단백질 상호작용 데이터 통합 및 자료 검색 시스템 설계

최지혜*, Bayarsaikhan Itgel*, 오세종*
*단국대학교 대학원 나노바의오의과학과
e-mail : jiheachoehihi@hanmail.net

Integration of Protein-Protein Interaction Data and Design of Data Search System

Ji-Hye Choi*, Bayarsaikhan Itgel*, Se-Jong Oh*
*Dept of NanoBioMedical Science, Dankook University

요 약

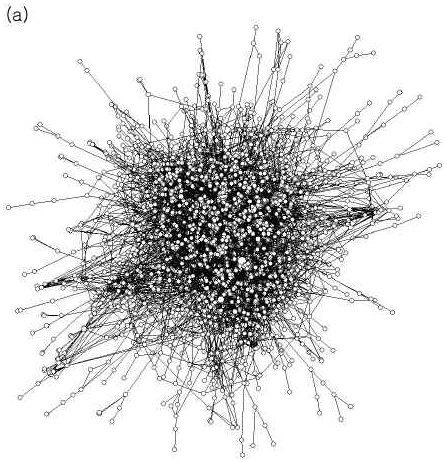
Post-genomic 시대에 접어들면서 단백질의 기능의 주석이 중요한 문제로 떠오르기 시작하였다. 이런 단백질 기능을 예측하기 위해 단백질 상호작용(Protein-Protein interaction) 데이터를 이용한 방법들이 지난 10여 년간 발표되어왔다. 단백질 상호작용(Protein-Protein interaction) 데이터는 단백질들 간의 서열 등의 특징을 이용해 상호간의 연결 관련성이 있는 단백질끼리의 관계를 네트워크로 나타낸 자료이다. 현재 이러한 단백질 상호작용(Protein-Protein interaction) 데이터들은 MIPS, DIP, BioGrid등 약 5~6군데에서 제공되고 있다. 각각의 데이터는 다른 형식을 가지고 있고, 중복되는 정보도 포함하고 있다. 여러 연구 방법에서 데이터를 사용할 때 한군데에서만 추출하기 보다는 여러 데이터에서 추출하는 경우가 많기 때문에 다른 형식의 데이터를 이용하는데 불필요한 수고가 들어가게 된다. 때문에 여러 군데의 데이터를 한 가지 형식으로 맞추어 통합적으로 구축하여 연구 시 데이터 사용에 용이하도록 설계 하였다. 또한 발표된 단백질 기능 예측 방법에 대한 정리를 통해 앞으로의 연구를 하는데 있어서 필요한 자료를 얻고 열람할 수 있도록 설계하였다. 이를 통해 관련 연구를 하거나 관심이 있는 사람들의 데이터를 검색하는데 많은 도움이 될 것이다.

1. 서론

최근 생물정보학 분야의 성장에 따라 생물학적 정보의 양은 급속도록 증가하게 되었다. 이런 정보들을 이용해 보다 빠르고 효과적으로 분석하기 위한 기술들이 요구, 개발되고 있다[1].

이러한 정보들을 이용해 기능이 알려지지 않은 단백질의 기능을 예측 하는 것이 중요시 되고 있다. 현재까지의 단백질 기능을 밝히기 위한 방법은 주로 생물학적 실험에 주로 의존하고 있다. 하지만 이러한 실험을 하기 위해서는 많은 비용과 시간이 요구된다. 따라서 연구자들은 다양한 실험적인 방법과 전산학적 방법들을 사용하여 단백질의 기능 예측을 하려고 하고 있다. 단백질의 기능 예측을 위해 사용하는 대표적인 데이터로는 단백질 서열, 단백질 상호작용, 유전자 발현 및 구조 데이터 등이 있다.

