단백질 결합정보 검증 도구

임상택* 이환구* 이석기** 김규원** 차재혁*
*한양대학교 정보통신대학원, **서울대학교 종합약학연구소
{ss009, prodog}@ihanyang.ac.kr, sekido@snu.ac.kr qwonkim@plaza.snu.ac.kr chajh@hanyang.ac.kr

Protein Interaction Information Verification Tool

Sangteak Lim*, Hwangu Lee*, Seok-Ki Lee**, Kyu-Won Kim**, Jaehyuk Cha*
*The Graduate School of Information & Communications, Hanyang University
**Research Institute of Pharmaceutical Sciences, Seoul National University

요 약

현재 방대한 생물학 관련 문헌 DB로부터 단백질 결합 정보를 찾아내어 데이터베이스를 구축하고 있다. 이 과정에서 자동적으로 추출된 단백질 결합 정보에 대해 생물학 전문가가 이를 보고 올바른 정보인지를 검증하는 작업이 매우 중요하다. 그러나 이러한 검증 작업은 많은 시간이 소모 되므로 이를 쉽게 검증할 수 있도록 지원하는 통합환경의 검증 도구를 개발하고 이를 소개한다.

1. 서 론

현재 국내는 물론 세계 각 나라에서 연구된 생물학 정보를 데이터베이스로 구축하여 인터넷으로 공개하고 있다. 생물학 정보로 쏟아지는 량은 실로 방대하다. 따라서 이 분야의 연구는 이미 공개된 생물학 정보를 효율적으로 수집하고 분석도구로 파악, 예측하고 그것을 검증된 자료로 시스템을 구축하는 것에 성패가 달려있다고 해도 과언이 아니다.

방대한 생물학 정보 중에 단백질 결합정보를 사람이 하나씩 수동적으로 수집하기엔 너무 많은 시간과 노력이 든다. 그래서 자동화된 문헌 추출 도구를 이용하여 수집을 하고 있다. 문제는 이러한 자동 추출 도구의 정확도가 한계가 있다. 또한 생물학 관련 전문가가 이를 검증하지 않고서는 의미 있는 단백질 결합 네트워크 데이터베이스를 구축할 수 없다. 뿐만 아니라 자동적으로 수집한 단백질 결합 정보를 검증하기 위한 통합환경의 검증 도구도 없다.

현재는 검증하기 위해서 생물학 전문가가 일일이 추출된 정보의 출처를 찾아서 그 원문을 검토하여 정확도에 따라 올바른 정도를 채점하고 있는 실태이다.

본 연구는 공개된 생물학 정보 중에 단백질 간의 상호작용에 관련한 문헌을 찾아 표준화된 데이터베이스로 구축하는 과정에 필요한 단백질 결합정보 검증 도구(Protein Interaction Information Verification Tool)를 개발하여 이를 소개한다.

2. 자료 수집과 검증 방법

일반적으로 단백질 결합 네트워크 데이터베이스(Protein Interaction Network Database)[1],[2],[3],[5],[7]를 구축하기 위해 자료를 수집하는 방법은 다음과 같다. 생물학 정보를 데이터베이스로 구축하고 이를 공개하고 있는 대표적인 PubMed, Medline, GenBank[1],[3],[5] 와 같은 시스템의 인터페이스를 통하여 단백질 결합정보(Interaction Information)의 의미를 나타내거나 표현하는 문헌 정보 수집한다. Medline의 경우 현재 문헌 정보 수가 200만 건이 넘는다. 이렇게 많은 자료에서 어떤 단백질이 어느 단백질에 어떻게 작용을 관계인지를 사람이 하나씩 수집하기는 너무 많은 비용과 수고를 필요로 하므로 대부분 자동 추출 도구를 사용하고 있다. 수집을 위해서 먼저 단백질 이름을 키워드로 검색하고 그 단백질의 작용을 나타내는 동사(Verb)로 문장 형식에 따른 패턴인식을 하는 형태로 자동 추출하여 Raw Information로 취합한다.[5] 현재는 단순한 패턴 인식 형태를 넘어서 보다 적중률을 높이기 위해 바이오 텍스트 마이닝 기술을 이용해 단백질 결합정보를 추출하는 연구가 활발히 진행되고 있다.[3],[6] 이렇게 수집된 단백질 결합 정보는 생물학 관련 전문가의 검증을 거쳐 데이터베이스로 구축하게 된다.

