제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

단백질 모티프간 연관성 탐사

이현숙*, 이도헌**, 최덕재*
*전남대학교 전산학과
** 한국과학기술원 바이오시스템학과

*hslee@dbcore.chonnam.ac.kr, **dhlee@mail.kaist.ac.kr,
*dchoi@chonnam.ac.kr

Association Discovery Among Protein Motifs

Hyun-suk Lee*, Doheon Lee**, DeokJai Choi*
*Dept of Computer Science, Chonnam University
**Departmenet of BioSystems, KAIST

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. 본 논문에서는 연관성 탐사 기법을 적용하여 Hits 데이터로부터 상당히 높은 연관성을 갖는 모티프 집단을 밝히고, 실제 자연현상에서 자주 나타나는 연관성을 교차타당성(cross-validation) 기법을 통해 입증하였다. 이렇게 밝혀진 단백질 모티프간 연관성을 트라이 탐색 기법을 통해 웹으로 제공함으로써 단백질의 기능유추에 쉽게 접근하고자 한다.

Ⅰ. 서론

인간의 특성을 담고 있는 유전정보의 염기배열 순서가 모두 밝혀진 가운데, 여러 유전자의 기능을 총체적인 관점에서 파악함으로서 인간의 모든 생물학적 현상을 담당하는 정상 유전자의 기능을 빠른 속도로 밝혀내고 있다. 생물정보학(bioinformatics)은 생물학적인 연구에 컴퓨터를 응용하여 서열들간의 비교를 효율적으로 할 수 있도록 하기 위한 것으로 생명현상에 복잡하고 다양한 문제를 풀고자 인간 유전체 프로젝트가 진행되면서 엄청나게 발생하는 생물 정보들을 효과적으로 처리하고 대용량의 데이터를 어떻게 효율적으로 이용하여 의미 있는 정보를 이끌어 낼 것인가 하는 것이 중요하게 대두되고 있다.

인간 유전에 관여하는 주요물질인 DNA는 전사, 번역을 통해 단백질을 만들게 되는데, 이러한 단백질의 기능을 밝혀냄으로써 생물학적인 현상을 밝혀낼 수 있다.

모티프(motif)란 유사한 기능을 가진 여러 단백질 서열에서 공통적으로 발견되는 패턴이다. 이러한 패턴은 알려지지 않은 서열이 어떤 기능을 하는지 밝혀내는 방법 즉 알려진 모티프와의 비교를 통해 알려지지 않은 단백질 기능을 예측하는데 활용될 수 있다. 여러개의 단백질에는 서로 다른 모티프들이 존재하기도 하며 공통된 하나의 모티프가 존재하기도 한다.

본 논문에서는 HITS에서 제공하는 모티프를 가지

고 IBM Intelligent Miner를 통하여 여러 모티프들 간의 연관성이 있음을 발견하고, 탐사된 결과가 실제 자연현상에서 자주 나타나는 연관성임을 입증하였으며, 트라이 탐색 기법을 통하여 웹으로 제공하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로서 여러 가지 모티프 분석 도구와 Prosite, pfam에서 제공하는 모티프 표현형태를 기술하였다. 3장에서는 Hits 데이터베이스로부터 탐사된 모티프 간의 연관 규칙과 이를 교차 타당성기법을 통해 연관성 입증하고 수행결과를 제시하였다. 마지막으로 4장에서는 결론과 향후 문제점을 언급하였다.

Ⅱ. 관련 연구

1. 모티프 분석 도구

PROSITE는 단백질 패밀리과 도메인의 데이터베이스로서, 지금까지 밝혀진 모티프를 이용하여 알려지지 않은 단백질이 속하는 군를 찾아낼 수 있고, Pfam은 단백질 모티프와 패밀리의 데이터베이스로서으로 서열 전체 가운데 유사성을 탐색한다.[3][4]

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

통계적 모델링 기법과 경험적 방법을 바탕으로 한 MEME은 중복되지 않은 여러 모티프들을 발견하고 모티프의 영역을 검색한다. MAST는 좀더 먼 상동체를 검색하고 MEME의 분석 결과를 통해 단백질 서열 데이타베이스에 질의하는데 사용된다.

단백질 도메인 데이터베이스인 Hits는 하나 이상의 단백질의 이름을 통하여 그와 관련된 모티프를 보여주며, 하나 이상의 모티프 이름을 통하여 모티프가 속한 단백질의 정보를 제공하고 있다.[7]

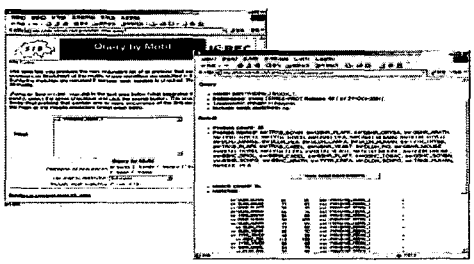

그림 1 Hits 데이타베이스에서 모티프 검색

그림 1에서는 Hits 데이터베이스에 수록된 RYRIDINE_REDOX_1 모티프를 입력하여 그와 관련된 단백질을 검색함으로써 단백질과 모티프의 일대일 연관성을 보여주고 있다.

