Pathway Database 통합 활용을 위한 웹 서비스 환경 구축

*이호일 ○ *유성준 **김민경 ***박현석
*세종대학교 컴퓨터공학부 ** 이화여자대학교 공학연구소 *** 이화여자대학교 컴퓨터학과
headil@hanmail.net sjyoo@sejong.ac.kr hspark@macrogen.co.kr minkykim@ewha.ac.kr

Web Service Environment Construction for Pathway Database

*HOIL LEE○ , *Seongjoon Yoo, **Minkyung Kim, ***Hyun Seok Park
*School of Computer Engineering, Sejong University, **Center for Engineering Research, Ewha University, ***Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 pathway 정보의 중요성이 점점 커지고 있다. 하지만 이런 정보를 이용하기에 많은 문제점이 발생하고 있다. 이런 문제점의 해결 방법으로 웹 서비스가 도입되고 있다. 이 논문에서는 주요 pathway 데이터베이스 중 하나인 BIND와 체계적인 개념으로 유전자 용어를 정리한 Gene Ontology(GO)에 대한 웹 서비스를 개발하였다. 개발자들은 이 웹 서비스를 이용하여 BIND와 GO 데이터를 보다 쉽게 이용할 수 있을 것이다.

1. 서 론

바이오인포매틱스 연구에서 정보의 통합은 단순히 정보의 양을 늘릴 뿐만 아니라 서로 다른 유형의 정보를 참조하여 정보의 질을 향상시키고 새로운 발견을 가능하게 한다. 기존의 바이오 데이터베이스를 통합하는 방법으로는 링크 기반, 뷰 기반, 데이터웨어하우스 기반 방법 등이 활용되었다. 그러나 이러한 통합 방법은 여러 데이터베이스에서 데이터를 추출/저장하기 위해 유사한 스크립트를 작성하는 중복된 노력이 요구되고, 유지/관리가 어려우며, 시스템에 의존적인 문제점들이 있다. 최근에는 이러한 문제를 극복하기 위하여 바이오인포매틱스 자원 통합 분야에도 웹 서비스 기술을 도입하고 있다.[1]

생물 정보학의 연구 분야는 크게 다섯 가지로 분류할 수 있는데, 그 첫째는 유전자 단백질의 서열을 다루는 분야, 두 째는 유전자와 단백질의 발현을 다루는 분야, 세 째는 단백질의 구조를 주제로 하는 분야, 네 째는 유전자, 단백질의 상호작용 및 네트웍을 다루는 분야, 마지막으로는 이러한 사실을 기초로하여 총체적으로 기능을 유추하려는 분야 등이다. 상호작용 및 pathway/network 데이터베이스는 systems biology를 위한 기본적인 자료들로 활용된다. 즉, 단백질, 유전자 등 생물체를 이루는 각각의 엔티티들에 대한 정보 못지 않게 그들의 연관성, 상호작용 등이 각종 질병, 약의 작용 등을 이해하는 데 중요하다. 그럼에도 불구하고, 이에 대한 정보를 통합 관리하는 데이터베이스가 존재하지 않으며 따라서 기존의 서로 다른 형식으로 저장된 데이터베이스들을 통합하고 기존의 개별적인 서열 정보등과도 연결해주는 시스템이 필요할 것으로 보인다.

본 연구에서는 pathway 관련 정보를 통합하고자 함에 있어 웹 서비스 방식을 적용하여 pathway/network 정보를 찾기 위해 관련된 데이터베이스를 통합 검색할 때 필요한 기능이 무엇인지 알아본다. 그리고, 아직 웹 서비스 기능을 제공하고 있지 않은 BIND[2]와 Gene Ontology[3] 데이터베이스에 대해 Pathway 데이터베이스 활용에 필요한 기능을 정의하고 이를 설계,구현하도록 한다.

2. 관련연구동향

웹 서비스 기술을 이용하여 바이오 데이터를 서비스하는 시스템으로는 일본의 KEGG[4]와 DDBJ(DNA Data Bank of Japan)[5], 캐나다의 BioMOBY[6] 등이 있다. 이 중에서 Pathway 연구에 직접적으로 필요한 것은 KEGG 데이터베이스이다. 기타의 시스템은 웹 서비스를 바이오인포매틱스 시스템에 도입한 방법을 연구하기 위하여 함께 검토하도록 한다.

