단백질 구조 예측을 위한 서열 연관 규칙 탐사

김 정 자[†] · 이 도 헌[††] · 백 윤 주[†††]

요 약

바이오정보학(bioinformatics)은 생물학 분야 특히 분자 수준의 유전체 연구에서 발생하는 데이터를 저장, 관리, 분석하여 실험 프로젝트를 지원함은 물론, 기능 예측 및 조절에 대한 실험 설계를 가능하게 하는 제반 컴퓨터 기술을 의미한다. 유전체 연구의 다양한 접근 방식 중 단백체학(proteomics)는 유전체의 최종 산물인 단백질을 직접적으로 다룬다는 측면에서 그 효용성에 대해 많은 기대를 모으고 있다. 본 논문에서는 단백질의 기능을 결정하는 가장 중요한 요소 중 하나인 단백질의 구조를 예측하기 위한 데이터 마이닝 기법을 제안한다. 단백질의 일차 구조인 아미노산 서열에 나타나는 부서열간의 연관성이 해당 단백질의 이차 혹은 삼차 구조를 결정하는 중요한 단서임을 설명하고, 아미노산 부서열간의 연관성을 표현하기 위한 모델로서 서열 연관 규칙을 정의한다. 서열 연관 규칙의 유용성을 평가하기 위한 지지도와 신뢰도를 새롭게 정의하고, 주어진 단백질 집단으로부터 유용한 서열 연관 규칙을 발견하기 위한 기법을 제안한다. 아울러, SWISS-PROT 단백질 데이터베이스로부터 입수한 단백질 서열 데이터를 이용하여 제안한 기법의 성능을 평가한다.

Discovering Sequence Association Rules for Protein Structure Prediction

Jungja Kim[†] · Doheon Lee[††] · Yunju Baek[†††]

ABSTRACT

Bioinformatics is a discipline to support biological experiment projects by storing, managing and analyzing data arising from genome research. It can also lead the experimental design for genomic function prediction and regulation. Among various approaches of the genome research, the proteomics have been drawing increasing attention since it deals with the final product of genomes, i.e., proteins, directly. This paper proposes a data mining technique to predict the structural characteristic of a given protein group, one of dominant factors of the functions of them. After explains associations among amino acid subsequences in the primary structures of proteins, which can provide important clues for determining secondary or tertiary structures of them, it defines a sequence association rule to represent the inter-subsequence associations. It also provides support and confidence measures, newly designed to evaluate the usefulness of sequence association rules. After it proposes a method to discover useful sequence association rules from a given protein group, it evaluates the performance of the proposed method with protein sequence data from the SWISS-PROT protein database.

키워드 : 바이오정보학(Bioinformatics), 기능 유전체학(Functional Genomics), 단백체학(Proteomics), 연관 규칙(Association Rule), 단백질(Protein)

1. 서 론

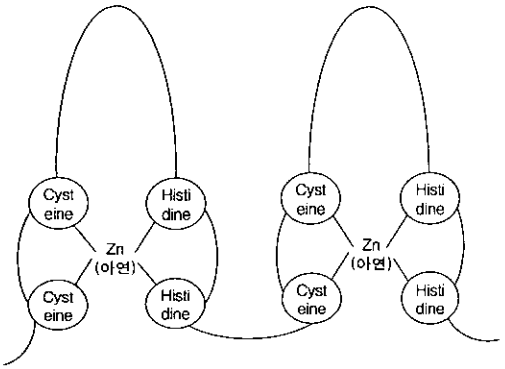
바이오정보학(bioinformatics)은 생물학 분야에서 발생하는 데이터를 저장, 관리, 분석하여 실험 프로젝트를 지원함은 물론, 예측 및 조절에 대한 실험 설계를 가능하게 하는 제반 컴퓨터 기술을 의미한다. 특히, 인간 유전체 프로젝트(Human Genome Project)와 같은 대형 유전체 연구사업을 통해 대량의 유전체 정보가 생성되면서, 분자 수준의 유전체 정보를 다루는 컴퓨터 기술에 대한 관심이 급격히 높아지고 있다[1-6].

† 준 회 원 : 전남대학교 대학원 전산통계학과
†† 정 회 원 : 전남대학교 전산학과 교수
††† 정 회 원 : 네이버컴주식회사 기술이사
논문접수 : 2001년 7월 27일, 심사완료 : 2001년 9월 7일

다양한 유전체 연구 분야 중, 단백체학(proteomics)은 신진대사 기능과 암호 체계들이 DNA나 RNA 수준보다 직접적으로 나타나기 때문에 가장 주목받는 분야의 하나이다[3, 6]. 현재까지 지구상에 존재하는 것으로 알려진 단백질의 종류는 수만종에 달하기 때문에, 그 기능을 모두 밝히는 데는 엄청난 시간과 비용이 필요하다. 따라서, 많은 바이오정보학 도구들이 효율적이면서 효과적으로 단백질의 기능을 밝히기 위해 개발되고 있다. 예를들어 BLAST[7]와 FASTA[8]는 유사한 단백질의 1차 구조를 찾기 위한 검색 서비스이며 PROSITE[9]와 MEME(Multiple EM for Motif Elicitation) [10]같은 다양한 모티프(motif) 라이브러리(libraries)들은 알려진 수많은 모티프를 유지하고 모티프에 기반한 검색과, 비교 서비스를 제공한다. 또한 알려진 염기나,

단백질 서열에대한 3차 구조를 제공해주는 PDB(Protein Data Bank)[11], MMDB(Molecular Modeling Database)[12] 등이 있다.

