Gene Ontology와 그 응용

유상원

1. 서론

온톨로지는 간단히 말하면 어떤 분야의 지식을 통일되고 일관성 있게 공유할 수 있는 형태로 만들어 놓은 것을 말한다. Gene Ontology(줄여서 GO)[1] 역시 생물학 분야의 지식 중 cellular component, molecular function, biological process 이라는 세가지 분야에 대한 용어를-GO 컨소시움에서는 생물학적 개념들에 대해 용어(term)라는 표현을 사용한다- 통일된 형태로 제공하고 이를 공개하여 원하는 연구자는 누구나 사용할 수 있도록 하고 있다.

GO project의 시작은 1998년에 FlyBase[2], Saccharomyces Genome Databases[3], 그리고 Mouse Genome Database[4]의 세 DB그룹들의 참여로 이루어졌다. 이 프로젝트의 가장 기본적인 목적은 파리, 효모, 쥐와 같은 서로 다른 생물종간에 일관성 있게 생물학적인 용어들을 기술하는 것이었다. 왜냐하면 생물학 분야의 연구를 수행하는데 있어서 용어의 통일이 큰 문제가 되었기 때문이다. 예를 들어 한 그룹에서 어떤 유전자의 기능을 밝혀낸 후 A라고 부르고 연구결과를 발표한 논문에서도 A라는 용어를 사용하는데 다른 그룹에서는 B라는 용어를 사용한다면 해당 연구결과를 검색하거나 찾는데 어려움이 따를 것이다. 이런 것들이 한 두 가지에 국한된 것이 아니라 꽤 많은 용어들이 의미는 같은데 서로 다르게 표현되고 있는 상태라고 하면 문제는 심각해질 것이다.

또한 생물학 분야에는 공개된 형태로 서비스를 하고 있는 다양한 생물정보DB들이 있는데 여기서 원하는 정보를 얻는데도 용어의 통일이 중요한 문제이다. 유전자의 서열정보나 단백질 정보 등을 검색하는데 있어서 생물 종마다 서비스를 하는 DB도 있고 여러 종에 대해 통합된 DB를 구축하고 검색기능을 제공하는 DB도 있다. 그러므로 한가지 유전자나 단백질정보가 여러 개의 DB에 흩어져 조금씩 다른 형태로 제공되는 환경이 일반적이다. 이 경우 용어가 통일되어 있지 않고 DB마다 다르게 사용된다면 사용자 입장에서는 경험을 통해 익숙해지기 전까지 혼란과 착오를 반복할 수 밖에 없게 된다.

현재 GO 컨소시움에는 16개의 기관들이 참여하여 참여하는 DB그룹들간에 용어 통일을 꾀하고 있다. 각 그룹이 가지고 있는 데이터베이스에는 관련 정보에 GO ID값이 들어가 해당 정보가 GO의 용어로 표현하면 어떻게 되는지 쉽게 알 수 있게 된다. 그러므로 만약 학습이나 기억에 관한 유전자 정보가 필요하다면 이를 GO에서는 learning and/or memory 라고 기술하고 있으므로 각 DB에 learning and/or memory라는 키워드를 이용하거나 여기에 부여된 ID값을 이용하면 원하는 결과를 얻을 수 있다.

GO를 만들고 사용하는 것의 가장 기본적인 요구사항이 용어통일이라면 여기에 GO는

이 용어들간의 관계도 표현하고 있다. 앞서 언급한 것처럼 생물학 개념에 해당하는 용어들을 크게 세 그룹으로 나누고 is_a 관계와 part_of 관계를 이용하여 이 용어들간에 상하관계를 표현하고 있다. 또한 각 용어들에 해당하는 실제 인스턴스들도 같이 표현되어 있다.

이에 대한 자세한 형식적인 면들은 다음 장에서 다루도록 한다.

2. Gene Ontology 의 형식과 내용

Gene ontology가 제공되는 형식은 OBO format, XML, MySQL, OWL 네 가지의 형태로 제공되고 있다.(FASTA 형식은 GO 컨소시움에서 가이드를 제공하지 않는다) OBO는 Open Biological Ontology의 약자로 자체적인 syntax를 가지고 flat file의 형태로 제공된다. XML 형식은 RDF에서 제공하는 syntax를 이용하여 정확한 RDF문서는 아니지만(RDF like XML) is_a관계나 par_of관계를 표현하고 있다. MySQL은 DB테이블의 dump형태로 제공되고 있으며 OWL은 표준 OWL형식을 따르고 있다.

다음은 MySQL dump형태로 제공되는 Gene Ontology의 여러 형태이다.

- termdb - ontologies, definitions and mappings to other dbs
- assocdb - the above, plus associations to gene products
- seqdb - the above, plus protein sequences for some of the gene products
- seqdblite - the above, with IEA associations stripped out (this is the version that drives AmiGO)

<MySQL dump 형태로 제공되는 Gene Ontology 의 종류>

각각의 DB가 포함하고 있는 내용을 살펴보면 termdb에는 생물학적인 용어들간의 관계를 표현한 온톨로지와 각 용어에 대한 정의 그리고 이 용어와 같은 의미를 가지는 개념들이 다른 DB에도 기술되어 있는 경우 외부 DB에 대한 참조값을 가지고 있다. assocdb에는 termdb에는 없는 유전자나 전사체, 단백질 등에 관한 추가적인 정보가 있다(GO에서는 이를 association이라 한다). 온톨로지는 개념이나 용어만을 나타낼 뿐 실제 인스턴스를 나타내는 것이 아니다. 따라서 온톨로지에 표현되어 있는 용어를 적용시킬 수 있는 실제 단백질이나 유전자에 관한 정보가 인스턴스가 되는 것이다. Seqdb는 여기에 추가로 단백질의 서열정보까지 포함하고 있는 DB이며 seqdblite는 이중에서 사람에 의해 검증되지 않은 정보들을 제외한 것이다. 정보의 양에 비해 전문가의 수는 적기 때문에 사람이 모든 정보가 GO에 제대로 대응이 되어 있는지 검증하기는 어려운 일이다. 따라서 서열의 유사성이나 다른 DB로부터 정보를 그대로 가져오는 기계적인 방법으로 GO를 대응시키기도 하는데 이러한 정보는 정보의 품질을 고려하여 제외되는 것이다.

