1998.

[5] Elizabeth Shoop, Paulo Casaes, Getiria Onsongo, Lisa Lesnett, Erla Osk Petursdottir, Edward Kofi Yeboah Donkor, Dennis Tkach, Michael Cosimini. Data exploration tools for the Gene Ontology database. *Bioinformatics* 20(18):3442-3454, 2004.

[6] J.W. Jung, H.W. Park, D.H. Lim, K.P. Lee, H.J. Kim. GOGuide: Browser for Gene Ontology.

KDBC 5:44-51, 2005.
[7] Porter MF. An algorithm for suffix stripping. *Program* 14(3):130-137, 1980.
[8] Diana Maynard, Sophia Ananiadou. Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology*, 1999.

박 형 우
정보과학회논문지 : 컴퓨팅의 실제
제 12 권 제 3 호 참조

정 준 원
정보과학회논문지 : 컴퓨팅의 실제
제 12 권 제 3 호 참조

김 형 주
정보과학회논문지 : 컴퓨팅의 실제
제 12 권 제 3 호 참조