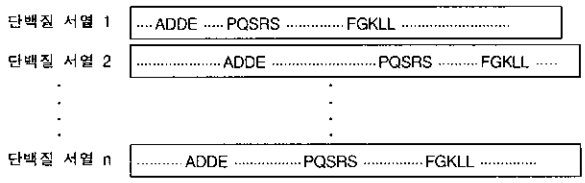
단백질 분자는 아미노산 레지듀(residue)의 선형 중합체이다. 따라서, 엄격히 따진다면 단백질 분자의 1차 구조는 아미노산 레지듀(residue) 선형 서열이라고 불러야 하지만, 바이오정보학 에서는 단순히 단백질 서열이라 부르는 것이 보편적이다. 단백질의 기능은 단백질의 2차, 3차 구조에 의해 결정되며 그들 중 상당수가 1차 구조로부터 예측된다. 예를 들어, (그림 1)과 같이 징크 핑거(zinc-finger)라부르는 잘 알려진 2차 구조는 징크(Zn) 이온을 중심으로 항상 연속된 두 개의 시스테인 (Cysteine) 레지듀들과 연속된 두 개의 히스티딘(Histidine) 레지듀들이 손가락 모양을 이루며 존재함으로써 예측할 수 있다. 이 영역이 DNA의 특정 염기 쌍과 결합하면서 3차 구조의 복합체를 형성하는 것이다[13]. 이와 같이, 단백질 서열에서 임의의 부서열의 동시출현(co-occurrence)은 그들의 2차, 3차 구조에 대한 유용한 단서를 제공하고 이를 통해 단백질의 기능예측을 가능하게 한다.



(그림 1) 징크 핑거 단백질

본 논문에서는 단백질 서열의 집합에서 동시 출현되는 부서열들을 탐사하기 위한 방법을 제안한다. 단백질 부서열의 동시 출현은 서열 연관 규칙의 형태로 표현된다. 서열 연관 규칙은 임의의 길이를 가진 단백질 부서열간의 연관성을 나타낸다. DNA결합(binding), 다른 단백질과의 상호 작용, 암호 체계 인자 등에 공통적인 기능을 가지는 단백질 서열의 집합이 (그림 2)에 주어졌다고 가정하면 하나의 단백질 서열에서 ADDE와 PQSRS 부서열이 나타나면 대부분 FGKLL 부서열도 함께 나타나는 것을 발견할 수 있다. 이러한 관찰에 근거하여 부서열 ADDE, PQSRS와 부서열 FGKLL 간에 상호작용과 해당 기능과의 연관성에 대한 정교한 생물학적 가설을 세우고, 실험을 통해 검증할 수 있게 된다.

이처럼 발견된 서열 연관 규칙은 실험 대상으로 적합한 가설을 선정하는데 유용할 뿐 아니라, 과거에는 전혀 고려하지도 않았던 새로운 발상의 전환을 가능케 할 수 있다.



(그림 2) 단백질 부서열 간의 연관성

데이터마이닝 분야에서 연관 규칙 발견 기법은 트랜잭션 데이터(transactional data)를 대상으로 지난 수년동안 활발히 연구되어 왔다[14-17]. 편의상 본 논문에서 제안하는 서열 연관 규칙과 구별하기 위해 기존의 연관 규칙을 트랜잭션 연관 규칙이라고 한다. 트랜잭션 연관 규칙이 원자적 항목간의 연관성을 나타내는 데 반해, 서열 연관 규칙은 부서열간의 연관성을 나타내므로 규칙의 형태는 물론, 발견 기법 역시 훨씬 복잡하게 된다. 또한 유용한 서열 연관 규칙을 평가하기 위한 지지도와 신뢰도 역시 트랜잭션 연관 규칙의 경우와는 달리 서열의 특성을 반영하기 위해 수정되어야 한다. 본 논문에서 제안하는 서열 연관 규칙 발견 알고리즘(Sequence Association Rule Discovery Algorithm : SARA)은 크게 네 가지 단계로 구성된다. 첫 번째 단계에서는 충분히 많은 수의 단백질 서열에서 공통적으로 발견되는 부서열(subsequence) 집합을 추출한다. 두 번째 단계에서는 역시 충분히 많은 수의 단백질 서열에서 공통적으로 발견되는 부서열 조합(subsequence combination)의 집합을 추출한다. 세 번째 단계에서는 규칙으로서 신뢰도가 높은 연관 규칙을 선별하고 마지막으로 네 번째 단계에서는 의미있는 규칙으로 집약하고자 규칙을 요약한다.

본 논문의 구성은 다음과 같다. 2절에서는 서열 연관 규칙을 정의하고 유용한 서열 연관 규칙을 판별하기 위해 새롭게 정의한 지지도-신뢰도 기준을 제안한다. 3절에서는 제안하는 서열 연관 규칙 탐사 알고리즘 (SARA)을 예제와 함께 설명한다. 4절에서는 SWISS-PROT으로부터 입수한 단백질 데이터를 대상으로 알고리즘을 적용하여 성능을 분석하고 마지막으로 5절에서 결론을 맺는다.

2. 서열 연관 규칙 (Sequence Association Rule)

본 절에서는 서열 연관 규칙(SAR)과, 발견 규칙의 타당성을 보장하기 위한 척도를 정의한다.

2.1 서열 연관 규칙 (Sequence Association Rule)

서열 연관 규칙은 임의의 길이를 가진 서열간의 연관성을 나타내는 규칙이다. 서열 들로부터 발견한 임의의 공통적인 서열들을 S_i, S_j, …, S_k라 할 때, 임의의 서열 S_i, S_j, S_k의 연관 규칙은 S_i, S_j 서열이 나타나면 S_k 서열이 존재하는 관련성을 의미한다.

정의 1. 서열 연관 규칙(sequence association rule)

주어진 기호의 집합 Σ의 원소로 구성된 모든 가능한 서열(sequence)의 집합을 Ψ라고 할 때, 서열 연관 규칙(sequence association rule)은 다음과 같이 정의된다.

$$\{s_1, s_2, \cdots, s_{m-1}\} \Rightarrow s_m$$

단, $1 \leq i \leq m$인 모든 i에 대하여, $s_i \in \Psi$. 이 때, s_i를 해당 서열 연관 규칙의 항목 서열(item sequence)이라고 부른다. □

단백질은 20가지 종류의 아미노산들로 이루어진 조합이다. 예를 들어 아미노산의 집합 Σ = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}이면, 단백질 서열 S = {A, AA, ···, B, BB, AB, ABC, BAC, CAB, ···}가 된다. (그림 3)과 같은 10종의 단백질 서열이 주어졌을 때 MMMDIL, APHT, LSRS 서열들은 여러 단백질에서 빈발하게 나타나며 이로부터 구성된 서열 조합(MMMDIL, APHT, LSRS)은 (MMMDIL, APHT) ⇒ (LSRS)라는 서열 연관 규칙으로 표현한다. 이는 MMMDIL, APHT 서열이 나타나면 LSRS서열이 존재한다는 것을 의미한다.