정리하면 GO 컨소시움에 참여하고 있는 각 기관들은 자신의 DB를 GO내의 통일된 용

어들로 기술하게 되는데 GO 컨소시움에서는 통일된 용어와 이들간의 관계만 제공하기도 하고 실제 이 용어로 기술되어 있는 인스턴스와 부가정보까지 함께 제공하기도 한다. 이는 형식적인 면에서도 나타나고 있다. OBO format과 OWL 형식은 termdb에 나타난 내용만을 기술하고 있고 XML은 termdb와 assocdb의 두 가지 형태가 있다. MySQL dump는 위의 네 가지 내용을 모두 다 제공하고 있다. 그러므로 자신이 필요한 응용분야에 따라 형식이나 내용을 선택하는 것이 필요하다.

각각의 자세한 내용은 Gene Ontology 컨소시움에서 제공하는 File Format Guide에 나와 있으므로 여기서는 그 중 XML형식과 OWL형식을 이용해 Gene Ontology가 어떻게 기술되어 있는지 살펴보기로 한다.

기본적으로 GO는 DAG(Directed Acyclic Graph)로 표현되며 OWL과 XML은 그래프로 표현되는 데이터를 나타내기에 적합한 표준이다.

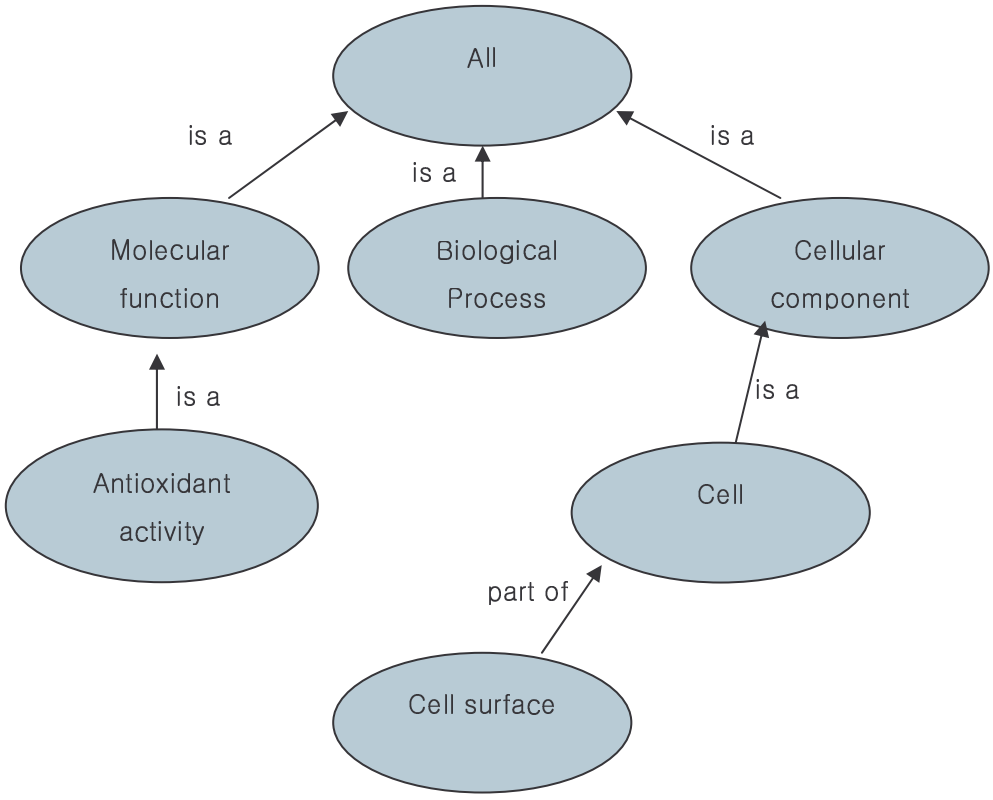

그림1. Gene Ontology 그래프

그림1은 GO에 나타난 용어들의 일부를 그래프 관계로 표현한 것이며 용어들은 노드가 되고 용어들 사이에 is_a관계나 part_of관계가 간선이 된다. 먼저 XML을 이용하여 어떻게

위의 관계들이 표현되어 있으며 부가적으로 어떤 정보들이 담겨있는지 살펴보자.

