생물 의료 정보의 효과적인 텍스트 시각화

Effective text visualization for biomedical information

김탁은, Tak-eun Kim, 박종철, Jong C. Park
한국과학기술원 전자전산학과 전산학전공

요약 생물 의료 분야에서 정보의 양이 아주 빠르게 증가하고 있다. 이러한 방대한 양의 정보에서 유용한 정보를 추출하기 위해 텍스트 마이닝 기법을 이용한 연구들이 많이 진행되어 왔다. 그렇지만 이렇게 뽑아진 정보조차 그 양이 방대하고, 또한 텍스트로 되어 있기 때문에 직관적으로 이해하기가 어렵다. 따라서 이러한 정보들을 좀 더 직관적으로 이해하기 위해서는 정보 시각화 시스템이 필수적이다. 최근 들어 이러한 정보 시각화에 대한 연구가 많이 진행되었으나 이러한 시각화 정보조차 너무나 방대하기 때문에 사용자가 필요로 하는 정보를 여과해 주는 방법이 필요하다. 그리고 시각화 시스템에서의 지식 발견을 위한 방법을 제공하여야 한다. 본 논문에서는 생물 의료 정보의 텍스트 시각화에 초점을 맞추어 생물 의료 정보의 효과적인 표현 방법과 지식 발견을 위한 직관적인 인터페이스를 제안하고자 한다.

핵심어: Text Visualization, Biomedical Information, Text Mining, User Customization, Knowledge Discovery

1. 서론

방대한 양의 데이터를 시각화를 통해 가공해서 보여주는 것은 정보의 소비자에게 빠르고 직관적인 이해를 가져다 준다. 대표적인 예가 지도라고 할 수 있는데, 지도는 방대한 양의 좌표 데이터를 잘 시각화하였기에 사용자는 별다른 배경지식 없이도 직관적으로 지도 정보를 이해할 수 있다. 이러한 데이터 시각화 시스템은 Information Visualization이라는 독립된 분야로서 많이 연구되고 있다[1].

최근 들어 생물 의료 정보학에 대한 관심이 높아지면서 정형화되지 않은 방대한 생물 의료 데이터로부터 텍스트 마이닝을 통해 의미 있는 정보를 추출하려는 연구들이 많이 있어 왔다[9]. 텍스트 마이닝의 목적은 가공되지 않은 정보들에서 필요한 정보들만 추출하는 것인데, 이렇게 추출된 의미 있는 정보 역시 텍스트 형태의 테이블로 나열될 수 밖에 없기 때문에[4] 사용자에게 또 다른 종류의 분석을 해야 하는 부담을 제공한다.

이러한 이유로 인해, 생물 의료 정보에서 텍스트 마이닝을 통해 추출된 정보들을 시각적으로 표현하여 사용자의 이해를 도울 뿐만 아니라 새로운 지식 발견을 위해 사용될 수 있는 시각화 시스템이 필요하다[1].

시각화 시스템이 사용자들의 직관에 많은 도움을 주는 반면 방대한 양의 데이터들을 잘 선별해서 보여주지 않으면 오히려 시각적으로 혼란스러워져 사용자의 이해를 방해한다는 문제가 있다. 이러한 문제를 해결하는 방법으로 [3]과 [4]에서는 정보의 여과를 제안한다. 그런데 이렇게 정보 여과를 하려면 정보 추출 시스템에서 추출된 결과를 종류별로 분류해주는 작업이 필요한데 시각화 시스템에서 이런 분류를 자동으로 제공해 주기가 어렵다.

그리고 사용자의 직관적인 이해를 도와 지식 발견을 유도하는 것이 시각화 시스템의 목적인데, 시각화 시스템의 입력으로 들어오는 데이터를 화면상에 어떻게 배치를 해야 시각적으로 잘 보일 수 있는지에 대한 layout 배치 알고리즘에 대한 연구는 그동안 많이 시도되었으나, 실제로 지식 발견을 위한 도구를 제공하는 연구는 잘 되어 있지 않다.

본 연구에서는 생물 의료 정보 도메인에서 정보 여과를 위해 추출된 결과를 종류별로 자동으로 분류해주고, 분류된 종류별로 시각화하는 방법을 제안하고, 지식 발견을 위한 pathway를 자동으로 찾아서 가시화하는 과정에 대해서 설명한다.

본 논문의 2절에서는 생물 의료 정보의 시각화에 관한 관련 연구를 살펴보고, 3절에서는 사용자 개별화와 지식 발견을 위한 방법을 제안한다. 4절에서는 구현을, 5절에서는 구현한 시스템의 성능에 대해서 논의한다. 마지막으로 6절에서는 결론과 향후 과제에 대해서 논의한다.

2. 관련 연구

생물 정보에서 pathway를 시각화하는 연구로는 [3], [5-8]이 있다.

