2002년 한국정보처리학회 춘계학술발표논문집 제 9 권 제 1 호

# 단백질 서열 연관 규칙 마이닝을 위한 효율적인 알고리즘 설계

김현민, 김지혜, R.S. Ramakrishna
광주과학기술원(K-JIST) 정보통신공학과
e-mail : hmkim@kjist.ac.kr

# Efficient Sequence Association Rule Mining for Discovering Protein Relations

Hyun-min Kim, Jihye Kim, R.S. Ramakrishna
Dept. of Information and Communication K-JIST

**요 약**

DNA 의 염기서열 탐색을 위한 유전체학의 다음 세대인 구조유전체학은 유전체 사업으로 인한 인간 게놈지도의 완성과 축적된 생물정보를 이용한 생물정보학의 발달과 함께 급속한 성장을 계속하고 있다. 포스트 게놈 시대를 맞이하여 생명현상에 대한 궁극적인 이해를 위한 노력으로 단백질의 구조와 기능에 대한 연구가 주목을 받게 되었다. 다양한 구조 규명을 위한 도구들과 단백질 정보를 관리하기위한 데이터베이스 구축에 따른 관련 기술의 발전은, 앞으로 다가올 생물정보의 방대함을 감안할 때, 가치 있는 지식정보를 얻기 위한 데이터 마이닝 기법들을 통해서만 가능하다.

본 논문은 데이터 마이닝의 근간 기술인 연관규칙 마이닝을 응용한 효율적인 서열 연관 규칙 알고리즘을 제안 하며, 단백질 구조의 예측을 위한 단백질 서열 및 DNA 서열 간의 패턴 비교 및 연관성을 목적으로 한다. 또한, 공간적 시간적 복잡성을 CMS-tree 라는 자료구조를 통해 알고리즘의 확장성 및 병렬화의 기본 알고리즘으로 사용하도록 개발하였다.

## 1. 서론

생물정보학(Bioinformatics)은 컴퓨터를 활용해 생물학적 데이터를 수집· 관리· 저장· 평가· 분석하는 기술을 말한다. 분자생물학의 급속한 발달과 유전체 사업(Human Genome Project)으로 인한 방대한 데이터의 축적은 데이터 저장 기술 뿐만 아니라 유용하지만 잘 알려져 있지 않은 패턴을 얻어내기 위한 여러 가지 데이터 마이닝 기법을 필요로 하고 있다. 이는 생물정보학의 연구 초점이 유전체 전체 구조를 밝히고 wet-lab 의 실험을 보조하는 초기의 유전체 정보학(Genome Informatics)에서 포스트 게놈 정보학(Post-genome Informatics)으로 전이되고 있음을 의미한다[1]. 포스트 게놈 정보학에서는 유전체학으로부터 얻은 데이터를 통해 생물의학의 실재적인 응용과 생명의 궁극적인 원리를 규명하고 생물학적 지식을 종합하기 위한 제반 기술을 다룬다.

생명현상은 궁극적으로는 단백질의 특이한 3 차 구조에 의하여 수행되므로 단백질의 구조를 얻어 기능을 유추하고자 하는 구조유전체학(Proteomics)이 각광을 받고 있다. 유전정보에 의하여 암호화되어 있는 단백질의 아미노산 서열 정보로부터 직접 단백질의 3 차 구조로 전환하는 과정은 현재 매우 미진한 단계이나, 단백질의 구조 정보 해석, 폴딩 경로의 추적, 컴퓨터 프로그래밍의 개발로 그 기술이 급성장 하고 있다. 이러한 의미에서 구조유전체학은 포스트 게놈시대의 중추적 학문으로서, 유전자의 산물인 단백질을 대상으로 이들을 분석하고 상호 기능관계 지도의 작성 및 구조분석을 통해 특정 단백질과 이를 만드는 유전자의 기능을 동시에 밝혀내는 기술이다. 단백질