2.1 KEGG

KEGG는 기능 유전체학 연구를 위한 대사과정 비교, 대사과정 재구성 및 대사과정 설계 등과 같은 새로운 생물정보학 기술을 개발하기 위한 데이터베이스를 가지고 있고, 2003년부터 웹 서비스로 일부 데이터를 서비스하고 있다. KEGG의 웹 서비스는 크게 SSDB(Sequence Similarity DataBase), PATHWAY, GENES, KEGG로 구분한다. SSDB, PATHWAY, GENES 는 각각 KEGG의 SSDB, PATHWAY, GENES 데이터베이스들을 위한 웹 서비스이고, KEGG는 버전 등과 같은 데이터베이스에 대한 정보를 제공하는 웹 서비스이다. [4]

KEGG는 pathway에 관한 대량의 정보를 가지고 있지만, 현재의 웹 서비스 API 만으로 모든 정보를 이용하는데 제한적이다. 예를 들면, 유전자에 대한 NCBI's Gene ID(gi)를 가져오는 서비스가 없다.

2.2 DDBJ

DDBJ는 미국의 NCBI(National Center for Biotechnology Information)[7], 유럽의 EMBL(European Molecular Biology Laboratory)[8]과 같이 DNA sequence 정보를 제공해 주는 곳이다. DDBJ의 웹 서비스는 Blast, ClustalW, DDBJ, ExClustalW, Fasta, GetEntry, Gib, Gtop, PML, SRS,

TxSearch로 구분하여 서비스하고, Blast, ClustalW, 등과 같은 여러가지 툴을 웹 서비스를 통해 이용 가능하도록 한다.[5] 이 서비스는 Pathway 연구에 직접적인 관계는 없지만, 유전자의 서열 정보를 비교 가능하게 한다.

2.3 BioMOBY

BioMOBY는 바이오 데이터를 Ontology개념을 적용한 웹 서비스를 제공한다. MOBY 시스템은 Object, Central, Server, Client로 구성되어 있다. MOBY-Central은 서비스 레지스트리 역할을 한다. 즉 서비스 제공자는 서비스 이름, 서비스 타입, 입/출력 Obejct, 서비스 제공자의 URI, 서비스 스크립트의 URL, 서비스 설명 등의 정보를 Central에 등록한다. 서비스 제공자를 역할을 하는 MOBY-Server는 제공할 서비스를 Central에 등록한다. MOBY-Client는 Central에서 원하는 서비스를 검색하여 서비스 제공자에게 서비스를 요청 한다. 클라이언트와 서버 사이에서는 Object라는 구조화된 데이터를 주고 받는다. [9]

2.5 관련연구분석

이상에서 볼 수 있듯이 Pathway 데이터베이스를 웹 서비스 기반으로 제공하는 곳은 KEGG가 유일하다. KEGG 웹 서비스의 문제점은 다른 데이터베이스와 연결이 어렵다. 바이오 데이터들은 서로 정보를 참조하여 정보의 질을 향상시키고 새로운 발견을 가능하게 해야 하는데, 다른 데이터베이스를 참조하기 위한 서비스가 없다.

Pathway 관련 연구에 필요한 바이오 데이터를 웹 서비스로 제공하는 곳이 부족하다. 특히 pathway 연구에 중요한 데이터베이스인 BIND와 유전자를 체계적으로 분류한 GO의 웹 서비스는 아직 개발되지 않고 있다.

바이오 데이터 특성에 알맞은 표준화된 웹 서비스 API 구조가 없다. 따라서 같은 특성의 정보들도 제공자마다 조금씩 다른 구조로 웹 서비스를 제공하고 있다.

BioMOBY는 Server, Client, Central의 구조를 가져 일반적인 웹 서비스 구조와 일치하지만, 사설 레지스트리 사용으로 인해 서비스를 검색 및 이용이 어렵다.

마지막으로, 충분한 서비스를 제공하지 못한다. 서비스 제공자가 보유하고 있는 정보는 많지만, 서비스 요청자가 원하는 모든 서비스를 제공하지 못하는 상태이다.

3. Pathway 데이터베이스 통합을 위한 웹 서비스

앞에서 웹 서비스 기반의 pathway 데이터베이스는 KEGG 뿐이다. Pathway에 대한 연구를 보다 포괄적으로 하고자 한다면 BIND와 GO에 대한 내용도 포함해야 하며 따라서 이들 데이터베이스 활용을 위한 웹 서비스를 개발할 필요가 있다고 언급한 바 있다. 이장에서는 BIND와 GO의 데이터 구조를 검토하고 통합을 위한 웹 서비스 기능을 구현하는 것에 대하여 기술한다.

