GO Guide : 생물학 온톨로지를 위한 브라우저 및 질의 변환

(GO Guide : Browser & Query Translation for Biological Ontology)

정 준 원 † 박 형 우 † 임 동 혁 † 이 강 표 † 김 형 주 ††
(Jun-Won Jung) (Hyoung-Woo Park) (Dong-Hhyuk Im) (Kang-Pyo Lee) (Hyoung-Joo Kim)

요 약 생물학 분야에서 유전자에 대한 연구가 활발하게 이루어지면서 유전자에 대한 정보 구축 및 통합에 대한 필요성이 대두 되었다. 그 결과 Gene Ontology Consortium은 W3C에서 제정한 온톨로지 기술언어인 OWL로 유전자에 대한 정보와 분류를 담고 있는 Gene Ontology를 구축하였다. 하지만 Gene Ontology를 위한 기존의 브라우저들은 키워드, 트리, 그래프 기반의 단순 검색만을 지원할 뿐 다양한 관계를 고려한 고급 정보 검색이 불가능하다.

본 논문은 실제 생물학 연구를 수행하는 사용자들이 Gene Ontology를 효과적이고 편리하게 사용할 수 있도록 하기 위해 다양한 온톨로지 검색 기법을 통합적으로 지원하는 방법을 제안하였다. 또한 질의어 입력대신 검색 중에 손쉽게 질의를 생성하는 기법과 생성된 질의를 SeRQL 질의로 변환하는 기법을 제안함으로써 온톨로지에서 지원하는 질의어에 독립적으로 손쉽게 질의를 생성하고 고급정보를 얻을 수 있도록 하였다. 그리고 이렇게 구축한 GO Guide 브라우저를 통해 Gene Ontology의 방대한 정보를 효과적으로 이용할 수 있음을 확인하였다.

키워드 : 온톨로지, 생물학 온톨로지, 브라우저, 생물정보학

Abstract As genetic research is getting more active, data construction of genes are needed in the field of biology. Therefore, Gene Ontology Consortium has constructed genetic information by OWL, which is Ontology description language published by W3C. However, previous browsers for Gene Ontology only support simple searching mechanisms based on keyword, tree, and graph, but it is not able to search high quality information considering various relationships.

In this paper, we suggest browsing technique which integratesvarious searching methods to support researchers who are doing actually experiment in biology field. Also, instead of typing a query, we propose query generation technique which constructs query while browsing and query translation technique which translate generated query into SeRQL query. It is convenient for user and enables user to obtain high quality information. And by this GO Guide browser, it has been shown that the information of Gene Ontology could be used efficiently.

Key words : Ontology, Gene Ontology, Browser, Bioinformatics

1. 서 론

최근 생물학 분야에서는 유전자의 역할과 기능을 규명하기 위한 연구들이 활발하게 이루어지고 있다. 이와 같은 연구에 있어 컴퓨팅 기술은 유전자의 염기서열을 분석하거나 비교하는 것뿐만 아니라 유전자 연구를 통해 발견한 방대한 데이타를 구현하고, 검색하는데 있어 중요한 역할을 한다. 특히 생물학 분야의 연구방식이 이미 밝혀진 유전자와 새로 발견한 유전자의 비교를 통해 이루어지는 등 기존의 결과를 이용해서 수행되는 경우가 많기 때문에 연구 데이타를 잘 조직하고 검색하는

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

† 학생회원 : 서울대학교 전기.컴퓨터공학부
jwjung@oopsla.snu.ac.kr
hwpark@oopsla.snu.ac.kr
dhlim@oopsla.snu.ac.kr
kplee@oopsla.snu.ac.kr
†† 종신회원 : 서울대학교 전기.컴퓨터공학부 교수
hjkim@oopsla.snu.ac.kr

논문접수 : 2005년 4월 22일
심사완료 : 2006년 3월 14일

것은 매우 중요하다[1]. 따라서 각 유전자 연구 그룹들은 유전자나 생물학 정보를 저장하는 데이타베이스를 구축하는 경우가 많다[2]. 하지만 각 연구 그룹들이 개별적인 데이타베이스를 구축했기 때문에 동일한 대상에 대한 명칭이나 분류, 정보 기술방식이 다르다. 따라서 기존 데이타를 잘 이용할 수 없고 통합된 정보 검색이 불가능하기 때문에 매우 비효율적이다.

이와 같은 문제를 해결하는 방법으로서 온톨로지를 통한 생물학 데이타의 통합을 생각해 볼 수 있다. 온톨로지란 데이타에 부가 정보를 기술하거나 연관 관계를 표현하여 데이타를 조직화 할 수 있는 기술로서, W3C (World Wide Web Consortium)에서는 OWL(Web Ontology Language)[3]이라는 온톨로지 표준을 제정하였다. 그리고 생물학 데이타의 통합에 대한 필요성에 의해 각 유전자 연구 그룹이 모여 결성한 Gene Ontology Consortium에서는 각 그룹의 유전자 데이타를 온톨로지로 통합한 Gene Ontology[4]를 구축하였다.

하지만 아직까지 Gene Ontology 뿐만 아니라 다른 생물학 온톨로지에 대해서도 효율적으로 데이타를 관리하고 검색하는 시스템이 부족한 상황이다. 기존의Gene Ontology 검색 시스템들은 단순한 키워드 매칭을 통한 정보검색 및 계층구조 표현, 정적인 그래프의 관계표현만을 지원하고 있으며 각 검색기법간의 연계가 부족하다. 또한 단순히 Gene Ontology의 데이타를 탐색하는 정도의 기능만을 제공할 뿐 온톨로지를 구축함으로써 얻을 수 있는 고급정보 검색을 수행할 수 없다. 예를 들어 '항산화작용과 합성작용에 하위 기능을 포함하여 공통으로 관련된 유전자를 찾아라'와 같이 A 작용과 B 작용의 하위 작용에 공통으로 관련된 유전자를 찾는 검색을 수행할 수 없다. 기존의 시스템에서 이와 같은 검색을 수행하기 위해서는 A 작용과 그 하위 분류의 작용들을 모두 찾고 각각에 대한 유전자 들을 검색해야 하며 이것을 B 작용의 하위 작용에 대해 찾은 유전자들과 각각 비교하는 소모적인 작업을 수행해야 한다. 이것은 비효율적일 뿐만 아니라 각 단계를 사용자가 하나씩 수행하면서 매번 비교해야 하기 때문에 정확도가 떨어지고 많은 시간이 소모된다.