검증하는 현재 방법은 소스 단백질과 타겟 단백질 그리고 그 상호작용 관계(Interaction Relation)를 나타내는 문장을 사람이 읽고 판단한다. 즉, 단백질

결합정보를 추출한 원문이 있는 사이트나 공개된 데이터베이스 시스템을 검증자(Curator)가 일일이 직접 접속하여 검증하려는 자료를 찾아서 문헌과 비교하고 정확도에 따라 등급을 구분하여 점수를 채점하고 있다.[3][7] 이러한 검증 방법은 서로 이질적인 인터페이스와 통일되지 못한 형식을 제공한다. 결국, 그 환경에 익숙하지 못한 검증자는 복잡한 단계의 절차를 배워야만 한다. 그리고 단백질 결합정보를 검증하는 시간 보다 검증을 위해 문헌을 찾는 시간과 노력이 더 들게 된다. 따라서 매우 비효율적이고 많은 비용이 허비된다. 이러한 문제점을 해결하기 위해서는 통합 환경(Integrated Data Environment)의 검증 도구가 요구되고 있다.

3. 단백질 결합정보 검증 도구(PIIVT)

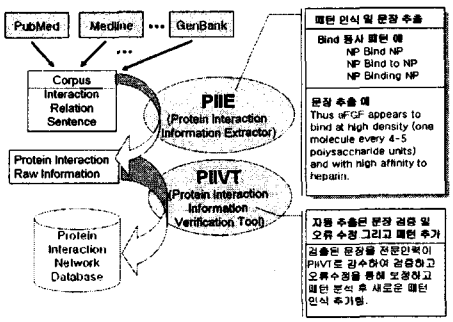
검증 도구는 기본적으로 상호작용 관계에 있는 단백질을 검색할 수 있어야 하며 그 관계가 어떤 상호작용 인지 알 수 있는 기능이 있어야 한다. 또한 그 관계를 검증하기 위해 참고 할 수 있는 문헌이나 문장을 제시하는 기능이 있어야만 한다. PIIVT는 이런 기능을 제공하기 위해 원문(Corpus) 검색, 자동 추출된 문장 검색(Sentence Window)을 제공하고 있다. 부가적인 기능으로 데이터베이스 질의와 통계 수치를 얻기 위한 질의 수행기(SQL Commander), PINDB의 탐색기(Viewer), 검색된 자료의 결과를 서브셋 형태의 XML로 변환 저장하는 기능, 자동 추출기(Protein Interaction Information Extractor)[5],[7] 패턴 추가 기능 등을 가지고 있다.

Interaction	Positive Regulation	Negative Regulation	Regulation
bind	activate	Inhibit	target
interact	induce	reduce	regulate
conjugate	stabilize	degrade	modulate
associate	..	suppress	control
..		destabilize	ubiquitinate
		..	..

표 1 PIIE 패턴 인식을 위한 동사 규칙표

자동 추출기 (PIIE: Protein Interaction Information Extractor)는 단백질간의 상호작용 관계를 나타내는 동사와 문장 형태에 따라 패턴 인식을 한다. 즉, 단백질 개체명, 상호작용 관계를 나타내는 동사들을 분석하여 미리 정해진 패턴에 의해 검색 후 인식하게 된다. PIIE에 의해 추출된 단백질 상호작용 정보들은 IRS (Interaction Relation Sentence) 및 IR(Interaction Relation) 정보로 세분화되어 PINDB(Protein Interaction Relation Database)에 저장된다.[5]

현재 PIIE의 자동 추출 기능은 70% 이상의 적중률이 나타나고 있으며 보다 지능적이며 높은 정확도를 만들기 위해 바이오 텍스트 마이닝 연구가 계속해서 진행되고 있다. 그리고 위에서 설명한 PIIE와 PIIVT의 유기적인 동작 관계가 [그림 1]의 동작 개요도에 잘 나타나 있다.