[3]은 정보 추출 시스템이 추출한 정보를 시각화하고, 수정할 수 있도록 하였으며 사용자 개별화를 위한 시각화를 제공하는 시스템이다.

정보 추출 시스템이 추출한 결과가 완벽하지 않다는 것에 주목하여, 색 밴드를 이용한 막대와 중재 시각화 기법을 통해 사용자가 틀린 결과를 수정할 수 있도록 하였다. 그리고, 사용자 개별화를 위해 분자간 상호 작용 데이터의 특성에 따라 데이터를 분류하는 방법을 제안하고, 층을 사용한 시각화 기법을 제안하였다.

[5]는 관계형 SQL 데이터베이스 시스템이 핵심을 이루고 있는데, 사용자가 데이터베이스 질의를 통해 pathway 정보를 수정할 수 있도록 하였으며, 데이터베이스의 정보를 바탕으로 분자간 상호 작용 pathway를 보여줄 수 있도록 시각화 하는 모듈을 포함하고 있다. 시각화 시스템은 주로 데이터베이스와 관련된 정보의 수정과 가공을 쉽게 하기 위한 보조기능으로서 이용되고 있기 때문에, 간단한 기능만을 제공한다.

[6]은 사용자가 biological pathway를 생성하고 수정할 수 있도록 하였는데, 이는 다른 시각화 시스템과는 차별화 되는 것으로, microarray 데이터와 결합할 수 있도록 설계되었다. 특히 상호작용의 edge에 decoration라는 자질을 통해서 두 vertex 간의 연결 edge의 특성을 정의할 수 있도록 하였다.

[7]은 단백질 간의 상호 작용 pathway를 추출, 수정, 관리하는 시스템으로, MEDLINE 요약문및 pathway, 단백질-단백질 상호작용을 관리하는 Kleisli, 단백질-단백질 상호작용 정보를 추출하는 BioNLP, 추출한 정보를 시각화하는 GraphViz로 구성되어 있다.

[8]은 전문가들의 수작업을 통해 만들어진 metabolic pathway에 대한 데이터베이스로, metabolism에 대한 상세한 pathway정보가 상세한 조건 정보와 함께 들어 있다. 특히 KEGG에서는 pathway maps라는 시각화 데이터를 제공하는데, 특정 모양의 다이어그램을 이용해 서로 다른 종류의 상호작용을 표현한다.

3. 방법

생물 의료 정보의 경우에는 생명 반응이 어떤 종(species)에서 발생하는지, 어떤 생체 조건에서 발생하는지, 실험적으로 발생한 것인지(in vitro), 생체 내에서 발생하는 것인지(in vivo)와 같은 정보들이 매우 중요하다. 이러한 조건 정보들을 고려하지 않는다면 생명 반응 정보가 다른 쪽에서는 전혀 무의미한 정보가 될 수도 있기 때문이다. 따라서 다른 분야의 시각화와는 달리 생물 의료 정보의 시각화에서는 이러한 조건 정보를 명시해 줄 수 있는 방법이 필요한데, 아주 쉬운 방법으로는 모든 정보들을 화면에 나타내는 것이 있다.

그러나 이렇게 할 경우에 정보의 양이 많아지면 많아질수록 화면은 복잡해지게 되고, 그 결과 시각화를 통해 얻으려고 했던 직관적인 정보 이해를 오히려 저해하게 된다.

반면 사용자가 필요한 정보와 필요 없는 정보를 잘 여과해 주면 사용자가 복잡함을 느끼지 않고 원하는 정보를 쉽게 찾을 수 있다.

생물 의료 정보를 시각화하면 좋은 점은 pathway를 쉽게 볼 수 있다는 것인데, pathway 역시 생명 조건 정보에 따라 inhibit/decrease/suppress 등과 같은 negative한 pathway, activate/increase와 같은 positive한 pathway로 구분될 수 있기 때문에, 사용자가 일일이 생명 조건 정보를 살펴 보면서 원하는 pathway를 찾기는 쉬운 일이 아니다. 따라서 생명 조건 정보를 바탕으로 사용자가 찾고자 하는 pathway를 쉽게 가시화할 수 있는 방법이 필요하다.

3.1절에서는 개체와 개체간 상호작용의 종류, 생명 반응 조건 정보의 여과를 통해 사용자 개별화를 할 수 있는 방법을 제안하고, 3.2절에서는 사용자가 찾고자 하는 pathway를 가시화하여 지식 발견에 도움을 줄 수 있는 방법에 대해 논의한다.