3.1 데이터 모델

BIND는 생체 분자들의 상호작용과 Pathway를 다루는 데이터베이스이고, 데이터를 Interaction, Molecular Complex, Pathway 형태로 분류한다. Interaction은 자신을 구성하는 개체들의 상호작용 정보와 실험 조건, PubMed 등의 정보를 포함하고 있다. 개체는 protein, DNA, RNA, ligand, molecular complex, gene, photon, unclassified billogical(?) 이 될 수 있다. Complex는 2개 이상의 Interaction들이 결합되어 특정 기능을 수행하는 집합체를 말한다. Complex 데이터는 자신을 구성하는 개체들과 Interaction들의 정보를 가지고 있으며, Complex에 대한 설명, PubMed 정보등을 포함하고 있다. Pathway는 2개 이상의 Interaction들을 규정된 순서대로 연결한 생체작용경로를 말한다. Pathway는 자신을 구성하는 여러 개의 Interaction에 대한 정보와 Pathway에 대한 설명, PubMed 등의 정보를 포함하고 있다.[9]

GO는 체계화된 언어로 다양한 생명현상을 정의하고 그 기능을 수행하는 유전자를 Gene Ontology 정의에 따라 Molecular Function, Biological process, Cellular Component로 구분하여 정의한다. Molecular Function은 유전자가 세포내에서 수행하는 분자수준의 기능에 따라 분류하고자 한다. Cellular Component는 cell의 구성원(component)을 말하는 것으로써 anatomical structure에 대한 명명이 주가 될 것이다. GO는 이러한 데이터를 관계형 데이터베이스로 제공한다.[10]

3.2 BIND 웹 서비스

[표 1]은 BIND의 웹 서비스에 대한 설명을 보여준다. BIND의 웹 서비스는 Bind_InteractionIF, Bind_ComplexIF, Bind_PathwayIF로 크게 3개로 구분한다. Bind_InteractionIF는 Interaction을 위한 웹 서비스 API를 정의하고, Bind_ComplexIF는 Complex를 위한 웹 서비스 API를 정의하며 Bind_PathwayIF는 Pathway를 위한 웹 서비스 API를 정의한다.

표 1. BIND 웹 서비스

객체 이름	객체 설명
Bind_InteractionIF	BIND에서 Interaction 관련 서비스 예) get_InteractionByInterId()
Bind_ComplexIF	BIND에서 Complex 관련 서비스 예) get_Sub_ObjectByComplexId()
Bind_PathwayIF	BIND에서 Pathway 관련 서비스 예) get_Sub_InteractionByPathwayId()

누구든지 웹 서비스를 이용하여 BIND 데이터베이스를 쉽게 이용 할 수 있다.

3.3 GO 웹 서비스

GO 웹 서비스는 임의의 유전자가 GO에서 분류한 계층에서 어디에 속하는지 구분하고, 서로 다른 유전자를 비교 가능해야 한다. 즉 서로 다른 유전자의 정보를 비교하여 연구에 활용 할 수 있다.

[표 2]는 GO 웹 서비스에 대한 설명을 보여준다. Gene_OntologyIF는 Gene Ontology 데이터 베이스를 위한 웹 서비스 API를 정의한다.

표 2. GO 웹 서비스

객체 이름	객체 설명
Gene_OntologyIF	Gene Ontology 관련 서비스 예) get_AncestorByTerm(Term)

이 웹 서비스를 이용한다면 누구든지 쉽게 Gene Ontology 데이터를 이용할 수 있다.

4. Pathway 데이터베이스 통합 활용

[그림 1]은 웹 서비스 기반의 Pathway 데이터베이스 통합 활용의 간단한 시스템 구조이다. Client가 Pathway 데이터베이스들을 통합하여 활용하고자 한다면, Pathway를 구성하는 유전자를 서비스 제공자에게 요청 가능해야 한다. KEGG와 BIND의 웹 서비스는 이와 같은 기능을 제공한다. 하지만 BIND에서는 유전자를 구분하는 키로 NCBI의 gi와 이름을 결합하여 사용하지만, KEGG에서는 그들이 부여한 키를 사용하여서 KEGG와 BIND를 통합 하는데 어려움이 있다. 하지만 KEGG의 데이터베이스에는 NCBI의 gi 정보를 가지고 있으므로 곳 이러한 문제는 해결 될 것으로 예상된다.