sequence

P1 : MMMD ILNTQQQKAAEGGRVLAPHT ISSKLVKRLSS HSSHKLSRSDLKALG
P2 : QLTFKDRYVFNESLYLKKLKKTALDDYYTRG IKLT NRYEEDDGD
P3 : HSGVKFFSTTPYCRKMRSDSDELAWNE IAT
P4 : KPGLNKEISDMMMD ILKAWLAPHTNGRTMQLSRSEM
P5 : ALEAMMMD ILNRYHSVVSYWPNLKAPHTDKP ITN TAEFT
P6 : ELDDW INRFSP ISSSDNCQEDFDGVP
P7 : MFKLKLHF IVYRETLTKMN IKYPYERLRSLLAFPV
P8 : DQMMMD ILAF IRLSVAPHTQLKYTLTKYCSVDF
P9 : SKQNFKAPDLLLKYWDHILKNTGHY INGAETV IP
P10 : FANMMMD ILWLSS IFEAPHTMKRKLSRSLNRFSN ILV

(그림 3) 단백질 서열

다음은 탐사된 규칙들이 어느 정도 유용한가의 타당성을 평가하기 위한 척도로서 지지도(Support)와 신뢰도(Confidence)를 정의한다. 이때 서열 연관 규칙에서는 기존의 트랜잭션 연관 규칙 탐사 기법과는 다르게 생물학적인 서열 데이터를 취급하므로 서열의 특성을 반영하여 수정되어야 한다. 서열 연관 규칙 SAR : $\{s_1, s_2, \cdots, s_{m-1}\} \Rightarrow s_m$에 대한 지지도는 다음과 같이 정의되며, 탐사된 규칙조합을 구성하는 서열 항목들이 전체 서열을 지지하기 위해서는 정의 2에 제시된 조건들을 만족하여야 한다.

정의 2. 서열간의 포함과 중첩관계

서열 s의 길이와 i번째 원소를 각각 len(s), s[i]라고 표시하자. 서열 s_1와 s_2에 대해,

$$s_1[1] = s_2[k+1] \wedge s_1[2] = s_2[k+2] \wedge \cdots \wedge s_1[len(s1)] = s_2[k+len(s1)]$$

를 만족하는 k가 하나 이상 존재하면, s_1은 s_2에 포함된다고 하고, $s_1 \subseteq s_2$라고 표시한다.

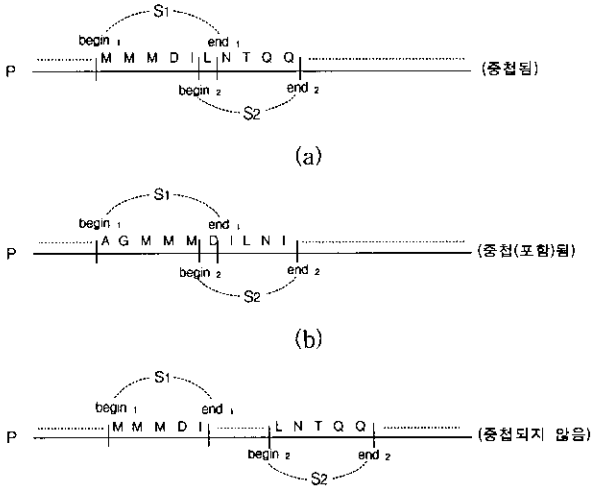
이 때, s_2 상에서 s_1이 나타나는 위치를 표시하기 위하여 다음과 같은 *오프셋 집합(Offset Set)*을 정의한다.

$$Offset(s_1 \mid s_2) = \{(begin, end) \mid s_1[1] = s_2[k+1] \wedge s_1[2] = s_2[k+2] \wedge \cdots \wedge s_1[len(s1)] = s_2[k+len(s1)] \wedge begin = k+1 \wedge end = k + len(s1)\}$$

$s_1 \subseteq p$, $s_2 \subseteq p$를 만족하는 세 개의 서열 s_1, s_2, p가 있을 때, $\exists (begin_1, end_1) \in Offset(s_1 \mid p)$, $\exists (begin_2, end_2) \in Offset(s_2 \mid p)$ $(end_1 < begin_2$ or $end_2 < begin_1)$

을 만족하면, *s_1과 s_2는 p에 대하여 중첩되지 않는다*고 정의한다. □

부 서열 *s_1과 s_2가 p에 대하여 중첩되지 않는다는 것*은 (그림 4)의 (a), (b)와 같이 s_1, s_2가 위치적으로 겹치거나 포함되지 않고 (c)와 같은 형태로 존재하는 경우가 있다는 것을 의미한다.



(그림 4) 서열 간의 중첩 관계

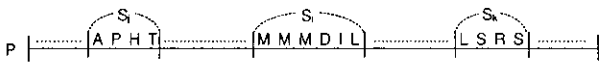
정의 3. 서열 연관 규칙의 지지도(support)

주어진 서열의 집합 P 에 대한, 서열 연관 규칙 $\{s_1, s_2, \cdots, s_{m-1}\} \Rightarrow s_m$ 의 지지도는 다음과 같이 정의된다.

$$Support(\{s_1, s_2, \cdots, s_{m-1}\} \Rightarrow s_m \mid P) = |\{p \mid p \in P \wedge \{s_1, s_2, \cdots, s_m\} \angle p\}| / |P|$$

단, $1 \leq i \leq m$인 모든 i에 대하여, $s_i \subseteq p$이고, $\{s_1, s_2, \cdots,$

s_m}에 속하는 모든 서열쌍이 p에 대하여 중첩하지 않으면, {s_1, s_2, …, s_m} ∠ p가 성립한다. □



(그림 5) {s_1, s_2, …, s_m} ∠ p 의 조건

즉 s_i ⊏ p임은 규칙을 구성하는 모든 서열 항목들이 모든 단백질에서 발생해야 함을 의미하며, (그림 5)와 같이 규칙을 구성하는 모든 서열 쌍이 p에 대하여 중첩하지 않아야 한다.