3.1 사용자 개별화

시각화해야 할 정보들이 아주 많을 경우에는 모든 정보들을 다 보여주는 것은 또 다른 혼란을 야기하기 때문에, 정보를 적절히 여과해야 한다. 특히 생물 의료 정보의 시각화에서는 개체의 이름이나, 상호작용의 종류, 상호작용이 일어나는 조건들을 기술하는 내용을 같이 나타내어야 의미가 있기 때문에 개체의 수가 많아질수록 시각화의 결과는 더욱 더 복잡해지므로 정보의 적절한 여과 과정이 필수적이다. 또한 단백질, 유전자, 질병 등과 같이 서로 다른 유형의 정보들이 섞여서 나타나기 때문에, 다른 유형의 정보들을 구분해서 나타내 줄 수 있어야 한다.

다른 유형의 정보들을 시각화 하는 방법을 설명하기에 앞서, 우선 개체들의 유형을 결정해 주는 방법에 대해서 알아본다. 그림 1은 본 논문에서 정보 추출 시스템으로 사용한 BioIE[10]의 정보 추출 예제이다. Reticulons RTN3 and RTN4-B/C와, BACE1이 개체로 인식되었고, 이 둘 사이의 상호작용 키워드는 interact이다. 하지만 RTN3과 RTN4-B/C, BACE1이 단백질인지, 유전자인지, 질병인지에 대한 정보는 얻을 수 없다. 이와 마찬가지로 거의 대부분의 정보 추출 시스템이 개체에 대한 유형 정보를 알려주지 않는다.

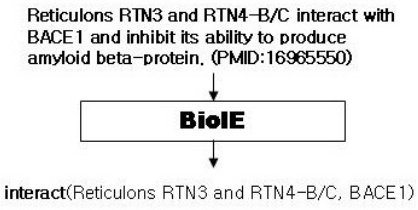

그림 1. BioIE에서의 정보 추출 과정

따라서 정보 추출 시스템의 결과를 바탕으로 시각화를 하기 위해서는 개체들의 유형을 자동으로 결정해 주는 것이 중요한 문제이다. 본 논문에서는 UMLS Semantic Network[13]과 GPSDB[12]를 이용하는 방법을 통해서 개체의 유형 구분을 한다. 이 분야의 연구는 독립된 연구 분야를 형성하고 있으므로[9] UMLS와 GPSDB 정보만을 이용하는 간단한 방식을 사용하였는데, 4절에서 보다 상세한 설명을 한다.

이렇게 정보 추출 시스템으로부터 추출된 개체들의 유형이 결정되면, 개체의 유형에 따라 다른 색상을 쓰거나, 다른 모양을 사용하거나, 아니면 다른 층(layer)에 위치하게 해서 유형의 다름을 구분한다.

그리고 화면에 출력되는 시각화 정보의 양을 조절하고, 각각의 사용자가 원하는 내용의 정보만을 볼 수 있도록 정보 여과를 할 수 있는데, 시스템 좌측의 컨트롤 메뉴에서 체크 박스를 선택하거나 취소함으로써 정보를 여과한다.

3.2 지식 발견

BRCA2는 breast cancer에 직접적인 영향을 주는 단백질인데, 병을 치료하기 위해서는 BRCA2를 inactivate 해야 한다. BRCA2 단백질을 inactivate하는 단백질을 찾기 위해서는 BRCA2와 인과관계가 있는 다른 물질들을 모두 찾아 보아야 한다.

생물 의료 분야에서 나타나는 다양한 개체간 상호작용을 pathway라고 하는데, pathway 정보를 통해서 두 개체간의 상호작용이 일어나기 위한 인과 관계 등을 알 수 있게 되므로 pathway를 잘 구축하는 것이 매우 중요하다. Pathway가 잘 구축되어 있다면 위에서와 같이 breast cancer를 유발하는 물질을 찾는 경우에 큰 도움이 된다. 이렇게 pathway 정보가 중요하기 때문에 관련 정보를 잘 시각화해줄 수 있다면 지식 발견에 큰 도움이 된다.

그림 2의 오른쪽에서와 같이, 정보의 양이 적다면 충분히 육안으로 pathway를 찾을 수 있지만, 왼쪽과 같이 정보의 양이 매우 많다면 쉽게 pathway를 찾기가 어렵다.

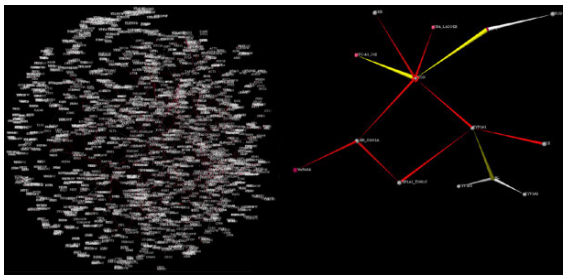

그림 2. 두 개의 시각화 예에서 Pathway 찾기

만약 그림 2의 왼쪽과 같이 매우 많은 데이터들에서라도 사용자가 찾고자 하는 pathway를 edge의 굵기 조절이나 색상 조절을 통해 강조해서 보여준다면, 사용자는 원하는 pathway를 좀 더 쉽게 찾아볼 수 있을 것이다. 이와 같은 이유에서 본 연구에서는 사용자가 찾고자 하는 pathway를 찾아주는 모듈을 추가하였고, 찾은 pathway를 가시화하는 방법을 사용하였다.