본 논문에서는 Gene Ontology를 효율적으로 검색하기 위한 브라우저로서 GO Guide를 구축하였다. GO Guide는 사용자로 하여금 Gene Ontology를 편리하게 검색할 수 있도록 기존의 검색기법을 통합한 검색 기법을 제안한다. 또한 사용자가 소모적인 작업 없이 복잡한 관계를 검색 할 수 있도록 질의를 통한 검색을 지원하도록 하였다. 이때 사용자가 복잡한 질의를 직접 기술하지 않고 탐색 과정에서 질의를 생성할 수 있는 인터페이스를 제안하였다. 그리고 이렇게 생성된 질의를 OWL 질의어인 SeRQL[5]로 변환하는 질의 변환기법을 제안하였다. 이와 같이 인터페이스에서 검색을 통해 생성한 기본 질의를 온톨로지에서 지원하는 질의 형태로 변환하도록 함으로써 사용자는 온톨로지에서 지원하는 질의어와 독립적으로 쉽게 질의를 생성하고, 고급 검색이 가능하도록 하였다.

논문의 구성은 다음과 같다. 먼저 2장에서는 OWL과 Gene Ontology에 대한 배경지식을 설명하고 Gene Ontology와 관련된 기존의 시스템에 대해서 살펴본다. 그리고 3장에서는 본 논문에서 구축한 시스템 구조에 대해 소개하고 4장에서 본 논문에서 제안한 브라우저와 질의변환 기법에 대해서 설명한다. 그리고 5장에서는 기존의Gene Ontology 검색 시스템과 비교해 보고, 마지막으로 6장에서 결론과 향후 연구 방향에 대해서 정리한다.

2. 배경지식 및 관련연구

2.1 온톨로지와 OWL

온톨로지란 초기에 철학분야에서 개념을 표현하기 위해 사용되는 여러 용어의 집합을 의미하거나 특정 분야에서 사용되는 용어의 집합을 의미하기도 하였다. 일반적인 의미에서의 온톨로지란 작게는 용어의 집합에서부터 크게는 용어에 대한 부가정보나 관계, 색인 등을 표현한 정보라고 할 수 있다[6]. W3C에서는 웹에 대한 부가 정보를 기술하기 위한 시멘틱웹으로서 RDF[7](Resource Description Framework)를 제정하였다. RDF는 Subject, Predicate, Object의 세가지 요소로 이루어지는 트리플[7]을 통해 정보들의 관계를 기술할 수 있는 기능을 제공한다. 그리고 W3C는 RDF를 기반으로 온톨로지 구축을 위한 언어로서 OWL을 제정하였다. OWL은 RDF에 상하위 관계, 동등 관계, 제약 등의 다양한 관계를 표현할 수 있으며 논리나 추론을 위한 정보를 표현할 수 있는 기반을 제공한다. 온톨로지는 데이타에 대한 부가 정보나 관계를 기술할 수 있기 때문에 여러 형태의 데이타를 통합하거나 다양한 목적을 위해 정보를 구축하는데 용이하다. 또한 XML을 기반으로 기술되므로 정보의 교환이 용이하고, 의미적 정보를 통한 추론이나 복잡한 관계에 대한 고급 정보 검색이 가능하다.

2.2 Gene Ontology

Gene Ontology Consortium은 1998년 FlyBase Consortium, SGD(Saccharomyces Genome Database), MGD (Mouse Genome Informatics)등 대표적인 생물체에 대한 데이타베이스 단체들에 의해 설립되었다. Gene Ontology Consortium의 기본적인 목적은 단백질이나 RNA와 같은 유전자 산출물(gene product)들이 각 생물체에서 수행하는 작용(term)을 기술하기 위한 공통의 용어를 생성하고 이들간의 관계를 기술하는 것이다[4]. 하나의 생물학적인

작용들이 하나의 종에만 나타나는 것이 아니라 여러 종에 나타날 수 있기 때문에 다양하게 쓰이고 있는 용어들을 정리하고 모든 종에서 공통적으로 사용할 수 있는 용어를 확립하는 것은 중요한 문제이다. 그리고 이 용어들이 각 데이타베이스에서 사용될 수 있는 형태로 만드는 것 또한 중요한 문제이다. 이를 위해 Gene Ontology Consortium은 유전자에 대한 정보를 Molecular function, Biological Process, Cellular com- ponent로 나누고 각 유전자와 유전자 산출물에 들에 대한 정보를 Gene Ontology로 구축하였다. 이 정보는 유전자에 대한 기본 정보뿐만 아니라 생물학적 작용들에 대한 계층구조와 유전자 산출물들의 포함 및 부분관계를 담고 있으며 해당 유전자에 대한 자세한 정보를 담고 있는 다른 데이타베이스에 대한 링크를 포함한다. Gene Ontology에서 제공하는 온톨로지 데이타는 OBO format [8], XML(OWL), MySQL dump의 형태로 제공되고 있다.

2.3 관련 연구

Gene Ontology에서 정보를 얻기 위해서는 온톨로지에 대한 검색, term과 term들간의 관계 그리고 gene product와 term의 관계에 대한 질의 및 브라우징이 필요하다. 이러한 기능을 제공하는 기존의 응용프로그램으로써 AmiGO[9]와 GoView[10] 그리고 GoGet[10]이 있다. 먼저 AmiGO는 Gene Ontology consortium내의 BDGP(Berkeley Drosophila Genome Project)에서 개발된 Gene Ontology 브라우저로서 키워드, 트리, 그래프형태의 탐색을 지원한다. AmiGO는 트리를 통해 계층관계만을 표현하고, 질의를 지원하지 않기 때문에 term이나 gene product간의 다양한 관계를 검색하기 힘들고 다양한 관계의 검색이 소모적이다. 또한 트리의 계층관계를 그래프로 생성해 주지만 단지 그래프로 가시화했을 뿐 그래프상에서 탐색이 불가능하다.

