서열 분석을 위한 연관 규칙 탐사

김정자[U] 이도헌
전 남 대 학 교 컴 퓨 터 정 보 학 부
{jjkim, dhlee}@dblab.chonnam.ac.kr

Association Rule Discovery for Sequence Analysis

Jung-Ja Kim, Doheon Lee
School of Computer and Information, Chonnam National University

요 약

최근 지놈(Genome) 프로젝트를 통해 핵산, 단백질 서열 정보가 밝혀짐에 따라 분자 수준의 유전자 정보를 다루는 기법들이 활발히 연구되면서 방대한 서열 정보를 데이터 베이스화하고, 분석하기 위한 효과적인 도구와 컴퓨터 알고리즘의 개발을 필요로 하고 있다. 본 논문에서는 여러 단백질에 공통적으로 존재하는 서열 정보간에 존재하는 연관성을 탐사하기 위한 서열 연관 규칙 알고리즘을 제안한다. 원자 항목을 취급하였던 기존 알고리즘과는 달리 중복을 반영해야 하는 서열 데이터의 특성을 고려하여야 한다. 실험은 단백질 서열 데이터를 대상으로 수행하였다. 먼저 여러 서열에 빈발하게 발생하는 부 서열 집합을 찾고, 부 서열 집합들간에 존재하는 관련성을 탐사한다. 본 연구의 결과는 탐사된 규칙으로부터 다른 단백질의 구조와 기능을 예측할 수 있고, 이 정보는 필요로 하는 생물학적 분석을 방향을 제시할 것이다. 이는 생물학적 실험 대상의 후보조합을 최소화함으로써 많은 시간과 노력 비용을 절감할 수 있다.

1. 서론

생물 정보학(Bioinformatics) 은 생물학 연구에 의하여 생성된 제반의 정보를 컴퓨터 및 전산 기법을 사용하여 저장, 검색, 분석, 가공하는 연구 분야를 말한다. 특히 최근 인간 지놈(Human Genome) 프로젝트를 통하여 대량의 유전자 정보가 산출됨에 따라 그 중요성이 날로 증가하고 있다 [1]. 현재 유전체 프로젝트에서 연구하고 있는 생물 정보학 기술들은 다음과 같다. 연구 정보의 저장과 검색을 위한 통합 데이터 베이스로서 데이터 웨어하우스의 구축과, 서열의 특성 및 진화적 관계를 파악하기 위한 서열 분석 알고리즘과 프로그램 등을 개발하고, 더불어 데이터 마이닝 기법을 통하여 새로운 지식들을 발견하고자하는 연구가 진행되고 있다 [2].

핵산(DNA, RNA)은 단백질을 생산하는데 필요한 정보를 암호화하여 다음 세대에 전달해주며, 생명체가 살아가는데 필요한 일을 수행하는 것은 단백질이다. 현재 유전체 연구 방식의 하나인 단백체학(Proteomics)은 여러 개의 단백질의 발현을 동시에 살펴보는 것이다. 그 중 한가지 예로 단백질의 구조 분석을 통하여 단백질의 기능 및 상호작용을 분석하는 문제는 현재 중요한 이슈가 되고있다. 이와 관련하여 핵산(DNA, RNA)이나 단백질 서열분석을 통하여 다른 유전자와의 상관관계를 연구하는 유전자 정렬이나, 유전자 재배치 문제들에 다양한 알고리즘들이 개발되어 많은 도움을 주고 있다 [2][3].

