다양한 명명 규칙을 가진 대용량 데이터베이스의 통합 방안

한범수, 이현주
광주과학기술원 정보통신공학과
e-mail : *gwhanbs@gist.ac.kr, hyunjulee@gist.ac.kr*

Integration of large-scale databases in various naming schemes

Beomsoo Han, Hyunju Lee
Information and Communications
Gwangju Institute of Science and Technology

Abstract

It is of importance to integrate several databases to improve its coverage and usage of large amount of biological information produced by diverse biological experiments. In this paper, we proposed a method to integrate the protein interaction databases with various naming schemes. The identifier (ID) mapping methods in the process of integration was also presented.

I. 서론

생물정보 분석을 위한 기본 요소는 다양한 실험집단에서 수행한 생물학적 실험결과로 구성된 데이터들이다[1]. 이러한 생물정보 데이터는 다양한 실험방법의 등장으로 그 양이 급격히 증가하고 있다. 그러나 각 실험집단의 데이터들은 그 실험집단의 연구 특성에 따라 다양한 구조와 명명 규칙을 사용하고 있어 효율적인 정보의 활용에 어려움이 있다 [2].

단백질 상호작용 (protein interaction)과 단백질 기능 예측의 연구에서도 많은 실험 집단들이 개별적인 실험결과를 자신들의 웹 페이지를 통해 제공하고 있다. 그러나, 미지의 단백질 기능을 예측하기 위해서는 가능한 많은 양의 상호작용 데이터를 활용할수록 예측의 정확성을 높일 수 있게 된다. 따라서 오랜 시간과 노력을 통해 구해진 다양한 단백질 상호작용 데이터를 하나로 통합하여 관리할 필요가 있다. 그리고 주로 플랫 파일이나 XML 형태로 제공되고 있는 대용량의 생물정보 데이터들을 효과적인 활용을 위해서는 데이터베이스 관리시스템 (DBMS)에 기반 한 통합 데이터베이스의 구축이 필수적이다.

본 연구에서는 다양한 단백질 상호작용 데이터들을 하나의 통합된 데이터베이스로 구축하기 위한 방안을 제시하였다. 특히, 주요 명명규칙에 따른 ID들을 매핑 (mapping) 정보를 활용하여 통합하는 과정에서의 문제점과 해결방안들을 살펴보았다.

II. 본론

단백질 상호작용 데이터는 이미 기능이 알려진 단백질들을 통해 관련된 미지의 단백질 기능을 예측하는데 유용하게 사용될 수 있다. 그러나 각 실험집단에서 자신들의 목적과 편리에 맞게 UniGene Cluster, Swiss Prot/UniProt, RefSeq, Gene Symbol, EntrezGene, UniGene 등과 같은 다양한 형태의 명명규칙을 사용하고 있어 이들 데이터를 쉽게 통합하여 활용하지 못하

는 어려움이 있다.

2.1 통합 데이터베이스 구축의 주안점

본 연구에서는 다양한 명명규칙을 가진 단백질 상호작용 데이터의 효율적인 검색을 위해 주로 사용되는 ID들을 기준 ID로 정하고 하나의 레코드에 동시에 포함되도록 하였다. 또한 기존의 데이터베이스들에서 한 필드에 여러 개의 ID들이 포함되어 있는 경우 검색의 효율을 높이기 위해 한 필드에 한 ID값만 가질 수 있도록 하였다. 그리고, ID 매핑 정보를 이용하여 기준 ID들에 널(null) 값을 최소화하도록 하였다.

2.2 세 가지 기준 ID

단백질 상호작용 데이터들과 관련 연구들에서 주로 사용하고 있는 다음 세 가지 명명규칙을 기준 ID로 정하였다.

- NCBI Reference Sequence (RefSeq)
- UniProtKB/Swiss-Prot
- Gene Symbol

2.3 단백질 상호작용 데이터베이스

단백질 상호작용과 기능분석에 관한 대표적인 관련 데이터베이스들 중에서 그림 1과 같이 다음 네 가지 데이터베이스를 이용하였다 [3].