그림2에 나오는 첫번째 예는 항산화 작용(antioxidant activity)이라는 용어에 대한 정보가 담겨있는 XML로 기술된 termdb ontology의 예이다.

```
<go:term rdf:about="http://www.geneontology.org/go#GO:0016209" n_associations="0">
        <go:accession>GO:0016209</go:accession>
        <go:name>antioxidant activity</go:name>
        <go:definition>
Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can itself be more easily oxidized than the substance protected. The term is often applied to components that can trap free radicals, thereby breaking the chain reaction that normally leads to extensive biological damage.
         </go:definition>
        <go:is_a rdf:resource="http://www.geneontology.org/go#GO:0003674" />
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>SP_KW</go:database_symbol>
            <go:reference>Antioxidant</go:reference>
        </go:dbxref>
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>HAMAP</go:database_symbol>
            <go:reference>MF_00269</go:reference>
        </go:dbxref>
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>HAMAP</go:database_symbol>
            <go:reference>MF_00401</go:reference>
        </go:dbxref>
 </go:term>
```

그림2. XML로 이루어진 termdb

태그의 시작부분마다 나오는 go는 name space에 해당하며 accession은 7자리로 이루어진 고유한 id 값이다. name은 해당 개념을 표현하는 용어이고 definition은 용어에 대한 정의를 담고 있다. is_a는 이 용어가 용어0003674(molecular function)와 is_a 관계에 있다는 것을 말한다. 즉 antioxidant activity is a molecular function이 된다. dbxref는 해당 개념이 나타나 있는 외부의 DB에 대한 참조를 나타내며 database_symbal은 어떤 DB인지를 나타낸다. 위의 예의 경우 SP_KW는 SwissProt KnowledgeBase를 나타내고 HAMAP은

HIGH-QUALITY AUTOMATED AND MANUAL ANNOTATION OF MICROBIAL PROTEOMES을 나타내며 이러한 약어와 대응 관계는 GO 컨소시움에서 별도로 제공하고 있다. Reference는 해당 DB에 접근하는 id를 나타낸다. 이렇게 별도의 reference를 제공하는 이유는 동일한 의미를 가지는 용어나 개념이 다른 DB에 존재하기 때문이다. 실제로 각 DB그룹에서 자체적으로 GO와 같은 분류체계를 사용하는 곳이 많기 때문에 GO와의 연동이나 GO의 확장을 위해서 필요한 정보라고 할 수 있다.

그림3에 나오는 두 번째 예는 assocdb에 나오는 예이다. 똑같은 항산화 작용에 대한 설명이지만 하나의 용어에 대한 정보가 훨씬 많다. 정보의 양에서 차이가 나는 이유는 용어 자체에 대한 설명뿐 아니라 각 용어가 적용되는 단백질이나 유전자에 대한 주석들이 붙어 있기 때문이다. 첫줄에 나오는 n_associations가 해당하는 단백질이나 유전자 주석의 수를 말해준다. 이 예에서는 322라고 되어 있는데 항산화 작용을 하는 유전물질로 참조가 가능한 것의 개수가 322개라는 의미가 된다.

```
<go:term rdf:about="http://www.geneontology.org/go#GO:0016209" n_associations="322">
        <go:accession>GO:0016209</go:accession>
        <go:name>antioxidant activity</go:name>
        <go:definition>Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can itself be more easily oxidized than the substance protected. The term is often applied to components that can trap free radicals, thereby breaking the chain reaction that normally leads to extensive biological damage.</go:definition>
        <go:is_a rdf:resource="http://www.geneontology.org/go#GO:0003674" />
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>SP_KW</go:database_symbol>
            <go:reference>Antioxidant</go:reference>
        </go:dbxref>
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>HAMAP</go:database_symbol>
            <go:reference>MF_00269</go:reference>
        </go:dbxref>
        <go:dbxref rdf:parseType="Resource">
            <go:database_symbol>HAMAP</go:database_symbol>
            <go:reference>MF_00401</go:reference>
        </go:dbxref>
**************************위부분은 termdb와 동일한 부분************************************
        <go:association rdf:parseType="Resource">
            <go:evidence evidence_code="ISS">
```

```
                    <go:dbxref rdf:parseType="Resource">
                        <go:database_symbol>MGI</go:database_symbol>
                        <go:reference>MGI:2429377</go:reference>
                    </go:dbxref>
                </go:evidence>
                <go:gene_product rdf:parseType="Resource">
                    <go:name>4930414C22Rik</go:name>
                    <go:dbxref rdf:parseType="Resource">
                        <go:database_symbol>mgi</go:database_symbol>
                        <go:reference>MGI:2444701</go:reference>
                    </go:dbxref>
                </go:gene_product>
            </go:association>
…생략
…생략
            <go:association rdf:parseType="Resource">
                <go:evidence evidence_code="ISS">
                    <go:dbxref rdf:parseType="Resource">
                        <go:database_symbol>PMID</go:database_symbol>
                        <go:reference>10952301</go:reference>
                    </go:dbxref>
                </go:evidence>
                <go:gene_product rdf:parseType="Resource">
                    <go:name>VC1350</go:name>
                    <go:dbxref rdf:parseType="Resource">
                        <go:database_symbol>tigr_cmr</go:database_symbol>
                        <go:reference>VC1350</go:reference>
                    </go:dbxref>
                </go:gene_product>
            </go:association>
</go:term>
```

그림 3. XML로 이루어진 assocdb

이제 association부분에 대해 설명을 하겠다. 각 DB group들은 자신들이 가지고 있는 단백질이나 유전자에 관한 정보를 Gene ontology에 대응시킬 때 근거를 명시하도록 되어 있다. 그래서 첫번째 association의 경우 evidence에 그 근거가 나와 있고 11가지의 코드

중 코드 값이 ISS라면 inferred from sequence similarity를 의미한다. 근거의 구체적인 내용을 알기위해 MGI(Mouse Genome Informatics) http://www.informatics.jax.org/ 에 가서 accession id에 해당하는 MGI:2429377 로 검색해 보면 “The FANTOM Consortium and The RIKEN Genome Exploration Resea, Nature 2002;420():563-573” 이라는 논문에 관한 정보를 얻을 수가 있다. 다음으로 실제 유전자에 관한 정보를 얻기 위해 유전자 이름에 해당하는 4930414C22Rik 이나 id인 MGI:2444701 로 검색을 해보면 “Prdx6-rs1, peroxiredoxin 6, related sequence 1, Chr 2” 과 같은 유전자에 관한 자세한 정보를 얻을 수 있다. 근거 또는 참조로 제시되는 DB는 외부 DB뿐 아니라 GO 자체일 수도 있고 책인 경우 ISBN이 사용되며 논문인 경우 PMID(PubMed ID)가 사용된다.