다음으로Gene Ontology database exploration project (Macalester college, Data exploration laboratory)에서 개발된 브라우저로서 GoView와 GoGet이 있다. GoView는 Gene Ontology를 DAG형태로 그래프 표현하고 검색하는 가시화 도구로서 그래프 상에서 선택된 term을 강조해서 보여준다. 많은 데이타를 브라우징 하기 위해 확대/축소기능을 제공하지만 한 term과 연관된 정보를 찾기 위해서는 강조된 부분을 term의 간선을 따라 여러 번 움직여야 하는 단점이 있다. GoGet은 Gene Ontology에 질의를 수행하는 도구로서 특정 term이나 gene product가 탐색하려는 데이타베이스에 포함되거나 포함되지 않는지를 조건으로 기술하여 질의를 생성한다. 하지만 GoGet은 'common ancestor', 'part Of'등 term들간의 복잡한 관계를 지원하지 못하며 단지 특정 term이나 gene product가 어떤 데이타베이스에 포함되는지 여부만을 질의로 생성하므로 매우 제한적인 질의가 수행된다.

반면에 GO Guide에서는 트리, 그래프, 키워드, 질의의 네 가지 기능을 복합적으로 지원하며 특히 질의 생성기능을 이용하여 다양한 관계의 정보를 얻을 수 있는 고급 질의를 지원한다.

3. 시스템 구조

다음 그림 1은 GO Guide의 구조를나타내고 있다. GO Guide 브라우저는 OWL 온톨로지 데이타를 저장하고 처리하는 기능을 제공하는 Sesame[11] API와 My SQL DBMS를 이용하여 구현되었다. Sesame는 RDF 및 OWL의 저장과 RDFS 수준 데이타처리를 위한 inference를 지원하며 RDQL[12], RQL[13] 등과 같은 query language뿐만 아니라 RQL을 확장한 SeRQL도 지원한다. 또한 확장성과 유연성을 갖고 있어 memory-based에서부터 RDBMS까지 다양한 하부 저장 구조를 사용할 수 있으며 본 논문에서는 MySQL RDBMS를 사용한다.

Gene Ontology에서 제공되는 XML 데이타가 OWL 형식으로 작성되었으나 OWL 표준을 완전히 준수하지 않아 본 연구에서는 위의 데이타를 OWL 데이타로 변환하고 트리플형태로 Sesame API를 통해 MySQL에 저장하였다. 현재 Gene Ontology의 데이타 크기는 600MB에 1만 7천 개 term과 1백 30만개의 유전자 정보로 이루어져 있다. 그리고 이와 같은 데이타는 앞으로도 계속 증가하게 될것이므로 OWL데이타를 트리플 기반으로 나누어 RDB에 저장함으로써 대용량의 온톨로지 데이타를 효과적으로 저장 및 검색할 수 있게 된다.

GO Guide는 크게 브라우저와 질의 변환기(Query Translator)의 두 부분으로 나뉜다. 먼저 브라우저부의 Tree generator는 사용자 키워드 입력을 받아 해당 정보를 검색한 뒤 이 정보들간의 관계를 트리로 생성한다.

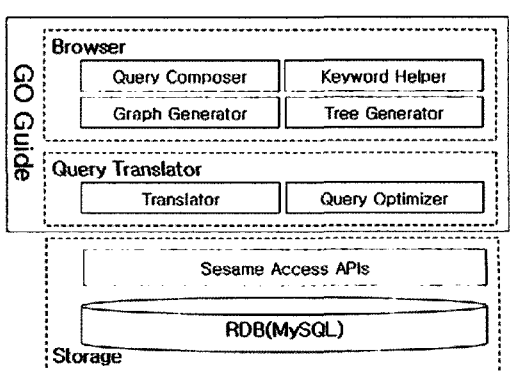

그림 1 GO Guide 시스템 구조

또한 생성된 트리에 대해 계층구조를 통해 탐색할수 있도록 한다. Gene Ontology 내의 데이타는 그래프 형태로 트리플이 연결되어 있기 때문에 Tree generator를 통해 트리형태의 계층구조 표현 및 검색을 가능하게 한다. Graph generator는 Gene Ontology 내의 트리플 연결관계에 따라 그래프를 생성함으로써 정보의 연관관계를 자연스럽게 표현할 수 있도록 한다. 또한 표현된 그래프에서 노드 선택을 통해 연관된 정보를 검색할 수 있는 기능을 지원한다. keyword helper는 생물학 관련 단어가 일반 단어에 비해 길거나 생소하기 때문에 입력에 어려움이 있을 수 있는 문제를해결하기 위해서 키워드를 입력할 시 입력되는 글자의 철자를 체크하여 해당 철자로 완성되는 단어를 보여주고 선택하여 쉽게 입력하는 기능을 지원한다. 마지막으로 탐색 중에 질의를 생성할 수 있도록 Query Composer가 존재한다. 사용자는 키워드, 트리, 그래프와 같은 다양한 탐색기법을 통해 정보를 검색해 나간다. 이때 별도로 질의어를 입력하는 것이 아니라 다른 탐색 기법의 결과나 정보를 질의의 요소로 선택함으로써 질의를 생성하는 기능을 제공한다. 이와 같은 각 모듈은 서로 상호작용 함으로써 유기적으로 동작한다. 즉 키워드, 트리, 그래프, 질의와 같은 검색이 개별적으로 이루어지는 것이 아니라 각 검색 기법의 탐색내용이 다른 검색기법에 반영되고 상호작용 할 수 있도록 한다.

질의 변환기 부분은 브라우저에서 생성된 질의를 시스템에서 지원하는 목적 질의어로 바꾼다. 여기서는 OWL 질의어인 SeRQL로 변환한다. Translator는 브라우저의 질의문을 저장된 SeRQL패턴으로 변환하고, Query Optimizer는 표현 경로를 하나로 통합한 SeRQL 질의를 생성하도록 한다. 이와 같이 함으로써 저장소나 온톨로지의 데이타에 독립적으로 질의를 수행할 수 있도록 하였고, 사용자가 특정 질의어에 대한 지식을 습득할 필요 없이 상위레벨의 질의를 쉽게 생성할 수 있게 된다. 또한 브라우저와 질의 변환기를 두부분으로 나눔으로써 추후 Gene Ontology 이외의 다른 온톨로지를 지원하는 기반도 마련할 수 있다. 따라서 해당 온톨로지를 지원하기 위한 질의 변환기를 구축하고 현재 GO Guide에서 질의 수행의 편리성을 위해 4가지로 정형화한 Predicate을 해당 온톨로지에 있는 Predicate을 지원하도록 함으로써 추후 다른 온톨로지에 대한 적용이 가능하다.

4. GO Guide

4.1 브라우저

GO Guide는 이전 시스템들의 단점을 극복하고, 개별적인 탐색 기법들의 단점을 보완하기 위해 기본적으로 트리, 그래프, 키워드, 질의기반의 검색을 지원하는 것은 물론 각 기능을 연동함으로써 각 기법의 장점을 수용하도록 하였다. 예를 들어 키워드 검색에 대한 결과를 계층 관계를 표현한 트리로서 동적으로 생성할 수 있으며 이 트리를 다시 그래프로 표현하여 검색이 가능하다. 또한 트리나 그래프 검색 시 표현되는 내용을 바로 질의 생성을 위한 키워드로 입력할 수 있도록 하고, 질의어의 유형을 Gene Ontology에 대한 주요 관계를 중심으로 정형화함으로써 질의를 편리하게 생성할 수 있다.

그림 2는 GO Guide의 전체 실행화면을 나타낸다. GO Guide는 크게 정보가 표현되는 (1)프레임과 키워드를 입력하는 (2)프레임, 질의를 생성하는 (3)프레임, 생성된 질의를 확인하고 실행시키는 (4)프레임으로 나뉘어 있다. 그림 2의 (5)번 메뉴는 검색한 결과와 관련된 동작을 나타내는 메뉴이다. 그림 3은 GO Guide를 통해 트리, 그래프 형태의 탐색을 수행하거나 'biosynthesis' 라는 term에 대한 세부 정보를 표현하는 것을 나타내고 있다. 그래프 탐색에서는 수 많은 노드가 표시되어 탐색에 어려움이 발생하는 것을 막기 위해 탐색이 되고 있는 노드를 중심으로 거리(한 노드와 다른 노드간의 간선의 최소 수) 3까지의 노드만 표현하며 탐색이 표현된 그래프의 범위를 넘어갈 경우 연결된 부분이 동적으로 생성된다.

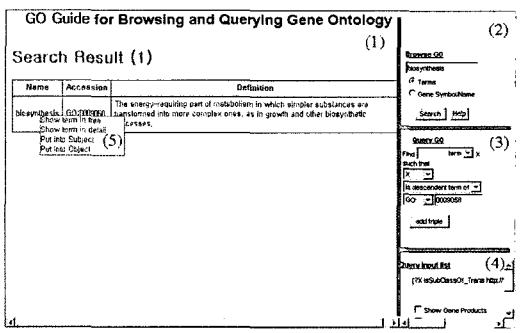

그림 2 GO Guide의 전체 화면

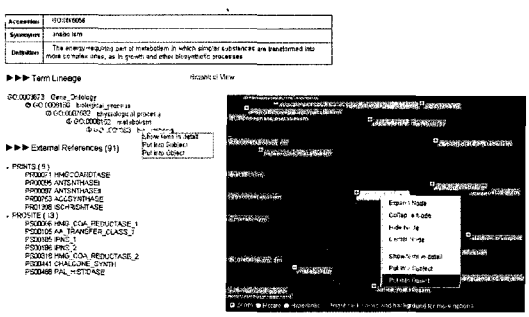

그림 3 GO Guide의 탐색화면

키워드, 트리, 그래프 탐색에서 공통적으로 각 정보개체에 대응하는 메뉴를 지원하고 있다. 이와 같은 메뉴는 각 정보에 대해 관련된 기능을 수행하거나 질의를 바로 생성할 수 있도록 함으로써 탐색을 편리하게 하고 탐색 중에 쉽게 질의를 생성할 수 있도록 한다.

질의 생성에 있어서 특정 질의문을 기술하는 것이 아니라, RDF의 트리플 관계에 따라 Subject, Object, Predicate 관계를 기존의 탐색 메뉴를 통해 기술하도록 하여 직관적으로 사용할 수 있으며, Predicate은 Gene-Ontology에 기술되는 4가지의 Predicate을 메뉴로 선택하도록 정형화 하였다. 따라서 사용자는 질의 문법에 구애 받지 않고 쉽게 질의를 생성할 수 있다. (3)프레임에서 트리플 관계를 모두 입력 한 후 Add triple버튼을 누르면 (4)프레임에 생성된 질의가 표현된다. (3)프레임에서는 여러 문장을 입력할 수 있으며 이와 같은 방법으로 복합적인 질의를 수행하는 질의문을 생성할 수 있고 이것은 모두 (4)프레임에 표현된다. 예를 들어 'a' term과 'b' term의 공통 자손이면서 'c' term의 조상인 term을 찾는 것과 같은 질의를 생성할 수 있고, 해당 질의를 (4)프레임에서 확인한 뒤 send query 버튼을 누름으로써 질의를 수행하고, (1)프레임에 결과가 표시된다.

그림 4는 Biosynthesis term과 lipid metabolism term의 공통 자손의 term에 대한 gene products를 질의 생성을 통해 찾은 결과이다. GO Guide에서는 질의를 통해한번에 결과를 찾지만 기존의 시스템에서는 Biosynthesis term과 lipid metabolism term의 하위 term들을 계속해서 비교하면서 공통 term들을 찾아내고 다시 찾아낸 공통자손 term에 대해 각각의 gene products를 찾아야 하는 매우 소모적인 작업이 필요하다.

