# 귀납적 논리 프로그래밍을 이용한 단백질 구조 분류 기법

안건태°, 김진홍, 윤형석, 박양수, 이명준
울산대학교 컴퓨터정보통신공학부
{java2u°,avenue,miracle,yspark,mjlee }@mail.ulsan.ac.kr

# A Method For Protein Structure Classification Using Inductive Logic Programming

Geon-Tae Ahn°,Jin-Hong Kim,Hyeong-Seok Yoon,Yang-Su Park,Myung-Joon Lee
School of Computer Engineering & Information Technology, University of Ulsan

요 약

정보의 급속한 확산과 더불어 체계적이고 효율적으로 정보를 분류하고 활용할 수 있는 방법에 대한 연구의 필요성이 증대되고 있다. 생물정보에 있어서도 기존에 축적된 많은 정보 뿐 만 아니라 새로 밝혀지는 정보들을 자동적으로 분류하고 재활용하는 방법의 일환으로 귀납적 논리 프로그래밍을 적용한 방법론이 채택되고 있다. 본 논문에서는 귀납적 논리 프로그래밍을 이용하여 단백질 구조 분류 데이터베이스를 생성하고 이를 기반으로 단백질 폴더에 내재된 공통의 규칙들을 발견하고, 새로운 단백질에 적용하여 구조를 예측할 수 있는 방법론에 대하여 기술한다.

## 1. 서 론

최근 단백질 서열정보를 비롯한 생물정보의 양이 기하급수적으로 증가함에 따라 서열의 구조와 기능을 실험을 통하여 규명하는 것이 쉽지 않게 되었다. 단백질 구조 분류 및 인식 분야는 크게 두 부류로 나누어 질 수 있는데, 단백질의 서열 *모티프(motif)* 분석을 통한 방법연구와 단백질 3차 구조의 유사성에 대한 연구가 있다. 전자는 단백질의 구조는 아미노산 서열에 의해서 결정된다는 사실을 기반으로 서열정보를 분석하여 단백질 이차 구조 혹은 삼차구조를 예측하는 분야이며, 후자는 삼차원 구조가 가지는 진화적인 유사성에 따라 단백질을 여러 단계로 분류화시킴으로써 구조가 가지는 공통성을 찾는 연구 분야이다. 단백질의 3차 구조 분류와 관련한 데이터베이스로는 *SCOP*[1], *CATH*[2], 그리고 *DALI*[3] 등이 있으며 모두 *PDB*[4]에 등록된 자료를 기반으로 특정 분류 기준 아래 생성된 것이다.

본 논문에서는 단백질 구조 분류를 효율적으로 지원할 수 있는 방법에 대한 연구로, *귀납적 논리 프로그래밍기법 (ILP:Inductive Logic Programming)*을 이용하여 단백질 3차 구조의 위상정보에 내재된 규칙을 발견함으로써 분류 작업을 자동화하고 새로운 단백질에 대한 구조의 예측을 가능하게 하는 방법론에 대하여 기술한다. ILP 기법은 *논리 프로그래밍 기법*과 *기계학습이론*을 결합한 형태로 예제 학습을 통하여 예제들이 내포하고 있는 일반적인 규칙을 찾아내는데 적합한 방법이다[5][6].

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 ILP 시스템의 개요와 적용방법에 대하여 설명하고, 3장에서는 단백질 구조 분류 데이터베이스와 단백질 구조를 구성하는 속성들에 대하여 기술한다. 4장에서는 ILP를 이용한 단백질 구조 분류 기법을 통한 실험 및 검증 결과에 대하여 설명하며, 마지막으로 5장에서는 결론과 향후과제에 대하여 기술한다.

† 본 연구는 한국과학재단 목적기초연구(R01-2001-00535) 지원으로 수행되었음.

## 2. ILP와 단백질 구조연구

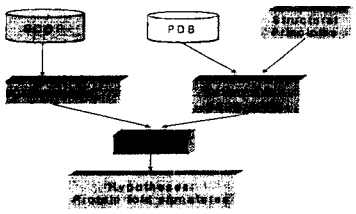
대량의 생물정보 데이터 뿐 만 아니라 오늘날 과학분야에서는, 생성된 정보들을 보다 체계적 원칙아래 분류하고 재활용하는 방법에 대한 연구가 많이 이루어지고 있다. 특히 단백질 구조와 관련된 정보들은 복잡하고 다루기가 쉽지 않아서 지금도 대부분이 관련분야 전문가들에 의해 수작업으로 이루어지고 있는 실정이다. 하지만, 구조가 밝혀진 단백질의 수가 점점 늘어나고 있어서 이런 분류작업에도 자동화된 방법론이 필요하게 되었다.