생명체 내에서 일어나는 대부분의 생명현상은 여러 단백질들이 복합적으로 상호작용함으로서 발생된다. 단백질들은 서로 매우 복잡한 상호작용 관계를 형성하는데 이들 전체 단백질들의 상호작용 관계를 연결하면 하나의 거대한 네트워크를 형성한다. [그림 1]은 단백질 상호작용 네트워크를 나타낸다[2]. 이러한 단백질 상호작용은 기능과 밀접한 관계가 있다. 따라서 이러한 단백질 상호작용(Protein-Protein Interaction) 네트워크를 이용한 단백질 기능 예측에 초점이 맞춰지고 있는 추세이다.



[그림 1] 단백질 상호작용 네트워크

현재 상호작용하는 단백질 쌍들에 대한 대표적인

데이터베이스들은 DIP (Database of Interacting Proteins), GRID (General Repository for Interaction Datasets), MIPS (Munich Information Center for Protein Sequences) 등이다. 이 중 DIP는 단백질 상호작용 데이터베이스 중에서 가장 대표적인 것으로 현재 15,114개의 단백질 쌍들에 대한 데이터를 갖고 있다. BIND는 상호작용하는 단백질 쌍 이외에도, 분자 복합체 및 대사경로 등에 대한 정보들도 포함하고 있는데, 현재 11,237개의 단백질 쌍 엔트리를 포함하고 있다[3]. MIPS는 Yeast에 대한 상호작용 데이터를 가지고 있으며 3,050개 쌍에 대한 정보를 갖고 있다.

이러한 단백질 상호작용 데이터들은 단백질 기능 예측에 주로 이용되고 있다. 예측 방법에 대한 개발 시 한가지의 데이터를 사용하기 보다는 여러 데이터들을 취합해서 이용하게 된다. 이는 각 데이터들을 도출해 내는 방법들도 각각 다르고 그 방법들에 의해 나온 데이터들은 부정확한 데이터들도 포함하고 있게 되는데 여러 가지의 데이터에서 중복되는 단백질들을 실험군으로 하게 되면 실험에 대한 정확도를 높일 수 있기 때문 이다. 그러나 이 여러 데이터들은 제각각 형식이 다르고 가지고 있는 내용 구성도 다르게 되어있다. 때문에 이를 실험에 사용하기 위해서는 각각마다 전처리를 해야 하는 불필요한 수고가 들어가게 된다. 또한 각 데이터베이스마다 사용하는 단백질 ID 체계가 다르기 때문에 이 또한 동일시 시켜주는 작업도 필요하게 된다. 이러한 작업은 본 개발 방법에 대한 실험을 하는데 있어 시간 낭비를 초래하게 된다. 따라서 본 논문에서는 이러한 데이터베이스들의 이용에 용이하기 위한 통합 설계를 제시하였다.

또한 기존보다 나은 개발을 하기 위해서는 기존에 나와 있는 방법에 대한 충분한 이해와 앞선 방법보다 나은 정확도를 가지고 있는지에 대한 비교분석이 필요하다. 이러한 기존에 나와 있는 대표적인 방법들에 대한 데이터들을 효율적으로 열람할 수 있고, 비교분석 할 수 있는 시스템을 제시하였다.

2. 본론

2.1 단백질 상호작용 통합 데이터베이스

단백질들 간의 상호작용 데이터를 제공하는 데이터베이스들 중 대표적인 DIP, GRID, MIPS를 이용하여 설계하였다. 설계의 목적은 각기 다르거나 중복된 단백질을 가지고, 다른 형태로 이루어진 이 데이터베이스들을 한 가지 형태로 통합하여 단백질 기능 예측 기술을 개발하는데 용이하도록 하는 것이다.