- DIP (Database of Interacting Proteins) [4]
- MINT (Molecular INTeraction database) [5]
- IntAct (protein INTerACTion data) [6]
- HPRD (Human Protein Reference Database) [7]

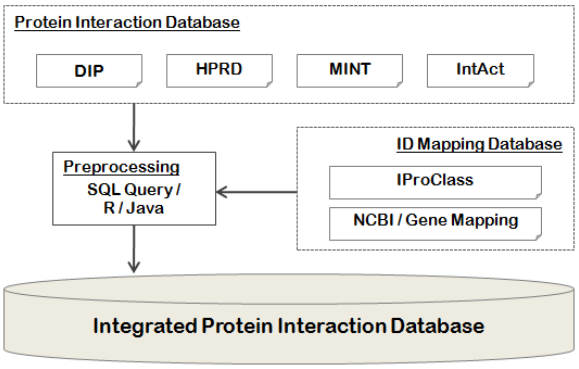

그림 1. 통합 데이터베이스의 구성

2.4 기준 ID 맵핑(mapping)

각각의 데이터베이스들은 각 연구기관의 특성에 따라 하나 또는 두 개 정도의 다양한 ID들을 사용하고 있다. 따라서 통합 데이터베이스의 효율적인 활용을 위해서는 ID 매핑을 통해 기준 ID 정보들을 지정해 주어야 한다. 이를 위해, iProClass (Integrated Protein Knowledgebase) [8]와 NCBI/gene mapping [9]의 매핑 정보를 이용하여 개별 데이터베이스들의 ID들에 매핑되는 세 가지 기준 ID값들을 구하였다. 여러 개의 ID들이 매핑되는 경우에는 그림 1과 같이 SQL Query, R, Java 등을 활용하여 적절히 변환하고 한 필드에 하나의 ID값만 가지도록 해당 레코드를 처리하였다.

Ⅲ. 요약 및 결론

다양한 실험집단에서 산출된 대용량의 생물정보 데이터를 효과적으로 활용하기 위해서는 통합 데이터베이스의 구현이 필수적이다. 본 연구에서는 단백질 상호작용에 초점을 맞춘 통합 데이터베이스를 구현하기 위한 방법을 제시하였다. 그리고 통합 과정에서 문제가 되는 명명 ID들의 매핑 방법에 대해 살펴보았다.

통합 데이터베이스는 단백질 상호작용에 대한 다양한 정보를 쉽고 빠르게 검색할 수 있어 미지의 단백질 기능 예측에 보다 효과적으로 활용될 수 있을 것이다.

끝으로, 생물정보 데이터는 지속적으로 갱신되고 있으므로 향후 새로운 데이터의 추가와 기존 데이터의 갱신이 정기적이며 지속적으로 이루어져야 할 것이다.

감사의 글

본 연구는 광주과학기술원의 GIST faculty start-up fund의 지원에 의한 것입니다.

참고문헌

[1] 최요한, 유성준, 김민경, 박현석, 웹 서비스 기반 바이오 정보 통합 분석 도구, 한국정보과학회 2004년도 봄 학술발표논문집, 31권 1호(B), 289-291, 2004

[2] 박성희, 류근호, 포스트 게놈시대의 생명정보 데이터베이스 연구: Issues & Challenges, 전자공학회지, 30권 10호, 29-42, 2003

[3] 정민철, 박완, 김기봉, 웹 기반의 단백질 상호작용 및 기능분석을 위한 보조 시스템 개발, 생명과학회지, 14권 6호, 997-1002, 2004

[4] http://dip.doe-mbi.ucla.edu

[5] http://mint.bio.uniroma2.it/mint

[6] http://www.ebi.ac.uk/intact

[7] http://www.hprd.org

[8] ftp://ftp.pir.georgetown.edu/databases/iproclass

[9] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA