단백질 구조 분류의 통합 검색을 위한 웹 정보시스템

신원준° 황의윤 김진홍 안건태 이명준
울산대학교 컴퓨터정보통신공학부
{mathpf°, heyoon, avenue, java2u, mjlee}@mail.ulsan.ac.kr

A Web-Based Information System for the Integrated Search for Protein Structure Classifications

Wonjoon Shin°, Euiyoon Hwang, Jinhong Kim, Geontae Ahn, Myungjoon Lee
School of Computer Engineering & Information Technology, University of Ulsan

요 약

단백질은 대부분 공간상의 특징을 고려할 때 유사한 부분을 기준으로 분류되는 경우가 많다. 단백질 구조 분류 데이터베이스는 단백질이 가지는 다양한 구조 정보를 바탕으로 단백질 구조 분류 정보를 제공하고 있다. 대표적인 단백질 구조 분류 데이터베이스에는 CATH와 SCOP 데이터베이스가 있다. 이들 데이터베이스는 서로 다른 구조 분류 기준으로 단백질 구조를 분류하고 있으며, 단백질 구조 분류 정보를 검색하는 웹 서비스를 개별적으로 제공하고 있다. 따라서 여러 종류의 단백질 구조 분류 정보를 하나의 웹 사이트에서 검색할 수 있으면 유용할 것이다.

본 논문에서는 CATH와 SCOP에서 정의한 단백질 구조 분류 정보의 통합적인 검색 기능 및 통계 정보를 체계적으로 제공하는 웹 정보시스템에 관하여 기술한다. 제안된 시스템은 CATH와 SCOP에서 제공하는 각각의 데이터를 가공하여 효과적인 구조 분류 검색을 지원하는 구조화된 데이터베이스를 구축하였다. 개발된 시스템은 PDB 식별자, CATH 식별자, 그리고 SCOP 식별자 또는 단백질 분류 이름으로 한번의 검색으로 두 데이터베이스에서 제공하는 계층적 구조 분류 정보를 제공한다. 또한, 단백질 구조에 대한 유용한 통계 정보를 제공한다.

1. 서 론

포스트지놈 시대에 있어서 가장 주된 연구는 단백질의 구조적 유사성이나 분류학적인 연관성을 밝히는 것이다. 유전체 정보의 최종산물인 단백질의 구조와 기능을 규명하는 연구에 관심이 집중되면서 단백질 관련 정보도 급속히 증가하게 되었다. 단백질은 대부분 분류학적으로 공통의 기원을 가지거나 구조적으로 유사한 경우가 많다. 단백질 사이의 구조와 분류학적인 관계 정보는 인간 유전체사업의 결과 생성된 대량의 서열정보를 번역하고 단백질의 기능을 밝혀내는 데 중요한 역할을 담당할 것이다. 이러한 구조적인 정보를 이해하고 접근할 수 있도록 하기 위하여 단백질 구조 분류 데이터베이스가 구축되었다.

CATH(Class, Architecture, Topology, Homologous Superfamily)[1]과 SCOP(A Structural Classification of Proteins)[2]데이터베이스는 3차원 구조가 알려진 단백질에 대한 구조적, 분류학적 관계에 대한 정보를 제공하고 있다. 각각의 단백질 구조 분류 데이터베이스는 그들이 제공하는 시스템에서 웹 기반으로 정보를 제공하고 있다. 이들 데이터베이스는 서로 다른 구조 분류 기준으로 단백질 구조를 분류하고 있으며, 단백질 구조 분류 정보를 검색하는 웹 서비스를 개별적으로 제공하고 있다. 따라서 여러 종류의 단백질 구조 분류 정보를 하나의 웹 사이트에서 검색할 수 있으면 유용할 것이다.

본 논문에서는 한 번의 검색으로 CATH와 SCOP에서 정의한 단백질 구조 분류 정보를 함께 보여 줄 수 있는 웹 정보시스템에 관하여 기술한다. 본 시스템은 CATH와 SCOP에서 제공하는 데이터를 가공하여 새로운 데이터베이스를 만들었다. PDB에서 정의한 PDBID, CATH와 SCOP에서 정의한 DomainID, 그리고 단백질 분류 이름으로 검색이 가능하며, 각각의 계층적 구조 분류정보를 제공하고 각 분류는 분류속의 하위분류가 몇 개인지를 알 수 있다. 또한 단백질 구조에 대한 유용한 통계 등 다양한 정보를 얻을 수 있도록 지원하다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 구조 분류 데이터베이스인 CATH와 SCOP을 소개 한다. 3장에서는 본 시스템의 구성을 살펴보고, 4장에서는 사용자 인터페이스에 대하여 설명한다. 마지막으로 5장에서는 결론으로 끝을 맺고자 한다.

