단백질 모티프간 연관성 탐사

이현숙*, 이도헌
*전남대학교 전산학과

e-mail:hslee@dbcore.chonnam.ac.kr

Exploring Association Among Protein Motifs

Hyun-suk Lee*, Doheon Lee
*Dept of Computer Science, Chonnam University

요 약

단백질 모티프(motif)란 유사한 기능을 가진 여러 단백질 서열에서 공통적으로 발견되는 패턴으로서 단백질의 기능을 예측하는 단서로 활용된다. 현재 Prosite, Pfam 등의 데이터베이스에서 정규식(regular expression), 가중치 행렬(weighted matrix), 은닉 마코프 모델(hidden Markov model)의 형태로 4천여종 이상의 모티프가 등록되어 있다. 하지만, 이러한 데이터베이스는 모티프와 단백질간의 일대일 관계만을 저장하고 있기 때문에, 모티프 간의 연관성을 파악하기는 어렵다. 본 논문에서는 모티프 간의 연관 관계를 연관 규칙의 형태로 발견하는 데이터 마이닝 기법을 제시한다. 아울러 HITS 데이터베이스로부터 입수한 단백질-모티프 데이터베이스에 본 기법을 적용함으로써 상당히 높은 연관성을 갖는 모티프 집단이 실제로 존재한다는 것을 밝힌다.

Ⅰ. 서론

인간 지놈 프로젝트는 완전한 유전자 정보를 알아내어 생명현상의 신비와 유기체의 구조, 진화과정를 밝히고자 한다. 이러한 결과로 엄청난 양의 데이타들이 쏟아졌으며, 이로부터 유용한 데이타들을 발견해 내고 좀더 효율적인 정보를 얻기 위한 연구 분야가 대두되었다. 기본적인 생명현상인 유전 물질을 포함하는 DNA로부터 전사(trascription)을 통해 RNA가 되고, RNA는 번역(traslation)을 통해 세포의 상태와 모양을 유지하는 단백질이 된다. 인간 유전의 최종 산물인 단백질의 기능을 발견하는 일은 매우 중요한 일이며 현재 이와 관련된 연구가 활발히 진행중이다[3]

모티프(motif)란 유사한 기능을 가진 여러 단백질 서열에서 공통적으로 발견되는 패턴이다. 이러한 모티프는 하나의 단백질 가운데 여러개가 존재할 수 있고, 하나의 단백질 안에 서로 다른 모티프가 존재할 수도 있다.

```
XXXA T HRTYW E L M Q W XXX
XXXD M HRCYW K L V Q F XXX
XXXC F HRTYW R L L Q W XXX
XXXY Q HREYW L L I Q Y XXX
-----  HR-YW-L-Q-  ---
```

그림 1 모티프 유형

그림 1은 단백질 서열에서 공통으로 나타나는 패턴이고, 이를 모티프라고 한다.

알려진 모티프는 알려져 있지 않은 단백질의 기능을 예측할 수 있는 단서로 활용된다.

본 논문에서는 HITS에서 제공하는 알려진 모티프를 가지고 모티프들간의 연관성을 발견함으로써 단백질 기능 분석에 필요한 중요한 정보를 밝히고자

한다

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로서 Prosite나 Pfam 등의 데이터베이스에서 모티프를 구성하는 패턴과 , 연관 규칙에 대해 논의하겠다. 3장에서는 Hits 데이터베이스에서 제공하는 데이터를 가지고 연관규칙을 적용하여 모티프들간의 연관성을 탐사한다 4장에서는 결론와 향후 연구되어야 할 문제점을 제시한다.

Ⅱ. 관련 연구

1. 모티프 분석 도구

PROSITE는 단백질 군과 도메인의 데이터베이스로서, 지금까지 밝혀진 모티프를 이용하여 알려지지 않은 단백질이 속하는 군를 찾아낼 수 있고, Pfam은 단백질 모티프와 패밀리의 집합으로 서열 전체 가운데 유사성을 탐색한다.[2][5]

지역적인 유사성 탐색할 수 있는 BLOCKS와 PRINT는 전체 유사성을 갖는 Homologous를 검색하기 위해 사용되며, PROSITE 데이타베이스에 수록된 단백질 그룹을 분석하여 단백질의 보존된 영역을 자동으로 검색해 준다.