서열 연관 규칙 SAR : {s_1, s_2, …, s_{m-1}} ⇒ s_m 지지도는 전체 단백질 서열 수에 대해 규칙을 만족하는 서열 수의 비율을 말한다. (그림 2)의 단백질 데이터의 경우 전체 서열에서 서열 MMMDIL은 5번, APHT도 5번, LSRS는 3번 나타나지만 (MMMDIL, APHT, LSRS) 규칙조합은 전체 10개의 단백질 서열중에서 3개의 서열(P1, P4, P10)에서 발생하기 때문에 서열 연관 규칙 (MMMDIL, APHT) ⇒ (LSRS)은 지지도 : 3/10 (30%)라고 할 수 있다.

정의 4. 서열 연관 규칙의 신뢰도(Confidence)

주어진 서열의 집합 P 에 대한, 서열 연관 규칙 {s_1, s_2, …, s_{m-1}} => s_m 의 신뢰도는 다음과 같이 정의된다

$$\text{confidence}(\{s_1, s_2, \cdots, s_{m-1}\} \Rightarrow s_m \mid P) = |\{p \mid p \in P \wedge \{s_1, s_2, \cdots, s_m\} \angle p\}| \,/\, |\{p \mid p \in P \wedge \{s_1, s_2, \cdots, s_{m-1}\} \angle p\}|$$

단, 1 ≤ i ≤ m인 모든 i에 대하여, s_i ⊏ p이고, {s_1, s_2, …, s_m}에 속하는 모든 서열쌍이 p에 대하여 중첩하지 않으면, {s_1, s_2, …, s_m} ∠ p 가 성립한다. □

서열 연관 규칙 신뢰도(Sequence Association Rule Confidence)는 서열 연관 규칙 {s_1, s_2, …, s_{m-1}} ⇒ s_m에서 {s_1, s_2, …, s_{m-1}}를 규칙의 조건 부(전제 부), s_m를 결론 부 라 할 때 규칙의 조건부를 만족하는 서열 수에 대해 결론 부까지를 동시에 만족하는 서열 수의 비율을 의미한다. (그림 2)의 단백질 데이터로부터 발생한 서열 연관 규칙 (MMMDIL, APHT) ⇒ (LSRS)의 신뢰도는 MMMDIL, APHT 서열을 포함하면서 LSRS 서열까지 발생한 비율을 의미한다. 5개의 단백질 서열(P1, P4, P5, P8, P10) 중에서 3개의 서열(P1, P4, P10)에서 발생하므로 서열 연관 규칙(MMMDIL, APHT) ⇒ (LSRS)은 신뢰도 : 3/5(60%)를 만족한다.

3. 서열 연관 규칙 발견 기법

SARA에 적용하는 입력 데이터는 특정 종들의 단백질 서열 데이터를 대상으로 한다. 단백질은 20종류의 아미노산으로 구성되므로, 특정 단백질은 20가지 알파벳의 문자 스트링의 조합으로 해석한다. SARA는 다음 4단계의 탐사과정을 거친다. 단계1에서 빈발 서열 집합을 찾는다. 단계2에서는 규칙을 구성하기 위한 빈발 부 서열 항목 조합을 구한다. 각 후보 서열 항목 생성 전에 부분집합과 중첩문제를 해결하여 선택된 빈발 서열 항목으로만 구성된 규칙조합을 추출한다. 단계3에서 일정 지지도와 신뢰도 이상의 규칙만을 생성하고, 마지막으로 단계4에서는 규칙들간에 포함관계에 있는 규칙들은 제거하여 요약된 규칙을 보인다. 본 논문에서는 4단계에서 임의의 규칙에 대하여 대응되는 규칙에서 전제부와 결론부 모두가 포함되는 경우 이를 부 규칙(sub rule) 이라 부른다.

3.1 제 1단계 : 빈발 부 서열의 추출

이 단계는 주어진 서열 집합으로부터 특정 빈도이상 빈발하게 발견되는 임의의 길이를 가진 부 서열들을 추출하는 단계이다. 새로운 후보 서열 조합을 만들어 가는 결합과정과, 지지도 미만의 서열은 제거해 나가는 전정과정을 거쳐 다음 후보 서열을 결정하는 과정을 반복한다. 추출된 빈발 서열 집합들은 각 단백질종별 빈발 서열 항목으로 재배열되며 알고리즘은 (그림 6)과 같다.

```
Σprotein = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V,
W, Y} Input : P = {Pi | Pi는 Σprotein 원소들의 조합, 0 <i ≤ n},
support /* 지지도 */
FrequencySet MakeCandidateSequence( FrequentSequence FS,
            FrequencySet cFS_Set){
            NewFS = Ø ;
            Each fs ∈ cFS_set
                /* FSi는 FS에서 i번째 단백질 */
                if( FS2≡fs1 ∧FS3≡fs2 ∧…∧FSn≡fsn-1 ) {
                 NewFS = { FS1FS2…FSnfsn} ∪ NewFS ; }
            end Each
            return NewFS ; }

 FrequencySequenceExtraction() {
   FS_Set1 = Σprotein ;
   /* FrequencyCount(fs)는 단백질 서열들에서 fs의 발생 횟수 */
(1)          Each fs ∈ FS_Set1
                    if( FrequencyCount(fs) < support )
                                FS_Set1 = FS_Set1 - fs ;
             end Each
             if ( FS_Set1 = Ø ) return ;
             m = 2 ;
             while (1) {
            FS_Setm = Ø ; /* 길이가 m인 빈발 서열들의 집합 */
            Each fs ∈ FS_Setm-1
(2)          /*  MakeCandidateSequence(fs, FS_Setm-1) : 가능한
                                            후보 서열 생성 */
               new_fs = MakeCandidateSequence(fs, FS_Setm-1)
               FS_Setm = FS_Setm ∪ new_fs
               end Each
               if ( FS_Setm == Ø ) break ;
(3)            Each fs ∈ FS_Setm
                        if( FrequencyCount(fs) < SS )
                                FS_Setm = FS_Setm - fs ;
               end Each
               if ( FS_Setm == Ø ) break ;
            m = m+1 ;
            FS_Set = FS_Set ∪ FS_Setm ;
            }
   }
```