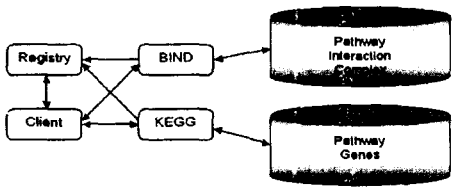

그림 1. Pathway 데이터베이스 통합 활용 시스템 구조

우리가 구현한 BIND의 웹 서비스를 이용하여 [표 3]에서는 pathway ID가 '10'인 pathway를 구성하는 interaction ID를 보여주고, [표 4]에서는 Interaction을 구성하는 개체에 대한 정보를 보여준다. 이와 같이 다른 pathway에 대해서도 Pathway를 구성하는 유전자 리스트를 쉽게 가져 올 수 있다.

표 3. pathway를 구성하는 Interaction 검색 결과

PATHWAY_ID	INTER_ID
10	116
10	118
10	145
10	148
10	167
10	1444
10	1448
10	1451

표 4. Interaction을 구성하는 개체 검색 결과

INTER_ID	A_LABEL	A_ID	B_LABEL	B_ID
116	EGF	4503491	EGFR	4885199
118	EGF-EGFR	12966	EGF-EGFR	12966
145	(EGF-EGFR) dimer complex	12970	ATP	
148	(EGF-EGFR) dimer-ATP	12971	(EGF-EGFR) dimer complex	12970
167	(EGF-EGFR) dimer P-Tyr1068	12970	Grb2-Sos1	12973
1444	Grb2-Sos	12973	H-Ras-GDP	13012
1448	Grb2-Sos-H-Ras complex	13013	GTP	
1451	Ras-GTP	13014	Raf	4506401

그리고 KEGG의 웹 서비스를 이용하여 [표 5]에서는 pathway ID가 'path:eco00100'인 pathway를 구성하는 유전자를 보여준다.

표 5. pathway를 구성하는 Interaction 검색 결과

path:eco00100	eco:b2515
path:eco00100	eco:b2746
path:eco00100	eco:b2747
path:eco00100	eco:b1208
path:eco00100	eco:b0173
path:eco00100	eco:b0420
path:eco00100	eco:b0347
path:eco00100	eco:b2889
path:eco00100	eco:b0421

이와 같이, 다른 데이터베이스에 존재하는 pathway를 비교하기 위해서, pathway를 구성하는 유전자를 찾고, 그 유전자들을 비교함으로써 pathway들 사이의 관계를 예측 할 수 있다. 이와 같이 데이터베이스의 통합은 정보의 질을 향상시키고 새로운 발견을 가능하게 한다.

5. 결론 및 향후 과제

웹 서비스 API가 개발되지 않은 중요한 데이터베이스인 BIND와 Gene Ontology를 KEGG의 웹 서비스를 참고하여 웹 서비스 API를 개발 하였다. 이 연구에서 개발한 웹 서비스는 이미 웹 서비스를 개발한 곳에서 발생하는 문제점을 보완하여 개발하였으므로, 이후에 BIND의 데이터를 사용하고자 하는 개발자들은 쉽게 웹 서비스를 이용하여 개발 시간과 노력을 많이 줄일 수 있을 것이다. 하지만 아직도 Pathway에 관련된 많은 데이터베이스가 웹 서비스로 데이터를 제공하지 않고 있다. 이러한 데이터베이스들도 웹 서비스로 데이터를 제공해야하고, 서로 다른 데이터베이스를 참조하기 위한 서비스도 개발하여, 쉽게 웹 서비스 기반의 바이오 데이터 통합이 가능하도록 해야 한다.

6. 참고문헌

[1] Lincoln D. Stein, INTEGRATING BIOLOGICAL DATABASES, NATURE REVIEWS:GENETICS, 337-345, MAY 2003

[2] www.bind.ca

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

[4] http://www.kegg.com

[5] http://xml.nig.ac.jp

[6] http://www.biomoby.org

[7] http://www.ncbi.nlm.nih.gov

[8] http://www.ebi.ac.uk

[9] Bader GD & Hogue CW, BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics , 465-477, MARCH 2000

[10] http://www.godatabase.org