통계적 모델링 기법과 경험을 바탕으로 한 MEME은 중복되지 않은 모티프들을 발견하고 모티프의 영역을 검색하는데 반해 MAST는 좀더 먼 상동체를 검색하고 MEME의 분석 결과를 통해 단백질 서열 데이타베이스에 질의하는데 사용된다.[6][7]

단백질 도메인 데이터베이스인 Hits는 하나 이상의 단백질의 이름을 통하여 그와 관련된 모티프를 보여주며 그와 반대로 하나 이상의 모티프 이름을 통하여 모티프가 속한 단백질의 정보를 제공하고 있다.[3]

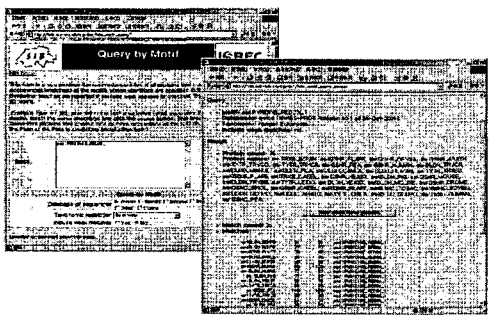

그림 2 Hits 데이타베이스에서 모티프 검색

그림 2에서는 RYRIDINE_REDOX_1 모티프를 입력하여 그와 관련된 단백질을 검색함으로써 단백질과 모티프의 연관성을 보여주고 있다.

2. 다음은 각 모티프 분석 도구에서 표현하는 모티프 형태이다

(1) 정규식(regular expression)

모티프가 등록된 데이터베이스에서 모티프의 패턴은 다음과 같다.

C-x-C-x(5)-G-x(2)-C

이는 prosite에서 제공하는 EGF_1 엔트리의 패턴을 설명한 것으로, C는 아미노산 Cysteine를, G는 아미노산 Glycine을 나타는 것처럼 아미노산의 첫글자로 표현한다. x는 아미노산 어떠한 값이든 나타낼 수 있고, x(N)은 x가 N개임을 나타낸다. 예를 들어, x(5)는 x-x-x-x-x를 의미한다.[2]

(2) 가중치 행렬(Weighted Matrix)

프로파일은 다중 서열의 정렬을 통하여 어떤 위치에 어떤 아미노산 잔기가 허용되는지의 여부, 삽입과 일치하는 부분 등을 종합하여 행렬로써 표현한 데이터베이스이다.

MA /I: MI=-32; MD=-32; IM=-32;

이는 ADAM_MEPRO 엔트리의 패턴으로, MA는 행렬(matrix)를 의미하고, /I는 삽입을 나타낸다. MI는 M부터 I까지, MD는 M부터 D까지 IM은 I부터 D까지의 상태변이 값을 의미한다.

(3) 은닉 마코프 모델(Hidden Markov Model)

Pfam 데이터베이스는 HMMER2 소프트웨어를 사용하여 모티프를 검색한다. 서열분석에서 이전의 아미노산 잔기가 다음 아미노산 잔기로 전이될 때 이전 아미노산 잔기에서 발생할 수 있는 모든 경우의 수를 계산하여 최대 확률 값을 가지는 경우를 선택한다. 이때 최대 확률 값을 행렬로 표현하였다.[8]

3. 연관규칙

거대한 양의 데이터를 분석하는 작업을 지원하는 기술이 데이터마이닝이다. 데이터마이닝은 여러 가지 방법을 통해 기존에 얻을 수 없었던 추가적인 정보들을 제공한다. 본 논문에서는 연관규칙을 사용하였다. 연관규칙(association)이란 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업을 말한다. 이러한 작업들 가운데 의미 있는 규칙들만을 뽑아내

는 기준이 근거확률(support), 신뢰확률(confidence) 그리고 리프트(lift)이다. 근거확률은 모티프 전체 가운데 모티프A와 모티프B를 포함하는 수를 말하며, 신뢰확률은 모티프A가 발생할 때 모티프 B가 발생할 확률을 말한다. 리프트는 모티프A 집단에서 B가 발생할 확률을 말한다.[10]

Ⅲ. 모티프간의 연관성 탐사

1. 관련 데이타베이스로부터 입수한 데이터 분석

본 논문에서는 Hits 사이트에서 입수한 데이타를 토대로 모티프들간의 관련성을 알아보았다.
(ftp://ftp.isrec.isb-sib.ch/pub/databases/htis/)