2. 단백질 구조 분류 데이터베이스

단백질 구조는 1차(primary), 2차(secondary), 3차(tertiary), 그리고 4차(quaternary) 구조로 표현된다. 3차 구조는 일반적으로 한 폴리펩티드에 있는 모든 원자들의 공간적 배열을 의미한다[4].

3차 구조가 알려진 단백질에 대한 구조적, 분류학적 관계에 대한 대표적인 데이터베이스로 CATH와 SCOP이 있다. 이들 데이터베이스는 이미 구조가 밝혀진 단백질들에 대한 자료가 저장되어 있어 미지의 단백질에 대한 구조를 밝히거나 기능을 예측하는 연구에 많이 활용되고 있다.

2.1 CATH Database

CATH는 PDB에 수록된 단백질 삼차원 구조를 분류한 데이터베이스이다. 분류 기준으로 DETECTIVE, PUU, 그리고 DOMAK과 같은 알고리즘을 사용하였다. CATH는 Class, Architecture, Topology, 그리고 Homologous superfamily의 첫 글자에서 유래된 약자이다. Class는 α-구조, β-구조, αβ-구조, 그리고 2차 구조의 조성이 낮은 "Few Secondary Structures"로 정의되었다. Architecture는 2차 구조의 공간상에서의 배열을 연결 상태(topology)를 배제한 개념으로 정의되며, Topology는 전체 구조와 2차 구조간의 연결 양상을 고려하여 분류되어 있다. Homologous superfamily는 서열 정렬과 구조의 비교를 동시에 수행하여 판단한다[2].

2.2 SCOP Database

SCOP은 단백질이 지닌 구조적인 유사성과 분류학적인 관계를 기반으로 단백질들을 체계적으로 분류해 놓은 데이터베이스이다. SCOP의 계층적인 구조 분류는 분류 기준에 따라 11개의 Class로 나뉘어 지며, 각각의 클래스는 2차 구조의 구성과 Topology에 의해 Fold로 나뉘어 진다. Fold는 다시 Superfamily로, Superfamily는 family로, 그리고 family는 Domain으로 나뉘어 진다. 이처럼 PDB의 모든 단백질은 다른 단백질들과 비교되어 구조적 유사성(Structural Similarities)을 가지는 그룹으로 분류된다[3].

3. 단백질 구조 분류의 통합 검색 시스템

단백질 구조 분류의 통합 검색을 위한 웹 정보시스템은 검색어(PDBID, CATH와 SCOP의 DomainID, 단백질 구조 분류 이름)를 통한 CATH와 SCOP에서 정의한 구조 분류 정보를 한눈에 볼 수 있다. 이는 각각이 정의한 계층적 구조에서 한 항목을 선택할 때 하위 계층의 내용을 볼 수 있으며, 각각의 계층적 구조에서 계층별 항목 선택을 통한 두 데이터베이스에 대한 AND또는 OR 연산을 수행 할 수 있도록 지원 한다. 또한 본 시스템은 웹 기반의 사용자 인터페이스를 제공함으로써 누구나 손쉽게 사용할 수 있게 하였다.

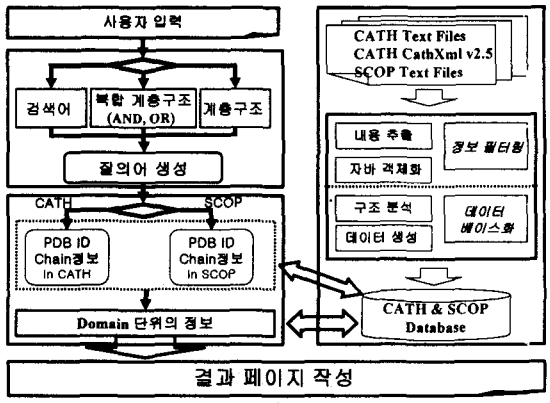
본 시스템은 크게 4개의 모듈로 나뉘어 진다. CATH와 SCOP에서 제공하는 텍스트파일과 XML(eXtensible Markup Language)파일에서 데이터베이스를 구축하는 모듈, CATH와 SCOP의 상호간 가장 유사한 도메인을 연결하는 모듈, 사용자의 질의를 수행하는 모듈, 그리고 질의에 관하여 결과를 보여주는 모듈로 구성되어 있다. (그림 1)은 개발된 웹 정보시스템의 전체적인 구조이다.