(그림 6) 빈발 서열 추출 알고리즘

(1)에서는 지지도 이상의 하나의 알파벳 서열로 구성된 초기 후보 서열 조합을 구성하고, 이로부터 각 단계별 모든 가능한 후보 서열을 (2)에서 생성시킨다. (3)에서는 생성된 각 서열 조합에 대한 빈발 수를 계산하여 아이템 지지도 이하의 서열들은 제거한다. 추출된 빈발 서열 조합들은 각 단백질별 존재하는 서열조합들로 재분류하여 2단계 알고리즘을 수행하기 위한 입력 형태로 변환한다. (그림 2)의 단백질 데이터를 받아들여 서열 지지도 30%일 때 생성된 빈발 부서열 집합들은 (그림 7)과 같다.

```
[[[ 1 's Frequence Set]]]
    A( 9),C( 3),D( 9),E( 9),F( 8),G( 6),H( 8),I( 10),K( 9),
    L( 10),M( 7),N( 9),P( 9),Q( 6),R( 9),S( 10),T( 9),V( 9),
    W( 6),Y( 6),
[[ 2 's Frequence Set]]]
    AE( 3),AL( 3),AP( 6),DI( 5),EL( 3),FK( 3),FS( 3),HS( 3),
    HT( 5),IL( 5),KA( 4),KL( 4),KY( 3),LA( 5),LK( 7),LN( 4),
    LS( 4),LT( 3),MD( 5),MM( 5),NR( 4),NT( 3),PH( 5),QL( 3),
    RL( 3),RS( 5),SD( 4),SL( 3),SR( 3),SS( 3),TN( 3),
[[[ 3 's Frequence Set]]]
    APH( 5),DIL( 5),LKA( 3),LSR( 3),MDI( 5),MMD( 5),
    MMM( 5),PHT( 5),SRS( 3),
[[[ 4 's Frequence Set]]]
    APHT( 5),LSRS( 3),MDIL( 5),MMDI( 5),MMMD( 5),
[[[ 5 's Frequence Set]]]
    MMDIL( 5),MMMDI( 5),
[[[ 6 's Frequence Set]]]
    MMMDIL( 5),
```

(그림 7) 빈발 부서열 집합

3.2 제 2단계 : 빈발 부서열 조합 발견

이 단계는 1단계의 결과를 이용하여 특정 빈도이상 발견되는 부서열 조합을 모두 추출하는 단계이다. 1단계의 빈발 서열 조합들은 (빈발 서열 ID, 단백질 ID, offset)로 분류함으로써 빈발 서열 조합을 다입 원자 하모으로 처리하였다. 부서열 조합을 추출할 때, 서열 데이터의 특성상 정의 2에 제시된 부서열간의 중첩 관계를 고려해야 한다. (그림 8)은 2단계의 알고리즘이다. (2)에서는 각 단계 후보 규칙 항목 생성 전에 중첩 관계를 가지는 항목들을 걸러낸다. 서열조합 항목의 시작과 끝 위치 값을 검사하여 포함되는 관계와 중첩되어있는 항목을 제거함으로써 정확하게 규칙을 지지하는 항목들만을 선택한다. (3)에서는 3단계에서 임계치 이상의 연관 규칙을 추출하기 위해 규칙 항목 조합을 포함하는 단백질 수를 계산한다.

```
Input : Database = {Pid, Pid_FS_Set = {Pid_FS_Setm | Pid_FS_Setm는 임의의 Pid
                   단백질 내에 존재하는 모든 빈발 서열 항목 조
                   합}}
        /*  k-itemset : k 항목을 갖는 항목집합  */
        /*  Ck : 후보 k 항목집합들의 집합  */

SubsetAndOverlapFiltering() {
        L1 = {Large 1-itemsets} ;
(1)     For (k = 2 ; Lk-1 ≠ Ø ; k++) {
            /*  Gen(Lk-1) : 새로운 후보 생성  */
            Ck = Gen(Lk-1) ;
            Each FID_Set ∈ Ck
(2)         Forall pairs fIDi, fIDj,∈FID_Set, i≠j {
            /* IsOverlap(fIDi, fIDj)는 두 항목간 overlap 검사
               IsSubset(fIDi, fIDj)는 두 항목간 subset 검사 */
               if(IsOverlap(fIDi, fIDj) || IsSubset(fIDi, fIDj)) {
                    Ck = Ck-FID_Set ;
                    break ; }
            end Forall
        end Each
(3)     Forall proteins P∈Database
    /* subset(Ck, P) : P에 포함되어있는 후보 규칙 항목조합들 */
        Cp = subset(Ck, P) ;
        Forall candidates c ∈ Cp
            c.count++ ;
                    end Forall
    end Forall}
```

(그림 8) 빈발 부서열 조합 발견 알고리즘

3.3 제 3단계 : 서열 연관 규칙 도출

이 단계는 2단계의 최종 선택된 후보 규칙 조합으로부터 일정 지지도와 신뢰도를 적용하여 타당성 있는 서열 연관 규칙을 선별한다. (그림 9)는 3단계의 알고리즘이고 (그림 10)는 서열 연관 규칙 지지도 30%, 신뢰도 90%를 적용하였을 때 결과로 생성된 규칙의 예이다.

```
Input : # of read records  /*  단백질 수  */,
        c.count  /*  규칙 후보 항목 조합수  */,
        SEQsupp  /*  서열 연관규칙 지지도  */,
        SEQconf  /*  서열 연관 규칙 신뢰도  */
        /*  Ck : 후보 규칙 항목 조합  */
            For each c∈Ck do
                if(c ≥ SEQsupp and c ≥ SEQconf) then
            /*  Gen_Rules(c) : 후보 조합을 이용한 연관 규칙 생
                                성  */
                Ass_Rules = Gen_Rules(c)
                            end
```

(그림 9) 연관 규칙 알고리즘

```
                .
                .
          G <- LA SD (30.0%, 100.0%)
       APHT <- W MMMDIL (30.0%, 100.0%)
    MMMDIL <- LA APHT (30.0%, 100.0%)
    MMMDIL <- APHT LSRS (30.0%, 100.0%)
    MMMDIL <- APHT LKA (30.0%, 100.0%)
```