### 2.1 ILP 시스템

귀납적 논리 프로그래밍은 여러 단백질 구조 사이에서 공통의 부분 구조나, 서열과 3차 구조 사이의 관계 같이 단백질 폴드 형성과정에 숨겨진 새로운 원리를 추론할 수 있는 방법을 제공한다. ILP는 SCOP 데이터베이스에 정의된 폴드와 그 지역구조사이의 관계를 학습한다. ILP 알고리즘은 예제들과 배경지식을 입력으로 자동적으로 데이터셋에 내재된 원칙들을 학습하게 된다. 이처럼 ILP는 단백질 구조 예측, 신약개발, 그리고 화학적인 변이 실험을 포함한 몇몇 구조 분자 생물학에서 적용되고 있다[7]. 본 논문에서는 단백질의 구조 분류 중 폴더에 내재되어 있는 규칙들을 발견하고 그 규칙들이 가지는 타당성을 검증하기 위하여 ILP 시스템을 사용하였다.

(그림 1)은 ILP를 이용한 단백질 구조 분류 및 폴더에 내재된 규칙을 발견하기 위하여 이용되는 정보의 흐름도이다. 단백질 구조 분류 데이터베이스인 SCOP와 PDB를 바탕으로 예제셋과 배경지식을 추출했으며 ILP 시스템을

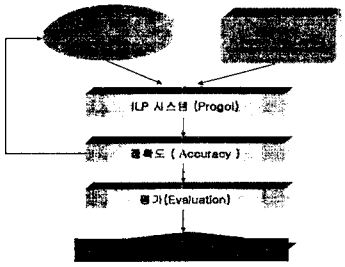
동작시켜 하나의 폴더에 내재된 공통의 규칙을 찾게 된다.




(그림 1) ILP 학습을 위한 정보의 흐름도

### 2.2 ILP를 통한 학습

(그림 2)는 단백질 폴더 규칙을 발견하기 위하여 특정 예제샘플을 학습시키는 과정을 나타낸다. 본 논문에서는 SCOP 데이터베이스에 정의된 단백질 폴더 중 *글로빈폴드(Globlin-like fold)*에 대한 규칙을 발견한다. 예제샘플들은 도메인 정보들의 집합이며 긍정예제와 부정예제들의 조합으로 구성된다. 배경지식은 각 도메인의 구조적인 특징을 설명하기 위한 세부 속성들의 집합이다. 예제샘플과 배경지식을 이용하여 ILP 시스템을 동작시키게 되면 최종적으로 해당 폴더에 대한 일반화된 규칙이 생성된다.



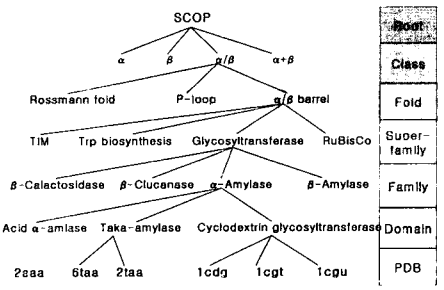

(그림 2) 단백질 폴더 규칙 학습

## 3. 단백질 구조 데이터

ILP 시스템을 이용하여 단백질 구조를 분석하기 위해서는 단백질의 구조를 논리 프로그램으로 기술한 데이터베이스가 필요하다. 본 논문에서는 단백질 구조 데이터베이스인 PDB로부터 구조정보를 추출하여 자동 번역해주는 *PROMOTIF*[8] 프로그램과 단백질 구조를 계층적으로 분류시켜 놓은 SCOP 데이터베이스를 기반으로 프롤로그 데이터베이스를 생성하였다. Prolog는 오래 전부터 단백질 구조를 기술하고 표현하기 위한 도구로 사용되어 왔으며, ***α-나선구조***나 ***β-판상구조***와 같은 이차구조에서부터 단백질 구조를 이루는 구성요소 전반에 대한 효율적인 접근방법을 제공해 준다.

### 3.1 단백질 구조 데이터베이스

단백질 구조에 대한 분류는 복잡하고 많은 시간을 필요로 하는 작업이다. 체계적인 계획을 기반으로 하여 분류의 기준을 세우고 구조가 가지는 특징들을 분석하는 과정이 필요하다. 현재 이러한 단백질 구조를 체계적으로 계층화하고 연관성을 규명해놓은 데이터베이스로는 SCOP가 대표적이다.