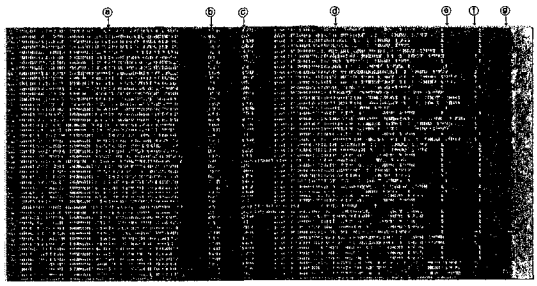

그림 3 Hits에 저장된 데이터 양식

그림 3의 데이터는 다음과 같다.
ⓐ는 sw|O00087|DLDH_SCHPO|511|E68350146A93AD56로 구성되어 있는데, sw는 SWISS_PROT의 데이터베이스 코드를 나타내며, O00087는 단백질의 AC, DLDH_SCHPO는 단백질의 ID, 511는 서열의 길이, E68350146A93AD56는 서열의 CRC64를 의미한다. ⓑ의 81은 단백질 서열의 첫번째 위치, ⓒ의 91은 단백질 서열의 마지막 위치를 나태낸다.
ⓓ는pat|PS00076|PYRIDINE_REDOX_1|-|-|OCT-1993로 구성되어 있는데, pat는 PROSITE의 데이터베이스 코드를 나타하며, PS00076는 모티프 AC, PYRIDINE_REDOX_1는 모티프 ID, OCT-1993는 모티프 마지막 갱신 날짜를 의미한다.
ⓔ의 1은 모티프 중 첫번째 위치를 나타내고, ⓕ의 -1은 모티프 중 마지막 위치를, ⓖ의 -는 일반화된 매치의 값을 나타내고 있다.

그림4는 Hits에서 입수한 데이터를 가지고 모티프들 간의 연관성를 알아보기 위한 형식으로 표현하였다.
①은 swiss_protein_info에 저장된 protein_index이고 ②는 protein_index를 count한 것이다. ③은 motif_pattern_info에 저장된 motif_index이고, ④는 하나의 단백질에서시작 위치와 끝위치만 다른 모티프들을 count 하였다.

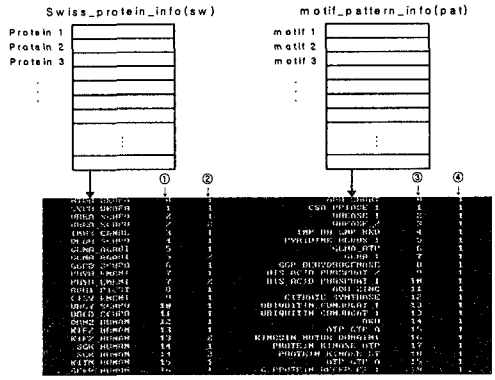

그림 4 연관성을 알아보기 위해 저장한 형태

그림 5의 데이터는 56000개 이상의 단백질 중 3번 이상 count된 protein_index와 모티프들을 행렬로 표현한 것이다. 예를 들어, 행렬을 통해 62번째 단백질에는 64, 25, 24,의 모티프들이 발견이 되었고, 각각의 모티프가 몇 번정도 발생하였는지를 보여주고 있다.

```
 38   33_1, 32_1, 32_2,
 41   18_1, 34_1, 35_1,
 46   24_1, 15_1, 17_1,
 62   64_1, 25_1, 24_1, 25_2, 24_2, 25_3, 24_3, 25_4, 24_4, 25_5,
 68   70_1, 25_1, 64_1, 25_2, 64_2, 25_3, 24_1, 64_3, 25_4, 64_4,
 69   73_1, 71_1, 72_1, 25_6, 23_1, 24_3, 64_6, 25_7, 64_7, 25_8,
 79   88_1, 73_1, 73_2,
 81   85_1, 81_1, 82_1, 83_1,
 83   15_1, 86_1, 87_1,
187   116_1, 89_1, 90_1, 91_1,
114   73_1, 90_1, 99_1,
122   17_1, 15_1, 73_1,
```

그림 5 단백질과 모티프간의 행렬

IBM DB2 Intelligent Miner for Data V6(http://www-3.ibm.com/software/data/iminer/fortext/index.html)를 이용하여 모티프들간의 연관성을 알아보고자 한다. 이에 사용될 데이터의 형태는 그림 6과 같다.

protein_index	*motif_name_motif_count*
46	ASX_HYDROXYL_1
68	VWFC_1
68	CTCK_1_1
79	CPSASE_1_1

그림 6 연관규칙에 사용될 데이타

그림 7은 motif_index_motif_count를 아이템 필드로 protein_index를 처리하여 이미지로 연관성을 보여주었다. IMB DB2 Intelligent Miner for data V6에서는 근거확률과 리프트는 고려하지 않고, 신뢰확률을 25%로 하였다.

