단백질의 기능 예측을 위한 도메인 검색 방법

허미영, 김홍기, 최진성
한국전자통신연구원

Search method of Domain for prediction of protein function

Meeyoung Huh, Hongkee Kim, Jinsung Choi
Electronics and Telecommunications Research Institute
E-mail : hmy63069@etri.re.kr

Abstract

모든 생명체는 유전자의 최종 산물인 다양한 단백질들이 각각의 복잡한 기능을 수행함과 동시에 그들 사이의 긴밀한 상호작용에 의해 생명을 유지한다. 도메인 (Domain)은 단백질의 기능적 단위로서 한 개 단백질은 최대 수십 개의 도메인을 가지는데 이들 도메인에 대한 정보는 단백질의 기능을 예측하는데 도움이 될 수 있다.

본 논문에서는 종양을 억제하는 기능을 가지는 단백질과 그러한 기능을 가질 것으로 추정되어지는 단백질의 아미노산 서열, 또 기능이 밝혀지지 않은 미지의 아미노산 서열을 가지고 이미 밝혀져 있는 도메인 서열과 비교 검색하여 이들 사이에 일치하는 도메인을 통하여 표적 단백질의 기능 동정에 관한 연구에 도움이 되며, 또한 기능이 밝혀지지 않은 아미노산 서열의 도메인을 검색하여 새로운 기능을 예측함으로써 다른 실험적 방법과 비교하여 시간과 비용을 절약할 수 있는 효과적인 방법을 얻었기에 제안하고자 한다.

I. 서론

생명체의 생명 유지와 관련된 정보는 유전자가 가지고 있다. 유전자의 DNA 서열은 전사라는 과정을 통해 아미노산 서열을 만들어 단백질로 유전정보를 전달하게 되고 단백질은 그 기능을 수행하게 된다. 단백질은 생체 내 유전자의 최종 산물로서 생명유지에 필수적이며 매우 다양한 기능을 담당하는 물질이다. 인간의 유전자의 수는 약 3-4 만개로 알려져 있다. 단백질은 단독으로 고유한 기능을 수행하기도 하지만 많은 단백질들은 다른 단백질들과 유기적으로 연관되어 그 기능을 나타낸다.

단백질 내부를 세부적으로 구분하게 되면 더 작은 기능적 단위로서 비교적 짧은 아미노산 서열을 가리키는 도메인(domain)으로 나누어질 수 있다. 도메인은 하나 혹은 그 이상의 구조적 의미의 패턴(pattern)이 모여 이루어진 독립적인 단위체로 독자적인 고유의 기능을 갖는다. 단백질은 모두 약 천개 미만 정도의 기능적 도메인으로 이루어져 있을 것으로 추정하고 있다. 개개의 단백질에는 최대 수십 개의 기능적 도메인이 있을 것으로 예상된다.

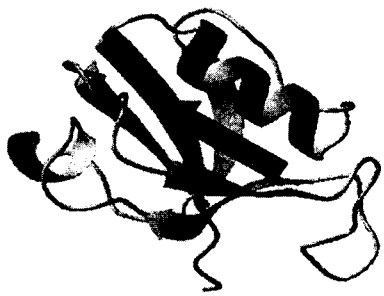

그림 1. 도메인 구조 (PDB ID: 1IU2). 각각 다른 색깔을 가지는 서열들이 별개의 도메인을 나타낸다.

최근 휴먼게놈 프로젝트의 완성 이후로 폭발적인 대량의 데이터가 생산되고 있다. 이런 데이터를 가지고 연구하는 목적은 궁극적으로 질환 연구 또는 신약 개발을 위해서인데 이를 위해서는 질환과 관련된 표적 단백질의 기능에 대한 연구가 선행되어야 한다. 단백질의 기능을 알기 위한 한 가지 방법이 도메인을 파악하는 것이다. 도메인에 대한 정보를 얻게 되면 단백질의 기능을 예측할 수 있기 때문이다. 그러나 실험적으로 도메인을 찾는 방법은 많은 시간과 비용이 소요되므로 이를 해결하기 위한 방법 중 하나가 표적 단백질의 서열 정보만을 가지고 이미 알려져 있는 도메인들을 검색하여 일치하는 도메인을 얻는 것이다. 본 논문에서는 항암 작용을 나타내거나 그러한 기능을 가지는 것으로 추정되는 단백질들의 서열을 가지고 이미 기능이 밝혀진 도메인을 검색하여 표적 단백질에 대한 추가적인 정보를 얻을 수 있는 방법을 얻었기에 소개하고자 한다.