2. 모티프 형태

Prosite는 genome 혹은 cDNA 서열로부터 전사된 알려지지 않은 단백질의 기능을 결정하는 방법을 제공한다. 이는 새로운 서열에 포함되어 있는 단백질의 군을 알아냄으로써 빠르고 유사한 패턴을 찾아낼 수 있다. 이러한 단백질의 기능을 결정짓는 단백질 서열 패턴은 서열 분석 시 중요한 단서를 제공한다.

(1) 정규식(regular expression)

prosite에서 사용하는 모티프 표현방식인 정규식은 변질된 위치 가운데 보존된 서열들을 나타내는 패턴이다. Prosite 패턴은 하나의 완전한 도메인이나 단백질을 나타내는 것이 뿐만 아니라 효소의 촉매에 의한 위치, 금속 이온을 굳히는 것에 포함된 아미노기를 갖는 산들, 이황화물에 포함된 시스테인과 같은 기능적으로 중요한 잔기들까지도 나타낸다. [3]

(2) 가중치 행렬(Weighted Matrix)

Profile은 다중 서열 정렬을 통하여 어떤 위치에 어떤 아미노산 잔기가 허용되는지의 여부와 보존된 지역, 삽입과 삭제 부분 등을 고려하여 유도된 결과에 갭값을 포함하는 테이블로 나타낸다. 또한 부분적으로 부정확한 서열 정렬에서 유사성을 찾을 수 있게 한다. [1][2]

(3) 은닉 마코프 모델(Hidden Markov Model)

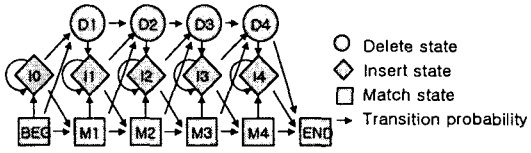

그림 2 Hidden Markov Model

Pfam은 하나의 jduf 대신 다중서열정렬의 탐색 데이터베이스인 profile에 HMM(Hidden Markov Model)를 접목시켜 단백질의 모티프를 정의하고 있다. PROSITE profile과 유사한 HMM은 상동성을 갖는 영역으로부터 출발하여 점차적으로 먼 집단들

의 그성원들을 포함시키면서 반복적으로 수행한다.[5]

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 의미있는 규칙들만을 뽑아내는 기준으로써 근거확률(support), 신뢰확률(confidence) 그리고 리프트(lift)가 있다. 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.[12]

Ⅲ. 모티프간의 연관성

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

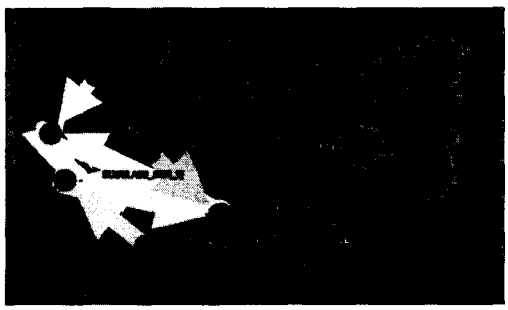

그림 3 연관성이 탐사된 모티프 집단

모티프간의 연관성을 탐사하는데 있어서 하나의 단백질 중 서로 다른 위치에서 발견된 동일한 모티프를 별개로 하여 근거확률로 나타내었다. 아래 t 는 각각의 단백질을, a,b,c는 모티프를 의미하며, 숫자는 단백질에서 발생하는 모티프 수를 의미한다. 이러한 자료를 토대로 지지도를 구하였다.

본 논문에서 탐사된 단백질 모티프간 연관성이 실제 자연현상에서 나타나는 연관성인가를 입증하기

t_1 a_1, a_2, b_1, c_1
t_2 a_1, b_1
t_3 $a_1, a_2, a_3, b_1, c_1, c_2$

$Supp(a_1) = \frac{3}{3}$, $Supp(b_1) = \frac{3}{3}$, $Supp(c_1) = \frac{2}{3}$

위해 C4.5 결정트리의 교차타당성(cross-validation) 기법을 적용하였다. 교차 타당성(cross-validation) 기법이란 학습 데이터를 N개의 같은 크기 블록들로 나누고 각 N번의 반복에서 한 개의 서로 다른 블록을 하나의 집합으로 하고 나머지 레코드들을 생성 집합으로 사용하여 두 집합 가운데 연관성이 나타나는 빈도를 구하면 된다. 즉 단백질의 개수가 10만개라면, 그것을 각각 만개씩 10개의 부분집합으로 분할한다. 분할기준은 랜덤 함수를 사용하였다. 이렇게 분할된 1번부터 9번까지 총 9만개를 학습데이타로 하여 연관성을 구하고 나머지 만개에서도 비슷하게 나타나는지 확인하는 방법으로 타당성을 증명하였다. 실험 결과 지지도가 60%이상에서 77.2%의 일치율을 보였다. 이는 자연현상에서 자주 나타나는 연관성임을 입증한 것이다.

