바이오데이타통합 미들웨어시스템 구조

나도균[O1] , 이필현[1] , 이서우[1] , 이도헌[1] , 이광형[1], 배명남*
[1]{blisszen[O], phlee, sean, dhlee, kwlee}@bioif.kaist.ac.kr
* mnbae@etri.or.kr

Middleware System Architecture for Bio-data Integration

Dokyun Na[O1] ,PhilHyoun Lee[1],Sean Lee[1],Doheon Lee[1],Kwanghyung Lee[1],MyungNam Bae*
[1]Dept. of Biosystems, KAIST
*Electronics and Telecommunications Research Institute

다양한 바이오 정보 데이터베이스와 분석 도구들을 효율적으로 검색하고, 개별 데이터베이스에서는 얻을 수 없는 새로운 지식을 생성하기 위해서는 통합된 형태의 정보 검색 시스템이 필수적으로 요청된다. 여기서 우리는 바이오 정보 시스템 통합을 어렵게 하는 요소들을 살펴보고, 다중 질의 수행과 확장성 등을 기준으로, 현재 서비스되고 있는 바이오 정보 통합 시스템들의 특성을 분석 비교하였다. 또한 이를 기반으로 바이오 정보 통합 시스템의 구조를 제시하였다.

1. 서 론

2003년 4월 현재 바이오 정보 데이터베이스의 수[1]는 500개가 넘으며, 목적 및 정보의 내용 역시 그 수만큼 다양하다. 생물학 지식의 특성상 각 정보는 서로 밀접하게 연관되어 있어 특정 목적을 위한 데이터베이스라 하더라도 독립적이지 않고 다른 데이터베이스와 연계되어 이용되는 것이 일반적이다. 이런 이유로 연구자들은 의미 있는 바이오 정보를 얻기 위하여, 여러 데이터베이스를 검색, 각 결과의 통합, 그리고 다양한 도구로의 분석이 필요하다. 그러나 현재와 같이 다양한 데이터베이스가 독립적으로 존재하는 경우, 이러한 통합적인 정보 검색과, 새로운 지식 추출이 쉽지 않다. 특히 바이오 정보 데이터베이스의 형식적, 의미적 이질성은 검색 결과간의 정보 공유를 어렵게 하여, 연구자는 수작업으로 각 데이터를 분석하고 변형, 통합하는데 많은 시간을 소요하게 된다. 이러한 어려움들을 해결하기 위해서는 바이오 정보의 검색, 분석, 관리 및 새로운 정보의 창출을 가능케 하는 바이오 정보 통합 시스템의 구축이 필수적이다.

2. 통합 정보 시스템 구축 시 문제점

첫째, 통합 대상이 되는 바이오 정보 시스템의 자료 구조가 매우 다양하다. 현재 바이오 정보 통합 시스템은 플랫 파일, HTML, XML, Relational Database와 Object-oriented Database, 분석도구에 의해 생성된 결과 등을 모두 통합해야 한다[2]. 현재 대부분의 바이오 정보의 저장 형태는 플랫 파일이며, 대부분 정보 분석 도구 역시 애플리케이션이나 web 서비스로 제공되고 있다. 둘째, 각 소스 시스템의 정보는 특성상 다중 중첩(nested) 구조를 갖거나, 다른 소스 시스템과 복잡한 개념적 연관 관계를 가지고 있어 모든 데이터를 표현하는 모델을 만들기가 어렵다. 또한 새로운 생물학적 정보의 발견에 따라, 기존 데이터 모델에 개념이 추가되거나 구조가 변경되는 경우 역시 빈번하게 발생한다. 마지막으로, 통합 대상이 되는 바이오 정보 시스템의 수준 및 범위가 매우 다양하다. 실험을 통해 얻은 데이터를 저장하는 1차 바이오 데이터베이스와, 이의 정보를 선별하여 기능 등의 주석 정보 등을 담거나, 종합적인 정보 제공을 목적으로 하는 2차 데이터베이스가 있다. 그리고 데이터베이스에서 얻은 정보를 분석

및 가공하는 도구 역시 다양하게 존재한다. 이런 통합 대상의 다양성이 바이오 데이터베이스 통합 시스템이라는 용어 대신 바이오 정보 통합 시스템이라는 명칭을 사용하는 이유이기도 하다[3].