II. 실험방법

프로그램에 사용된 도메인 데이터베이스는 SIB (Swiss Institute of Bioinformatics)에서 제공하는 것으로 기존의 데이터베이스에 있는 단백질 서열들로부터 생물학적으로 의미가 있는 부위나 패턴들을 찾아내 패턴별로 모아 만든 데이터베이스인 PROSITE 를 이용하였다. 단백질의 아미노산 서열 데이터는 미국 국립 생명공학 정보센터(NCBI)의 데이터베이스 GenBank 와 SIB 의 SWISS-PROT 에서 제공하는 데이터를 사용하였다. 실험데이터로는 결장암의 종양형성 억제 단백질 Serine-threonine protein kinase (SWISS-PROT accession number: Q9P114), 세포 주기 조절과 종양 억제 단백질의 작용시 중요한 역할을 하며 유전자의 전사에 관여하는 단백질 E2F Transcription factor 1 (GenBank accession number: NP_005216), 종양 괴사에 관여하는 단백질 TNF receptor-associated factor 2 (NP_663770), 백혈병과 관련하여 종양을 억제하는 단백질 Ret finger protein 2 (060858), 종양 억제 기능을 가지는 것으로 추정되는 단백질 Cadherin-related tumor suppressor homolog (SWISS-PROT accession number: Q14517)와 RNA-binding protein 5 (P52756)을 사용하였으며 추가로 단백질의 2 차 전기영동으로부터 얻은 기능이 전혀 알려지지 않은 단백질 P39173 과 XP_141934 의 서열을 가지고 실험하였다. 그리고 본 논문에서는 문자열 일치 모듈 구현을 위해 정규 표현 방법을 이용하였다.

DB accession number	Protein name
SWISS-PROT Q9P114	Serine-threonine protein kinase
GenBank NP_005216	E2F Transcription factor 1
GenBank NP_663770	TNF receptor-associated factor 2
GenBank 060858	Ret finger protein 2
SWISS-PROT Q14517	Cadherin-related tumor suppressor homolog
SWISS-PROT P52756	RNA-binding protein 5
GenBank P39173	Unknown
GenBank XP_141934	Unknown

표 1. 실험에 사용된 기능이 알려진 데이터들과 기능이 알려지지않은 서열의 데이터베이스 출처와 단백질 이름.

III. 결과 및 고찰

아래 표 2 와 표 3 에서 6 개의 항암 작용을 나타내는 단백질의 아미노산 서열에 대한 도메인 검색 결과를 나타내었다. 실험 데이터들 사이에 공통적으로 가지고 있는 것으로 나타나는 도메인들의 기능이 각각의 항암 단백질의 종양 억제 기전과 관련이 있을 것으로 사료된다. 표 2 에서 Ret finger protein 2 와 RNA-binding protein 5 이 Cell attachment sequence 도메인을 공통적으로 가지는 것으로 나타났는데 이는 암세포의 특징중 하나인 조절되지 않는 세포 증식 억제 기작과 관련이 있을 것으로 예측된다.

Protein name / Domain name	Cadherin-related tumor suppressor homolog	Ret finger protein 2	RNA-binding protein 5
Cell attachment sequence	Not found	**Found**	**Found**
EGF-like domain signatures	Not found	**Found**	Not found
Zinc finger RING type signature and profile	Not found	**Found**	Not found
Zinc finger RanBP2 type	Not found	Not found	**Found**

signature and profile			
Endoplasmic reticulum targeting sequence	**Found**	Not found	Not found
Microbodies C-terminal targeting signal	**Found**	**Found**	**Found**

표 2. Cadherin-related tumor suppressor homolog, Ret finger protein 2, RNA-binding protein 5 단백질의 도메인 검색 결과.