데이터베이스 구축 모듈은 텍스트파일과 XML파일에서 필요한 정보를 추출하게 된다. 추출된 정보는 객체화를 거쳐 데이터베이스의 각각의 테이블에 추가 된다.

CATH와 SCOP의 상호간 가장 유사한 도메인을 연결하는 모듈은 PDB에 제공하는 코드와 체인정보를 이용하여 둘 사이의 도메인을 연결해 둠으로써 필요할 때 상대 데이터베이스 구조 분류 정보를 알 수 있게 된다.

사용자의 질의를 수행하는 모듈은 검색 조건별로 어떤 일을 수행 할 것인지를 결정하게 된다. 검색어가 정확한 것인지, 검색어의 내용을 가지고 적적한 검색이 될 수 있도록 질의어를 만들어 내는 과정이다.

사용자 질의에 관한 결과를 보여주는 모듈은 검색의 결과나 사용자 요구 사항의 결과를 정리하여 실제 사용자 화면을 구성 하는 모듈이다.



(그림 1) 단백질 구조 분류의 통합 검색 시스템 구조

4. 사용자 인터페이스

본 시스템의 인터페이스는 크게 검색, 계층적인 표현 보여주기, 복합검색으로 이루어 져 있다. 각 부분의 설명은 다음과 같다.

4.1 검색

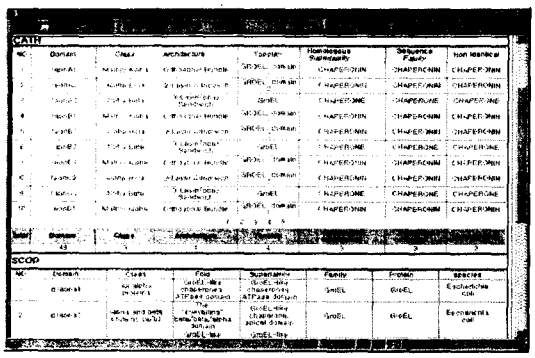
검색에는 PDB에서 정한 PDBID로 검색, CATH와 SCOP의 DomainID로 검색, 그리고 단백질 구조 분류 이름으로의 검색이 있다. 검색의 결과는 CATH와 SCOP에서 정의한 계층적인 표현과 통계 수치를 테이블 형태로 보여 준다.

"*PDBID로 검색*"은 CATH와 SCOP에서 정의하고 있는 도메인 중 PDBID를 가지고 있는 것을 모두 찾아 보여준다. 도메인의 계층정보도 함께 보여지며 PDB코드가 속해있는 도메인의 개수와 기타 정보도 함께 보여 준다.

"*DomainID로 검색*"은 CATH의 도메인을 사용하여 검색할 경우 CATH의 정보 뿐 만 아니라 상호간 가장 유사한 도메인을 찾는 모듈을 이용하여 SCOP의 정보도 함께 찾아 보여 주게 된다. SCOP의 도메인으로 검색 할 경우에도 위와 같은 과정을 거쳐 CATH의 정보를 함께 보여주게 된다. 서로 간의 정보를 한 페이지에서 볼 수 있다.

"*단백질 구조 분류 이름으로 검색*"은 단어 검색으로 되어있다. 한 단에 대해서 후보가 될 수 있는 이름들을 먼저 보여 주게 된다. 후보들은 CATH의 Superfamily 와 SCOP의 Superfamily 레벨의 이름으로 사용된다.

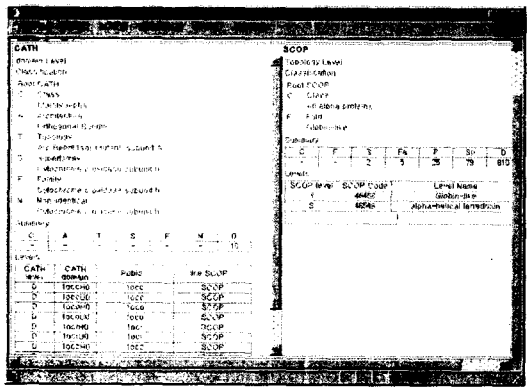
Superfamily를 후보로 선택한 이유는 하위 레벨의 이름들은 중복이 많아서 후보로서는 적합하지 않았다. 그리고 그 후보들 중에 하나를 선택하면 그 이름을 가진 도메인들을 모두 보여 줌으로 그 분류에 속한 단백질들을 모두 찾을 수 있게 된다. (그림2)는 PDB 코드(예를 들어 '1aon')로 검색 했을 때의 결과 화면 이다. 결과 화면에서 특정계층의 한 항목을 선택하면 계층 분류 정보로 보여주게 된다.