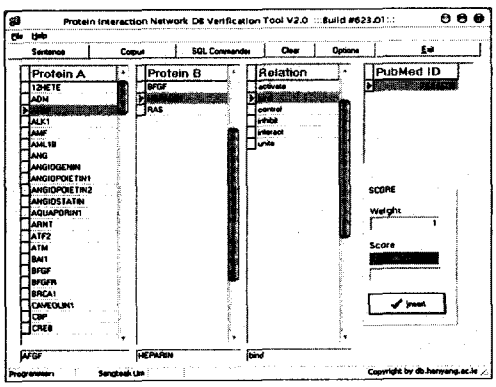


[그림 1] 검증 도구(PIIVT)와 자동 추출기(PIIE)의 동작 개요도

3.1 효율적인 사용자 인터페이스

공개된 생물학 관련 데이터베이스 시스템들의 다양한 인터페이스로부터 단백질 결합정보를 수집하여 이를 검증하기 위해선 직관적이고 효율적인 사용자 인터페이스를 제공해야 많은 시간과 비용을 줄일 수 있다. 뿐만 아니라 검증자(Curator)들이 대부분 컴퓨터 사용에 익숙하지 못한 생물학 관련 전문가 많다. 그러므로 데이터베이스 접근과 설정이 용이하면서 사용이 직관적이고 간편한 형태로 사용자 인터페이스를 제공해야 한다.

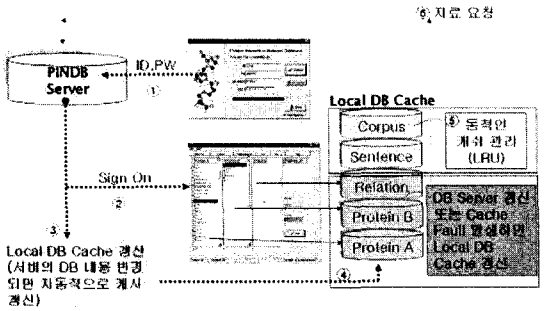
PIIVT는 3단계 내에서 모든 기능이 수행되도록 설계했다. 메인 창을 기준으로 기능별 윈도우로 나뉘어 구동한다. 사용자가 필요로 하는 자료(인자의 이름, 염기서열, 세포내 위치, accession 번호, 키워드 항목, 자동 추출된 문장, 문헌) 검색과 검증에 필요한 이벤트가 모두 수행된다.



[그림 2] Protein Interaction Information Verification Tool

3.2 서버자원 최소사용 설계

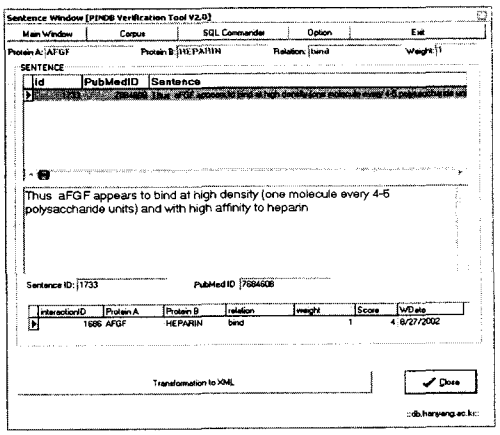
PIIVT는 Short Transaction Life 통한 데이터베이스 서버의 자원 최소 사용과 클라이언트 사이드에서 비연결 지향적인 접속 방법을 택있다. 그리고 Data Set을 Client 메모리 안에서 유지시켜 빠른 정렬과 검색이 가능하다. 즉, 빠른 검색을 위하여 PIIVT를 구동하는 시스템에 Local DB Cache를 운영한다. 캐시 운영 방식은 동적인 것과 정적인 것으로 구분하며 설계했다. 데이터 단위가 커서 메모리에 Data Set을 모두 유지하기 힘든 단백질 결합정보의 원문(Corpus)[1],[5]와 IRS(Interaction Relation Sentence)는 LRU(Least Recently Used) 방식으로 동적인 캐시 운영을 한다. 그리고 작은 용량이면서 빈번한 검색이 발생하는 "단백질 A" 와 "단백질B" 그리고 "Relation" 등은 정적인 캐시로 동작하며 Cache Fault 또는 데이터베이스 서버의 내용이 변경될 경우만 캐시가 갱신된다. 이러한 캐시 운영은 결과적으로 데이터베이스 서버에게 질의 요구를 최소화하며 검증도구 자체가 빠른 동작과 검색이 가능하게 된다.



