연관적 분류기법을 이용한 단백질 구조예측

조경환*, 이헌규*, 이범주*, 정광수*, 류근호*
*충북대학교 전자계산학과
e-mail : khcho@dblab.chungbuk.ac.kr

Protein Structure Prediction Using Associative Classification

Kyung Hwan Cho*, Heon Gyu Lee*, Bum Ju Lee*, Kwang Su Jung*, Keun Ho Ryu*
*Dept. of Computer Science, Chung Buk National University

요 약

단백질 구조로부터 단백질 기능을 예측하고자 하는 일은 생명정보학 에서 중요한 이슈 및 연구 과제가 되어 왔다. 그 중 단백질의 3 차 구조를 이해하고 분류하는 데에는 계층적인 분류방법을 이용하는 CATH database 가 사용되고 있다. 이 논문에서는 CATH database 의 계층적 분류의 특성을 이용하되, 단백질의 3 차 구조가 아닌 단백질 서열로부터 데이터마이닝 기술을 적용, 마이닝 기법 중 순차패턴과 연관적 분류 기법을 이용하여 CATH database 의 계층별 구조 분류 기법을 제안 하였다.

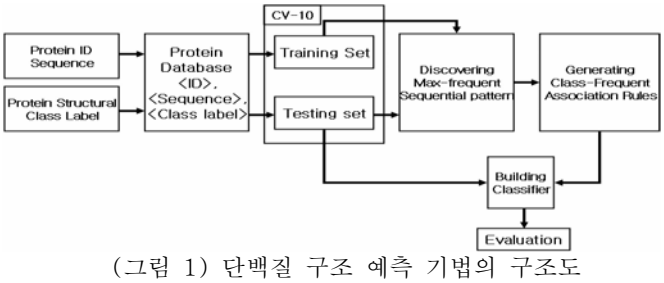
1. 서론[1]

현재까지 단백질 구조로부터 단백질 기능을 예측하고자 하는 연구는 생명정보학 에 있어서 중요한 이슈가 되어왔다. 단백질 구조는 크게 단백질 서열, 3 차원 단백질 구조 그리고 단백질 표면으로 나타낼 수 있다. 이 중 단백질의 3 차 구조의 분석은 진화적인 관계뿐 아니라 단백질간의 기능을 이해하는 데에도 많은 도움을 준다. 그러나 아직까지는 단백질 3 차 구조의 복잡성으로 인해 아직 알려지지 않은 단백질 구조의 분류 및 기능 예측은 어려운 실정이다. 단백질 3 차 구조를 이해하고 분류하는 데 도움을 주는 CATH database 는 PDB(Protein Data Bank)에 나와있는 서열데이터베이스로써 크게 4 가지의 계층으로 분류된다. 각 계층은 각 도메인의 2 차 구조 구성을 묘사하고 있는 C (Class)-Level 과, 단백질의 구조(Barrels or Sandwiches)등의 2 차 구조, Units 의 모양을 나타내는 A (Architecture)-Level, 2 차 구조의 연결성과 전체적인 형태(모양)등을 나타내는 T (Topology)-Level, 마지막으로 높은 구조의 유사성과 기능의 유사성을 나타내 주는 H (Homologous Super family)-Level 로 분류 되어진다.[1, 2]

이 논문에서는 CATH database 의 특성을 이용하여, 단백질의 3 차 구조가 아닌 단백질 서열로부터 데이터 마이닝 기술을 이용, CATH database 의 계층별 구조 분류 기법을 제안하며, 다음과 같은 연구를 수행한다.

첫째, CATH database 와 PDB 로부터 Protein ID, 아미노산 서열 그리고 해당 서열의 CATH 계층 구조별 클래스 정보를 추출하여 마이닝 작업을 위한 데이터 전처리 작업을 수행한다. 둘째, 전처리된 전체 서열 데이터에서 빈발한 서열 패턴을 발견하기 위한 순차 패턴 마이닝을 수행한다. 셋째, 두 번째 단계에서 발견된 모든 서열 패턴은 각 Protein ID 의 서열의 훈련, 테스트 데이터 셋을 생성하며, 이 데이터 셋으로부터 단백질 구조 예측을 위한 연관적 분류기법을 적용한다. 마지막으로 제안된 분류 기법의 성능 평가를 수행한다.