3. 기존 통합 바이오 정보 시스템 분석

현재 다양한 통합 시스템이 존재하나 그 방법이 상이한 3개의 시스템을 위주로 분석하였다. 다양한 자체 데이터베이스를 기반으로 하고 있는 Entrez[4], 확장성에 중심을 둔 통합 방식인 SRS[5], 소스 정보 시스템들의 투명성(transparency)을 중시하는 TAMBIS[6] 등을 전체 구조, 전역 글로벌 모델의 유무, 소스 정보 시스템으로의 사상을 위한 질의 변환 및 질의 결과의 통합 등을 중심으로 다음과 같이 분석하였다. Entrez는 NCBI에서 구축한 통합 바이오 정보 검색 시스템으로 통합 시스템의 전역적인 데이터 모델은 존재하지 않으며, 내부 데이터베이스들을 하이퍼 링크 방식으로 통합하고 있다. 여러 데이터베이스를 동시에 검색할 수 없으며, Boolean 방식에 기반한 단순 질의만이 가능하다. 자체적으로 운영되는 데이터베이스만을 링크로 연결하는 제한된 방식의 통합 시스템이라는 단점을 갖는다. SRS는 LionBioscience에서 상업적으로 판매중인 소프트웨어이다. 각 소스 정보 시스템의 구조, 자료형식 등을 통합한 전역 데이터 모델은 존재 하지 않는다. 질의 수행시 인덱스 테이블을 이용하여 검색하므로, 검색 성능이 우수하다. 사용자가 데이터베이스를 여러 개 선택할 경우, 선택된 데이터베이스의 공통 필드만을 대상으로 한 제한된 형태의 다중 질의가 가능하다. 다른 통합시스템과 달리 사용자가 SRS 소프트웨어를 설치, 커스터마이징을 할 수 있으며, 자체wrapper 제작 언어를 제공해 새로운 데이터베이스를 쉽게 추가할 수 있다. TAMBIS는 영국의 멘체스터 대학교에서 데이터베이스 검색 투명성(transparency)을 목적으로 개발한 시스템이다. 생물학적인 개념을 포괄적으로 표현하기 위하여 온토로지(ontology)를 전역적인 데이터 모델로 정의했다는 것이 가장 큰 특징이다. 이 시스템은 각 데이터베이스 스키마를 온토로지와 대응시켜 놓고 있으며, 온토로지에 의해 같은 의미의 항목들을 하나로 식별할 수 있어 강력한 다중 질의가 가능하다. 그리고 온토로지를 이용한 직관적인 사용자 인터페이스를 제공하고 있다. 그러나 온토로지의 제작이나, 새로운 데이터베이스 추가시 전문가가 오랜 시간 작업해야하는 어려움이 있다. 이로 인해 현재까지 통합된 정보 시스템의 수가 5개에 그칠 정도로 확장성이 떨어진다는 단점이 있다.

이런 분석의 결과, 바이오 정보 통합 시스템은 다중 질의와 데이터베이스의 확장성을 모두 만족시키는 것이 매우 중요하다.

4. 바이오 정보 통합 시스템 구축

시스템의 구조는 크게 Application, Mediator, Wrapper layer로 나뉜다. 여기서 언급되는 시스템은 다양한 application이 질의하여 원하는 결과를 얻을 수 있는 기능을 제공하는 Mediator에 초점을 두고 있다. 즉, 단순히 다중 데이터베이스 검색만을 지원하는 시스템이 아니라 개발자들이 만든 여러 application은 이 시스템을 통해 원하는 정보를 얻고, 그것을 스스로 처리할 수 있는 형태로 결과를 제공하는 것이다.

