# 이질형 바이오 데이터베이스 통합을 위한 게이트웨이 시스템

정진희* · 정민아

## Bio-Gateway System Architecture for Integrating Heterogeneous Bio-Databases

Jinhee Jung* · Mina Jung

요 약

이질적인 생물 데이터베이스의 통합은 데이터간의 연계 분석의 필요성이 높아짐에 따라 중요한 문제로 대두되고 있다. 그러나 이러한 데이터베이스들은 초기에 이질적 환경에서 각기 다른 목적에 의해 생성되므로 포맷, 설계자가 불일치하는 등 여러 가지 문제점으로 인해 통합하는데 어려움이 따른다. 그러므로 이질적인 데이터베이스의 통합을 위해서는 초기단계의 설계가 무엇보다도 중요하다. 본 논문에서는 대표적인 핵산 데이터베이스인 Genbank와 단백질 데이터베이스인 Swiss-Prot을 통합하기 위해 ER 모델을 사용하여 개념적 모델을 보인 후, 이를 합병하여 통합모델을 제시한다. 또한, 핵산-단백질 자료로 연계되는 정보를 통합 서비스할 수 있는 시스템 구조를 제안한다. 제안된 바이오 게이트웨이 시스템은 개념적 설계 단계에서 가장 원자적인 단위로 분할하여 모델링 함으로써 정교한 질의 처리가 가능하고, 사용자가 상세 조건을 알고 있을 경우에 기존의 검색시스템과 달리 여러 번의 검색 과정을 거치지 않고, 단시간에 원하는 결과를 얻을 수 있다는 장점을 지닌다.

ABSTRACT

The Integration of biological databases is critically important because of the interconnectedness of biological research. But it's not easy to integrate these databases for the different formats and designers in heterogeneous environments. So initial design is indispensable to integrate heterogeneous databases. In this paper, after we performed conceptual modeling on a popular nucleotide database, GenBank and a protein database, Swiss-Prot and integrated them by considering cross-reference. we also propose the integration system architecture called Bio-Gateway System, which can help users query closely linked information between two biological databases within one system differently from existing systems as well as query easily on condition that user knows fine conditions for less effort.

키워드

바이오 데이터베이스 통합, ER 모델, 개념적 설계

## Ⅰ. 서 론

유전자 프로젝트를 기반으로 수많은 서열 데이터베이스가 생성되었고 현재도 기하급수적으로 늘어나고 있다[1]. 이러한 데이터베이스들은 구조가 복잡하고 이질적 성격을 가지며, 또한 데이터양이 증가함에 따라

* 한국생명공학연구원 국가유전체 정보센터, 목포대학교

접수일자 : 2005. 6. 2

이를 저장 및 분석하고 정보를 표현하는데도 많은 어려움이 따른다. 그럼에도 불구하고 이러한 정보들을 연계하여 분석할 필요성이 높아짐에 따라 통합하려는 연구도 활발히 진행되고 있다[1][2].

생물 데이터의 특성을 고려할 때 통합작업은 결코 쉬운 일이 아니다. 그러므로 통합 작업을 하기 위해서 초기의 설계 과정이 특히 중요한데, 이는 요구사항 명세로부터 데이터베이스의 개념적 스키마를 산출하는 과정이다. 이러한 초기 설계 단계를 통하여 설계자와 사용자간의 의사소통이 활발히 이루어지고, 초기 모델링으로부터 오류도 쉽게 발견할 수 있으며, 언제든지 확장이 가능하다는 등의 장점들이 있다. 이러한 이유로 인해 생물적 데이터의 구조를 개념적 설계를 통해 보인 예들이 많이 있다. 예로, Extended Entity-Relationship(EER)모델을 사용하여 DNA 데이터베이스의 개념적 스키마를 제시하거나[3], 유전자 정보를 UML을 통하여 제시한 경우[4], ER모델을 확장한 EAV(Entity Attribute Value) 모델을 제안하여 생물 데이터를 모델링한 경우가 있다[5].

이러한 모델링을 기반으로 생물정보를 통합하여 정보를 제공할 수 있는 통합 시스템에 관련된 연구도 활발히 진행되고 있다. 대표적인 시스템으로 TAMBIS[6]], SRS[7][8], BioKleisli[9] 등이 있고, 여러 가지 문제점들로 인해 지금도 연구가 활발히 진행 중이다[10][11][12]13]. 본 논문에서는 바이오 데이터베이스 통합을 위한 초기의 개념적 모델로 Chen의 Entity-Relationship모델을 사용하여 데이터의 구조를 원자적 단위까지 표현하고, 이를 바탕으로 실제 정보를 처리하고 검색할 수 있는 바이오 게이트웨이 시스템의 구조를 제안한다. 제안한 시스템을 통해 사용자가 검색 조건을 알고 있는 경우에 기존의 검색시스템에 비해 빠르고 정확하게 결과를 얻을 수 있다.