본 논문에서는 pathway를 찾기 위하여 다음 두 가지 방법을 제안한다.

첫번째 방법은 사용자가 관련된 pathway를 알고 싶어하는 두 개체를 선택하면 시스템이 pathway를 찾아주는 것이다. 이 방법은 두 개체간의 상호작용을 중재하는 중간 물질이 무엇인지 찾고자 할 때에 도움을 줄 수 있다.

두번째 방법은 사용자가 한 개체를 선택하면, 그 개체와 관련 있는 pathway를 모두 찾아주는 것인데, 이 방법은 앞의 breast cancer 예와 같이 특정 현상을 일으키는 모든 원인이 되는 물질들을 찾는 데에 이용될 수 있다.

그래프에서 경로를 찾는 것이라면, 첫번째 경우에는 그래프의 경로를 찾아주는 알고리즘으로 풀 수 있고, 두번째 경우에는 spanning tree를 찾아주는 방식으로 풀 수 있다. 그렇지만 생물 의료 정보에서는 조건 정보가 매우 중요하기 때문에, 이러한 pathway를 찾아줄 때에는 조건 정보를 고려해 주어야 한다. 예를 들어 A activate B와 B activate C의 경우에는 인과관계 상으로 A activate C가 성립하지만, A inhibit B와 B activate D의 경우에는 A와 D의 관계가 애매하다. 따라서 이런 경우에 조건 정보들을 고려한 pathway finding을 해야 한다.

표 1

Interaction 1	Interaction 2	Transitive
A activate B	B activate C	A activate C
A increase B	B increase C	A increase C
A activate B	B inhibit C	A inhibit C
A increase B	B decrease C	A decrease C
A inhibit B	B activate C	-
A decrease B	B increase C	-

A inhibit B	B inhibit C	-
A decrease B	B decrease C	-

표 1에서 일부 보인 것과 같이, transitive한 관계가 확실하지 않은 것은 pathway상으로는 연결되어 있더라도, pathway finding시에 제외하고 pathway를 찾는다. 여기에서는 도메인에 종속적이고, logic과 생명 반응이 일어나는 조건 등에 대한 고려를 해 주어야 하는데, 본 연구에서는 간단히 active, inhibit, induce, increase, decrease, suppress에 대한 관계만 고려하였고, 나머지에 대해서는 향후 연구에서 해결하려고 한다.

지식 발견에 도움을 줄 수 있는 또 다른 방법으로는 조건 정보를 고려하지 않고 pathway를 재구성하는 것이 있다.

예를 들어서 효모에서의 분자간 상호작용 반응과 사람에서의 분자간 상호작용 반응이 비슷한 경우가 종종 나타나는데, 효모에서의 분자간 상호작용은 많이 알려져 있는 반면에 사람에서의 분자간 상호작용은 많이 알려져 있지 않다. 이렇게 정보 추출 시스템의 성능 문제나, 알려진 생물학적 사실의 부족으로 정보 추출 시스템의 입력이 되는 데이터의 부족으로 인해 pathway가 완벽하게 구축되지 않는 경우가 많다.

따라서 사람에서의 분자간 상호작용을 알기 위해서는 효모에서 알려진 분자간 상호작용 데이터를 참고한다면 새로운 분자간 상호작용을 찾아 내는 데에 도움이 될 수 있다.

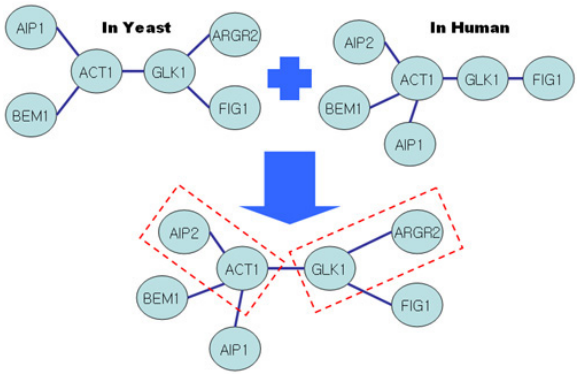

그림 3. 조건 정보에 따른 pathway 재구성

그림 3과 같이 일반적으로 pathway는 조건 정보를 포함하게 된다. 하지만, 그림 3과 같이 조건 정보를 고려하지 않고 pathway를 재구성한 결과, 점선 박스 내 상호작용과 같은 새로운 관계가 생겼음을 볼 수 있다. 실제로 human에서도 GLK1과 ARGR2가 상호작용을 하는데, 우리는 이러한 방법을 통해 사전에 파악하지 못했던 GLK1과 ARGR2의 상호작용이 사람에서도 일어남을 추측할 수 있다.