데이터 마이닝은 방대한 양의 데이터로부터 흥미 있는 패턴이나 특성을 발견하는 기법이다. 발견 대상 패턴에 따라서 탐사기법 또한 일반화(generalization), 분류(classification), 군집화(clustering), 연관 규칙(association rule), 순차 패턴(sequential pattern), 신경망(neural network)등이 있다. 특히 통계학과 관련하여 분류나 군집화 기법 등은 생물 정보학 에서도 이미 응용되고 있다 [2][3][4]. 연관 규칙이란 동시에 발생하는 사건들을 규칙의 형태로 표현하는 것으로 특정 사건이 발생하면 동시에 혹은 일정한 시간 간격사이에 다른 사건이 일어나는 관련성을 의미한다 [5][6][7][8]. 본 논문에서는 이를 단백질 구조 분석 문제에 적용하고자 한다. 단백질(Protein)을 구성하는 1차원적인 아미노산 서열로부터 단백질의 입체 구조를 예측할 수 있고 이로부터 발견된 공통적인 패턴을 통하여 다른 단백질의 기능 및 상호 작용을 분석할 수 있다. 서열 연관규칙은 서열을 구성하는 임의의 부 서열 조합들간의 연관성을 의미한다. 예를 들어 임의의 단백질 서열을 분석해 보았더니 adde서열과 pqsrs란 서열이 나타나면 fgkll의 서열이 반드시 나타나는 것이다. 이는 {adde, pqsrs} ⇒ {fgkll}라는 연관규칙 으로 표기할 수 있다. 유전자 연구에 위의 연관 규칙 알고리즘을 적용한다면 규칙에 의해 발견된 특정 단백질을 구성하는 부 서열 군이 반드시 다른 서열군의 출현을 초래한다는 사실을 예측할 수 있다. 발견 패턴 자체가 특정 단백질 구조가 될 수 도 있고, 더 나아가서는 발견된 규칙 패턴과 동일한 서열 조합을 갖는 다른 단백질을 기능을 유추할 수 있다. 물론 실제적인 생물학적 실험을 거쳐 증명하여야 한다 하더라도 실험에 소요되는 많은 문제점이나 노력들을 배제할 수 있다.

본 논문에서는 임의의 단백질 서열 데이터를 대상으로 이들에 빈발하게 출현하는 공통적인 서열집합들을 찾고, 그들간에 존재하는 관련성을 탐사하는 서열 연관 규칙 알고리즘을 제안한다. 기존의 트랜잭션 데이터를 대상으로 한 연관 규칙 알고리즘과 비교하여 본 논문에서 제안한 알고리즘은 첫째, 데이터의 형태가 다르다. 가장 널리 알려진 Apriori 연관 규칙 알고리즘의 경우만 보더라도 원자 항목의 데이터를 취급하였지만, 부

서열 알고리즘 에서는 20개의 알파벳의 조합으로 된 단백질 서열 데이터를 대상으로 한다. 둘째, 생성된 규칙 항목의 조합 또한 부 서열 데이터군의 조합이라는 점이다. 이를 해결하기 위해서는 먼저 여러 단백질에 빈발하게 발생하는 부 서열 집합을 찾는데 Apriori 와는 다르게 데이터의 중복을 고려해 주어야 하며, 다음 단계에서 부 서열 집합들간에 존재하는 관련성을 탐사해내는 것이다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문에서 제안한 서열 연관 규칙 알고리즘을 정의하고 3장에서는 실제 단백질 데이터를 축소한 예로써 알고리즘의 수행 과정을 보이며 마지막 4장에서 결론을 내리고 문제점과 추후 연구방향을 논한다.

2. 서열 연관 규칙 알고리즘

2.1 서열 연관 규칙 (Sequence Association Rule)

정의 1. 서열 연관 규칙(SAR)
기호(symbol)의 집합 Σ의 원소로 구성된 서열들의 조합을 S 라고 할 때 서열 연관 규칙 은 $R : \{s_1, s_2, \ldots\ldots, s_{m-1}) \Rightarrow s_m$ 으로 표시되고 이때 모든 i에 대해서 $s_i \in S$ 이 성립한다. □

규칙 타당성 척도로서 지지도(Support)와 신뢰도(Confidence)의 정의는 다음과 같다.

정의 2. 지지도(support)
기호(symbol)의 집합 Σ의 원소로 구성된 임의의 서열의 전체 집합을 P 라고 할 때 서열 연관 규칙 R에 대한 지지도는 다음과 같이 계산되고, 아래의 성립 요건들을 만족해야 한다.