(그림 3) SCOP 데이터베이스의 분류

(그림 3)에서 보는 바와 같이 SCOP는 PDB에 저장되어 있는 3차원 구조를 여러 단계별로 계층화하여 체계적으로 분석하여 작성한 데이터베이스이다. SCOP는 PDB정보를 구조의 특성에 따라 그림에서 표시한 4개의 클래스(*class*)를 포함하여 11개의 클래스로 나뉘어지며, 각각의 클래스는 2차 구조의 구성과 위상정보를 나타내는 폴드(*fold*)로 그룹화되고, 폴드는 다시 *슈퍼패밀리(superfamily)*로, 슈퍼패밀리는 *패밀리(family)*로, 그리고 패밀리는 SCOP 데이터베이스의 최종 단위인 *도메인(domain)*으로 분류된다. 이처럼 SCOP 데이터베이스는 단백질 구조의 계통 분류학적 분석을 체계적으로 할 수 있도록 지원한다[1].

### 3.2 단백질 구조 모티프

*단백질 구조 모티프*란 몇 개의 이차 구조가 특정한 모양으로 배열되어 이룬 구조로 여러 단백질에서 공통으로 발견되며 기능 혹은 구조적 역할을 수행하는 구조적인 패턴을 의미한다. 기본 이차구조(***β***-판상구조, ***α***-나선구조)를 포함하여 ***β***-hairpins, ***β***-병풍구조, ***β***-와 ***γ***-turns, 그리고 E-F hand 등이 있다. 단백질 구조 모티프 프로그램인 PROMOTIF 시스템을 이용하면 PDB 포맷 파일에서부터 단백질의 구조 모티프 정보를 추출할 수 있다. PROMOTIF 결과물을 기반으로 단백질 구조 속성 정보들을 논리 프로그래밍으로 재기술함으로써 논리 데이터베이스를 생성한다[8].

## 4. 실험 및 검증

PDB, SCOP, 그리고 PROMOTIF 프로그램에 의해서 생성된 프롤로그 데이터베이스를 이용하여 ILP 시스템을 동작시킨다. ILP 시스템은 *Progol5.0*[9]을 사용하였으며 실험결과 생성된 규칙들에 대한 검증은 데이터의 량을 고려하여 *교차타당성(cross-validation)*검증방법의 하나인 *leave-one-out*을 이용하였다. leave-one-out은 데이터의 양이 비교적 작은 경우 많이 사용하는 기법이다.

### 4.1 실험데이터

실험에 사용된 단백질 폴드는 SCOP 데이터베이스 분류중 $\alpha/\beta$ 클래스에 속하는 4개의 폴드를 선정하여 테스트하였다. 예제들의 배경지식은 단백질 구조 속성 중 구조의 지역속성 값과 2차 구조에 대한 상호관련성, 그리고 단백질 구조 모티프 정보($\beta$-turn, $\gamma$-turn, helix-turn-helix)를 기반으로 구축하였다.

### 4.2 검증방법

실험결과 생성된 단백질 폴드 규칙이 가지는 정확도를 검증할 수 있다. 결과 규칙의 정확도는 예제 데이터의 양에 따라 다양하게 나타날 수 있다. 따라서, 예제 데이터셋이 소량이므로 leave-one-out을 이용하여 예측정확도를 검증하였다. leave-one-out 방법은 예제데이터가 N개일 경우 N-1개의 예제로 규칙을 추론하고 나머지 한 개를 이용하여 생성규칙의 타당성을 검증하는 방식을 N번 반복하는 방법이다.

### 4.3 실험결과

<표 1>은 ILP 시스템을 통하여 예제셋을 학습시킨 결과 생성된 결과를 나타낸 표이다.

<표 1> 폴더 규칙에 대한 예측정확도 검증 결과

| Fold | Examples | Families | Super-families | Rules | %Accuracy |
|---|---|---|---|---|---|
| $\alpha/\beta$ | | | | | |
| Rossmann-fold | 17 | 5 | 1 | 2 | 82.7%+/-5.3% |
| P-loop | 11 | 4 | 1 | 1 | 73.5%+/-7.6% |
| Periplasmic II | 11 | 2 | 1 | 3 | 76.5%+/-7.3% |
| $\alpha/\beta$-Hydrolases | 12 | 10 | 1 | 3 | 75.7%+/-8.0% |
| Other folds(70) | | | | | |

폴더는 SCOP의 $\alpha/\beta$클래스에 속하는 4개의 폴드(*Rossmann, P-loop, Periplasmic II, $\alpha/\beta$-Hydrolases*)를 나타내고, familiy, superfamiliy는 각각 선택된 예제셋이 속한 familiy의 수와 superfamily의 수를 나타낸다. Rules는 각 폴드가 학습과정을 거친 결과 생성한 폴드 규칙의 개수를 의미한다. 끝으로 Accuracy는 규칙이 가지는 예측정확도이다.