(그림 10) 발견 규칙 예

3.4 제 4단계 : 중복 규칙 제거

이 단계에서는 3단계에서 얻어진 서열 연관 규칙의 전제부와 결론부를 분석하여 중복된 규칙을 제거함으로써 규칙을 요약한다. 2단계에서는 하나의 규칙 내 에서 규칙을 지지하는 각 빈발 항목 조합간의 부분집합 관계를 제거하였다. 3단계에서 발견된 규칙에는 규칙들간에 부분집합 관계

에 있는 많은 수의 규칙이 또한 존재한다. 많은 수의 규칙은 생물학적으로 다양한 모든 경우의 수를 보임으로써 여러 관점으로 해석할 수 있겠지만, 반복되고 포함되는 무의미한 규칙들은 의미 있는 규칙의 의미를 약화 할 수도 있다. 4단계에서는 의미적으로 함축된 규칙으로 나타냄으로써 규칙들을 집약한다. 본 논문에서 사용하는 요약방법은 하향식 요약 방식에 근거하여 다른 규칙을 포함하는 최대 항목과 최대 서열을 갖는 규칙 항목조합으로 요약한다. 예를 들어, (그림 10)의 세 번째 규칙의 경우 그 이전에 MMM <- LA, APHT와 MMMDIL <- LA 또는 MMMD <- L, PHT 등의 규칙들이 모두 생성되어 있다. 이는 MMMDIL <- LA, APHT 규칙에 전제부와 조건부 모두가 포함되어지는 관계이므로 MMMDIL <- LA, APHT만 남겨두고 나머지 규칙은 제거한다. 포함되는 부 규칙들이 가시적으로 보이지 않더라도 의미적으로는 대표규칙으로부터 유추할 수 있다. 이의 알고리즘은 (그림 11)과 같다.

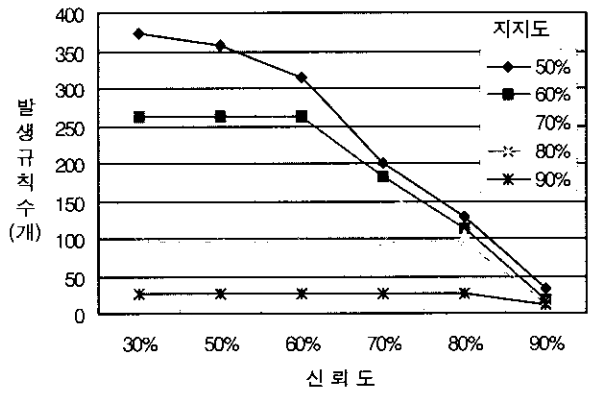
```
Input : SAR = {SARi| SARi : (s1, s2, ·····, sm-1) ⇒ sm ∀i, si∈S}
              /* 규칙 집합 */

RedundencyRulesFiltering() {
            FileteredSAR = Ø ;
        while( SAR != Ø) {
   /* tRule_Set : (전제부, 결론부)가 최장 길이, 개수를 갖는 SAR
                원소,
      sRule_Set : tRule_Set의 조합에 포함되어지는 SAR 원소 */
          tSAR = SAR - tRule_Set ;
          Each sRule_Set ∈ tSAR
          /* Antecedent(Rule_Set) : 규칙으로부터 전제부 추출,
           Consequence(Rule_Set) : 규칙으로부터 결론부 추출 */
            if(Antecedent(tRule_Set)⊂Antecedent(sRule_Set) ∧
             Consequence(tRule_Set)⊂Consequence(sRule_Set) ) {
                SAR = SAR-tRule_Set ;
                      break ; }
            if(Antecedent(tRule_Set)⊃Antecedent(sRule_Set) ∧
              Consequence(tRule_Set)⊃Consequence(sRule_Set)) {
                SAR = SAR - sRule_Set ;
                      continue ; }
                            }
          end Each
          FilteredSAR = FilteredSAR ∪ tRule_Set ; }
}
```

(그림 11) 중복 규칙 제거

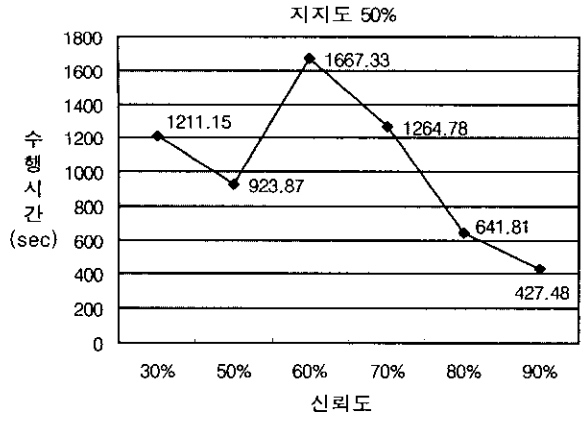
4. 성능 분석

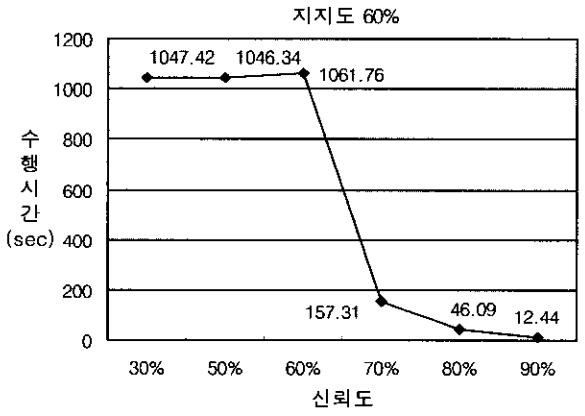
본 절에서는 제안한 알고리즘의 성능 분석 결과를 제시한다. 알고리즘은 C언어로 구현하였으며 실험 환경은 RAM 512M, HDD 40G, Pentium 933 프로세서를 탑재한 Linux server에서 실험하였다. 데이터는 가장 많은 서열 정보를 보유하고 있는 SWISS-PROT 데이터 베이스로부터 추출한 10종의 단백질 서열 데이터를 가지고 지지도와 신뢰도를 변화하면서 실험을 수행하였다. <표 1>은 지지도의 변화에 따르는 발생 규칙 수의 분포이다. 일반적으로 10% 미만을 지지도 임계값으로 부여하는 트랜잭션 연관 규칙의 경우와는 달리 서열 연관 규칙의 경우에는 주어진 단백질 집단에 속하는 상당히 많은 수의 단백질에서 공통성이 나타나야만 생물학적으로 유의할 만한 패턴이라고 간주할 수 있다. 따라서 본 실험에서는 최소 지지도 임계값을 50% 이상으로 설정한다. (그림 12)의 결과를 보면 규칙 신뢰도가 60이상에서의 발생 규칙의 수는 3단계 후보 규칙들의 개수가 매우 큰 차이를 보이지만, 4단계에서 포함관계에 있는 부 규칙들을 제거하고 나면 비슷한 분포를 보인다. 이는 많은 수의 규칙들이 그들간에 포함되는 부 규칙의 형태로 되어 있음을 알 수 있다.