GO는 분자수준의 역할(molecular function), 생물학적 작용(biological process), 세포구조(Cellular component) 이렇게 세가지로 이루어져 있다. 따라서 위에서 예로 든 쥐의 유전자 4930414C22Rik가 GO내에서 대응이 이루어지는 용어들은 하나가 아니라 여러 가지가 될 수 있다.

Gene Ontology (GO) classifications	Process	lipid catabolism
	Component	lysosome
	Function	antioxidant activity, catalytic activity...
	All GO classifications(7)	

그림 4. GO와 유전물질의 대응

그림4는 하나의 유전물질이 GO내에서 어떻게 대응되는지를 표로 정리한 것이다. 표를 보면 porcess에 하나 component에 하나 그리고 function에 5개가 대응되어 모두 7개에 대응됨을 알 수 있다. 또 용어 사이에는 is_a관계나 part_of 관계가 있기 때문에 하위 용어에 대응된다면 상위 용어에도 대응된다고 볼 수 있다. 표에 나타난 내용을 말로 풀어쓰면 “이 유전자는 리보좀 내에서 지질분할의 생물학적 역할을 하는데 항산화 작용과 연관이 있다” 와 같이 쓸 수 있을 것이다.

다음은 OWL로 표현한 예에 대해 살펴보도록 하겠다.

```
<owl:Class rdf:ID="GO_0016209">
    <rdfs:label>antioxidant activity</rdfs:label>
<!-- molecular_function -->
    <rdfs:subClassOf rdf:resource="#GO_0003674"/>
 </owl:Class>

<owl:Class rdf:ID="GO_0009986">
```

```
    <rdfs:label>cell surface</rdfs:label>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:about="#part_of"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#GO_0005623"/>
      </owl:Restriction>
<!-- cell -->
    </rdfs:subClassOf>
 </owl:Class>
```

그림 5. OWL로 이루어진 termdb

OWL 형식의 GO데이터는 termdb의 내용만을 제공하기 때문에 클래스(용어)와 클래스 사이의 관계만을 기술하고 있다. 위에서는 두 가지 클래스를 예로 들었는데 첫번째 클래스는 항산화 작용이 molecular function의 subclass임을 나타내고 있고 두 번째 예는 cell surface가 cell 의 part of 로 이루어져 있음을 나타낸다. OWL형식에는 용어와 용어 사이의 관계 이외의 부가정보는 제공되지 않고 있다. 하지만 OWL이 W3C의 표준 온톨로지 언어이고 앞으로 온톨로지의 확장이나 통합 환경을 생각할 때 GO의 OWL을 이용한 기술은 적절한 선택이라고 할 수 있다.

3. Gene Ontology의 응용

Gene Ontology의 응용분야는 크게 나누어 세가지 분야로 나누어 볼 수 있다. 첫째는 Ontology자체를 브라우징하고 질의하는 것이고 둘째는 각 데이터베이스를 GO를 이용해 주석함으로써 일관된 검색 및 질의 환경을 제공하는 것이다. 세번째는 생물학적 실험의 결과물을 GO내에 대응시켜봄으로써 좀 더 의미 있는 결과를 얻어내는 것이다. 이 장에서는 각각의 응용에 대한 대표적인 연구들을 통해 GO의 응용분야에 대해 알아보도록 한다.

3.1 GO 브라우저

GO의 크기는 2005년 3월 현재 17680개의 용어로 이루어져 있고 이중 obsolete term을 제외하면 9247개의 biological_process 관련용어, 1484개의 cellular_component 관련 용어, 6949개의 molecular_function 관련용어가 전체 GO를 이루고 있다. 또 각 용어에 해당하는 주석 정보는 기관별로 수천개에서 수십만개에 이르고 있다. 따라서 온톨로지 관련 정보를 쉽게 검색하기 위한 툴들이 필요한데 GO컨소시움에서는 기본적인 브라우저로 AmiGO[5]를

제공한다.

그림6을 보면 AmiGo 브라우저의 모습이 나와 있다. 왼쪽 상단에 키워드 검색메뉴가 있고 그 아래에는 원하는 생물종이나 DB 또는 evidence code를 설정하여 원하는 값을 찾는 것을 도와주고 있다. 오른쪽은 검색 결과나 브라우징을 위한 트리를 제공하며 Graphical View를 통해 그래프 형태의 모습도 보여준다.