구조 및 기능에 대한 실험적 검증을 위해서는 X-ray, 회절법이나 NMR 분광법을 통한 구조 규명이 필요하다. X-ray 나 NMR 구조 규명에서 구조를 가시화 시키는 computer modeling/graphics 는 구조 규명 기법이라기보다는 구조를 가시화 해주는 보조적 성격이 강하지만 최근 이미 알려진 단백질 구조 데이터베이스 등을 이용한 생물정보학 기술이 급성장하고 있어 computer modeling(homology modeling, threading, ab. initio 등) 만에 의한 신규 단백질의 high resolution 구조 규명이 가능한 시대가 도래하였다[2].

본 논문에서는 단백질 구조 및 기능 예측을 위한 단백질 상동성 모형화(homology modeling)에 초점을 맞추어 효율적인 서열 연관 규칙을 위한 알고리즘을 제안한다.

## 2. 관련연구

실험을 통하지 않고 단백질 3 차원 구조를 예측하는 방법 중에 하나인 상동성 모형화는 지금까지 실험으로 밝혀진 단백질데이터은행(PDB)에 수록된 약 14000 개의 단백질 구조를 검색하여 아미노산 서열이 유사한 단백질의 경우 그 구조가 서로 비슷하다는 데 근거를 둔 방법이다. 이 방법의 단점은 아미노산 서열 상동성이 매우 높을 때에도 단백질의 고리에 해당되는 부분의 예측이 매우 어렵다는 것과 상동성이 높지 않은 경우에도 단백질의 구조가 비슷한 단백질들이 많다는 점이다. 이처럼 단백질 서열상의 유사성이 불분명한 경우(*Twilight zone*), 기능 및 구조상의 중요한 부분인 모티프(motif)를 탐색하여 구조 예측을 할 수 있다. 이러한 모티프를 탐색하기 위한 기존의 기본적인 도구들은 다음과 같다. 유사한 단백질의 1 차 구조를 찾기 위한 검색 서비스인 BLAST[3]와 FASTA[4] 등과 PROSITE[5]와 MEME[6]과 같은, 알려진 수많은 모티프에 대한 검색 및 비교 서비스를 제공하는 다양한 모티프 라이브러리, 그리고, 단백질 서열에 대한 3 차 구조 검색을 위한 PDB (Protein Data Bank)[7], MMDB(Molecular Modeling Database) [8] 등이 대표적이다.

### 2.1 단백질 구조 예측과 서열 분석

단백질은 그 기능에 따라 Carrier protein, Enzyme, G-protein 등 10 여가지로 나뉜다. 이러한 다양한 단백질의 기능은 20 개의 아미노산(amino acid)의 선형 조합에 따른 단백질 구조의 다양성에서 기인한다. 단백질의 구조적 특성이 단백질의 기능을 결정하는 중요한 요인이며, 폴딩에 관여하는 모티프는 이러한 구조적 특성을 결정하는데 중요한 역할을 한다. 대표적인 예로, DNA binding 모티프를 살펴보자. DNA-binding 단백질은 조절 단백질(regulatory protein)이라 하며, 전사(transcription)에 관여하여 유전자의 표현(gene expression)을 활성화(activate) 또는 비활성화(inactivate)하는 기능을 수행한다. 이에 관련하여 세 가지 종류의 모티프(helix-turn-helix, zinc finger, leucine zipper) 중에서 zinc finger 에 대하여 서술한다.