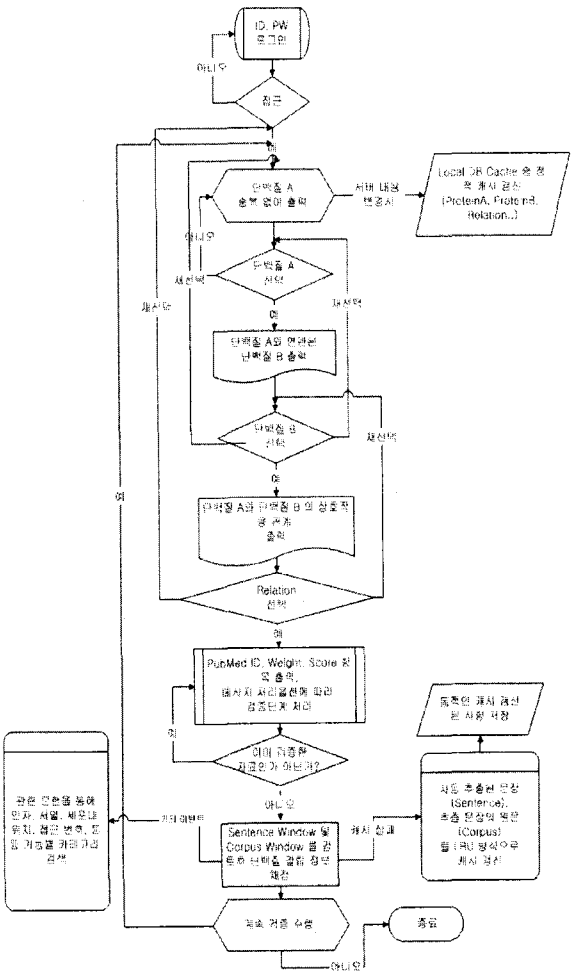
[그림 3] PIVT의 Local DB Cache 구성도

3.3 동작

PIIVT가 실행되어 정상적인 로그인이 이루어지면 기본적으로 PINDB의 전체 약 200만개가 넘는 튜플 중 Source에 해당하는 Protein A를 중복 없이 별개로 처리(동의어 포함)하여 출력한다. 오름차순 정렬한 단백질 개수는 약 189개로 Protein A에 표시된다. 단백질 개수는 새로 발견되는 문헌에 따라 개수가 증가될 수 도 있다. 사용자는 해당하는 항목을 Protein A에서 선택한다. 자동적으로 Protein A에서 선택된 항목에 상호작용 관계(Protein Interaction)를 가지고 있는 단백질들 리스트가 Protein B에 오름차순으로 표시된다. 다시, Target에 해당하는 단백질을 Protein B에서 선택하면 Relation에 그와 관련된 항목이 표시된다.



[그림 4] 추출된 문장을 검증하는 Sentence Window



[그림 5] PIIVT 흐름도

Relation까지 선택하면 단백질 간의 상호 관련된 자료 수가 Weight 값으로 나타나며 이때의 자료는 자동 추출된 문장 수를 의미한다. Score 값은 검증을 통해 정량화 된 수치를 입력한 것을 나타낸다. 수동 검증을 하지 않은 자료일 경우 Score 값을 비어있다. 이것을 생물학 관련 전문가가 추출된 문장과 문헌을 검토하여 채점을 한다. 채점 점수는 8점 high correct, 4점 low correct, 0 not match, -4점 low correct (the passive voice), -8점 -8 high correct (the passive voice)이다. 이때 PubMed ID는 문장을 추출한 문헌의 고유한 ID로 해당 문장의 원문을 검색하고 보여준다. 즉, 문헌을 통해 관련 인자의 이름, 서열, 세포내 위치, accession 번호, 키워드 항목을 이용하여 검색할 수 있고, 기능별 카테고리별로 검색이 가능하다.