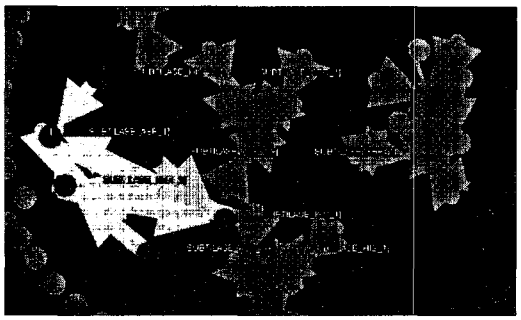

그림 7 연관된 모티프들

그림 7과 같이 많은 모티프들이 연관되어 있으며 그중 일부를 선택한 구체적인 결과를 그림 8에서 보여주고 있다. 이러한 결과를 토대로 단백질과 모티프간의 연관성을 밝힘으로서 단백질의 기능을 유추하는데 더욱더 활발히 진행될 수 있다.

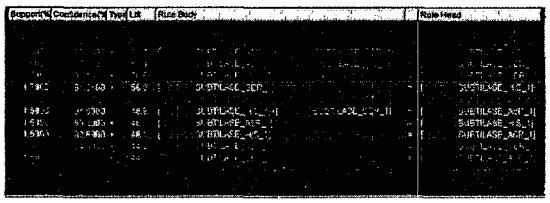

그림 8 연관된 모티프 집단

Ⅳ. 결론 및 향후 연구

인간 유전에 중요한 역할을 담당하는 단백질은 하나 이상의 서로 다른 모티프들이 존재한다. 이러한 모티프는 알려지지 않은 단백질의 기능을 유추하는데 사용되고 있다. 본 논문에서는 연관규칙을 통해 모티프간의 관련성이 높은 모티프 집단이 있음을 탐사하게 되었다. 따라서 기존의 단백질과 모티프간의 일대일의 기능유추가 아닌 하나의 모티프가 발견되어졌다면, 그와 관련된 모티프를 토대로 단백질의 기능을 유추하는데 빠르고 쉽게 접근할 수 있다.

향후 과제로는 데이터베이스에서 제공하는 많은 모티프간의 연관성들을 그림 7과 같이 가시적인 결과를 웹으로 제공하고자 한다.

참고문헌

[1] Y.-j. Hu, et al.,Combinatorial Motif Analysis and Hypothesis Generation on a Genomic Scale ,Bioinformatics,2000, Vol. 16, No. 3, pp. 222-232

[2] Kay Hofmann,Philipp Bucher,Laurent Falquet and Amos Bairoch, The PROSITE database, its status in 1999, Nucleic Acides Research ,1999, Vol.27,No.1 pp.215-219

[3] Marco Pagni, Christian Lseli,Thomas Junier, Laurent Falquet, Victor Jongeneel and Philipp Bucher,trEST,trGEN and Hits:access to databa ses of predicted protein sequences, Nucleic Ac ides Research,2001,Vol.29,No.1,pp.148-151

[4] Roman L. Tatusov, Michael Y. Gaperin, Darren A. Natale and Eugene V. Koonin, The COG database : a tool for genome -scale analysis of protein functions and evolution, Nucleic Acides Research,2000,Vol.28,No.1, pp.33-36

[5] Arne Elofsson and Erik L.L. Sonnhammer,A comparison of sequence and structure protein domain families as a basis for structural genomics,bioinformatics,Vol.15,no.6,1999,pp.480 -500

[6] Timothy L. Bailey and Charles Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymersProceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

[7] Timothy L. Bailey and Michael Gribskov, Combining evidence using p-values: application to sequence homology searches, Bioinformatics,Vol. 14, pp. 48-54,1998.

[8] Leonid Peshkim and Mikhail S. Gelfand, Seg mentation of yeast DNA using hidden Markov models,Bioinformatics,Vol.15,pp.980-986

[9] 김정자, 단백질 기능 분석을 위한 연관 규칙 탐사,2001,pp.29-36

[10] 정남식, 홍성완, 정재호,데이터마이닝,대청,1997