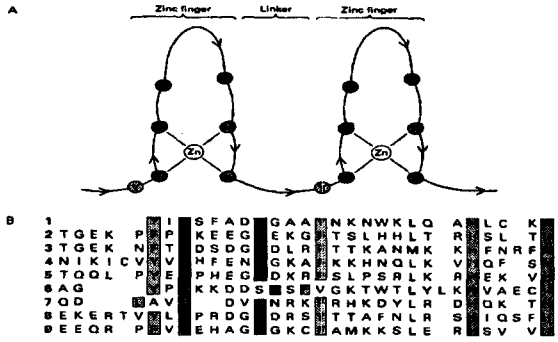



**그림 1 Zinc Finger Motif**

그림 1-(A)는 DNA binding 에 관여하는 zinc finger 의 모식도 이고 그림 1-(B)는 zinc finger 모티프를 포함한 여러 단백질들의 서열이다. x(2.4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3.5)-H 형식의 아미노산서열의 정규표현(regular expression)으로 모델링된 zinc finger C2H2 모티프는 두 개의 시스테인(Cysteine)과 두 개의 히스티딘(Histidine) 잔유기(residue)들이 아연 이온($Zn^{2+}$)을 중심으로 독특한 손가락(finger) 구조를 이루며 존재한다.

Query archea3 virus 3

- Protein: sw:YY06_METJA, sw:YZ27_METJA, sw:Y041_THEAC, sw:ZNFP_LYCVP, sw:Y5K6_SSV1, sw:ZNFP_LYCVT.
- Databases: pattern [Wed Nov 21 00:30:10 2001], profile [PROSITE Release 16.48 + Prerelease Mon Oct 22 14:07:13 200], pfam [Tue Jun 19 00:32:33 2001].

Result

- Motif count: 5.

Motif names: pat:ZINC_FINGER_C2H2_1, pfam:DUF133, pfam:ZF_C2H2, prf:ZINC_FINGER_C2H2_2, prf:LYS_RICH.

**그림 2 Motif 검색 툴(PROFILE)을 이용한 질의 결과**

그림 2 는 achea 3 종과 virus 3 종에 대한 DNA-binding 특성을 지진 단백질 서열이다. 이들은 모두 zinc finger 모티프를 가지고 있다. 이 밖에 특이한 것은 이러한 zinc finger 구조가 상동성이 거의 없는 생물종 사이에서 동일하게 발견된다는 점이다. 초기에 발견된 Aaron Klug 의 Xenopus 의 transcription factor(IIIA)에서만 아니라 진핵세포(Sp1, estrogen, glucocorticoid receptor, Drosophila development regulator)들에서도 발견된다. 또한 이들 중 leucine rich (LYS_RICH)구조를 가지는 모티프는 zinc finger(ZINC_PING ER_C2H2_1 등)와 빈발함을 볼 수 있다[http://hits.isb-sib.ch/].

### 2.2 서열 연관 규칙

서열 연관 규칙은 임의의 길이를 가진 서열간의 연관성을 나타내는 규칙이다. [8]에서 제안된 이 규칙은 Apriori[9]알고리즘을 바탕으로 구현되었다. 기존의 서열규칙(Sequential Patterns)[10]과는 달리 부서열

사이에 존재하는 순서와 빈공간 그리고 윈도우크기를 배제하여 연산의 복잡성을 크게 낮추었다. 단백질 서열의 도메인(Domain) 및 모티프가 가지는 특성을 고려한 기존의 알고리즘은 그림 3 과 같이 네 단계로 구성되어 있다.

1단계: 빈발 부서열의 추출
2단계: 빈발 부서열 조합 발견
3단계: 서열 연관 규칙 도출
4단계: 중복 규칙 제거

**그림 3 SARA algorithm**

## 3. CMS-Tree

### 3.1 CMS-Tree Algorithm

본 논문에서 제안하는 서열연관규칙 알고리즘은 기본 자료구조로서 개발된 Compressed Multi-Sequence-Tree (이하 CMS-Tree)를 이용한다. CMS-Tree 를 이용한 알고리즘은 기존의 Hash Tree 를 이용한 알고리즘[8]과 달리 서열 검색으로 추출한 아미노산의 조합으로 후보서열을 만들지 않고, 단백질 서열 데이터 자체를 압축하는 방법에 의해 생성하여 빈발 부서열(frequent sub-sequence)을 찾아낸다.