## Ⅱ. 관련연구

### 2.1 Entity-Relationship 모델

개념적 설계의 대표적인 방법으로 실세계를 개체와 그것들의 관계를 통해 표현한다. ER 모델은 관계형 모델로 매핑이 쉽고, 단순하고 이해가 용이하여 설계 단계에서 설계자와 사용자간의 의사소통을 하기에 적합하다.

### 2.2 데이터베이스 통합 접근 방법

2.2.1 데이터 웨어하우스

데이터 웨어하우스는 질의나 데이터 분석 작업등을 지원하기 위한 통합된 정보의 대용량 저장소이다. 데이터 웨어하우스에 저장된 데이터는 여러 지역 데이터베이스들에 존재하는 데이터 중 특정 조건에 만족하는 데이터를 추출하여 별도의 공간에 저장한 실체 뷰(materialized view)로 간주된다. 따라서 사용자의 질의에 대한 정확한 질의 결과를 제공하기 위해서는 하위 정보 자원들의 데이터 변경에 따라 데이터 웨어하우스 내의 실체 뷰를 일관성 있게 유지해야 한다.

2.2.2 연합 데이터베이스

연합 데이터베이스는 다수 개의 데이터베이스를 통합하는 방법 중 가장 단순한 방법으로서 모든 정보 자원 또는 데이터베이스 간에 일대 일의 연결 고리를 가지고 있는 형식을 취한다. 따라서 전체 데이터베이스의 연결을 위해서 연결을 위한 코드를 작성해야 하고 각각의 코드가 데이터베이스에 의존적으로 작성되어 있기 때문에 재사용성이 떨어진다는 단점이 있다.

2.2.3 미디에이터

미디에이터는 어떠한 모듈로부터 원하는 정보를 추출하여 상위 계층의 모듈에서 그 정보를 사용할 수 있도록 도와주는 소프트웨어 모듈이다. 미디에이터는 데이터 웨어하우스와 같이 통합된 데이터를 저장하지 않으므로 일관성에 대한 고려를 할 필요가 없으며, 멀티 데이터베이스와는 달리 통합의 대상이 DBMS만으로 한정되어 있지 않다. 또한 유사한 스키마 뿐만 아니라 서로 상이한 스키마의 통합이 용이하며, 특히 기존의 인터넷 응용과 같이 스키마가 명시적으로 결정되지 않은 정보 자원과 데이터베이스와 같이 정형화된 스키마를 가지는 정보 자원을 서로 통합하는데 용이한 특성을 가진다.

## Ⅲ. 바이오 데이터베이스의 분석 및 ER 모델링

생물 데이터는 자치적, 이질적이면서, 급변한다는 자체의 특성으로 인해 통합 과정에 특히 어려움이 따

른다.

이 장에서는 GenBank, Swiss-Prot 서열 데이터베이스를 각각 분석하고 초기의 개념적 설계 산출물로 ER 모델을 보인다. 모델은 우선 엔터티를 추출하고 엔터티를 설명해주는 속성, 엔터티 간의 관련성을 나타내는 관계로 표현한다. 여기서 표현되는 모델들은 시스템으로 구현될 때 정교한 질의처리를 가능하게 하도록 원자적인 단위까지의 모델을 포함한다는 특징이 있다. 좀 더 자세한 내용은 세부 절에서 보인다.

### 3.1 GenBank 데이터베이스와 ER 모델

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

또한, GenBank는 현재 769개의 플랫 파일로 구성되었으며, 파일은 헤더 정보와 서열 엔트리 정보로 구성된다. 표 1은 엔트리 필드의 정보를 간단히 기술한 것이다.

그림 1은 GenBank 플랫파일의 일부와 그 부분의 모델링을 예로 보인 것으로 플랫파일에서의 Features는 key name과 그에 따른 location과 qualifier로 구성된다.