· 지지도 $(R_p) = \frac{\text{규칙을만족하는서열수}(R \cup P)}{\text{전체단백질서열수}(N)}$ □

임의의 서열 P가 Σ에 속하는 $a_1, a_2, \ldots\ldots, a_n$의 서열로 구성될 때 규칙을 지지하기 위한 각 빈발 서열 아이템에 대한 필요 충분 조건은 다음과 같다.

(1) 규칙을 구성하는 임의의 서열 조합 si와 전체 서열 P는 $s_i \le P$를 만족해야 한다.
si = <$a_1, a_2, \ldots\ldots, a_p$>의 서열로 구성되고
P = <$b_1, b_2, \ldots\ldots, b_q$>의 서열로 구성된다고 할 때
이때 si ≤ P를 만족하기 위한 서열 원소들 간의 필요 충분 조건은
$a_1 = b_k \wedge a_2 = b_{k+1} \wedge \ldots\ldots \wedge a_p = b_{k+p-1}$
k = begin (si|P)
k+p−1 = end (si|P) 이다.
이때 begin은 임의 서열 조합의 시작위치부분이고 end는 끝 위치 부분이다.

(2) P에 속하는 서로 동일하지 않은 모든 s_k와 s_l에 대하여, $begin(s_k) > end(s_l)$ 나 $end(s_l) < begin(s_k)$를 만족해야 한다.

정의 3. 신뢰도(confidence)
서열 연관 규칙 SAR의 신뢰도는 규칙의 조건부를 만족하는 서열들의 수에 대해 결론 부까지를 동시에 만족하는 서열 수의 비율을 의미하며 다음과 같이 정량화 된다.

· 신뢰도$(R_p) = \frac{\text{지지도}(P \cup R)}{\text{지지도 } P}$ □

그림 1의 예를 보면 서열 연관 규칙 알고리즘에 의해 발견된 규칙 (mmmdil, apht) ⇒ (lsrs) 는 지지도 : 3/10 (30%), 신뢰도 : 3/5 (60%)를 만족한다.

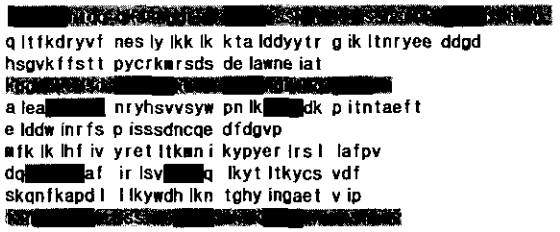

그림 1. 연관 규칙에 의해 발견된 단백질 서열

3. 서열 연관 규칙 탐사 단계

서열 연관 규칙 알고리즘에 적용하는 입력 데이터는 특정 종들의 단백질 서열 데이터를 대상으로 한다. 단백질은 20종류의 아미노산으로 구성되므로, 특정 단백질은 20가지 알파벳의 문자 스트링의 조합으로 해석한다. 서열 연관 규칙 탐사 알고리즘은 다음의 탐사단계를 거친다. 단계1에서 빈발 부 서열 집합을 찾고 알고리즘은 그림 2와 같다. 단계2에서는 공집합이 아닌 모든 빈발 서열의 부분집합에 대하여 각 빈발 서열간에 존재하는 연관 규칙을 추출하는 것이다.

3.1 빈발 부 서열 집합 추출

빈발 서열 집합은 단계마다 새로운 후보 항목 조합을 만들어 가는 결합(Join)과정과, 전 단계에서 포함되지 않은 후보 항목은 제거 해 나가는 전정(Prune)과정을 거쳐 다음 후보 항목을 결정하는 과정을 반복한다.

```
Frequent Sequences(){
    MakeFirstFrequenceSet();
    while( ResList[VCount] != NULL ) {
        MakeCandidateItems(VCount+1);
        /* 모든 가능한 후보를 생성 */
        CheckFrequenceItems(VCount+1);
        /* 각 항목에 대한 빈도 계산 */
        FilteringCandidateItems(VCount+1);
        /* Support 이하 제거 */
        InitialDanDataDCount();
        /* 다음 단계를 위한 초기화 */
        CheckRightAttachItem(VCount+1);
        /* 빈발서열에 대해 추가 항목 검사 */
        VCount ++;}}

/* 단백질 종별 빈발 아이템 분류 */{
        MarkingID();
        /* Frequence Item에 숫자 ID를 할당 */
        OutputResult(); /* 결과 파일을 생성함 */ }
```