GO Guide는 다음과 같은 점에서 기존의 시스템보다 개선된 특징을 가진다. 먼저 그래프 기반의 검색에 있어 단순한 그래프 구조표현이 아닌 그래프상에서의 브라우징을 지원함으로써 정보들간의 연관성을 쉽게 파악할 수 있도록 하였다. 또한 그래프 구현에 있어 가시성을 향상시키기 위해 정보의 거리에 따라 동적으로 그래프를 생성하고 확대 및 축소하는 기능을 지원한다. 그리고, 키워드, 트리, 그래프상의 탐색과 질의 입력을 연계하였다. 즉 탐색중에 트리나 그래프상의 노드(term or gene product)를 질의 입력부에 추가할 수 있도록 지원함으로써 탐색 중에 편리하게 질의를 생성 할 수 있다. 또한 키워드 입력을 돕는helper 제공뿐만 아니라 트리나 그래프 탐색 시 해당 term으로 어떤 연산을 하고자 할 때 각 term과 관련된 연산만을 표시하도록 하는 등 각 기능을 긴밀하게 연계하고 편의성의 최대화 하도록 구축하였다. 특히 질의를 통한 검색을 지원함으로써 기존 브라우저에서의 부정확하고 소모적인 작업을 질의입력을 통해 한번에 수행할 수 있음은 물론 복합적인 정보를 찾는 고급 정보 검색을 지원할 수 있도록 하였다. 또한 질의 결과 표현에 있어서도 질의가 수행되는 중간결과를 표현함으로써 어떤 데이타들이 결과 생성에 관련되었는지 중간 과정을 살펴보고 정보 파악에 도움이 될 수 있도록 하였다.

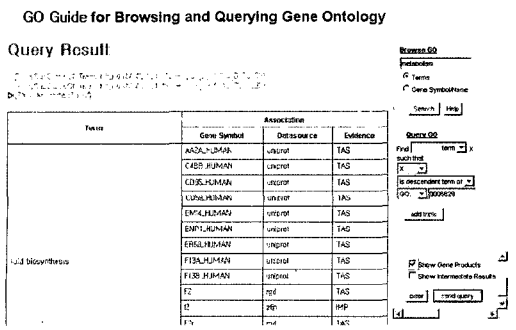

그림 4 질의 수행결과

4.2 질의 처리

4.2.1 SeRQL

GO Guide에서는 데이타를 Sesame를 통해 데이타베이스에 저장을 하기 때문에 이 저장된 데이타에 맞는 질의가 이루어져야 한다. Sesame에서는 RDF와 관련된 대표적인 질의어들을 지원하고 있다. 이러한 질의어에는 RDQL, RQL, SeRQL등이 있다. GO Guide에서는 SeRQL을 사용한다. 그 이유는 Sesame에서 지원하는 SeRQL 질의처리기가 다른 언어에 비해 좋은 성능을 보이며, RQL에 비해 표현력이 뛰어나기 때문이다[14]. SeRQL의 표현을 살펴보면 다음과 같다.

아래 그림 5는 SeRQL의 구조와 그 표현을 나타내고 Subject의 값이 "John"을 포함하는 스트링을 가지며 이 Subject의 Predicate값으로 각각 pre1과 pre2를 가지는 Object인 Obj1, Obj2를 찾는 질의문이다. SELECT문에서는 자기가 찾고자 하는 변수를 넣고FROM 절에는 찾고자 하는 데이타의 구조가 들어가게 된다. 그리고 WHERE 절에서는 특정 조건이 들어가고 namespace가 설정되도록 된다. 특히 FROM 절에서는 expression 경로가 사용되어 질의가 처리된다.

4.2.2 GO에서의 SeRQL

Gene Ontology에서의 질의는 크게3가지로 구분된다. term과 term사이의 is-a 관계, part-of 관계 그리고

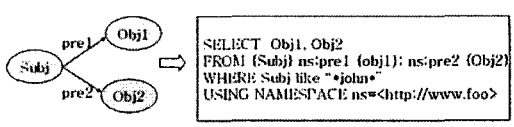

그림 5 SeRQL의 예

gene product와 term사이의 관계이다. 각각의 관계와 그 관계를 질의하기 위한 SeRQL은 다음과 같다. 우선 is-a 관계는 그림 9에서와 같이 표현된다. 즉 이름이 molecular function이면서 그 term의 자식 관계에 있는 term을 찾는 질의가 된다. 이 질의의 결과는 그림 6에서 보듯이 GO:0016209가 된다. Serql:directSubClassOf는 자식 관계를 나타내어주는 Predicate이며 rdfs:subClassOf의 Predicate은 자식이 아닌 모든 자손들을 다 나타내어 줄 수 있는 Predicate이다. Part-of 관계는 OWL에서 Restriction으로 묶여 있는 term을 찾는 관계가 된다. 그림 7은 이런 part-of 관계를 나타내어 준다. 이 예제는 GO:0003674의 term과 part-of 관계에 있는 term을 찾는 질의이다.

gene product와 term의 관계는 그림 8과 같이 나타내어 질 수 있다. 즉 그래프에서 자기 term에 속하는 여러 gene product들을 찾는 질의이다. 그림 8에서는 "4030414C22Rik"이라는 gene product에 관계하는 term을 찾는 질의가 된다.