이 논문에서 제안한 단백질 서열을 이용한 구조 예측 기법의 전체 프레임워크는 (그림 1)과 같다.



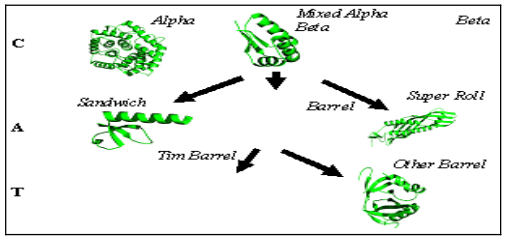
(그림 1) 단백질 구조 예측 기법의 구조도

본 논문은 2006 년도 교육 인적 자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음.

2. 관련연구

2.1 CATH database

단백질의 3 차 구조 정보 데이터베이스인 PDB (Protein Data Bank)는 많은 단백질이 구조적으로 유사성을 갖고 있다는 것을 착안하여 분류체계를 가진 데이터베이스로 CATH 데이터베이스를 연계하였다. CATH 데이터베이스는 계층적인 도메인에 따라 분류하여 크게 네 가지의 Level 인 Class, Architecture, Topology, Homology 를 이용한 정보검색을 지원하며 그 계층구조는 (그림 2.)과 같고, CATH database 의 단백질 구조 계층에서의 각 레벨에 대한 클래스 분포는 <표 1>과 같다.



(그림 2) CATH database 의 단백질 계층 구조

<표 1> CATH database 의 단백질 구조 계층에서의 각 레벨에 대한 클래스 분포

Each level of Protein Structural Class	Number of Class Labels
C-Level	4
A-Level	30
T-Level	379
H-Level	637

2.2 순차패턴 마이닝

순차패턴이란 동시에 발생될 가능성이 큰 항목 집합을 찾아내는 연관성 측정에서 시간이라는 개념이 포함되어 순차적으로 발생될 가능성이 큰 항목 집합을 찾아낸다.[4, 5, 6, 7]. 순차패턴 마이닝은 고객의 구매 패턴에서 자연 재해, 주식 가격의 변화, 질병에 관한 데이터나 DNA 시퀀스에 이르기까지 대부분의 데이터가 시간 순서를 갖는 데이터로서 순차패턴 마이닝의 대상이 되며, 또한, 항목 사이의 연관성을 측정하는 연관규칙(association rule)에 순서를 고려하여 유용한 지식을 찾는 기법이라고 정의할 수 있다. 예를 들어, 연관규칙에서 항목 A, B 에 대하여 A 가 먼저 발생했다고 가정하였을 때, A→B, B→A 라는 규칙이 발견될 수 있지만 순차 패턴에서는 A➔B 라는 규칙만이 발견될 수 있다.

2.3 연관적 분류

연관적 분류란 서로 독립적인 연관규칙과 분류규칙을 일부분 통합시킨 새로운 분류 방식으로 클래스 라벨을 예측하기 위해 연관규칙을 사용한다.[7, 8] 연관적 분류는 기존의 분류기법(결정트리, 베이지안 분류 등)에 비해서 다음과 같은 장점을 가진다. 첫째, 기존의 의사 결정 트리 가 데이터 객체의 분류를 위해 단지 수 백 개의 속성들만을 조작하는 것으로 제한되는 반면, 연관적 분류는 수 천의 속성 차원을 조작 할 수 있다는 것 이다. 둘째, Naïve Bayes 과 유사하게 항목들을 독립적으로 고려할 수 있다. 또한 연관적 분류는 다양한 변수들 사이에서 높은 신뢰도를 갖는 규칙을 탐사하므로 한 번에 하나의 변수만을 조사하는 의사결정트리에 의한 제약사항을 극복하였다. 마지막으로 탐사된 규칙은 단순성(simplicity)을 가지므로 사용자들은 규칙을 쉽게 이해할 수 있다.

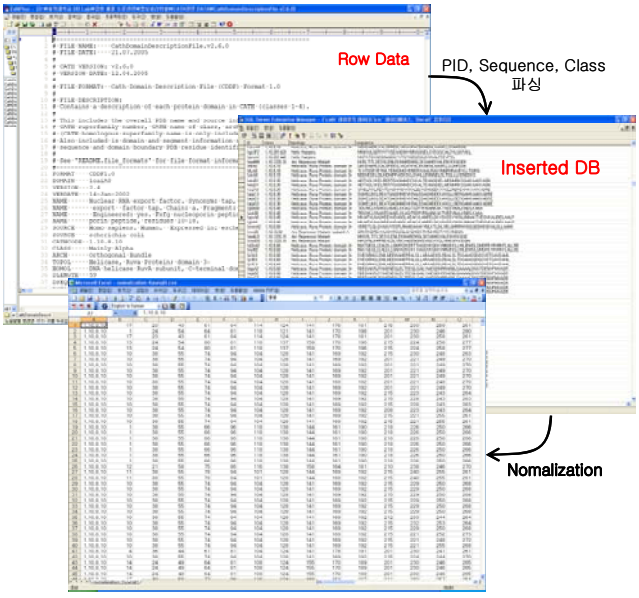
3. CATH 단백질 구조 데이터베이스에서의 순차패턴 발견

이 절에서는 CATH 단백질 구조 데이터베이스에서 추출한 각 계층별 서열 데이터로부터 순차패턴을 탐사하는 과정을 설명한다.

3.1 PID, 서열의 전처리 및 정규화

아래의 (그림 3)은 본 절에서 기술하려 하는 순차패턴의 탐사과정 중 CATH database 로부터 받은 CATH database Flat File 을 보이고 있다. 이렇게 받아진 데이터 파일 중에서 필요로 하는 CATH ID, Class, Topology 그리고 서열만을 뽑아내는 파싱작업을 한다.

정규화 과정은 순차 패턴 알고리즘의 입력을 정수 데이터로 변형을 하며, 이는 단지 아미노산 서열의 인덱스를 의미한다.



(그림 3) CATH data Flat File 의 전처리 과정 및 그 결과

3.2 서열 데이터에서의 순차패턴 마이닝

3.1 절에서의 정규화 된 데이터를 가지고, 순차패턴 마이닝 수행을 위해 기존의 AprioriAll 알고리즘을 적용, 각 빈발한 패턴의 서브 패턴들의 중복을 허용하지 않는 최대 서열 패턴(Maximal Sequence Pattern)을 탐사한다. 최대 서열 패턴 탐사는 H-Level 에서 수행

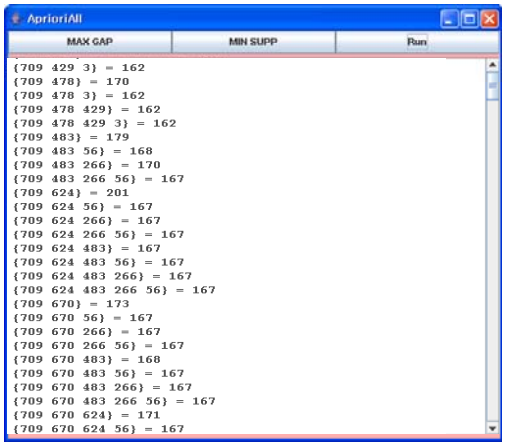
되며, C-Level, A-Level, T-Level 의 서열 패턴은 탐사된 H-Level 의 최대 서열 패턴으로 유도될 수 있다. (그림 4)는 H-Level 의 서열 데이터에서 최대 서열 패턴을 탐사하는 알고리즘이며, (그림 5)는 탐사된 패턴의 예를 보여준다.

```
Input: H-Level 에서의 단백질 전체 서열, D_H
Output: 빈발한 최대 서열 패턴 MSP.
MSP_1 = {1-frequent sequence patterns};
for (k=2 ; MSP_k-1 != ø ; k++) {
        C_k = Candidates generated from MSP_k-1;
        forall (protein sequences s∈D_H) {
                Increment the count of all candidates in C_k
                that are contained in s.
        }
        MSP_k = candidates in C_k >= minimum support.
        A set of MSP = Max-sequences in ∪_k MSP_k.
}
```

(그림 4) 최대 서열 패턴 탐사 알고리즘

Protein Index	Protein-Sequence
1	E,A,G,E,D,C,B
2	D,A,E,C,G,A,B,G
3	E,A,G,D,C,B
4	A,E,D,G,C,D,B,A
5	G,D,E,C,A,G,B

MSP
A,G,B
D,C,B
E,A



(그림 5) 탐사된 패턴의 예

4. 연관적 분류 기법을 이용한 단백질 구조 예측

이 절에서는 3 절에서 탐사된 빈발 서열 패턴들로부터 연관적 분류 규칙을 적용한 단백질 구조예측 기법을 기술 한다.

4.1 연관적 분류 규칙의 생성

연관적 분류 규칙의 생성을 위해 H-Level 에서의 최대 서열 패턴을 특징 벡터로 사용한다. MSP 와 단백질 구조 클래스들 사이의 관계 추출을 위해서는 먼저, 클래스 연관규칙 탐사를 위한 훈련데이터 집합을 생성한다. 훈련데이터 집합에는 Protein ID, MSP, 구조적 클래스로 구성되며, (그림 6)과 같은 형태를 갖는다.

(그림 6)에서의 테이블에서와 같이 각 Protein ID 에 해당되는 서열에서 순차패턴 알고리즘으로부터 생성된 MSP 포함 유무에 따라, 이진 값으로 표현한다. 또한, 모든 최대 서열 패턴들의 식별을 위해 넘버링 하여 나타낸다. 따라서, 연관적 분류 기법의 클래스 연관 규칙 탐사가 가능한 트랜잭션 테이블이 생성된다.

Protein ID	MSP_1	MSP_2	…	MSP_{419}	Structural Class
1	~1~	~0~	…	~1~	1.10.8.10
2	~1~	~0~	…	~1~	1.10.8.30
3	~0~	~1~	…	~0~	3.30.60.10
4	~0~	~0~	…	~1~	4.10.80.10
⋮	⋮	⋮	⋮	⋮	⋮
4,215	~0~	~1~	…	~0~	2.20.25.10

(그림 6) 연관적 알고리즘 (Class Based on Association)

연관적 분류 규칙들의 생성을 위한 클래스 연관규칙의 형태는 MSP_1, MSP_2, … MSP_k ➔ Class [support, confidence] 이며, 지지도와 신뢰도는 식(1), 식(2)와 같다.

$$Support = \frac{RuleSupCnt}{|D|} \qquad 식(1)$$

$$Confidence = \frac{RuleSupCnt}{ItemSupCnt} \qquad 식(2)$$

RuleSupCnt 는 RuleItem 의 개수 값으로 ItemSet 을 포함하면서 클래스 c_i 로 라벨 된 D 안의 항목들의 개수이다. ItemSupCnt 는 ItemSet 을 포함하는 D 안의 항목들의 개수이다.

(그림 7)은 MSP 트랜잭션 테이블에서 클래스 연관규칙 탐사 알고리즘이다.

```
Input: Protein D, minSup, minConf.
Output: <CAR_k>{
    L_1={large 1-RuleItems in D}
    }
For (k=2; L_k-1 ≠ ø; k++){
     // PHASE I: Generate RuleItems.
        C_k = candidateGen(L_k-1);
        For each data case d∈D{
            C_d = subset (C_k, d);
            For each candidate c∈C_d{
                c.ItemSupCnt++;
                if d.class = c.class then
                   c.RuleSupCnt++;
                }
            }
        }
        L_k = {c∈C_k | c.RuleSupCnt≥ minSup};
     // PHASE II: Generate class association
        CAR_k = genRule(L_k);
  Answer = ∪kCARk = From_calUpdatei();
```

(그림 7) 클래스 연관규칙 탐사 알고리즘

단계 1: 최소지지도와 최소신뢰도를 만족하는 모든 large RuleItem 집합을 생성한다. 빈발한 k-RuleItem 은 k 개의 ItemSet 을 가지는 RuleItem 을 나타내고 L_k 는 k-RuleItem 들의 집합이다. 집합 L_k 의 각 요소는 <(ItemSet, ItemSupCnt), (ClassLabel, RuleSupCnt)>이다. 따라서, 첫 번째 단계에서의 수행은 세 가지의 주요 연산을 거친다.

(1) (k-1) 패스에서 탐사된 L_{k-1} 은 candidateGen() 에 의해 C_k를 생성하기 위해 사용되어진다.
(2) 생성된 C_k 는 데이터베이스 스캔을 거쳐 후보 k-RuleItem 들의 지지도를 갱신하게 된다.
(3) 지지도 카운트가 진행 된 후, 임계값 minimumSupport 를 만족하는 RuleItem 들만을 선택하게 된다.

단계 2: 생성된 k-Ruleitemem 들의 집합 L_k 로부터 genRules()에 의해 k 개의 ItemSet 과 클래스 라벨을 가진 CAR_k 를 생성한다. 생성된 L_k 들은 신뢰도가 계산되어지고, 임계값, minimumConfidence 를 만족하는 규칙들만을 생성한다. 그러나 만약, ItemSet 이 같고 클래스 라벨이 다른 RuleItem 들이 존재할 경우, 높은 신뢰도를 가지는 RuleItem 을 선택한다. 예를 들어 ItemSet 이 같은 다음의 두 RuleItem 이 존재한다고 할 때,

1. $\{MSP_1, MSP_4\}$ = Class: 1.10.8.10,
2. $\{MSP_1, MSP_4\}$ = Class: 1.10.8.30

minimumConfidence = 0.3 이고, ItemSet 의 지지도가 3 이라 하자. 첫 번째 RuleItem 의 지지도는 2 이고, 두 번째 RuleItem 의 지지도는 1 이라 가정하자. 그러면, 첫 번째 RuleItem 의 신뢰도는 67%이 되고, 두 번째 RuleItem 의 신뢰도는 33%가 된다. 따라서 생성되는 RuleItem 은 $\{MSP_1, MSP_4\}$ = Class: 1.10.8.10 이 된다.

5. 실험 평가

이 논문에서 제안한 단백질 서열 데이터로부터의 구조 분류 기법에 대한 성능평가를 위해서 10-fold cross-validation(CV-10)를 적용하였으며, CATH database 의 각 계층 별 분류 결과를 평균자승오차제곱근(RMSE)과 F-Measure 를 이용하여 분석하였으며, <표 2> 에 요약하였다.

<표 2> CATH database 의 각 계층별 분류 결과

Level	Precision	Recall	F-Measure	RMSE
C-Level	0.889	0.941	0.914	0.2788
A-Level	0.921	0.939	0.93	0.2532
T-Level	0.88	0.889	0.884	0.334
H-Level	0.814	0.576	0.675	0.4825

<표 2>에서 precision 과 recall 에 동등한 중요도를 부여하여 하나의 평가 방법으로 사용하는 F-Measure 를 식 (3)으로 구하였다.

$$F = \frac{2 \times precision \times recall}{precision + recall} \qquad \text{식(3)}$$

F-Measure 는 분류 모델의 정확도를 평가하는 방법으로 precision 과 recall 의 조화평균으로 계산되어진다.

6. 결론 및 향후 연구

본 연구에서는 CATH database 특징인 계층적 구조를 이용하여 flat file 의 전처리 및 정규화를 수행하고, 데이터들을 이용한 분류방법으로, 마이닝 알고리즘의 하나인 순차패턴과 연관규칙 분류 기법, 클래스 연관규칙 탐사 알고리즘 등을 적용하여 각 클래스 별 분류를 실행해 본 결과, F-Measure 는 C-Level, A-Level, T-Level, H-Level 에서 각각 0.914, 0.93, 0.884, 0.675 의 결과로 H-Level 을 제외한 나머지 Level 에서 높은 정확도를 보였으며, 평균자승오차제곱근(RMSE)에서는 하위 Level 인 H-Level 을 제외한 나머지 Level 에서 낮은 오차를 가지므로 다소 높은 정확도를 가지는 것을 알 수 있다.

참고문헌

[1] CA Orengo, AD Michie, S Jones, DT Jones, MB Swindells and JM Thornton, “CATH – a hierarchic classification of protein domain structures”

[2] Frances Pearl, Annabel Todd, “The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis”

[3] N. Rattanakronkul, T. Wattarujeekrit, and K. Waiyamai, “Predicting Protein Structural Class from Closed Protein Sequences”

[4] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”

[5] Rakesh Agrawal, Ramakrishnan Srikant, “Mining Sequential Patterns”

[6] Qiankun Zhao, Sourav S. Bhowmick, “Sequential Pattern Mining: A Survey”

[7] R. Srikant, R. Agrawal, “Mining Sequential Patterns: Generalizations and Performance Improvements”

[8] Bing Liu, Wynme Hsu, Yiming Ma, “Integrating Classification and Association Rule Mining”

[9] R. Agrawal, T. Inielinski, A. Swami, “Mining association rules in large databases”