4. 구현

그림 4는 구현한 시스템의 개략적인 구조와, 실행 순서를 나타낸 것이다. 시스템은 크게 데이터 생성 부분과, 시각화 처리 부분으로 나뉜다. 본 논문에서 제안하는 부분은 데이터 생성 부분의 일부와 시각화 처리 부분을 포함한다.

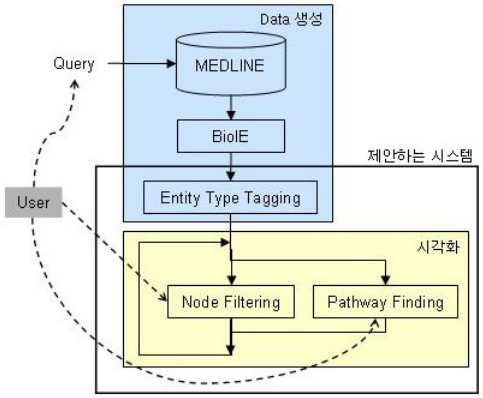

그림 4. 시스템 구현도

PubMed에서 사용자가 찾고자 하는 키워드로 MEDLINE 검색을 하면, PubMed는 질의한 내용이 포함된 MEDLINE a요약문을 돌려준다. 이렇게 수집된 문헌정보들은 BioIE의 입력으로 제공되는데, BioIE는 activate, inhibit, bind 등의 키워드를 먼저 찾아낸 다음 단백질, 유전자 이름이 포함된 명사구를 찾아내는 양방향 점진 파싱 기법을 사용하여 개체간 상호 작용을 추출한다.

그림 1에서 볼 수 있듯이, BioIE의 추출 결과에서 상호작용의 종류와 상호작용에 관여하는 개체의 이름은 알 수 있지만, 그 개체가 단백질인지, 유전자인지, 단순한 화학물질인지 알 수 없다. 주로 단백질-단백질 상호작용이 주를 이루므로 단백질일 가능성이 많지만, 그룹으로 나누기 위해서는 이것의 유형을 결정해 주는 문제가 본 연구에서는 아주 중요하다.

단백질, 유전자 등과 같은 개체는 많은 동의어를 가지고 있기 때문에, 해당 유형을 찾기에 앞서서 동의어 문제[9]를 해결해 주어야 하는데, 대부분의 경우에는 UMLS Metathesaurus를 이용하면 되지만 UMLS에 나타나지 않는 개체들도 많이 있기 때문에 이 때에는 GPSDB를 이용한다.

동의어 문제를 해결한 후에는 개체가 어떤 타입인지를 알아내야 하는데, 역시 UMLS Metathesaurus를 검색해 보면 알 수 있다.

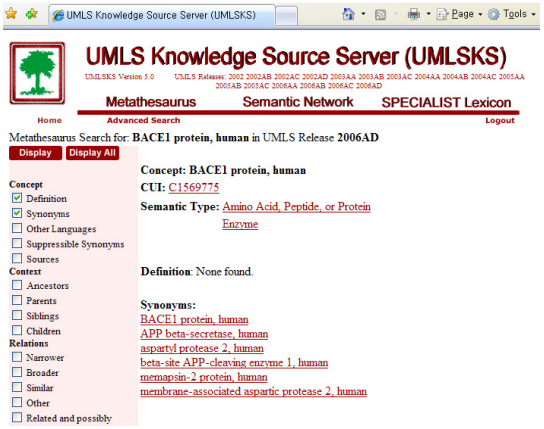

그림 5. UMLS 검색 예제

앞의 그림 1에서 BioIE에서 얻어온 개체 BACE1이 어떠한 유형인지 UMLS Metathesaurus에서 검색한 결과의 스크린샷은 그림 5와 같다. 그림에서 Semantic Type을 보면 BACE1은 Protein임을 알 수 있다. 마찬가지로 RTN3과 RTN4-B/C 역시 UMLS로 검색해 보면 단백질임을 알 수 있다. 따라서 BioIE에서 뽑은 결과는 단백질-단백질 상호작용이라는 것을 알 수 있게 된다.

이렇게 얻어진 상호작용 관계를 시각화의 좌표 계산 모듈의 입력으로 넣으면, 3차원 공간에 배치할 수 있도록 좌표값을 내어 준다. 3차원 좌표값을 계산하는 것은 Fruchterman Reingold algorithm을 사용한다[11]. 이 알고리즘은 Spring-embedder 모델을 기반으로 한 것으로, 개체들을 질량으로, 개체들 간의 관계를 용수철로 생각하고 이들간의 힘이 최소화 되는 배치를 한다. 이렇게 계산된 좌표값을 바탕으로 Open Inventor를 이용해 3차원으로 보여주게 된다. 그리고 사용자의 선택에 따라 실시간으로 좌표 값을 재계산한다.