4.2.3 GO Guide에서의 질의 변환

GO Guide에서는 앞서 설명한Gene Ontology의 관계를 처리하며 브라우저에서 입력한 질의형태, 즉 트리플 패턴을 SeRQL로 변환한다. 이 때 브라우저에서 사용자

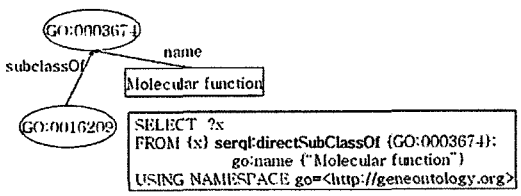

그림 6 is-a 관계와 SeRQL의 표현

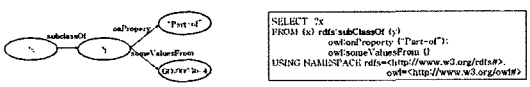

그림 7 part-of 관계

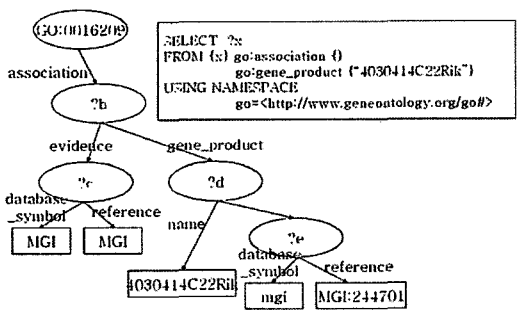

그림 8 term과 gene product 관계

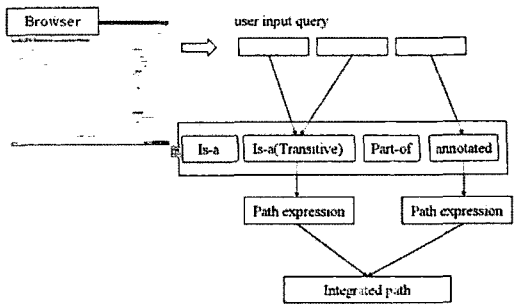

그림 9 질의 변환 과정

가 입력하는 질의는 OWL-QL의 형태이다. 즉 트리플의 패턴이고 must-binding 변수를 포함시켜 사용자가 찾고자 하는 변수를 찾을 수 있게 해 준다. SeRQL 질의 변환은 그림 9의 과정으로 이루어진다.

그림 9에서 사용자는 브라우저에서 질의를 입력한다. 이렇게 입력된 트리플 패턴은 해당되는 Predicate에 대한 SeRQL로 변환된다. 예를 들면 part-of 관계이면 part-of 관계를 기술할 수 있게미리 정의된 SeRQL을 사용한다. 이러한 패턴이 계속해서 입력이 되면 여러가지 SeRQL 로 표현이 된다. 본 논문에서는 이러한 SeRQL을 하나로 통합해서 사용하도록 한다. 즉 SeRQL에서must-binding 변수와 경로를 따로 분리해서 각각의 사용자 입력에 대해 맞는 패턴을 찾아서 계속해서 경로를 통합해서 하나의 경로로 만들어서 SeRQL에서 사용하게 된다.

표현 경로를 하나의 경로로 통합하는 알고리즘은 그림 10과 같다. 이렇게 하나의 경로로 통합이 되면 SeRQL의 FROM 절 뒤에 이 경로를 써서 사용하게 된다. 따라서 사용자는 트리플로 구성된 질의를 입력해도 GO Guide에서는 이 트리플 질의를 SeRQL로 바꾸어서 사용하게 되는 것이다. 이와 같은 과정을 통해 GO Guide는 브라우저에서 생성된 질의를 SeRQL질의로 변환해서 Sesame에 전달한다. 이 같은 분리를 통해 사용자에게는 질의어의 복잡한 기술방식이나 시스템에서 제공하는 질의어와 독립적으로 브라우징을 통한 트리플 패턴을 생성하게 함으로써 사용하기 쉽도록 하며, 시스템적인 측면에서는 SeRQL을 통한 효과적인 처리가 가능하고 필요에 따라 다른 질의어를 지원하는 온톨로지에 사용될 수 있다.

그림 11은 이러한 변환을 사용해서 질의를 처리하는 과정을 나타내고 있다. 사용자가 입력한 질의를 각각의 패턴에 맞는 경로를 만들어서 그 각각의 경로를 합쳐서 하나의 SeRQL로 표현한다. 즉 is-a 관계, part-of 관계, annotated(term과 gene product)의 관계를 그림 11에서는 하나의 질의어로 만들어서 사용한다. is-a 관계일 때

```
Algorithm : Integrated path
Require : Input : triple statement (Subject, Property, Object)
While (Input)
{
  /* find pattern
if (property is "is-a") predicate = is-a pattern
      else if (property is "part-of") predicate = part-of pattern
      else if (property is "annotated") predicate = annotated pattern
 /* build path expression
if (input is first)  path_expression = predicate
  else
       {  /* find common matching string
          matching_string = find (path_expression, predicate)
           /* integrate path expression
             if (path_expression starts matching_string)
             path_expression = path_expression + ";"
             + predicate's unmatching string
             if (path_expression ends matching_string)
             path_expression = path_expression
             + predicate's unmatching string
             else  path_expression =
             path_expression + ","+ predicate
} }
```

그림 10 표현경로 통합 알고리즘

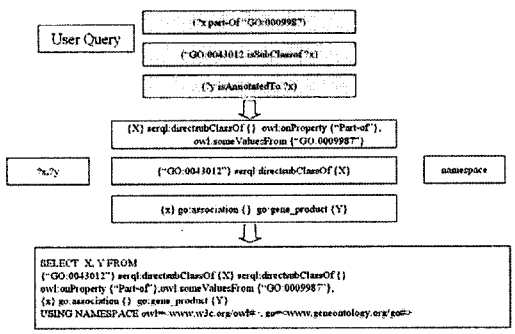

그림 11 GO Guide에서의 질의 변환 과정의 예

사용되는 경로, part-of 관계일 때 사용되는 경로, annotated 관계일 때 사용되는 경로가 하나의 경로로 합쳐지고 사용자가 주는 must-binding 변수에 대해 질의 처리를 수행하게 된다.

위의 질의처리 과정은 "GO:0009987"의 ID를 가지는 term과 "part-Of" 관계이면서 "GO:0043012"의 부모인 term에 해당하는 gene product를 찾는 질의이다. 최상단에서 브라우저 에서 생성된User Query가 SeRQL문으로 변환되는 과정을 나타내고 있다. 즉 3개의 트리플 경로가 하나의 경로 표현으로 나타내어지는 것이다.