그림 2. 빈발 서열 추출 알고리즘

<표 1>. 단백질 데이터 베이스

PID	sequence
P1	mmmdilntqq qkaaeggrvl aphtisskiv krlsshsshk lsrsdlkalg
P2	qltfkdryvf neslylkklk ktaldayytr gikltnryee ddgd
P3	hsgvkffstt pycrkmrsds delawneiat
P4	kpglnkelsd mmmdilkawl aphtngrtmq lsrsem
P5	aleammmdil nryhsvvsyw pnlkaphtdk pitntaeft
P6	elddwinrfs pisssdncqe dfdgvp
P7	mfkiklhfiv yretltkmni kypyerlrsl lafpv
P8	dqmmmdilaf irlsvaphtq lkytltkycs vdf
P9	skqnfkapdi llkywdhlkn tghyingaet vip
P10	fanmmmdilw lssifeapht mkrklsrsln rfsnilv

<표1>의 단백질 데이터 베이스를 입력 데이터로 취하여 수행 결과로 발생한 빈발 서열은 그림 3과 같다.

```
Total Read Record    10

[[[1 's Frequence Set]]]
        a(20),c(3),d(26),e(16),f(17),g(13),h(12),i(18),k(30),l(42),m(22),
        n(18),p(17),q(9),r(19),s(31),t(26),v(15),w(6),y(15),
[[[2 's Frequence Set]]]
        ae(3),al(3),ap(6),dd(3),di(5),el(3),fk(3),fs(3),hs(3),
        ht(5),il(6),ka(5),kl(7),ky(4),la(5),lk(9),ll(3),ln(4),
        ls(7),lt(4),md(5),mm(10),nr(4),nt(3),ph(5),ql(3),rl(3),
        rs(5),ry(3),sd(5),sl(3),sr(3),ss(6),sv(3),tw(3),
[[[3 's Frequence Set]]]
        aph(5),dil(5),lka(3),lsr(3),mdi(5),mmd(5),mmm(5),
        phl(5),srs(3),
[[[4 's Frequence Set]]]
        aphl(5),lsrs(3),mdi(5),mmdi(5),mmmd(5),
[[[5 's Frequence Set]]]
        mmdi(5),mmmdi(5),
[[[6 's Frequence Set]]]
        mmdil(5),
```

그림 3. 단계1 수행 후 빈발서열집합

3.2 빈발 부서열간의 연관규칙 발견

PID	FID
P1	1, 2, 3, 5, 10, 3 5 11, 12, 13, 14, 16, 19, 20, 22, 26, 25, 26, 28, 29, 31, 33, 34, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53
P2	2, 4, 7, 14, 17, 21, 24, 27, 30, 32, 36
P3	6, 8, 9, 10, 16, 29, 31
P4	3, 5, 6, 11, 12, 13, 16, 17, 19, 20, 22, 26, 26, 27, 29, 33, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53
P5	1, 2, 3, 5, 10, 3 5 11, 12, 13, 17, 19, 22, 26, 24, 25, 26, 30, 35, 36, 37, 38, 39, 41, 42, 43, 44, 46, 48, 49, 50, 51, 52, 53
P6	4, 6, 8, 9, 24, 31, 34, 51, 52, 53
P7	7, 14, 15, 16, 17, 18, 21, 28, 29, 32
P8	3, 5, 11, 12, 15, 16, 17, 20, 21, 22, 26, 26, 27, 28, 35, 37, 38, 41, 42, 43, 44, 46, 48, 49, 50
P9	1, 3, 7, 13, 15, 17, 18, 25
P10	3, 5, 8, 11, 12, 14, 19, 20, 22, 26, 24, 26, 29, 32, 33, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53