5. 실험

본 절에서는, 3절에서 제안한 방법에 대해 시스템을 구현하고, 실험을 하였다. 시각화 데이터로는 MEDLINE에서 protein, gene, disease, interaction, drug로 검색해서 658개의 MEDLINE Abstract를 얻어왔고, BioIE를 통해서 추출한 1,155개의 interaction 정보를 사용하였다. BioIE에서 뽑아낸 데이터가 정확하지 않을 수 있으므로 수작업으로 일부 데이터 보정을 하였다.

그림 6과 그림 7은 사용자 개별화에 대한 스크린샷이다. 사용자는 프로그램 좌측의 체크 박스를 선택해서 정보 여과를 통해 자신이 보고자 하는 유형의 정보만을 볼 수가 있다. 그림 6은 정보 여과를 하지 않은 상태인데, 너무 많은 정보가 한꺼번에 화면에 나타나기 때문에 원하는 정보를 찾기가 매우 힘들다. 그림 7은 그림 6과 동일한 데이터에서, 사용자에게 필요한 정보만 여과해서 보여주는 예이다. 분자간 상호작용의 관계가 “activate”, “bind”, “inhibit”인 것들만 시각화 하였는데, 그림 6과는 달리 사용자의 관심 밖의 정보들이 많이 제거되어 훨씬 간결해졌다는 것을 알 수 있다. 이렇게 간결화된 정보를 바탕으로 사용자는 자신이 원하는 정보를 찾을 수 있다. 이렇게 실시간 정보 여과를 통해서, 시각화 시스템이 안고 있는 정보의 출력 문제점[2]을 어느 정도 해소할 수 있고, 따라서 사용자의 이해도를 더 높일 수 있다.

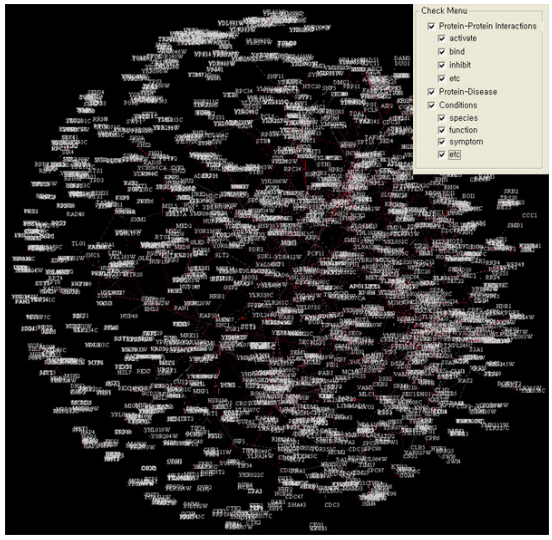

그림 6. 사용자 개별화 1

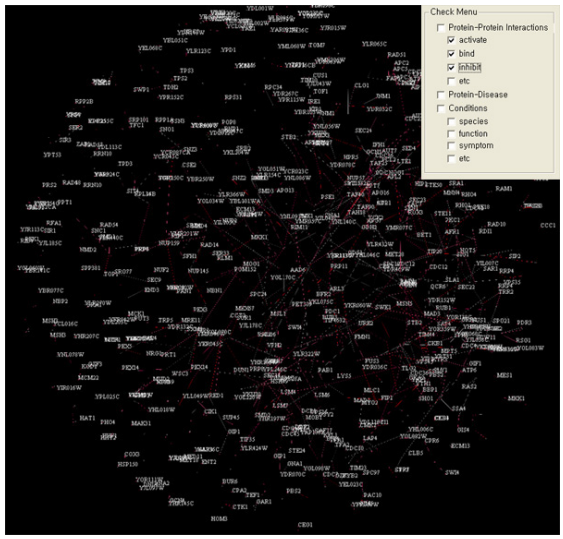

그림 7. 사용자 개별화 2

생물 정보에서는 사용자가 정보 탐색의 범위를 좁혀감에 따라 세부적인 내용을 표시해 주어야 하는 경우가 있는데, 대표적인 예가 단백질-단백질 복합체와의 상호작용이다.

본 논문에서는 E3 ubiquitin ligase의 하나인 SCF complex protein으로 실험을 하였는데, SCF complex protein은 3개의 sub-protein (Skp1, Cullin, F-box)으로

구성되어 있다. 특정 sub-protein이 도메인[1]을 형성하므로 이러한 도메인 정보를 표현해주는 것이 매우 중요하다.

즉, 문헌상에 어떤 물질 A 와 F-box protein이 상호작용을 한다고 나와 있다면, A ↔ F-box protein으로만 나타내기 보다는 F-box protein이 SCF complex protein의 sub-protein이라는 정보도 함께 나타내는 것이 좋다. 다음 그림 8과 그림 9에서 이러한 것들을 구현한 예를 보여 준다.