5. 기존 시스템과의 비교

Gene Ontology를 이용하기 위한 기존의 대표적인 시스템인 AmiGO 그리고 GoView와 GoGet은 관계형 데이타모델로 기술된 Gene Ontology 데이타를 사용하며, RDB를 바로 접근하는 어플리케이션이다. 반면 GO Guide는 OWL로 기술된 트리플 모델의 Gene Ontology 데이타를 Sesame를 이용해서 RDB저장소에 저장하고 접근하기 때문에 질의 수행시간의 비교가 어렵다. 특히 기존의 시스템은 질의를 지원하지 않거나 검색할 범위를 선택하는 매우 제한적인 질의만을 지원하기 때문에 실질적으로 질의를 지원한다고 할 수 없고, GO Guide에서는 한번에 수행되는 질의들을 기존의 시스템에서는 사용자가 여러 번 검색을 수행해야 하기 때문에 질의 시간 비교가 불가능하다. 물론 표 3에서 기존 브라우저가 지원하는단순 질의에 대해서 비교할 경우 트리플 모델의 이행적 폐포(transitive closure) 연산에 대한 비효율성과 데이타베이스 접근에 있어 질의 변환 단계를 거치기 때문에 관계형 데이타모델로 기술된 데이타를 사용하는 기존의 시스템이 나은 성능을 보일 수 있지만 이 질의들은 실제 생물학 연구에 있어 매우 비중이 적거나 무의미하기 때문에 본 장에서는 기존 시스템과의 기능별 비교와 주요 질의에 대한 지원여부를 비교한다.

표 1은 Gene Ontology 검색과 관련된 기능에 대해 각 시스템을 비교한 결과이다.

표 2에 나열하는 질의는 Gene Ontology Consortium에서 제시한 Sample Query로서 Gene Ontology에서 발생하는 주요 검색에 대한 예제이다. 총 12개의 질의가 제공되는데 이중 8개가 주요 질의이고, 나머지 4개의 질의는 8개 질의의 조합을 통해 수행 가능한 질의이므로 8개의 주요 질의에 대해 표 3에서 각 시스템 별 지원 여부를 나타내고 있다. 괄호가 없는 질의는 term 간의 관계를 검색하는 질의이며 괄호 안의 질의는 term과 gene product의 관계까지 검색하는 질의이다. 아래 표 2와 3에서 term은 'T'로, gene product는 'GP'로 표기한다.

표 3에서 보면 GoView와 GoGet은 특정 term과 관련된 gene product를 검색하는 기능만을 제공하고 특정 term에 속하는 하위 term은 찾지 못한다. 또한 Go Guide가 제한 없이 복합 질의를 생성할 수 있는 반면 GoView와 GoGet은 2개까지의 복합 질의만을 지원한다. 예를 들어 GoGuide는 A와 B와 C의 공통 자손에 대한 gene product 검색이 가능하지만 GoView와 GoGet은 A와 B의 공통 자손에 대한 gene product 검색만 가능하다.

표 1과 3에서 보듯이 GoGuide는 검색에 대한 다양한 기능과 편의성뿐만 아니라 Gene Ontology와 관련된 모든 유형의 질의를 지원한다는 점에서 기존 시스템보다

표 1 기능별 시스템 비교

	Go Guide	AmiGO	GoView & GoGet
데이타형식	OWL format	RDB format	RDB format
검색방식	동적 그래프, 질의 트리, 키워드	정적 그래프, 트리, 키워드	정적 그래프, 트리, 키워드
질의	SeRQL	없음	없음
질의 생성기능	모든 질의 생성가능	없음	단순, 단일 질의만 지원
그래프탐색	자유로운 탐색 가능	구조만 표현, 탐색 불가	구조만 표현, 탐색 불가
입력 도움기능	지원	없음	없음
질의 중간결과 표시	지원	없음	없음

표 2 Gene Ontology에 대한 주요 질의문

질의	유형	예제
Q1(Q9)	한 키워드에 해당하는 T 검색 (한 키워드에 해당하는 GP가 속하는 모든 T 검색)	Germination에 해당하는 T 검색 (ACC1이라는 GP에 대한 검색)
Q2(Q10)	특정 T의 child T 검색 (특정 T의 child T에 대한 GP 검색)	Germination의 child T 검색 (Germination의 child T에 대한 GP 검색)
Q3(Q11)	특정 T의 parent T 검색 (특정 T의 parent T에 대한 GP 검색)	neuron differentiation의 parent T 검색 (neuron differentiation의 parent T에 대한 GP 검색)
Q4(Q12)	특정 T의 descendent T 검색 (특정 T의 descendent T에 대한 GP 검색)	antibiotic biosynthesis의 descendent T 검색 (antibiotic biosynthesis의 descendent T에 대한 GP 검색)
Q5(Q13)	특정 T의 ancestor T 검색 (특정 T의 ancestor T에 대한 GP 검색)	adult behavior의 ancestor T을 검색 (adult behavior의 ancestor T에 대한 GP 검색)
Q6(Q14)	두 T의 공통 descendent T 검색 (두 T의 공통 descendent T에 대한 GP 검색)	Biosynthesis와 lipid metabolism의 common descendent T 검색 (Biosynthesis와 lipid metabolism의 common descendent T에 대한 GP 검색)
Q7(Q15)	두 T의 공통 ancestor T 검색 (두 T의 공통 ancestor T에 대한 GP 검색)	adult behavior와 cell communication의 common ancestor T 검색 (adult behavior와 cell communication의 common ancestor T에 대한 GP 검색)
Q8(Q16)	한 T의 descendent이면서 다른 T의 child인 T 검색 (한 T의 descendent이면서 다른 T의 child인 T에 대한 GP 검색)	Proplastid의 일부분이면서 plastid stroma에 속하는 T 검색 (Proplastid의 일부분이면서 plastid stroma에 속하는 T에 대한 GP 검색)

표 3 질의 지원여부

	Go Guide	AmiGO	GoView & GoGet
Q1(Q9)	지원(지원)	지원(지원)	지원(지원)
Q2(Q10)	지원(지원)	불가	불가(지원)
Q3(Q11)	지원(지원)	불가	불가
Q4(Q12)	지원(지원)	불가	불가(지원)
Q5(Q13)	지원(지원)	불가	불가
Q6(Q14)	지원(지원)	불가	불가(지원)
Q7(Q15)	지원(지원)	불가	불가
Q8(Q16)	지원(지원)	불가	불가

유용하다. Gene Ontology의 내용이 매우방대하고 그 안에서 중요한 검색이 관계에 따른 검색인 만큼 단순히 키워드나 트리의 검색에 따라 데이타를 나열하는 시스템들은 연구 도구로서 매우 미흡하다. 반면 GO Guide는 기존의 검색기법들뿐만 아니라 연구에 필요한 모든 관계를 질의로 생성, 검색하는 기능을 제공한다.