Protein name / Domain name	Serine-threonine protein kinase	E2F Transcription factor 1	TNF receptor-associated factor 2
Cell attachment sequence	**Found**	Not found	Not found
ATP/GTP-binding site motif A	Not found	**Found**	Not found
Zinc finger RING type signature and profile	Not found	Not found	**Found**
Microbodies C-terminal targeting signal	**Found**	**Found**	**Found**

표 3. Serine-threonine protein kinase, E2F Transcription factor 1, TNF receptor-associated factor 2 단백질의 도메인 검색 결과.

Protein name / Domain name	P39173	XP_141934
Gram-positive cocci surface proteins LPxTG motif profile	**Found**	**Not found**
ATP/GTP-binding site motif A	**Not found**	**Found**
Microbodies C-terminal targeting signal	**Found**	**Found**

표 4. 기능이 알려지지 않은 단백질 P39173 와 XP_141934 의 도메인 검색 결과.

표 4 는 GenBank 로부터 얻은 기능이 전혀 알려지지 않은 아미노산 서열 데이터에 대한 도메인 검색 결과이다. 이들 서열들이 특정 도메인 서열을 가지고 있는 것을 알 수 있다. 일치된 도메인의 기능에 대한 정보로부터 이 아미노산 서열들의 기능을 간접적으로 예측할 수 있을 것으로 사료된다.

사용된 모든 실험 데이터가 도메인 검색 결과 Microbodies C-terminal targeting signal 도메인을 가지는 것으로 나타났다. 이는 이 도메인의 서열이 3 개 잔기로 길이가 매우 짧고 각 잔기마다 여러 조합을 허용하므로 위와 같은 결과가 나온 것으로 사료된다.

IV. 결론

항암작용을 나타내는 것으로 알려진 단백질뿐만 아니라 그러한 기능을 가질 것으로 추정되는 단백질의 아미노산 서열을 가지고 이미 밝혀진 도메인들의 서열을 검색하여 일치하는 도메인을 찾음으로써 간접적으로 단백질이 항암 기능을 가지는지 여부를 밝히는데 도움이 되는 도메인 정보를 얻었으며 또한 기능이 알려지지 않은 아미노산 서열에 대해서도 동일한 실험을 시행한 결과 그 기능을 예측할 수 있는 정보를 얻었다. 이 논문에서 제안한 방법이 시간과 비용을 절약할 수 있는 효율적인 방법이 될 수 있을 거라 사료된다.

이미 정보를 가지고 있는 도메인들을 그 기능에 따라 또는 도메인이 갖는 특정 구조에 따라 분류한 다음 검색하게 되면 표적 단백질의 기능 또는 구조를 연구할 때 도움이 될 수 있을 것으로 예상되며 더 나아가 기능과 구조를 통합적으로 검색함으로써 이들 사이에 존재하는 의미있는 관계를 추출함으로써 유효한 결과를 얻을 수 있을 것으로 사료된다.

V. 참고문헌

[1] L. Falquet, M. Pagni, P. Bucher, N. Hulo, CJ Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002," *Nucleic Acids Research*, vol. 30, pp. 235-238, 2002.

[2] P. Bucher, A. Bairoch, "A generalized profile syntax for biomolecular sequence motifs and its function in automatic

sequence interpretation," *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 53-61, 1994.

[3] RF. Sewell, R. Durbin, "Method for calculation of probability of matching a bounded regular expression in a random data string", *Journal of Computational biology*, vol, 2, pp. 25-31, 1995.

[4] LF. Kolakowski, JA. Leunissen, JE. Smith, "ProSearch: fast searching of protein sequences with regular expression patterns related to protein structure and function", *Biotechniques*, vol, 13, pp. 919-921, 1992.