1. Scan the *DB*. Collect the set of frequent items $F_1$
2. Create the root of a CMS-Tree, T and label it as "null".
3. *While reading i many item(s) and point to next* tem in each transaction, add 1 to support if the items are already exist or insert those items into CMS-Tree in *i*th depth.
*After reading the DB, delete the leaf node when their* supports don't satisfy the minimum support count and add 1 to *i* and do 3 until last leaf node satisfies the minimum support.

**그림 4 CMS-Tree 생성 알고리즘**

그림 4 는 CMS-Tree 를 생성하는 알고리즘이다. 실제적으로 빈발 부서열을 위한 후보 생성 및 빈발 여부에 대한 탐색 과정이 생성 단계에서 동시에 이루어지므로 탐색 시간을 줄일 수 있다(SARA 의 1 단계 2 단계에 해당, 그림 3). 1 단계에서 입력된 단백질 서열(아미노산 서열)들을 읽어 들여 크기가 1 인 빈발 아미노산을 얻은 후, 크기가 2 인 부서열부터 단백질 서열 데이터에서 그림 5 와 같이 차례로 읽어 들여 CMS-Tree 에 저장한다. 부서열에서 아미노산 하나는 하나의 노드를 차지하며, 마지막 아미노산이 다른 부서열의 경우 다른 가지로 생성한다. 이 때 이전 크기(depth)의 부서열이 최소 지지도(minimum support)를 만족하지 않는 경우는 가지를 확장하지 않는다. 그림 3 은 CMS-Tree 를 생성하는 알고리즘이다.

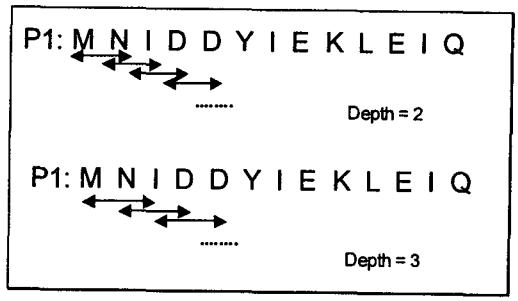



**그림 5 단백질 서열 탐색 과정**

sw:YY06_METJA
MNIDDYIEKLEIQKEGFFYKCPYCNYTNADVKAIKKHIKSKHYDIAKEVENLNK
QNKPQRKPMKKQPKKKDDDYKDYMLLFAHKKKCKIYLDNGMIVEGTVKAKD
RFNIMVLDAKVDDKEVERIIQKGHIVALIPLEE

sw:YZ27_METJA
MDILGEIVNVDEYVEKLELQKNDIGFYKCPFCDYTNADAKVVRKHVKSKHLEE
IEKELKKLESQKSKNNGKKQTGQKKQGKGKKQPKRVRETCVSTQERKDYVL
FFCHNHKVRLHLANGEVLEGKCCCKDPYTVLVDVGKGDVVIVNKAYIVKYVP
LDLEKL

sw:Y041_THEAC
MPWVGSQFHKGDAQTIMALSNVLYQVLIALGMIYVIDTSAIISRNLNLLEGDLM
FPSSVIGEIKKGKLRYMIDVLLPMIRVASPDHEYLKIVEETAAKTGDLMNLSQT
DKDVLALALQYDATIVTDDYSIQNVASYLNLGFLNANIKRIDKQIAWIYRCTGC
KKVFPGPVKVCDICGHEVKRHYDKRKSMIRKV