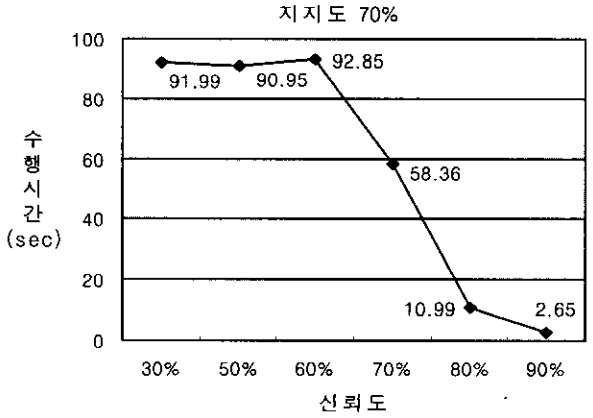
알고리즘의 수행에 있어서 매개변수의 역할은 여러 관점으로 해석할 수 있다. 지지도에 따르는 발생 규칙수 의 변이들은 낮은 지지도일수록 많은 후보 항목의 생성으로 인해 생성된 규칙은 많지만 더 많은 경우의 서열의 연관성을 보인다. 서열문제의 경우 완전한 매칭이 되지 않더라도 같은 실험 결과를 보인다면 동일 종이나 같은 기능을 수행할 것이라 예측한다. 이러한 측면으로 본다면 낮은 지지도가 다수의 규칙 조합을 생성한다 하더라도 생물학 적으로는 중요한 의미를 가질 수 있다. 현재 실제의 전체 서열을 가지고 실험중인 상황에 의하면 평균 길이 500서열, 지지도 30으로 주어졌을때 1단계 수행후 전체 10개의 단백질 중 3개의 단백질에서 68길이의 서열 조합이 추출되었다. 이는 생물학적으로 이 68길이의 서열 조합 자체가 한 단백질의 부 단위이며 소 단백질임을 의미한다. 지지도가 50이되면 이 서열은 발견되지 않는다. 또한 낮은 지지도의 생성 규칙 결과들은 더 많은 생물학적인 실험 후보 조합을 선택할 수 있게 한다. 더 나아가서는 설사 규칙으로 생성되지 않았더라도 생성된 규칙 조합과 유사성을 보인다면 기능적인 연관성을 예측할 수 있을 것이다. 높은 지지도와 신뢰도에 의해 생성된 규칙 조합은 타당성이 있는 의미있는 규칙임을 의미한다. (그림 13)은 지지도별 신뢰도에 따르는 수행시간의 분포이다. 지지도 60%이하에서는 큰 편차를 보이며 70%~90%의 변이는 거의 유사한 분포를 보인다.

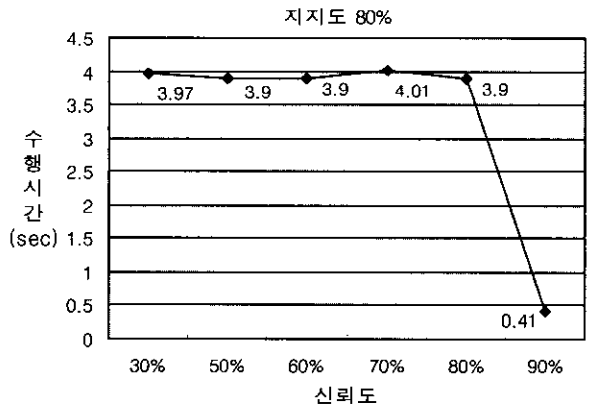


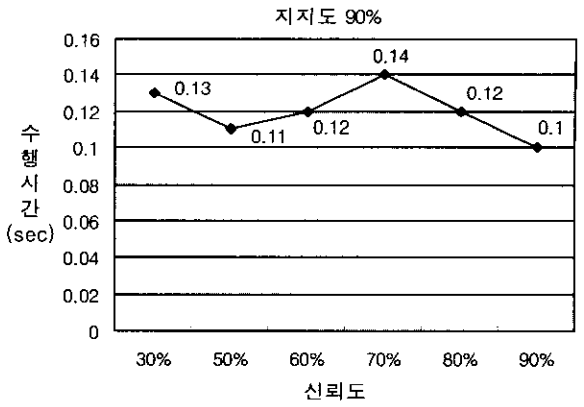
(그림 12) 신뢰도에 따르는 지지도별 발생규칙수











(그림 13) 지지도별 (신뢰도수행시간)

알고리즘의 수행시간은 각 단계를 분석해본 결과 2,3단계에서 규칙 후보 조합을 생성하는 것과 4단계에서 규칙을 요약하는 데 전체 수행시간의 98%정도가 소요되었다.