표 1. GenBank 데이터 필드
Table 1. GenBank data fields

| Line Code | contents |
|---|---|
| Locus | short mnemonic name |
| Definition | concise description |
| Accession | unique, unchanging code |
| Version | compound identifier |
| keyword | short phases describing gene products |
| source | common name of organism |
| organism | formal scientific name of organism |
| reference | citations for all articles |
| authors | list of authors |
| medline | medline unique identifier |
| pubmed | pubmed unique identifier |
| remark | the relevance of citation |
| comment | cross-reference |
| features | table containing information |
| origin | specification of sequence location within genome sequence data |
| (blanks) | sequence data |
| // | entry termination symbol |

실제 개체-관련성 모델에서는 features부분에 key name과 location을 속성으로 넣고 location부분은 복합속성을 갖게 함으로써 더 세분화 하였다. 또 하나의 속성인 qualifier부분을 분리해 name과 value의 속성을 주었다. 이와 같은 과정은 질의처리에 있어서 손실될 가능성이 있는 유용한 정보를 DBMS에 저장하고 기존의 시스템과 달리 사용자가 검색 조건들을 알고 있을 때, 여러 번의 과정을 거치지 않고 단일 플랫폼에서 검색이 가능하다는 장점을 제공한다. 이러한 과정을 거쳐 완성된 전체의 모델은 그림2와 같다.

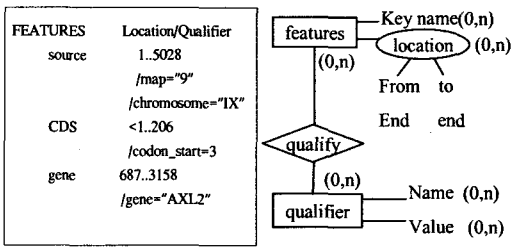


그림 1. features의 모델링
Fig 1. modeling of features

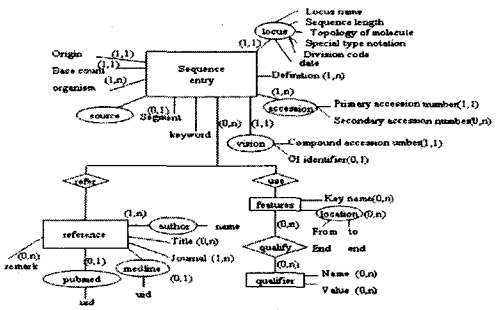


그림 2. GenBank의 ER 모델
Fig 2. ER model of GenBank

### 3.2 Swiss-Prot 데이터베이스와 ER 모델

SWISS-PROT는 스위스의 SIB(Swiss Institute of Bioinformatics)에서 운영하는 단백질 서열데이터베이스로 다른 단백질 데이터베이스에 비해 세 가지 장점을 지닌다. 즉, 쉽게 이해할 수 있게 주석을 제공하고, 데이터베이스 사이의 중복을 최소화시키며, 다른 데이터베이스와 통합이 쉽다.

SWISS-PROT 데이터베이스 역시 GenBank의 경우처럼 서열 엔트리들로 구성되고 서열 엔트리들은 서열에

대한 여러 정보를 가지고 있다. 표2는 서열 엔트리의 역할을 간단히 나타낸 것이다.

표 2. Swiss-Prot 데이터 필드
Table 2. Swiss-Prot data fields

| Line Code | contents |
|---|---|
| ID | identification |
| AC | accession number |
| DT | data |
| DE | description |
| DN | gene name |
| OS | organism species |
| OG | organelle |
| OC | organism classification |
| OX | taxonomy cross-reference |
| RN | reference number |
| RP | reference position |
| RC | reference comment |
| RX | reference cross-reference |
| RL | reference location |
| DR | database cross reference |
| FT | features |
| KW | keywords |

특히, Swiss-Prot 플랫파일에서의 DR라인이 cross-reference부분으로서 어떤 데이터베이스로 참조를 하느냐의 연관성을 보이므로 ISA 관계를 고려하여 세분화하고 모델링 한다. 그림 3에서와 같이 cross-reference 부분은 단백질 데이터와 핵산 데이터의 두 관계를 일반화하여 도식화한 것이다.

이 두 데이터들 역시 다른 데이터로의 연계과정의 중요한 역할을 하는 accession number나 ID같은 속성을 포함하고 있다. 그림 4는 완성된 전체 모델이다.