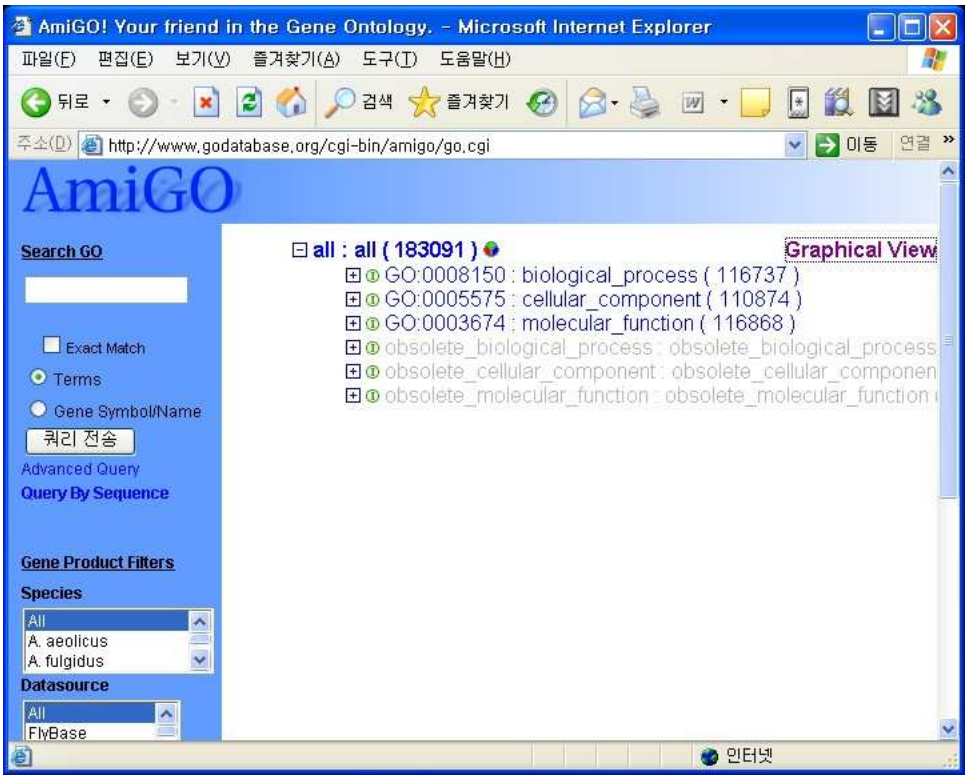

그림 6. AmiGO 브라우저

AmiGO브라우저는 웹 인터페이스를 기반으로 이루어져 있기 때문에 질의 결과의 제시와 2차 검색이 링크를 통해 이루어진다. 하부구조로 사용하는 DB는 GO에서 제공하는 MySQL의 dump중 seqdblite를 사용하고 있다.

AmiGO이외에도 GO를 지원하는 많은 브라우저들이 나와있다[6]. 브라우저의 기능은 대동소이 하지만 검색기능을 지원하기 위한 사용자 인터페이스를 어떻게 구성하느냐에 따라 복잡한 검색이 가능할 수도 있고 여러 번의 스텝을 한번으로 단축시킬 수도 있다. 즉 트리나 그래프를 이용한 브라우징과 키워드 검색, 질의 생성 등을 어떻게 결합시키느냐에 따라 여러 가지 가능성을 생각해 볼 수 있다. GO가 여러 형태 중 OWL형식으로 저장되어 있다면 온톨로지 질의어를 이용해 복잡한 검색을 수행하도록 할 수 있을 것이다. 이 경우는 사용자 인터페이스를 통한 질의 변환이 주요한 이슈가 될 것이다. 또한 읽어오는 데이터의 저장형태가 DB테이블의 형태인지 파일 시스템인지에 따라서 검색수행속도의 개선여부를 고려해

볼 수 있을 것이다.

3.2 GO annotation

GO 콘소시움에서 GO를 만든 기본적인 목적은 참여하고 있는 DB그룹들 간에 용어의 통일을 꾀하는 것이다. 따라서 각각의 DB그룹들은 자신들이 가지고 있는 데이터들을 GO에서 정의된 용어들을 이용해 주석할 필요가 있다. 예를 들어 Flybase[2]에서 catalase라는 효소에 관한 유전자 정보를 검색하면 그림7과 같다.