4. 결론과 향후 연구 방향

현재 방대한 생물학 관련 문헌 DB로부터 단백질 결합 정보를 찾아내어 데이터베이스를 구축하고 있다. 이 과정에서 자동적으로 추출된 단백질 결합 정보에 대해 생물학 전문가가 이를 보고 올바른 정보인지를 검증하는 작업이 매우 중요하다. 그러나 이러한 검증 작업은 많은 시간이 소모 되므로 이를 쉽게 검증할 수 있도록 지원하는 통합환경(Integrated Data Environment)의 검증 도구(PIIVT)를 제안하였다.

PIIVT는 단백질 결합정보를 검증하기 위해서 생물학 전문가가 일일이 추출된 정보의 출처를 찾아야 하는 노력과 시간 비용을 줄여준다. 뿐만 아니라 원문(Corpus) 검색, 자동 추출된 문장 검색 그리고 데이터베이스 질의와 통계 정보를 얻기 위한 질의 수행기, PINDB 탐색기(Viewer), 검색된 자료의 결과를 서브셋 형태의 XML로 변환 저장하는 기능, PIIE의 패턴 추가 기능 등을 갖춘 통합환경 툴이다. 실제로 PIIVT를 통해 명확하게 검증된 자료로 데이터베이스를 구축함으로써 단백질 결합정보 인프라를 구축하는 성과가 나타나고 있다. 그 사례로 AngioDB System[1],[5]의 경우를 보면 혈관신생 분야의 연구자들에게 혈관신생 인자에 대한 기본적인 정보와 활용성을 제공하고 있다. 또한 여러 제약회사 및 연구기관들에 질병 치료를 위한 약물 후보물질의 발굴 및 기능을 이해하기 위한 솔루션으로 매우 유용하게 활용되고 있다. 앞으로 본 연구팀은 혈관신생에 관련한 단백질뿐만 아니라 모든 단백질간의 결합 네트워크를 추출하는 것으로 연구 범위를 확대하고 있다.

참고문헌

1. Seok-Ki Lee, Yong S. Choi, Jaehyuk Cha, Eun-Joung Moon, Sae-Won Lee, Moon-Kyung Bae, Tae-Kwon Sohn, Youjip Won, Sangback Ma, Eun Bae Kong, Hwangu Lee, Sangteak Lim, Daejin Chang, Yung-Jin Kim, Chul Woo Kim, Byoung-Tak Zhang and Kyu-Won Kim. Identification of novel anti-angiogenic factor by in silico functional gene screenign method. Journal of biotechnology 2003 (revised).

2. Lee, Y.M., Jeong, C.H., Koo, S.Y., Son, M.J., Song, H.S., Bae, S.K., Raleigh, J.A., Chung, H.Y., Yoo, M.A. and Kim, K.W. Determination of hypoxic region by hypoxia marker in developing mouse embryos in vivo: a possible signal for vessel development. Developmental Dynamics 2001; 220; 175-186.

3. Bader GD, Betel D, Hogue CW.
BIND: the Biomolecular Interaction Network Database.
Nucleic Acids Res. 2003 Jan 1;31(1):248-50.

4. Kim, M.S., Kwon, H.J., Lee, Y.M., Baek, J.H., Jang, J.E., Lee, S.W., Moon, E.J., Kim, H.S., Lee, S.K., Chung, H.Y., Kim, C.W. and Kim, K.W. Histone deacetylases induce angiogenesis by negative regulation of tumor suppressor genes. Nature Medicine 2001; 7; 437-443.

5. 이석기, 차재혁, 임상택, 이환구, 김규원, 김성훈 "AngioDB: 혈관신생 인자에 대한 데이터베이스 구축 및 활용연구", 대한의료정보학회 하계종합학술대회 , 2002 11월

6. 임해창, 황영숙, 박경미, "바이오 텍스트 마이닝 시스템 개발" , 정보과학회지, 2003. 6 pp 60-68.

7. Blaschke, C., Andrade, M.A., Ouzounis, C. and Valencia, A. Automatic extraction of biological information from scientific text: protein-protein interactions. Proc. Int. Conf. Intell. Syst. Mol. Biol. 1999; 60-67.