2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

이질형 바이오 데이터베이스 통합을 위한 개체-관련성 모델링

정진희*, 이도헌**
전남대학교 컴퓨터 정보학부

e-mail:{jhjung,dhlee}@dbcore.chonnam.ac.kr

Entity-Relationship Modeling for Integrating Heterogeneous Bio-databases

Jin-Hee Jung*, Doheon Lee**
School of Computer and Information, Chonnam National University

요약

유전체 연구를 위해 구축된 바이오 데이터베이스는 해당 프로젝트의 목적에 따라 서로 다른 주체에 의해 독립적으로 구축되어 왔다. 그러나 바이오 데이터의 효과적인 활용을 위해서는 그러한 이질적인 바이오 데이터베이스의 정보를 상호 연계하여 분석할 필요성이 높아지고 있다. 본 논문에서는 대표적인 핵산 데이터베이스인 GenBank와 단백질 데이터베이스인 SWISS-PROT, 문헌 데이터베이스인 PubMed의 데이터 구조를 개체-관련성 도표로 각각 모델링한 후 합병하여, 핵산-단백질-문헌자료로 연계되는 정보를 통합 서비스할 수 있는 모델과 시스템 구조를 제시한다.

1. 서론

유전체 연구를 위해 구축된 바이오 데이터베이스는 해당 프로젝트의 목적에 따라 서로 다른 주체에 의해 독립적으로 구축되어 왔다. 그러나 급속도로 증가하는 데이터베이스들은 수많은 각기 다른 형태의 서열 데이터를 저장하므로, 바이오 데이터의 효과적인 활용을 위해서는 그러한 이질적인 바이오 데이터베이스의 정보를 상호 연계하여 분석할 필요성이 높아지고 있다[1].

일반적으로 데이터의 통합은 정보가 분산되어 있고, 저장구조가 다르다는 등의 이유로 어려운 일로 알려져 있다. 이러한 통합 작업을 하기 위해서 개념적 설계의 과정이 필요한데, 이는 요구사항 명세로부터 시작해서 데이터베이스의 개념적 스키마를 산출한다. 개념적 모델이란 개념적 스키마를 기술하는데 사용하는 언어라고 할 수 있다. 이러한 과정은 설계자와 사용자 간의 의사소통이 활발히 이루어지고, 초기 모델링으로부터 오류도 쉽게 발견할 수 있으며, 언제든지 확장이 가능하다는 등의 장점을 지니고 있다. 특히 최근에는 생물 데이터들을 위한 개념적 혹은 객체기반의 모델들이 다양하게 제시되고 있다. 예를들어, Extended Entity-Relationship(EER) 모델을 사용하여 DNA 데이터베이스의 개념적 스키마를 보인 경우가 있고[2], 유전자 정보를 UML (Unified Modeling Language)을 사용하여 개념적 데이터 모델을 제시한 예가 있다[3]. 또 데이터베이스 통합을 쉽게 하도록 Entity-Attribute-Value (EAV)모델을 제안한 경우도 있다[4].

본 논문에서는 바이오 데이터베이스 통합을 위한 개념적 모델링을 Chen의 Entity-Relationship모델[5]을 사용하여 제안하고, 구현중인 Bio Gateway시스템의 구조를 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 바이오 데이터베이스의 분석 및 각각의 ER모델을 제시한다. 3장에서는 통합된 ER 모델과 구현중인 시스템의 구조를 제시한다. 4장에서는 본 논문에 대한 결론 및 향후 연구방향을 제시한다.

2. 바이오 데이터베이스의 분석 및 ER 모델링

GenBank, SWISS-PROT, PubMed의 데이터베이스를 분석하고 각각의 ER 모델을 보인다. 모델은 우선 엔터티를 추출하고 엔터티를 설명해주는 속성, 엔터티 간의 관련성을 나타내는 관계로 표현한다.

2.1 GenBank

GenBank는 NCBI(National Center Biotechnology Information)에서 운영중인 염기서열 데이터베이스로서 최신의 포괄적인 DNA 서열 정보를 제공한다.

2.1.1 GenBank data

GenBank는 233개의 파일로 구성된 flat file형태이다. 파일은 헤더 정보와 서열 엔트리정보를 가지고 있다. 표1은 서열 엔트리의 일부를 나타낸 표이다.

표1.GenBank 데이터

Locus	엔트리 이름
Definition	서열의 간단한 정보 설명
Accession	유일 식별자
Version	복합 accession number+GI(GenInfo)
Source	생물유기체의 이름
References	참조,인용
Features	서열의 부분 정보를 포함하고 있는 테이블

2.1.2 GenBank의 ER 모델

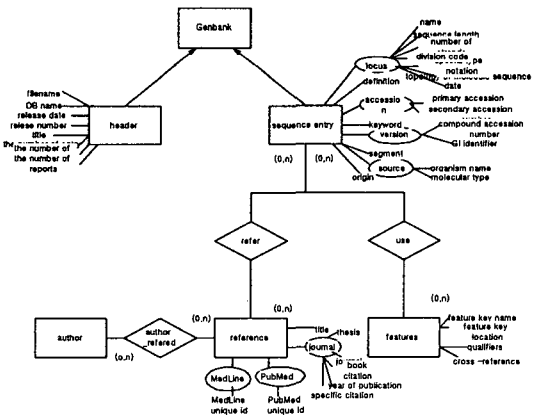

그림1.GenBank의 ER 모델

2.2 SWISS-PROT

SWISS-PROT는 스위스의 SIB(Swiss Institute of Bioinformatics)에서 운영 중인 단백질 서열 데이터베이스로 다른 데이터베이스과 비교해서 세가지 장점을 지니고 있다. 첫째, 쉽게 이해 할수 있게 주석을 제공하고 둘째, 데이터베이스 사이의 중복을 최소화 시키며 셋째, 다른 데이터베이스와 통합이 쉽다

2.2.1 SWISS-PROT data

SWISS-PROT 데이터 베이스 역시 GenBank의 경우처럼 서열 엔트리들로 구성되고 서열 엔트리들은 서열에 대한 여러 정보를 가지고 있다. 표2는 서열 엔트리의 일부를 나타낸 표이다.

표2.SWISS-PROT 데이터

ID	식별자
AC	accession 번호
GN	염색체 이름
OS	생물 유기체의 종
OG	세포기관
RN/RP/RC/RX/RA/RT/RL	참조,인용
DR	데이터베이스 교차참조
FT	feature 테이블 데이터
SQ	서열 헤더

2.2.2 SWISS-PROT 데이터베이스의 ER 모델

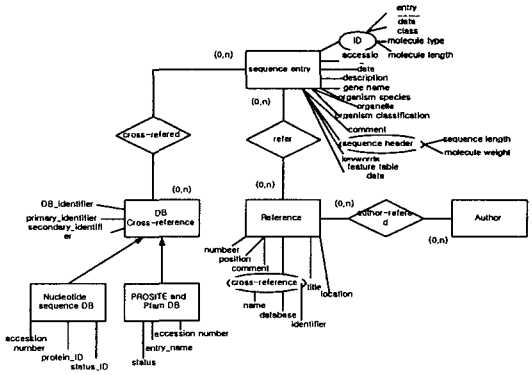

그림2.Swiss-Prot의 ER 모델

2.3 PubMed

문헌 검색에서 이용하는 PubMed는 MEDLINE에 있는 1,000 만개의 reference와 초록을 포함하고 있으며, 웹에서 이용할 수 있는 500여개의 저널들에 대한 링크도 제공하는 문헌 데이터베이스이다.

2.3.1 PubMed data

PubMed 데이터는 위의 두 데이터와 다른 형태로써, 웹기반의 정보 소스 즉, 검색에 필요한 저자,저널 제목,발행날짜 등의 참고문헌 정보를 포함한다.

2.3.2 PubMed 데이터 베이스의 ER 모델

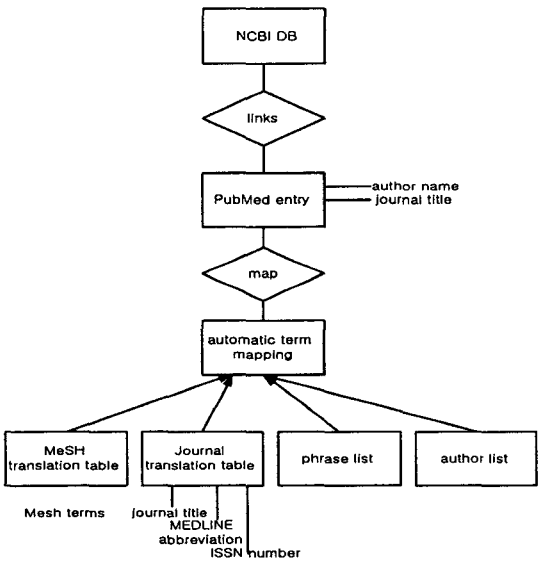

그림3.PubMed 데이터베이스의 ER 모델

3.바이오 데이터베이스의 통합

데이터베이스 통합 작업의 주목적은 실세계의 동일한 부분을 나타내는 모든 개념적 입력 스키마들을 찾아서 그들의 표현을 통일화하는 것으로 스키마의 통합이라고 한다. 스키마의 통합에 있어 가장 어려운 점은 합병할 스키마들의 차이점을 발견하는 것이다. 그림4는 스키마 통합의 접근 방법을 표현한 것이다. 충돌분석을 하는 동안에는 충돌을 찾는데 주의를 기울이게 되는데 이는 충돌의 조기 발견이 매우 중요하기 때문이다. 문제 해결 동안에는 한쪽 또는 양쪽 모두의 스키마에 각각의 충돌을 해결하거나 제거하기 위하여 수정이 가해진다. 스키마를 합병하는 동안 스키마들이 덧붙여지고 예비 통합 스키마를 얻게 된다.

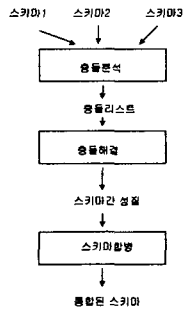

그림4.뷰통합의 접근 방법

3.1 통합 데이터 베이스의 ER 모델

그림4의 절차에 따라 실제로 위의 3개의 데이터베이스를 통합하는 모델을 구축하였다.

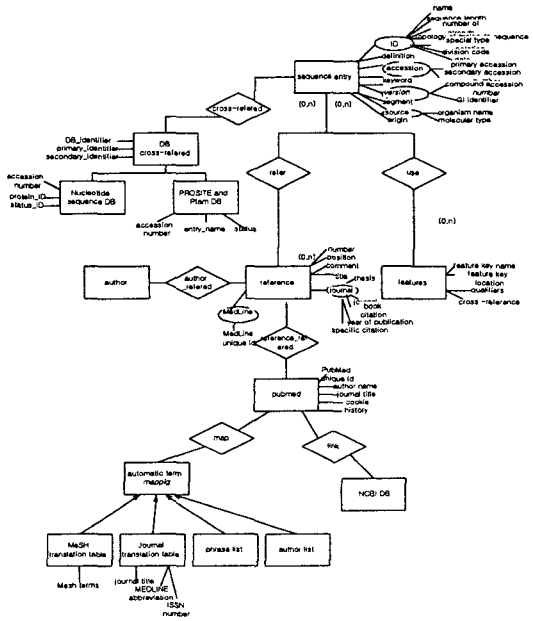

그림5.통합된 ER 모델

3.2 시스템 구조

그림5 는 현재 구현 중인 Bio Gateway시스템의 구조를 제시한 것이다. 이 시스템은 핵산-단백질-문헌

자료로 연계되는 정보를 통합 서비스 할수 있는 구조이다. 간단히 각각의 역할을 살펴 보면 다음과 같다. Metadata repository는 데이터 자원내의 데이터 요소들에 대해 정의한 집합 즉, 메타데이터를 관리하는 지속적인 저장공간으로 사용하고, 임시저장공간으로 캐쉬를 사용한다. 에이전트는 사용자를 대신하여 사용자의 작업을 수행하는 소프트웨어이며 대리인 개념으로 사용자의 행동 양식을 관찰하고 학습하여 정보공간에서 사용자를 대표하고 학습된 사용자의 행동 양식을 기반으로 사용자가 해야 할 작업을 자동으로 수행해 주는 소프트웨어이다. 통합 질의프로세서(IQP)는 질의가 발생했을 때 처리 방법을 찾고 그에 따라 요구하는 결과를 사용자에게 보여주는 일을 담당한다.

Bio Gateway 시스템은 웹로봇이 지능적으로 캐싱 작업을 해준다는 장점을 지닌다.

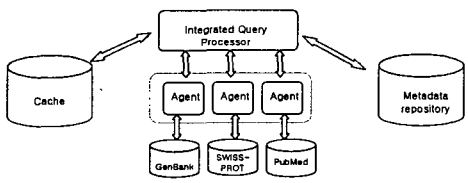

그림6.Bio Gateway 시스템 구조

4.결론

본 논문은 바이오 데이터베이스들을 통합하기 위한 개념적 모델을 제시하였다. 이질적이고 복잡한 특성을 지닌 여러 데이터베이스를 통합하는 것은 어렵고도 중요한 일이며, 개념적 설계를 필요로 한다. 개념적 모델은 개념적 스키마를 기술하는데 사용하는 하나의 언어로서 여기서는 가장 광범위하게 사용되는 Chen의 ER 다이어 그램을 사용하였다. 또 현재 구현중인 Bio Gateway시스템의 구조를 간단히 보였다. 이 시스템은 핵산-단백질-문헌자료로 연계되는 정보를 통합 서비스 할수 있는 모델이다.

현재 제안한 방법에 따라 시스템을 구현 중이며, 향후 연구로는 시스템의 성능 향상을 고려한 새로운 시스템을 설계하려고 한다.

참고 문헌

[1] Andreas D.Baxevanis "The Molecular Biology Database Collection:an online compilation of relevant database resources"Oxford University Press 2000

[2]Okayama,T.,Tamura,T.,Gojobori,T.,Ikeo,K., Miyazaki,S.,Fukami-Kobayashi,K.and Sugawara,H."Formaldesign and implementation of an improved DDBJ database with a new object-oriented library." Bioinformatics.

[3] Norman W. Paton, Shakeel A. Khan, Andrew Hayes, Fouzia Moussouni, Andy Brass, Karen Eilbeck, Carole A. Goble, Simon J. Hubbard, and Stephen G. Oliver "Conceptual modelling of genomic information"Bioinformatics 2000

[4] KH Cheung, PM Nadkarni, and DG Shin "A metadata approach to query interoperation between molecular biology databases" Bioinformatics 1998

[5] Chen,P.S."TheEntity-RelationshipModel:Toward a Unified View of Data".ACM trans.Database sys.

[6] NCBI-GenBank Flat File Release 124.0. ,2001 http://ncbi.nlm.nih.gov/genbank

[7] The SWISS-PROT Protein Sequence Database User Manual. , 2000 http://www.expasy.ch/sprot

[8] NCBI PubMed Manual. , 2001 http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html