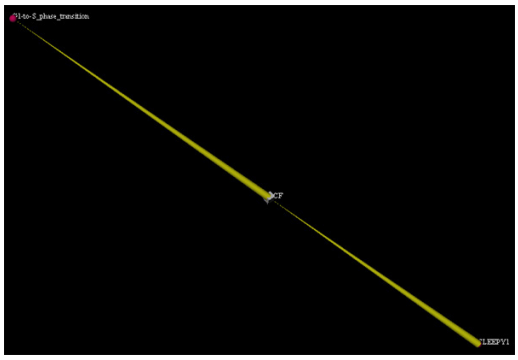

그림 8. 단백질 복합체와의 상호작용

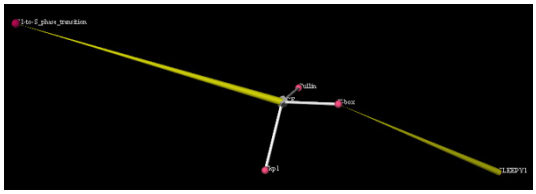

그림 9. 단백질 복합체와의 상호작용 확대

그림 8은 SCF complex protein과 다른 것들과의 상호작용을 나타낸 것이다. 웹 지도에서 확대를 하면 상세한 정보가 보이듯이, 여기에서도 확대 요청에 따라 보여주는 정보의 상세한 정도를 다르게 하였다. 그림 8은 단백질 복합체의 구성이 어떻게 되어 있는지와 같은 상세한 정보까지는 다루지 않는다.

그림 9는 그림 8에서 더 확대를 하여, 보다 상세한 정보를 표현한 예이다. 다음의 MEDLINE 요약문의 일부분 “The Arabidopsis SLEEPY1 gene encodes a putative F-box subunit of an SCF E3 ubiquitin ligase. [PMID:12724538]”에서, SLEEPY1 gene이 F-box protein을 encode한다고 했다. 따라서 그림 9는 SLEEPY1과 SCF complex protein의 subcomponent인 F-box protein과 연관이 있음을 나타낸다.

[1] 다른 물질과의 결합 부위

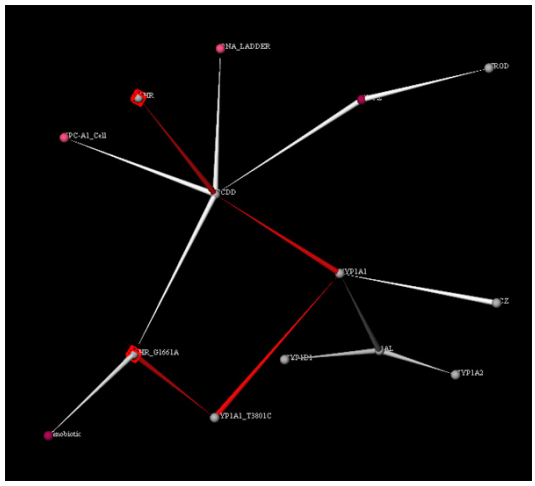

그림 10. 두 개체간의 Pathway 계산

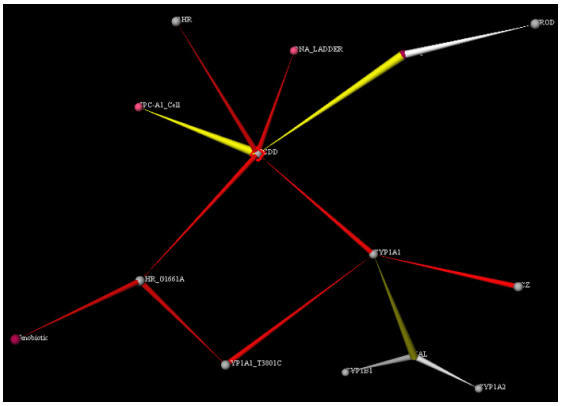

그림 11. 선택한 한 개체와 연결된 Pathway 계산

그림 10과 그림 11은 지식발견을 돕기 위해 사용자가 찾고자 하는 개체간의 pathway를 계산해 주는 과정을 보여준다. 3절의 표 1에서 논의한 기준을 바탕으로 transitive rule을 통해 가능한 pathway를 찾아주는데, 그림 10은 사용자가 선택한 두 개체의 pathway를 보여주는 예이다. 그림 10에서 붉은색 사각형 박스로 둘러 쌓여 있는 두 개의 노드가 사용자가 선택한 노드이고, 붉은색 연결선은 activate/bind와 같은 positive한 pathway를 나타내고, 푸른색은 inhibit/suppress와 같은 negative한 pathway를 나타내도록 하였다. 그림 10에서는 AHR과 AHR G1661A 사이의 pathway를 찾으라는 요구를 하였는데, AHR 1661A → CYP1A1 T3801C → CYP1A1 → TCDD → AHR로 positive pathway가 찾아졌음을 알 수 있다.