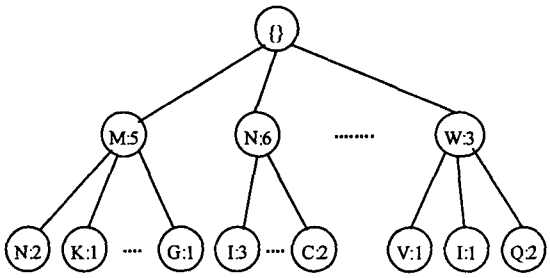
sw:ZNFP_LYCVP
PDTTYLGPLNCKSCWQKFDSLVRCHDHYLCRHCLNLLLTSSDRCPLCKYPL

sw:Y5K6_SSV1
MYQCLRCGGIFNKRREVVEHLLVGHKHKDRLTLDFYYIYFRVRGQ

sw:ZNFP_LYCVT
MGQSKSKEEKGISGTSRAEILPDTTYLGPLNCKSCWQKFDSFSKCHDHYLC
RHCLNLLLTSSDRCPLCKYPL

**그림 6 단백질 서열 데이터**

그림 7-(A)는 그림 6 의 Swiss-prot 데이터 베이스[12]에서 추출한 6 개의 단백질(sw:XXXX ,Swiss-prot 의 단백질 ID XXXX 를 의미함, 그림 2 참조)에 대하여 (그림 4)에 나타나는 빈발 부서열을 CMS-Tree 를 이용해 찾아 본 예이다. 첫 번째 전체 서열 탐색으로 찾아낸 빈발 아미노산 18 개는 depth 가 1 인 노드에 저장되고, 이들을 포함한 크기가 2 인 부서열들이 depth 가 2 인 자식 노드로 생성된다. $k$ 크기의 빈출 부서열을 참조하기 위해서는 depth 가 $k$ 인 노드로부터 depth 가 1 인 노드 까지 탐색해 나간 후 거꾸로 읽으면 된다. 예를 들어 그림 7-(A)에서 지지도가 3 이상인 2 크기의 빈발 부서열은{NI}가 된다. 각 크기의 빈발 부서열의 결과 값은 그림 7-(B)와 같다. CMS-Tree 는 읽어 들인 데이터 자체가 Tree 의 노드를 구성하여, 존재하지 않은 서브서열을 위한 노드를 위한 공간은 할당되지 않는다.




(A) CMS-Tree

1-sized : M N I DY E K L Q G F C P T A V S R W

2-sized : NI NA NL NK IY IV DY DV DI DR DA DT DH YI EK EI EG EV EE KL KE KC KK KH KS KD KG KR KY LE LN LL LD LG LV LT QK GH GQ GP FY CP CK CH CL PD TS AK AL VK VE VL VR VG SK SS RK RV RC RH

3-sized : NLL KSK

(B) 빈발 부서열 검색 결과

**그림 7 CMS-Tree 를 이용한 빈발 부서열 탐색 결과**

### 3.2 Hash Tree 와 CMS-Tree 의 비교

#### 3.2.1 단백질 서열 데이터 특성에 따른 CMS-Tree 중복제거효과

Hash Tree 와 CMS-Tree 의 기본적인 차이는 후보서열을 조합하는 부분에서 발생한다. 예를 들어 빈발서열 {{L},{M},{N}}을 가지고 후보서열 {{LL},{MM},{NN}, {LM},{ML},{LN},{NL},{MN},{NM}}만드는 Hash Tree 를 기본으로 하는 알고리즘과 달리 CMS-Tree 를 기본으로 하는 알고리즘은 실제 전체 단백질 서열을 탐색하며 후보서열을 만들어 나간다. 만약 모든 서열에서 {{LL}, {NN},{MN}}이 존재하지 않는다면 이들을 노드로 가지는 후보를 생성하지 않는다. 단백질 서열의 상동성을 비교할 때 같은 구조를 가진 단백질 간의 유사정도가 50%미만이고 30%만 넘어도 유사하다고 여겨지므로(i.e. 희소데이터), CMS-Tree 를 통한 후보서열의 중복제거 효과는 크다. 또한, 그림 8 와 같은 Hash Tree 구조에서는 후보노드를 저장하기위해 매 $k$ 번째 단계마다 뿌리 노드부터 잎노드까지를 생성해야 하는 반면 CMS-Tree 는 $k$-1 번 째 단계에서 이미 생성한 자료구조를 그대로 이용하여 $k$ 번째 노드만 생성한다.