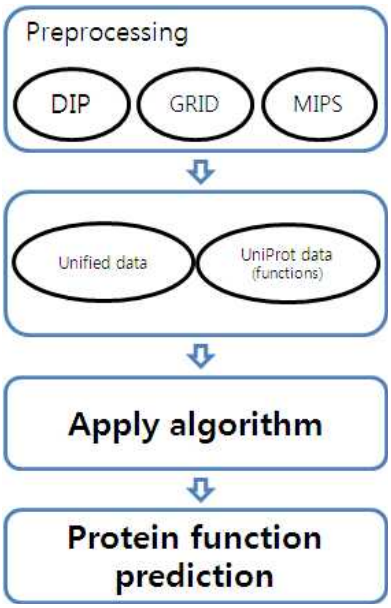
단백질 기능 예측을 위해 필요한 정보는 어느 단백질끼리 상호작용을 하는지와 연결되어 있는 각 단백질들이 가지고 있는 알려진 기능들이다. 상호작용 네트워크 데이터에서는 어느 단백질끼리 연결되어 있는지에 대한 정보를 가져오면 된다. DIP는 단백질 상호작용 네트워크는 단백질을 나타내는 node와 상호작용을 나타내는 edge가 고유의 ID를 가지고 구성하고 있다. edge의 정보에 어떤 단백질이 연결되어있는지 나와 있기 때문에 이것을 이용하게 된다. 부수적으로 다른 단백질 정보 데이터베이스들에서 사용하는 ID를 제공하고 있다. DIP는 고유의 ID를 사용하기 때문에 통합을 위해 공통으로 사용하는 ORF 형식으로 바꾸어 전처리를 해야 한다. GRID와 MIPS는 곧바로 어떤 단백질들이 연결되어 있는지 그리고 이 연결을 알아내기 위해 사용된 방법 등으로 구성이 되어 있다. 이 두 데이터베이스에는 단백질ID에 ORF형식을 제공하고 있다.

이러한 데이터베이스들의 구성을 이용하여 만들 통합 데이터베이스의 구조는 [표 1]과 같다.

[표 1] 통합 데이터의 구조

Protein 1	Protein 2	Database
P1	P2	DIP
P1	P2	MIPS
P1	P4	DIP
P2	P3	GRID
⋮	⋮	⋮

Protein 1과 Protein 2는 서로 상호작용 하고 있고 이 정보가 어느 Database에서 제공 되었는지를 표시하고 있다. 중복된 상호작용은 제거하지 않고 표시해 줌으로써 한가지의 데이터베이스에서만 나온 정보인지 그 이상의 곳에서 중복이 되는 것인지를 알 수 있다. 기본적으로 Protein 1과 Protein 2는 오름차순으로 통일한다. 이렇게 데이터베이스를 통합해 줌으로써 최종 상호작용 개수는 각 데이터베이스들과의 카디널리티 곱의 수과 상응한다. 이 데이터베이스를 이용해 보다 신뢰도 높은 데이터셋을 손쉽게 구성할 수 있다. 또한 기능 예측을 위해서는 각 단백질들이 어떠한 기능을 가지고 있는지 알아야 한다. 이러한 정보를 가지고 있는 uniProt 데이터베이스와 연동 하여 사용할 수 있다.



[그림 2] 단백질 기능 예측 전체 과정

단백질 상호작용 데이터를 이용하여 단백질 기능 예측을 하기 위해 필요한 정보인 어느 단백질끼리 상호작용 하고 있는지와 각 단백질이 가지고 있는 기능들을 알면 된다. 때문에 통합 데이터베이스에서 정확도 높은 상호작용 데이터를 효율적으로 추출하여 uniProt 데이터베이스에서 단백질이 가지고 있는 기능 정보를 추출하여 사용할 수 있다. 통합 데이터베이스를 이용하여 단백질 기능 예측을 하는 전체 과정이 [그림 2]에 있다.

2.2 단백질 기능 예측 자료 검색 시스템

단백질 상호작용 네트워크를 이용해 단백질의 기능 예측 분야는 2000년도의 neighbor counting 방법부터 시작하여 현재까지 주목되고 있다[4]. 약 10년간 기능 예측으로 제시된 방법도 많고 관련 자료들 또한 방대하게 존재하고 있다. 새로운 방법을 도출하기 위해서는 무엇보다 기존의 방법들에 대한 파악이 중요하다. 이를 위해 방대한 자료들을 사용자 기반으로 축적하여 효율적인 자료 검색을 도와주는 시스템을 설계하였다.