먼저 global schema server는 하나의 작은 규모의 Ontology이다. 그러나 기존의 ontology는 생물학적인 개념을 기반으로 계층구조를 이룬 것인 반면, 이 시스템의 경우 각 Database에서 검색 가능한 field를 기반으로 하고 있다. 앞서 설명된 것처럼, TAMBIS는 모든 schema를 수용할 수 있는 ontology를 제작하여 강력한 다중 질의를 제공하는 반면, SRS는 global schema없이 데이터베이스 간의 대응 정보만 가져 확장성이 높다. 그러나 ontology를 유지, 관리하는게 힘들며, 반대로 그것 없이 다중 질의를 하기가 어렵다는 단점이 있다. 이런 이유로 여기서는 이 둘의 장점을 수용할 수 있도록 유지 관리가 쉬운 범위의, 그리고 global schema의 기능을 하는 ontology를 제작하는 것이다.

Application은 global schema server를 통해 가능한 검색의 범위를 알 수 있으며, 이를 통해 질의를 생성하고 mediator로 전달한다. 그리고 global schema형태로 표준화된 결과를 제공 받거나 원래 데이터베이스의 형태로 받을 수 있다.

Mediator는 application layer에서 온 질의를 각 데이터베

이스에 맞게 분해, 재생성하며 이를 해당 wrapper로 보낸다. Wrapper에서 온 결과 중 cross-reference로 명시하여 다른 database일지라도 같은 정보를 의미한다면 이를 하나의 entry로 보고 mediator는 wrapper에서 온 결과에서 다중질의를 실행한다.

끝으로, 각 데이터베이스의 정보를 자체적으로 저장하고 있는 virtual integration 방식은, 비록 각 데이터베이스의 성능이나 신뢰도에 질의 수행 결과를 의존하게 되고 다중 데이터베이스 검색이 어려운 단점이 있으나 빠른 속도로 축적되고 갱신되는 생물학 정보의 특성을 감안할 때 적절하다고 여겨진다. 그래서 wrapper는 별도로 저장된 data를 검색하지 않고 각각의 해당 database가 제공하는 web interface(SOAP 등)를 통해 질의를 하고 결과를 받는다.

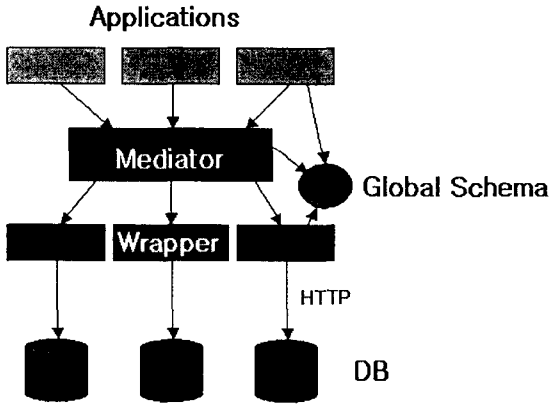

그림 1 제안된 통합 시스템의 구조

5. 결론

방대한 바이오 데이터를 효율적으로 검색하고, 다양한 데이터베이스 및 생물 정보학 분석 도구들을 효과적으로 활용하기 위해서는 통합된 형태의 바이오 정보 시스템이 필수적이다. 이 논문에서는 바이오 정보 시스템 통합을 어렵게 하는 요소들을 살펴보고, 현재 제공되고 있는 바이오 정보 통합 시스템들의 특성을 분석하였다. 또한 이를 기반으로 새로운 통합 시스템을 구성하였다. 새로 구현될 바이오 정보 통합 검색 시스템은 다중 질의 수행과 확장의 용이성 문제를 어느 정도 해결해 주리라 기대한다.

6. 참고 문헌

[1] Andreas D. Baxevanis "The Molecular Biology Database Collection 2003 update", *Nucleic Acids Research* 31(1), 1-12. 2003

[2] Francois Bry et al. "A Molecular Biology Database Digest"

[3] Ulf Leser. "Designing a Global Information Resource for Molecular Biology" In 8th GI Fachtagung: Datenbanksysteme in Buero, Technik und Wissenschaft, Freiburg, Germany, 1999

[4] Entrez Online Documentation http://www.ncbi.nlm.nih.gov/Database/index.html

[5] Etzold, T. *et al.* "SRS: Information Retrieval System for Molecular Biology Data Banks." *Methods in Enzymology* 266, 114-128, 1996

[6] Stevens R *et al*, " TAMBIS: transparent access to multiple bioinformatics information sources." *Bioinformatics* 16(2):184-185, 2000