#### 3.2.2 병렬화 효과 비교

서열 연관 규칙 알고리즘의 병렬화를 고려할 때 CMS-Tree 는 Hash Tree 에 비하여 이점이 있다. 단백질 서열 데이터를 병렬컴퓨팅 시스템의 여러 노드에 저장할 때 Hash Tree 를 Concept 으로 사용하는 기존의 병렬 알고리즘은 두 가지 병렬화 기법을 사용할 때 저장공간의 낭비와 노드간 통신시간에 의한 성능저하

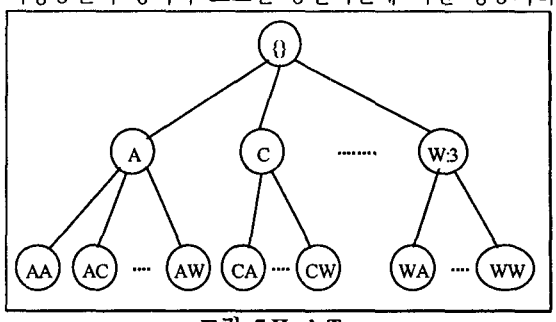



**그림 7 Hash Tree**

와 같은 문제가 발생한다[11]. CMS-Tree 에서는 Data set 으로부터 형성되는 Concept 자체가 Dataset 의 압축형태 이므로 각 각의 노드의 독립적인 연산이 가능하여 병렬화 시에 발생하는 동기화 문제를 피할 수 있다.

## 4. 결 론

본 논문은 단백질 서열분석을 위한 효율적인 서열 연관규칙 알고리즘을 제안하였다. 서열 연관 규칙 알고리즘은 서열간에 존재하는 상동성에 대한 패턴을 얻거나, 상이한 패턴을 가지는 서열간에 존재하는 특정 기능의 모티프를 구하기 위한 전처리로서 사용이 가능하다. CMS-tree 를 이용한 알고리즘은 기존의 Apriori 의 Hash Tree 를 기반으로한 알고리즘에 비해 저장공간 및 탐색시간의 효율성을 보일 뿐 아니라 생물 데이터의 특성을 고려할 때 확장성 및 병렬화를 고려한 방법이다. 방대한 실제 생물 데이터에 대한 서열 연관 규칙을 적용으로 실험으로 규명해야 데이터의 양을 줄여 주는 효과는 물론 인지하지 못했던 중요 데이터의 패턴의 발견에도 기여 할 것으로 예상된다.

## 5. 참고문헌

[1] Minoru Kanehisa, Post-genome informatics, Institute for chemical research koyto university, japan, oxford university press (2000)
[2] Proteins: Struct. Funct. Genet., Suppl. 3, Whole issue(1999).
[3] http://www.ncbi.nlm.nih.gov/BLAST/.
[4] http://www.ebi.ac.uk/fasta3/.
[5] http://www.sdsc.edu/MEME/meme.2.2/website/ meme.html.
[6] http://kr.expasy.org/prosite/
[7] http://www.rcsb.org/pdb/
[8] 김정자, 이도헌, 백윤주, " 단백질 구조 예측을 위한 서열 연관규칙 탐사" , 한국 정보처리학회 논문지, 8-D(5), pp. 553-560, 2001.
[9] R. Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. (1994).
[10] R. Agrawal and R. Srikant. Mining sequential patterns. In Proc. 1995 Int. Conf. Data Engineering(ICDE' 95), pages 3-14, Taipei, Taiwan, March 1995.
[11] D. Skillcorn, " Strategies for Parallel Data Mining," IEEE Consistency, OCT-DEC, (1999).
[12] http://kr.expasy.org/sprot/