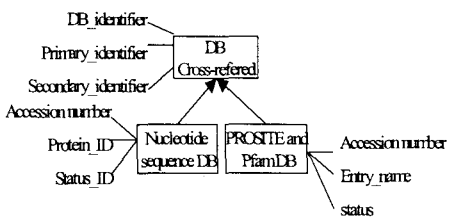


그림 3. cross-reference의 모델링
Fig 3. modeling of cross-reference

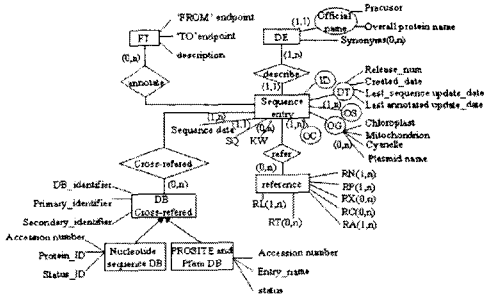

그림 4. Swiss-Prot의 ER 모델
Fig 4. ER model of Swiss-Prot

## Ⅳ. 바이오 데이터베이스의 통합

생물데이터의 통합은 앞서 언급한 것처럼 데이터 자체의 특성상 통합에 많은 어려움이 있다. 그러므로 통합의 문제는 오랫동안 연구가 진행 중이고, 아직도 많은 풀어야 할 문제들을 안고 있다. 특히 최근 들어 활발한 연구 주제중의 하나가 의미의 중복을 지니는 데이터 간의 통합 문제이다.

본 논문에서 발생하는 의미적인 충돌에 대해서는 GenBank를 기준으로 표준화하고 통합하였다. 즉 Swiss-prot의 DR라인은 GenBank Cross-reference와 동일시하며 이로 표준화하였다.

또한, 통합을 하는데 주요한 역할로 핵산-단백질 데이터베이스간의 연계부분은 cross-reference를 고려하여 표현한다. 그림 5는 두 데이터베이스를 통합하여 모델링 한 것이다. 그림 5에서 보는바와 같이 두 데이터베

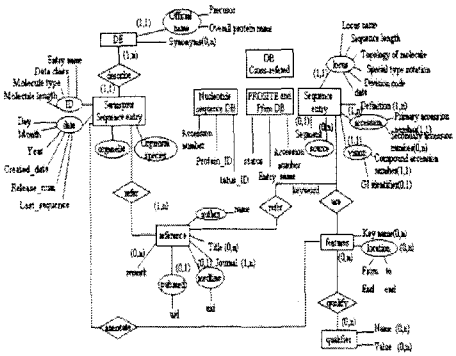

그림 5. 통합 ER 모델
Fig 5. Ingegrated ER model

이스의 sequence를 엔터티로 refefence를 관계로 표현하였으며, 그림 1과 3에서이미 제시한 세분화된 원자 단위의 모델을 포함한다.

## V. 바이오 게이트웨이 시스템

이질적인 데이터 통합을 위해 개념적 모델링을 수행한 후, 핵산-단백질 데이터간의 통합적인 질의를 할 수 있는 바이오 게이트웨이 시스템의 구조를 제시한다. 이 시스템은 핵산-단백질 자료로 연계되는 정보를 통합 서비스 할 수 있고 기존의 시스템에서는 처리가 쉽지 않았던 정교한 질의 처리가 가능하다. 예를 들어 "2001년 2월 이후에 등록된 2000bp 이상의 설치류(ROD)에 속하는 mRNA 중에서 GI:694699를 참조하고 CDS를 포함한 엔트리의 등록번호와 전체 서열을 찾고, 아데닌(A)과 티민(T)의 전체 개수를 찾아라". 라는 질의가 있다고 가정하자. 기존의 검색 시스템을 이용해서 검색을 할 경우, 단일 시스템에서 검색이 불가능하고 여러 번의 검색을 거쳐 결과를 조합해야만 원하는 결과를 얻을 수 있기 때문에 시간과 노력이 많이 필요했다. 그러나 본 논문에서는 사용자가 위와 같은 상세 조건을 알고 있다면 최소한의 검색으로 상세 질의에 따른 원하는 결과를 얻도록 원자적 단위의 모델링을 시도하였고 이를 바탕으로 시스템의 구조를 제안한다. 그림 6에서와 같이 웹 브라우저나 응용 프로그램 같은 사용자 인터페이스를 통해 질의가 입력되면 통합 프로세서가 통합 데이터베이스에 접근하여 질의를 수행하여 결과를 반환한다. 여기서의 통합 데이터베이스는 기존의 대표적인 관계형 데이터베이스인 오라클에 두 데이터를 포함하는 형태이다. 메타 데이터 저장소는 메타 데이터를 관리하는 지속적인 저장 공간으로 사용하고, 임시 저장 공간으로 캐쉬를 사용한다. 결과적으로 제안한 시스템을 통하여 개념적 모델인 ER모델을 바탕으로 원자적인 단위의 정보까지 표현함으로써 보다 쉽게 정교한 질의가 가능하다.