마지막으로 입증된 연관성을 웹을 통하여 제공한다. 모티프 이름과 단백질의 이름을 통하여 검색하게 되는데, 모티프의 이름으로 검색을 하는 경우에는 다른 모티프와의 신뢰도, 지지도 등의 결과를 보여준다. 단백질의 이름으로 검색하였을 경우에는 검색할 때마다 8만 라인 이상의 단백질과 모티프가 담겨진 파일을 검색할 수 없으므로 단백질 인덱스를 중심으로 subfile을 생성하였다. 따라서, Hits에서 입수한 플랫파일을 subfile로 분할하여, 트라이 인덱싱 기법을 통해 빠른 탐색을 하였다.

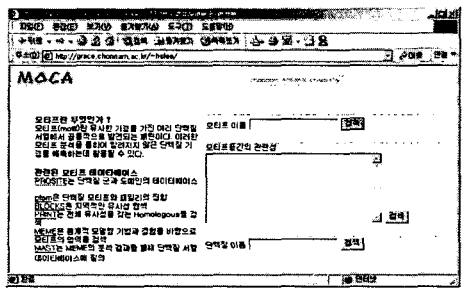

그림 4 웹으로 제공하는 연관규칙 홈페이지

예를 들어, 아래와 같은 정보를 갖는 EGF_2 모티프를 검색해 보았다. 아래의 내용은 Prosite에 나타나

는 EGF_2의 정보를 기술하였다.

ID EGF_2; PATTERN.
AC PS01186;
DT NOV-1997 (CREATED); NOV-1997
(DATA UPDATE); JUL-1998 (INFO UPDATE).
DE EGF-like domain signature 2.
PA C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C.

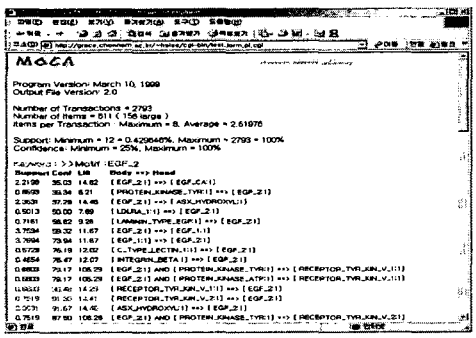

그림 5 연관규칙 결과화면

Ⅳ. 결론 및 향후 연구

핵산과 단백질의 데이터들이 대량으로 축적되는 가운데 앞으로도 더욱 많은 모티프들이 발견될 것이다. 본 논문에서 밝혀낸 모티프들간의 연관성을 탐사하고,이러한 연관성을 웹을 통해 제공하고자 함으로써 단백질의 기능을 유추하는데 빠르고 쉽게 접근할 수 있다.

향후 연구 과제로 논문 연구에 사용된 모티프는 단순히 모티프들의 연관성만을 제공하였다. 하지만 앞으로 이렇게 발견된 모티프 연관성이 생물학적으로 어떤 의미를 갖는지 해석하고자 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] Philipp Bucher, Kevin Karplus, Nicolas Moeri and Hofmann,A Flexible Motif Search Technique Based on Generalized Profiles, Computers Chem, Vol.20,No.1,pp.3-23,1996

[3] Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C.J, Hofmann K. and Bairoch A,"The PROSITE database, its status in 2002"Nucl. Acids Res. 30(1):235-238

[4] Erik L.L. Sonnhammer, Sean R. Eddy, Richard Durbin. (1997) Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. Proteins28:405-420. Original reference for the PFAM database.

[5] Bucher P., Bairoch A,A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation In "ISMB-94; Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology."(Altman R., Brutlag D., Karp P., Lathrop R., Searls D., Eds.), pp53-61, AAAIPress, Menlo Park,1994

[6] Sean R. Eddy, Profile Hidden Markov model, Bioinformatics,Vol.14,No.9,pp.755-763,1998

[7] 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

[8] 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.

[9] 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.

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

[11] Leonid Peshkim and Mikhail S. Gelfand, Segmentation of yeast DNA using hidden Marko -v models,Bioinformatics,Vol.15,pp.980-986

[12] 정남식, 홍성완, 정재호,데이터마이닝,대청,1997