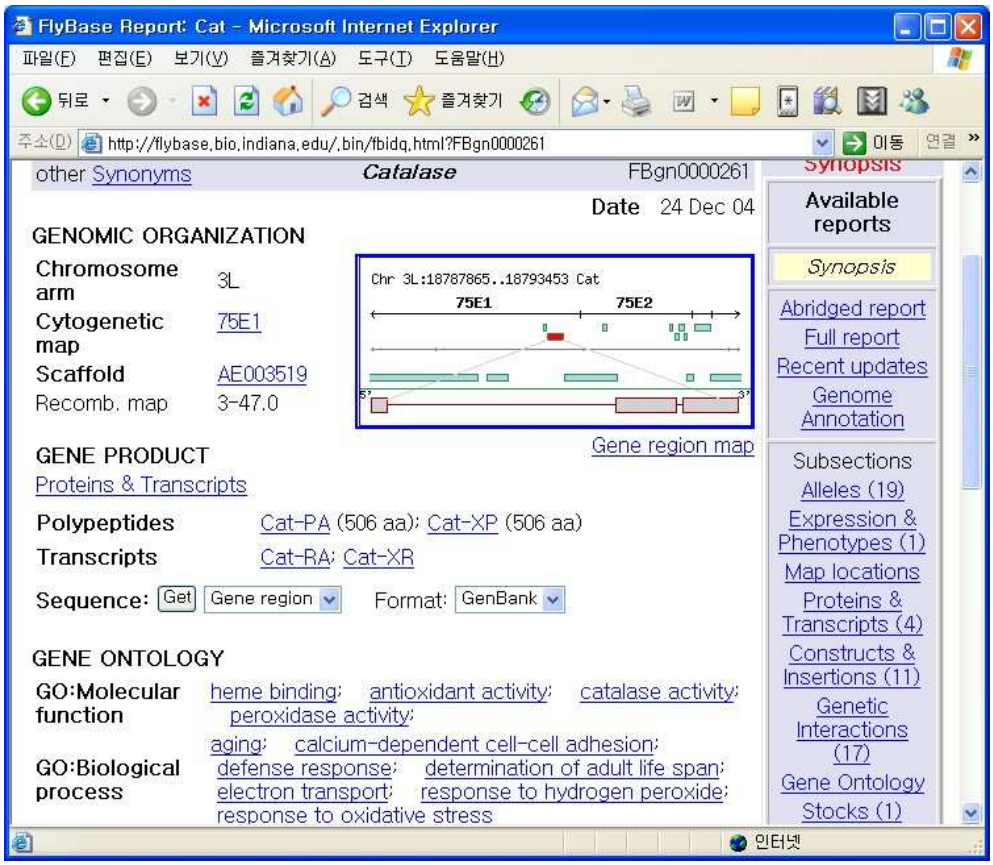

그림 7. GO참여 그룹의 GO annotation

그림 7을 보면 데이터베이스에 GO를 이용해 주석을 달고 있는 것을 알 수 있다. 이 정보를 이용하면 검색한 결과가 GO내의 어떤 용어에 대응되는지를 알 수 있고 다른 DB에서 검색한 결과도 동일한 GO의 용어를 사용한다면 두 데이터는 같은 세포내의 위치에서 발견되거나 같은 생화학적 작용을 하는 물질이라고 생각할 수 있게 된다. 하지만 새로운 유전물질이 계속 발견되고 최근 들어 대량의 실험데이터들이 쏟아지면서 각각의 데이터마다 모두 사람이 주석을 달기는 어려워지고 있다. 그래서 DB내의 데이터들을 주석하는데 도움을 주

는 툴들이 개발되어 쓰이고 있다.

DB내의 데이터들이 GO내의 어떤 개념에 대응되는지를 판단하는데 도움을 주는 대표적인 방법은 서열정보의 유사성과 문헌 정보를 이용하는 것이다. 먼저 유전자내의 염기서열이나 단백질 서열의 유사도를 비교하는 것은 서열이 유사할 경우 그 기능도 유사할 것이라는 가정에서 출발한다. 따라서 자신이 가지고 있는 데이터 내에 서열에 관한 정보가 포함되어 있다면 이미 GO로 기술된 정보와 서열의 유사성을 비교하여 일정한 유사도 값을 가지면 해당 데이터는 같은 GO의 용어로 매핑하는 것이다.

다음은 서열의 유사도를 이용하는 GoFigure[7]를 이용하여 얻어낸 결과이다.

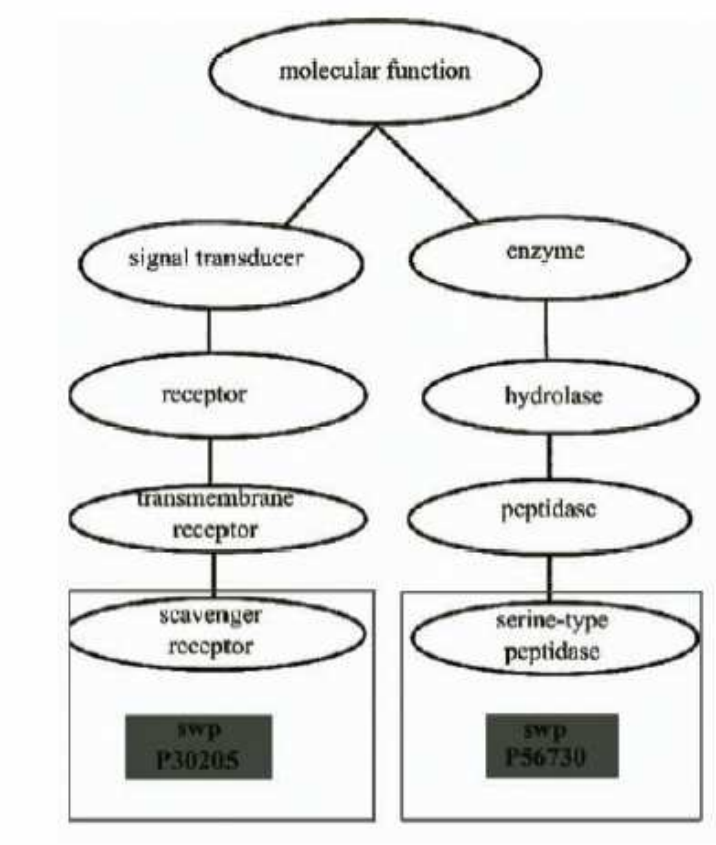

그림 8. GoFigure를 이용한 유사도 검색의 결과

위 그림8을 보면 서열을 입력한 후 유사도 검색의 결과로 GO graph가 반환되는 것을 알 수 있다. GOFigure는 먼저 유사도 검색을 하여 대응되는 GO의 용어들을 찾는다. 두번째 단계에서 GO를 DAG로 보고 매칭이 일어난 용어들을 모두 포함할 수 있는 서브그래프를 찾는다. 세번째 단계에서는 이 그래프내에서 매칭이 일어난 용어들을 대상으로 유사도를 기준으로 점수를 매긴다. 루트 노드는 이 모든 점수의 합이 되고 정규화 과정을 거치면 각 용어가 가진 점수가 나오게 된다. 마지막 단계에서 이 점수를 바탕으로 어떤 용어를 선택할지

표시를 해주게 된다. 이와 같은 과정을 거치게 되면 어떤 용어로 데이터에 주석을 달 것인가에 대해 사람의 판단에 시간과 노력을 절약해 줄 수 있다.