그림 11은 지식발견을 돕기 위해 pathway를 찾아주는 또 다른 방식으로, 사용자가 원하는 개체를 선택하고, 그 개체에 영향을 주는 모든 pathway를 찾고자 할 수가 있다. 이 경우 표 1의 규칙에 따라서 spanning tree를 구성한다. positive pathway인 경우에는 붉은색으로, negative pathway인 경우에는 노랑색으로 연결선을 표시한다. 이

렇게 함으로써 사용자는 선택한 개체에 negative하게 영향을 주는 pathway와positive하게 영향을 주는 pathway를 찾을 수 있게 된다.

6. 결론 및 향후 계획

방대한 양의 데이터를 시각화해야 하는 필요성에 대해서는 문제에 대해서는 많은 사람들이 동의하고 있지만, 문제는 이 방대한 양의 데이터를 어떤 시각화 기법을 이용하여야 하느냐는 것이다. 이렇게 필요한 데이터들만 걸러서 보여주는 것을 정보 여과라고 하는데, 본 연구에서는 각 데이터들의 타입을 자동으로 결정하고, 타입별로 사용자가 메뉴 선택을 해서 여과를 할 수 있도록 하였다.

그리고, 본 연구에서는 이러한 시각화 시스템이 단순히 pathway를 보여주는 것에 그치는 것이 아니라 어떻게 지식 발견에 도움을 줄 수 있는지에 대해서 activate/inhibit과 같은 interaction keyword를 통해서 가능한 pathway를 찾아주는 방법에 대해서도 논의하였다.

향후 계획으로는 서로 다른 종류의 생명 조건 정보에 대해서 어떤 시각화 기법을 사용해서 표현 하는 것이 좋은지에 대한 연구와, 지식 발견을 위한 pathway를 찾아주는 데에 있어서 생물학적인 지식과 logic을 통해서 좀 더 정확한 pathway를 찾아주는 연구가 있다.

Acknowledgments

본 연구는 첨단정보기술 연구센터와 특정기초연구개발사업을 통하여 과학재단의 지원을 받았음.

참고문헌

[1] Ben Shneiderman, “Inventing discovery tools: combining information visualization with data mining,” Information Visualization, Vol.1, Issue 1, pp. 5-12, 2002.

[2] Chaomei Chen, “Top 10 Unsolved Information Visualization Problems,” IEEE CG&A, Vol. 25, Issue 4, pp. 12-16, 2005.

[3] Changsu Lee, Jinah Park and Jong C. Park, “Mediatory Visualization for Structured Data and Textual Information,” The 3rd IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP 2003), pp. 926-932, Benalmadena, Spain, 2003.

[4] Peter Uetz, Trey Ideker and Benno Schwikowski, “Visualization and integration of protein-protein interactions,” in Protein-Protein Interactions: A Molecular Cloning Manual, E. Golemis ed., CSHL Press, Cold Spring Harbor, N.Y., 2002.

[5] Alexander Lüdemann, Daniel Weicht, Joachim Selbig and Joachim Kopka, “PaVESy : Pathway Visualization and Editing System,” Bioinformatics, Vol. 20, No. 16, pp. 2841-2844, 2004.

[6] Matthew Holfold, Naixin Li, Prakash Nadkarni and Hongyu Zhao, “VitaPad: visualization tools for the analysis of pathway data,” Bioinformatics, Vol. 21, No. 8, pp. 1596-1602, 2005.

[7] Limsoon Wong, “PIES, a Protein Interaction Extraction System,” Proceedings of Pacific Symposium on Biocomputing, pp. 520-531, 2001.

[8] Minoru Kanehisa and Susumu Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” Nucleic Acids Research, Vol. 28, No. 1, pp. 27-30, 2000.

[9] Sophia Ananiadou and John McNaught, Text Mining for Biology and Biomedicine, Artech House, 2006.

[10] Jung-jae Kim and Jong C. Park, “BioIE: Retargettable Information Extraction and Ontological Annotation of Biological Pathways from the Literature,” Journal of Bioinformatics and Computational Biology (JBCB), Vol. 2, No. 3, pp. 551-568, 2004.

[11] Thomas M. J. Fruchterman, Edward M. Reingold, “Graph Drawing by Force-directed Placement,” Software-Practice and Experience, Vol. 21, pp. 1129-1164, 1991.

[12] Violaine Pillet, Marc Zehnder, Alexander K. Seewald, Anne-Lise Veuthey and Johann Petrak, “GPSDB: a new database for synonyms expansion of gene and protein names,” Bioinformatics, Vol. 21, No. 8, pp. 1743-1744, 2005.

[13] http://www.nlm.nih.gov/research/umls/