<표 2> 생성 폴더 규칙에 대한 프롤로그 표현 및 의미

| | |
|---|---|
| 규칙 1.<br>**(Rossmann)** | fold('Rossmann-fold', X) :-<br>adjacent(X,A,B,1,e,h), adjacent(X,C,D,6,e,h),<br>length_loop(A,B,0). |
| 의미 1. | 첫번째 위치에 있는 판상구조A는 나선구조B와 인접해있고, 여섯번 위치에 있는 판상구조C는 D나선구조와 인접해 있으며, 2차구조 A와 B사이의 루프의 길이는 0이다. |
| 규칙 2.<br>**(P-loop)** | fold('P-loop', X) :-<br>adjacent(X,A,B,1,e,h), length_loop(A,B,5). |
| 의미 2. | 첫번째 위치에 있는 판상구조A는 나선구조B와 인접해 있으며, 이들 사이의 루프길이는 5이다. |
| 규칙 3.<br>**(Periplasmic II)** | fold('Periplasmic II', X) :-<br>adjacent(X,A,B,6,e,h), adjacent(X,C,D,11,e,h). |
| 의미 3. | 여섯번째 위치에 있는 판상구조A는 나선구조B와 인접해있고, 여섯번 위치에 있는 판상구조C는 D나선구조와 인접해 있다. |
| 규칙 4.<br>**(a/b-Hydrolases)** | fold('Rossmann-fold', X) :-<br>adjacent(X,A,B,9,e,h), length_loop(A,B,5). |
| 의미 4. | 아홉번째 위치에 있는 판상구조A는 나선구조B와 인접해 있으며, 이들 사이의 루프길이는 5이다. |

<표 2>는 실험에 사용된 폴드가 생성한 내재 규칙들과 그들의 의미를 설명하고 있다. <표 2>에서 보는 바와 같이 많은 구조 속성(서열의 길이, 이차구조의 구성, 소수성, 인접관계, 루프의 길이, 모티프 정보) 중에 이차구조의 인접관계와 루프의 길이 정보가 추론결과 우세하게 나타나는 것을 볼 수 있다. 따라서, 이차구조의 인접관계와 루프의 길이는 단백질 3차 구조를 특징짓는 주요한 구성 요소라고 할 수 있다.

## 5. 결 론

본 논문에서는 귀납적 논리 프로그래밍 기법을 적용하여 단백질 정보를 체계적으로 분류하고 새로운 단백질 데이터의 구조 및 기능을 예측할 수 있는 기법에 대하여 기술하였다. 단백질 구조 분류 데이터베이스인 SCOP 정보를 기반으로 ILP 학습을 통하여 같은 분류의 폴더가 가지는 일반적인 구조적인 특징을 유도해낼 수 있었고 새로운 데이터에 적용하여 그 단백질의 구조적인 특징을 예측할 수 있게 되었다. 실험결과 단백질 구조 분류를 위한 폴더 규칙을 생성하는데 단백질 이차구조 구성요소 사이의 연관관계가 가장 높은 영향을 주며, 다른 모티프 정보나 아미노산의 구성 같은 요소는 구조적인 특징을 구별하는데 비교적 낮은 영향력을 가진다고 할 수 있다.

향후 과제로는 특정 단백질 3차 구조 모티프 정보들을 분석하여 특정 구조가 가지는 규칙들을 발견하고 논리프로그래밍 데이터베이스화함으로써 특정 기능을 가지는 단백질에 대한 검색 및 부분구조 검색에 효과적인 방법을 제공하고자 한다.

## 6. 참고문헌


[1] Loredane Lo Conte, Bart Ailey, Tim J. P. Hubbard, Stevern E. Brenner, "SCOP:a Structural Classification of Proteins database", Nucleic Acids Research, Vol. 28, No. 1, p257-259, 2000

[2] "http://www.biochem.ucl.ac.uk/bsm/cath_new/", CATH

[3] "http://www2.ebi.ac.uk/dali/domain/3.0/", DELI

[4] T. N. Bhat, Philip Bourne, Zukang Feng, Gary Gilliland, "The PDB data uniformity project", Nucleic Acids Research, Vol. 29, No. 1, 2001

[5] Stephen. H. Muggleton, "Inductive Logic Programming", Academic Press, London, 1992

[6] Marcel. Turcotte, Stephen .H. Muggleton, Michael. J. E. Sternberg, "Automated Discovery of Structural Signatures of Protein fold and function", Journal of Molecular, 2000

[7] Geoffrey J. Barton, Christopher J. Rawlings, "A PROLOG Approach to Analysing Protein Structure", Tetrahedron, Computer Methodology, 3, No. 6C, p739-756, 1992

[8] Hutchinson E. Gail, Thornton Janet. M., "PROMOTIF - A program to identify and analyze structural motifs in proteins", Protein Science. 5, p212-210, 1996

[9] Stephen. H. Muggleton, John Firth, "CProgol4.4: a tutorial introduction", Univerity of York, United Kingdom