두번째 DB내의 데이터들을 GO를 이용하여 주석하는데 사용할 수 있는 방법은 문헌정보를 이용하는 것이다. 문헌정보를 이용한다는 것은 생물학과 관련된 다양한 논문들을 검색하여 GO에서 사용되는 용어와 대응되는 논문을 찾는 것을 말한다. GO내에 나타난 개념에 해당하는 논문들을 찾으면 논문의 내용을 통해 데이터베이스를 주석하는데 도움을 얻을 수 있다. 논문에서 언급하고 있는 실험 결과나 데이터등을 통해 해당 물질이 GO내의 어떤 용어에 대응되는지를 알 수 있기 때문이다.

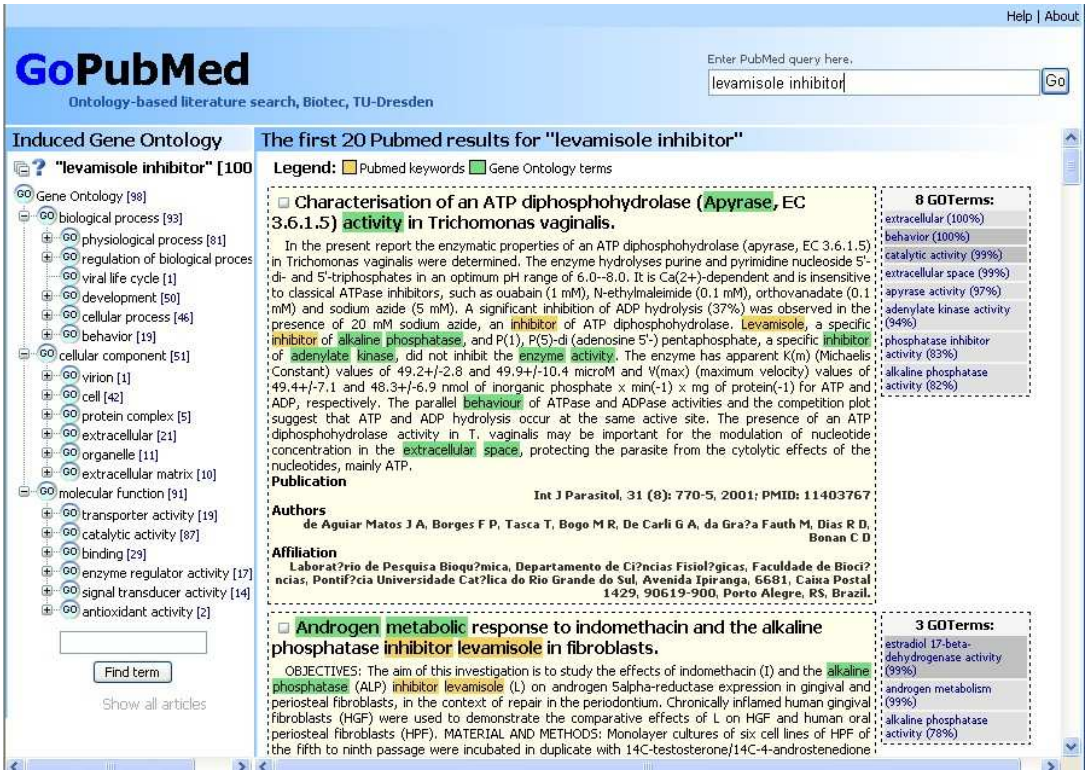

그림 9. GO를 이용한 문헌 정보의 검색

그림 9는 GOPubMed[8]를 이용하여 PubMed [1] 데이터베이스를 검색한 모습니다. GOPubMed는 사용자가 검색하고자 하는 키워드를 입력했을 때 그 결과를 온톨로지의 분류체계에 맞게 제공하여 사용자가 좀 더 쉽게 원하는 문헌을 찾을 수 있도록 하고 문헌에 나타난 정보 사이의 관계를 GO를 이용해 파악할 수 있도록 하고 있다. [9]와 같은 연구는 각 문헌에 나타난 유전자들을 GO의 용어들에 대응시키고 그 정확도를 실제 전문가들이 분류

[1] PubMed: PubMed® 는 의학연구 정보를 담고 있는 MEDLINE® 데이터베이스를 검색하는 시스템이다. MEDLINE 데이터베이스는 4600여 개의 의료,생물학 저널로부터 1200만개가 넘는 참조를 가지고 있다.

한 문헌에 근거에 검증하였다.

GO의 evidence code는 앞서 설명한 것처럼 해당 유전물질을 GO에 대응시키게 된 근거를 표기하는데 여기에는 앞에서 설명한 두 가지 기법인 서열의 유사도 검색과 문헌 정보에 근거한 경우를 표기하도록 하고 있다.

3.3. Experimental analysis

최근 들어 생물학실험에서 개별 유전자를 다루는 실험보다 마이크로어레이와 같은 기술을 통해 대량의 유전자를 한꺼번에 다루는 실험들이 보편화되고 있다. 따라서 실험결과로 나오는 데이터들을 개별적으로 다루는 것보다 전체적인 의미를 파악하는 것이 중요하게 되었다. 기존 연구들은 발현된 유전자들에 관한 수치 정보를 통계적으로 처리하고 클러스터링한후 그 의미를 분석하는 작업을 해왔다. 이러한 연구와는 다른 측면에서 실험 결과를 Gene Ontology를 이용해 분석해 보자는 연구들이 시작되고 있다.