6. 결론 및 향후 연구

본 논문은 생물학 연구에 있어 유전자에 대한 데이타를 효과적으로 이용하기 위한 브라우저와 질의변환기법을 제안하였다. 브라우저에 있어서는 키워드, 트리, 그래프 검색기법이 통합되어 유기적으로 탐색이 이루어지도록 하도록 제안하였으며 사용자 편의를 위한 키워드생성을 지원하도록 하였다. 특히 질의를 통해 기존의 시스템에서는 반복적으로 사용자가 수행해야 할 작업을 한 번의 질의로 처리할 수 있도록 하였다. 그리고 다양한 형태의 질의를 생성할 수 있게 함으로써 기존의 시스템과 달리 Gene Ontology 검색과 관련된 모든 유형의 질의를 지원한다. 이와 같은 질의의 생성에 있어서도 사용자가 복잡한 질의를 직접 기술할 필요 없이 탐색하는 과정에서 질의를 생성하는 기법을 제안하였다.

질의 변환에 있어서는 브라우저에서 생성된 질의를

OWL 질의어인 SeRQL로 변환하는 기법을 제안하였다. 이를 통해 사용자 수준에서는 브라우저를 통해 질의를 손쉽게 생성하고 처리에 있어서는 SeRQL을 통해 OWL 데이타를 효과적으로 다룰 수 있게 된다. 또한 온톨로지 독립성을 지원할 수 있는 기반이 제공되므로 추후 온톨로지의 종류나 기술언어가 달라져도 목적 질의어에 대한 변환기를 구축함으로써 사용자에게 일관된 검색환경을 제공할 수 있게 된다. 본 논문에서는 이와 같은 기법으로 구축된 시스템을 Gene Ontology 데이타를 통해 구현해봄으로써 효과적으로 Gene Ontology를 검색할 수 있음을 확인하였다.

추후 연구 과제로는 브라우징 시 연구 분야별로 관련된 데이타를 구분하여 사용자에게 제공하는 필터링 기법과 추론을 통한 데이타 검색과 같은 더욱 고차원의 검색 기법을생각해 볼 수 있다. 또한 다양한 온톨로지를 위한 범용 질의 변화기술 및 질의 처리에 있어 저장구조와 연동하여 질의 수행 성능을 향상시키는 기법에 대한 연구도 필요하다.

참 고 문 헌

[1] Robert Stevens et al., "Ontology based knowledge representation for bioinformatics," Briefings in Bioinformatics, 1(4), pp.398-414, 2000.

[2] P.D. Karp et al., "The EcoCyc and MetaCyc Databases," Nucleic Acids Research, 28, pp.56-59, 2000.

[3] Deborah L. McGuinness, Frank van Harmelen, "OWL Web Ontology Language Overview," W3C Recommendation, http://www.w3c.org/TR/owl-features/, 2004.

[4] The Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation," Genome Research, 11(8), pp.1425-33, 2001.

[5] Jeen Broekstra, Arjohn Kampman, "SeRQL: An RDF Query and Transformation Language," open RDF.org, http://openrdf.org, 2004.

[6] M. Uschold et al., "The Enterpnrise Ontology," The Knowledge Engineering Review, 13(1), 1998.

[7] O. Lassila, R. Swick, "Resource Description Framework(RDF) Model and Syntax Specification," W3C Recommendation, World Wide Web Consortioum, 1999.

[8] Open Biomedical Ontology, http://obo.sourceforge.net/, 2005.

[9] Berkeley Drosophila Genome Project, "AmiGO," http://www.fruitfly.org/, 2005.

[10] Eliazbeth Shoop et al., "Data Exploration Tools for the Gene Ontology Database," Bioinformatics, 20(18), pp.3442-3454, 2004.

[11] Jeen Broekstra et al. "Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema," ISWC, 2002.

[12] RDQL - "A Query Language for RDF," http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/, 2004.

[13] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, "RQL: A Declarative Query Language for RDF," WWWC, 2002.

[14] Peter Haase, Jeen Broekstra, Andreas Eberhart, Raphael Volz, "A Comparison of RDF Query Language," ISWC 2004.

정 준 원

2000년 동국대학교 컴퓨터공학과(학사) 2003년 서울대학교 전기.컴퓨터공학부(석사). 2003년~현재 서울대학교 전기.컴퓨터공학부 박사과정 재학 중. 관심분야는 시멘틱웹, 온톨로지, XML, 데이타베이스

박 형 우

2004년 서울대학교 컴퓨터공학부(학사) 2004년~현재 서울대학교 전기.컴퓨터공학부 석.박사 통합과정 재학 중. 관심분야는 시멘틱웹, 온톨로지, XML, 데이타베이스

임 동 혁

2003년 고려대학교 컴퓨터교육과(학사) 2005년 서울대학교 전기.컴퓨터공학부(석사). 2005년~현재 울대학교 전기.컴퓨공학부 박사과정 재학 중. 관심분야는 데이타베이스, XML, 시멘틱웹, 온톨로지

이 강 표

2004년 연세대학교 컴퓨터과학과(학사) 2004년~현재 서울대학교 전기.컴퓨터공학부 석사 과정 재학 중. 관심분야는 데이타베이스, 시멘틱웹, 생물정보학

김 형 주

1982년 서울대학교 전산학과(학사). 1985년 Univ. of Texas at Austin(석사) 1988년 Univ. of Texas at Austin(박사). 1998년 5월~1988년 9월 Univ. of Texas at Austin(Post-Doc). 1988년 9월~1990년 12월 Georgia Institute of Technology(부교수). 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이타베이스, XML, 시멘틱웹, 온톨로지