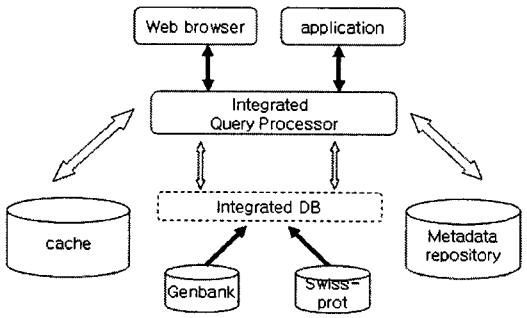


그림 6. 시스템 구조
Fig. 6. System Architecture

## VI. 결 론

이질적인 바이오 데이터베이스의 통합은 데이터 연계분석의 필요성에 의해 활발히 연구가 진행 중이다. 본 논문에서는 이질적인 바이오 데이터베이스들을 통합하여 정보의 제공을 목적으로 개념적 모델을 제시하고, 이를 적용한 바이오 게이트웨이의 구조를 제안하였다. 이질적이고 복잡한 특성을 지닌 바이오 데이터베이스 통합은 쉽지는 않지만 중요한 일로, 특히 초기 단계의 모델링이 중요한 과정으로써 본 논문에서는 Chen의 ER 다이어그램을 사용하여 개념적 모델링 과정을 수행함으로써 통합과정을 보였다. 특히 데이터 특성을 고려하여 원자적 단위로의 세분화된 모델링을 시도함으로써 중요한 정보의 손실을 막고 정보를 정확히 표현 하였다. 이러한 과정을 통해 얻어진 세분화된 각각의 모델은 통합의 과정을 통하여 두 데이터베이스 간의 연계정보를 포함하며 쉽게 정보를 얻을 수 있고, 이를 기반으로 한 바이오 게이트웨이 시스템은 원자적인 단계의 모델링 과정으로 인해 정교한 질의 처리가 가능하다는 장점을 지닌다.

## 참고문헌

[1] Andres D.Baxevanis " The Molecular Biology Database Collection : an online compilation of relevant database resources " oxford university press2000

[2] Okayama, T.,Tamura, T., Gohobori,T., Ikeo, K., Miyazaki,S., Fukami-Kobayashi, K. and Sugawara,H. "Formal design 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 modeling 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. " The Entity-Relationship Model:Toward a Unified View of Data". ACM trans. Database sys.

[6] Baker,P., Brass,A., Bechhofer,S., Goble,C., Paton, N. and Stevens,R :TAMBIS-transparent access to multiple biological information sources. In proceedings of International Conference on Intelligent Systems for Molecular Biology. AAAI Press.

[7] T.Etzold , A.Ulyanov, and P.Argos : SRS: Information Retrieval System for Molecular Biology Data Banks, Methods in Enzmemolog 226(1996)

[8] T.Etzold and P.Argos : SRS an indexing and retrieval tool for flat file data libraries, Appl. Biosci(1993) 49-57

[9] S.Davidson, C.Overton, V. Tannen, and L.Wong : Biokleisli : A digital library for biomedical researchers, Journal of Digital Libraries (1996)

[10] J.Hammer, H.Garcia-Molina, J.Cho, R.Aranha, and A.Crespo : Extracting Semistructured Information from the Web, Workshop on Management of Semistructured Data(1997)

[11] Chen, I. M.A.,Kosky, A., Markowitz,V. and Szeto,E. : Constructing and maintaining scientific database views in the frame work of the Object Protocol Model. In Proceedings of SSDBM. IEEE Press

[12] N.Ashish and C.Knoblock : Wrapper Generation for Semi-Structured Internet Sources, Workshop on Management of Semistructured Data(1997)

[13] A.Y.Levy, A.Rajarman, and J.J.Ordille : Querying Heterogeneous Information Sources Using Source Descriptions , Proc. Of the 22nd Conf. On Very Large Data Bases(VLDB'96

## 저자소개

**정진희(Jinhee Jung)**

1999 전남대학교 전산학과 (이학사)
2002 전남대학교 전산학과 (이학석사)
2002~현재 한국생명공학연구원 국가유전체 정보센터
※관심분야 : 데이터베이스, 데이터마이닝, 통합시스템, 생물정보학, 지식발견,기계학습

**정민아(Min-A Jung)**

1992년 전남대학교 전산통계학과 (이학사)
1994년 전남대학교 대학원 전산통계학과(이학석사)
2002년 전남대학교 대학원 전산통계학과(이학박사)
2002년 ~ 2003년 광주과학기술원 정보통신학과 Post-Doc.
2003년 ~ 2004년 전남대학교 전자통신기술연구소 Post-Doc.
2005년 ~ 현재 목포대학교 컴퓨터 교육과 교수
※관심분야 : 데이터베이스, 데이터마이닝, 생물정보학, 정보보호