GO는 생물학적인 개념들간의 관계를 표현한 것이기 때문에 GO내에서 유전자들의 분포를 알아내면 해당 유전자들의 대표적인 기능들을 추측할 수 있게 된다. 예를 들어 마이크로어레이 실험의 결과로 발현된 유전자들 수백 개를 GO의 각 용어에 대응시키고 그 분포를 파악한다면 어떤 클러스터가 어떤 기능 또는 생화학적 작용에 관여하는지 예측할 수 있게 된다.

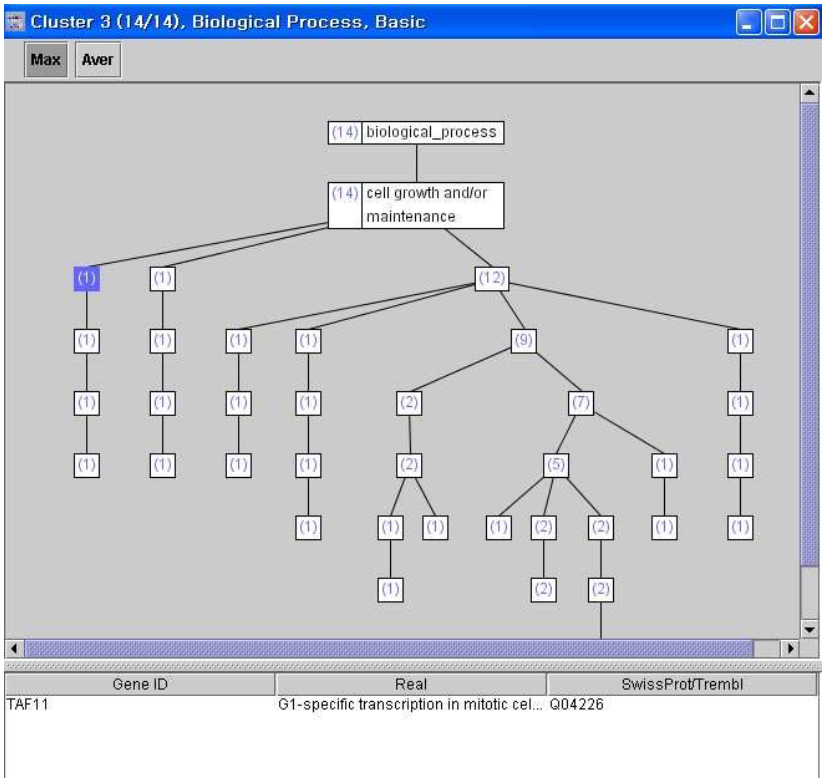

그림 10. Gene Ontology내의 유전자들의 분포

그림 10은 [10]의 연구결과로 얻어진 GOODIES라는 제품의 데모화면이다. 이 데모화면을 보면 마이크로어레이 실험의 결과로 얻어진 효모유전자들이 GO내에 어떻게 분포하는지를 트리로 보여주고 있다. 이와 관련된 연구의 주된 관심사는 GO내의 흩어진 유전자들의 대표 노드를 무엇으로 정할 것인가에 맞추어져 있다. [10]의 경우는 GO를 트리로 보고 유전자가 대응되는 각 노드의 LCA(Least Common Ancestor)를 대표노드로 선택하였다. 그 결과는 기존에 다른 기법을 이용해 클러스터링하고 그 기능이 밝혀진 데이터와 비교하여 GO를 이용한 클러스터링과 대표노드 선택이 타당함을 보였다. 다른 연구들에서는 GO를 그래프로 모델링하거나 대표노드 선택에 있어서 다른 알고리즘을 제안하기도 하였지만 전체적인 틀은 [10]과 같은 패턴을 크게 벗어나지 않는다.

4. 결론

GeneOntology와 관련된 응용분야는 앞에서 언급한 범주를 크게 벗어나지 않을 것으로 보인다. DB분야와 관련 있는 부분은 온톨로지를 이용한 시스템통합이나 문헌정보 검색, 데이터 마이닝 분야 등이 있다. 실제 생물학 실험결과의 의미를 분석하는 것은 어려운 만큼 우리의 역할은 실험을 위해 사용되는 기술을 제공하는 것인데 현재 생물정보학쪽의 연구들은 사용되는 기술보다는 그 결과의 생물학적 의미에 초점이 맞추어져 있다. 따라서 제공되는 기반 기술의 개선이나 정량적 분석이 가능한지를 따져보는 것이 GO관련 응용분야에 접근할 때 우선적인 고려의 대상이라고 생각된다.

<참고문헌>

[1] Gene Ontology Consortium, http://www.geneontology.org/

[2] FlyBase, http://flybase.bio.indiana.edu/

[3] Saccharomyces Genome Databases , http://www.yeastgenome.org/

[4] Mouse Genome Database , http://www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml

[5] AmiGO, Berkeley Drosophila Genome Project, http://www.godatabase.org/dev/

[6] Gene Ontology Tools, http://www.geneontology.org/GO.tools.shtml

[7] Khan S, Situ G, Decker K, Schmidt CJ., “GoFigure: automated Gene Ontology annotation”, Bioinformatics. 2003 Dec 12;19(18):2484-5

[8] Delfs, Doms, Kozlenkov, Schroeder. GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed. German Conference on Bioinformatics 2004: 169-178, Springer Verlag.

[9] Raychaudhuri S, Chang JT, Sutphin PD, Altman RB, “Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature”, Genome Res. 2002 Jan;12(1):203-14.
[10] Lee SG, Hur JU, Kim YS, “A graph-theoretic modeling on GO space for biological interpretation of gene clusters”, Bioinformatics. 2004 Feb 12;20(3):381-8