5. 결 론

본 논문에서는 단백질 서열 분석을 통하여 그들 부서열간에 존재하는 관련성을 탐사하는 서열 연관 규칙 알고리즘(SARA)을 제안하였다. 특정 단백질 서열에 거의 항상 함께 나타나는 부서열의 조합은 단백질 기능 분석의 중요한 단서가 된다. 즉 단백질 부서열 간의 동시출현(co-occurrence)은 특정 단백질의 2, 3차원 구조, 특정 핵산 서열과의 바인딩(binding), 특정 단백질과의 상호 작용 등의 기능을 밝히는 출발점으로 생각할 수 있다. 기존 Apriori 알고리즘은 단일 원자 항목 단위의 트랜잭션 데이터를 대상으로 그들간의 연관성을 탐사하였다. 그러나 제안하는 서열 연관 규칙 알고리즘에서는 단일 항목에 대응하는 한 문자 이상의 조합으로 구성된 서열 데이터이기 때문에 서열의 중복을 인정해야하며, 규칙을 구성하는 부 서열 항목들간의 중첩 문제들이 존재한다. 이는 규칙 생성시 많은 규칙 항목 조합을 생성시킴으로 인하여 수행시 시스템에 과부하와 실제 생물학적으로 의미 있는 정보의 추출에 대한 신뢰성을 보장할 수 없다. 본 논문에서 제안한 서열 연관 규칙 알고리즘은 위에 언급한 문제들을 각 단계별 후보 항목 생성 단계 전에 걸러냄으로써 생물학 연구에 필요한 정확한 규칙 생성을 유도하였다. 마지막 단계에서는 최종 탐사된 규칙들에 대해서 규칙간에 포함되어지는 부 규칙들은 제거함으로써 의미있는 규칙들로 집약하였다.

SARA는 각 탐사 단계별 모듈들이 서열 분석과 관련 분야에 적용 가능한 도구 박스와 같은 역할을 수행 할 것이다. 의미있는 서열을 탐사하는데는 낮은 서열 지지도로 1단계의 수행만으로 발견이 가능하며 의미있는 서열간의 연관성은 나머지 단계 수행결과로써 도출할 수 있다. SARA의 전체적인 알고리즘은 서열간의 연관성을 탐사하지만 1단계

알고리즘 만으로는 종별 빈발 서열 집합을 추출함으로써 의미 있는 서열 집합을 추출하는데 적용할 수 있다. 2, 3단계 프로그램은 서열간의 연관성을 탐사한다. 실제 적인 서열 연관 규칙은 3단계까지의 수행 결과로써 알 수 있지만 4단계에서는 중복된 규칙을 제거하여 요약된 규칙만을 보인다. 최장 길이의 서열 항목으로만 요약된 규칙만을 보임으로써 규칙의 손실 문제를 발생시킬 수 있지만, 가시적으로 보여지지는 않더라도 의미적으로는 최종 선택된 대표규칙에서 유추할 수 있다. 또한 3단계의 결과를 시각화함으로써 해결할 수 있다. 실험 수행결과 3단계의 결과에서 대부분의 규칙간에 포함 관계가 성립하였으므로 규칙의 요약 문제는 타당성이 있다고 본다. SARA는 실험 데이터의 대상을 단백질 서열을 취급하였지만 핵산 서열에도 적용이 가능하다. 생물학적으로 발견된 규칙은 규칙을 구성하는 빈발 서열 자체가 특정 단백질이 될수도 있고, 다 나아가서 발견된 패턴을 통하여 단백질의 구조나 상호 작용 등을 예측함으로써 실험 연구의 방향을 계획할 수 있다. 이는 실제 현장에서 실험 후보 조합의 수를 감소시킴으로써 많은 시간과 비용, 노력을 절감할 수 있다.

참고 문헌

[1] R. Hofestaedt, "Computer science and biology," *BioSystems* 43, pp.69-71, 1997.

[2] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, "GeneCards : a novel functional genomics compendium with automated data mining and query reformulation support," *Bioinformatics*, 14(8), August, pp.656-664, 1998.

[3] Setubal J, Meidanis J, *Introduction to Computational Molecular Biology*, Boston, MA : PWS Publishing Company, July, 1997.

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

[5] Luke Alphey, *DNA SEQUENCING from experimental methods to bioinformatics*, School of Biological Sciences, The University of Manchester, Manchester, UK, BIOS Scientific Publishers, 1997.

[6] Steven L. Salzberg, David B Searls and Simon Kasif, *Computational Methods in Molecular Biology*, Elsevier Science B.V., 1998.

[7] http : //www.ncbi.nlm.nih.gov/BLAST/.

[8] http : //www.ebi.ac.uk/fasta3/.

[9] http : //www.sdsc.edu/MEME/meme.2.2/webs-ite/meme.html.

[10] http : //www.rcsb.org/pdb/.

[11] http : //www.ncbi.nlm.nih.gov/Structure/.

[12] http : //Pfam.wustl.edu.

[13] C. Pabo, E. Peisach, and R. Grant, "Design and Selection of Novel CYS2HIS2 Zinc Finger Proteins," Annu. Rev. Biochem, 70, pp.313-340, 2001.

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

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

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

[17] Mohammed J. Zaki, "Scalable Algorithms for Association Mining," *IEEE Transactions on Knowledge and Engineering*, 12(3), May/June, 2000.

김 정 자

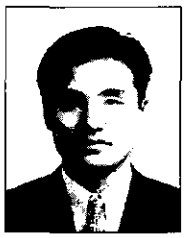
e-mail : jjkim@dbcore.chonnam.ac.kr

1985년 전남대학교 자연과학대학 계산통계학과(이학사)

1988년 전남대학교 자연과학대학 계산통계학과(이학석사)

1997년~1999년 전남대학교 자연과학대학 전산통계학과 박사수료

관심분야 : 데이터 마이닝, 바이오 인포매틱스



이 도 헌

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

1990년 한국과학기술원 전산학과(공학사)

1992년 한국과학기술원 전산학과(공학석사)

1995년 한국과학기술원 전산학과(공학박사)

1999년~2000년 Univ of Texas at Austin 방문교수

1996년~현재 전남대학교 전산학과 조교수

관심분야 : 데이터 마이닝, 바이오 인포매틱스, 워크 플로우 관리

백 윤 주

e-mail : yunju@mail.naver.com

1990년 한국과학기술원 전산학과(공학사)

1992년 한국과학기술원 전산학과(공학석사)

1997년 한국과학기술원 전산학과(공학박사)

200년~현재 네이버컴주식회사 기술이사

관심분야 : 멀티미디어 정보처리, 멀티미디어 시스템, 웹응용시스템, 인터넷 비즈니스, 전자상거래