그림 4. 단백질 종에 따른 빈발 서열 분류

그림 4는 1단계의 빈발 부 서열 집합을 단백질 종별로 분류 한 것이며 항목에 대해 연관 규칙 탐사과정을 수행한다. 이의 알고리즘과 수행 결과는 그림 5와 같다.

```
Generate Rules(){
for each frequent sub sequence set F do
    for each subset of F do
        if(support(F)support(F-f)  ≥  minimum  confidence)
        then
        output the rule (F-f) ⇒ fi
          with confidence = support(F)/support(F-f)
            and support = support(F);}
```

```
11 <- 3 44 49 22  (50.0%, 100.0%) : HT <- AP PHT MMDI MD
49 <- 3 44 11 22  (50.0%, 100.0%) : MMDI <- AP PHT HT MD
44 <- 3 49 11 22  (50.0%, 100.0%) : PHT <- AP MMDI HT MD
3 <- 44 49 11 22  (50.0%, 100.0%) : AP <- PHT MMDI HT MD
```

그림 5. 규칙 생성 알고리즘과 발견 규칙 예

4. 결론

본 논문에서는 단백질 서열 분석을 통하여 그들간에 존재하는 관련성을 탐사하는 알고리즘을 논하였다. 기존 Apriori 알고리즘은 단일 항목 단위의 트랜잭션 데이터를 대상으로 하지만 서열 연관 규칙 알고리즘의 대상은 단일 항목에 대응하는 한 문자 이상의 조합으로 구성된 서열의 집합이다. 그러므로 단백질 부 서열 알고리즘에서는 중복을 고려하여야 한다. 실험 데이터를 수행해본 결과 규칙에 나타난 빈발 서열들의 부분집합 문제와 빈발 서열들간의 중첩문제들로 인하여 많은 규칙이 발생하였다. 이는 수행 시 시스템에 과부하와 실제 생물학적으로 의미 있는 정보의 추출에 대한 신뢰성을 보장할 수 없다. 이는 지지도 적용단계에서 제약조건으로서 제한 할 수 있다. 연관규칙의 중요성을 평가하기 위한 여러 측정치들에 대한 문제도 향후 과제로 남아있다.

본 논문에서 제안한 서열 연관 규칙 알고리즘은 실험 데이터의 대상을 단백질 서열을 취급하였지만 핵산 서열에도 적용이 가능하다. 생물학적으로 발견된 규칙은 규칙을 구성하는 빈발 항목 자체가 특정 단백질이 될 수도 있고, 다 나아가서 발견된 패턴이 어떠한 기능을 하는지 등의 정보를 예측하고 실험 연구의 방향을 계획할 수 있다. 이는 실제 현장에서 실험 후보 조합의 수를 감소시키므로 써 많은 시간과 비용, 노력을 절감할 수 있다.

참고 문헌

[1] R. Hofestaedt, "Computer science and biology-the German Conference on Bioinformatics(GCB'96), BioSystems 43 (1997) 69-71

[2] Setubal J, Meidanis J. Introduction to Computational Molecular Biology. Boston, MA:PWS Publishing Company, 1997[7].

[3] Alvis Brazma, Inge Jonassen, Ingvar Eidhammer, David Gilbert, Approaches to the automatic discovery of pattern biosequences, In the Journal of Computational Biology, November 5, 1997

[4] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, "GeneCards:a novel functional genomics compendium with automated data mining and query reformulation support, Bioinformatics, V.14 N.8, pp.656-664, 1998080]

[5] Brachman, R. J. and Anand T. (1996), The Process of Knowledge Discovery in Databases. Advance in knowledge Discovery in Database and Data Mining(37-57). Menlo Park:AAAI/MIT Press

[6] R. Agrawal, T. Imielinski and A. Swami. "Mining Association Rules between Sets of Items in Large Database", Proc, ACM SIGMOD, 1993, pp.207-216

[7] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", Proc, VLDB, 1994, pp.487-499

[8] Mohammed J. Zaki, "Scalable Algorithms for Association Mining," IEEE Transactions on Knowledge and Engineering, V.12, N.3, May/June 2000