축척 시스템의 기본 데이터는 단백질 기능 예측을 다루는 논문들과 추가적으로 논문에 기제 된 방법에 사용된 개념들의 백과사전 지식으로 이루어진다. 검색엔진과 같이 키워드를 입력하면 관련 자료들이 검색 된다. 이 과정에서 가산점 부여 방법을 이용해 보다 효율적인 검색을 할 수 있다.

가산점 부여 방법은 보다 비중이 있게 다루어지는 자료에 높은 가산점을 부여하여 검색 시 높은 가산점의 자료 우선으로 찾아 볼 수 있는 방법이다. 높은 가산점을 부여하게 되는 기준은 세 가지가 있는데 다음과 같다. 첫째로 현재까지 나온 단백질 기능 예측을 위한 논문들 은 이 분야의 나무 기둥이라고 할 수 있는 방법들이 있고 이를 중심으로 잔가지 같은 방법들이 펼쳐져 있다. 이러한 구조를 고려하여 주요 방법을 다루고 있는 논문에는 높은 가산점을 부여하고 부수적으로 나온 방법들을 다룬 논문에는 보다 낮은 가산점을 부여하게 된다. 이는 각 논문의 참고문헌에 등록 된 것을 기준으로 판단하게 된다. 둘째로 어떤 분야의 기둥이 되는 방법이나 개념은 다른 많은 논문들의 참고문헌에 올라오게 된다. 이러한 성질을 가산점 부여의 기준으로 이용하였다. 또한 검색된 데이터 중 열람된 논문이나 자료는 더 중요도가 있다고 판단하여 가산점을 부여하게 된다. 마지막으로 부여하는 기준은 논문의 제목과 요약 글에 키워드와 일치하는 단어가 많은 순서대로 가산점을 부여하게 된다. 일반적으로 논문의 핵심을 다루는 내용은 제목과 요약 글 내에 축약되어 있기 때문에 이를 이용하였다. 이러한 세 가지 기준으로 가산점이 부여하고 중요도 순서대로 검색결과가 나타나게 된다.

3. 결론

최근 10년 동안 단백질과 단백질간의 상호작용을 나타내는 데이터를 이용하여 단백질의 기능을 찾아내는 방법이 연구되어 왔다. 새로운 방법을 찾아내는데 겪는 불편한 점을 줄이기 위한 두 가지 시스템을 설계하였다. 첫째로 여러 군데에서 제공하는 단백질 상호작용 데이터베이스들의 형식과 구성이 다른 점과 필요한 데이터만을 손쉽게 이용할 수 있도록 통합 데이터베이스를 구축하였다. 그리고 수많은 단백질 기능 예측 방법들을 축적하여 효율적으로 검색할 수 있도록 하였다. 이를 이용해 앞으로 단백질 기능 예측 방법을 찾는 분야의 연구를 도울 수 있다.

참고문헌

[1] Tae Ho Kang, Jea Woon Rye, "Protein Function Finding Systems through Domain Analysis on Protein Hub Network", 한국콘텐츠학회논문지, Vol. 8 No. 1, 2008.

[2] Jea Woon Ryu, Hak Yong Kim, “prediction of unannotated proteins from protein interaction network filtered by using localization and domains in yeast”, Journal of the Korean Physical Society, Vol. 51, No. 5, November 2007.

[3] Ki Bong Kim, "Preotein Function Analysis System Using Protein Interaction and Domain Information", NuriMedia Co., 2005.

[4] Benno Schwikowski, Peter Uetz, “A network of protein - protein interactions in yeast”, Nature Biotechnology, Vol 18, December, 2000.