(그림 2) PDB 코드 '1aon'의 검색 결과 화면

4.2 계층적인 표현 보여주기

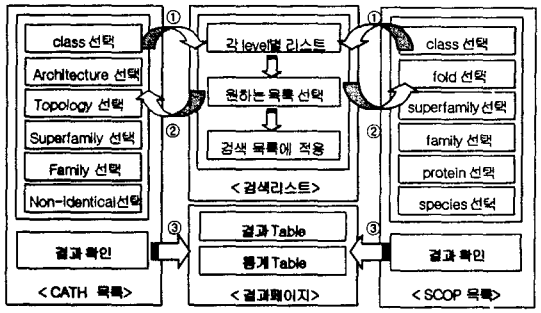
CATH와 SCOP의 단백질 구조 분류 정보를 계층적 표현으로 보여주는 기능이다. 각각의 계층에서 하나의 항목을 선택하면 하위 계층의 목록들을 보여 준다. 하위 계층의 목록과 더불어 현재의 계층이 정보와 계층별 개수와 이름을 볼 수 있다. 각 항목의 상위 계층의 항목이 무엇인지 정의 되어 있는 데이터베이스의 테이블을 참조하여 확인 할 수 있으며, 항목의 개수를 계산 할 수 있다. 마지막 도메인 단계에서는 상대되는 데이터베이스에서 가장 유사한 도메인으로 갈수 있는 항목이 있다. 이 항목은 상호간 가장 유사한 도메인을 연결하는 모듈을 사용한다. (그림 3)은 CATH의 마지막 도메인까지 선택된 모습이다.



(그림 3) CATH의 도메인까지 선택된 모습

4.3 복합 검색

복합 검색은 사용자가 작성한 조건을 바탕으로 CATH와 SCOP에서 검색하여 나오는 결과물을 합하거나 공통부분을 찾아 주는 것이다. 복합검색을 하기 위해서 별도의 검색을 사용한다. (그림 4)는 검색프레임으로부터 검색 결과가 나오는 결과테이블까지의 과정을 설명한 그림이다. 사용자는 각 level에서 검색을 원하는 데이터를 선택하는 과정(①, ②)이 있고, 선택이 끝난 후 검색 결과를 보는 과정(③)이 있다.



(그림 4) 복합 검색 기능의 수행 절차

4. 결론

본 논문에서는 CATH와 SCOP의 단백질 구조 분류 데이터베이스를 통합하여 검색 할 수 있는 웹 정보시스템에 관하여 기술하였다. 본 시스템은 웹 기반 사용자 인터페이스를 제공하여 사용자가 쉽고 효과적으로 단백질 구조 분류정보를 얻을 수 있다. 본 시스템을 서열정보의 번역으로 생성되는 단백질에 대한 연구 및 새로운 단백질 구조가 밝혀졌을 경우 단백질간의 연관관계를 규명하기 위한 도구로 널리 활용될 수 있으리라 생각 되며, 단백질 구조 및 도메인 연구에 본 시스템을 활용할 경우 단백질 구조나 도메인에 대한 용이한 검색 및 다양한 통계자료를 산출할 수 있어서 보다 효율적인 연구를 진행할 수 있을 것으로 기대된다.

[참고문헌]

[1] Orengo, C. A., Pearl, F. M. G., J. E. and Todd, A.E., "The CATH Database provides insights into protein *structure/function relationships.*" Nucleic acids research, Vol.27, No.1, pp.275-279, 1999

[2] Alexey, M., Steven, B., Time, H. and Cyrusm Ch., "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," J. Mol. Biol., 246, pp.536-540, 1995

[3] John, W., Zukang, F. and Helen. M., "The Protein Data Bank : unifying the archive", Nucleic Acids Research, Vol 30, No.1, pp245-238, 2002

[4] Branden, C and Tooze,J. 1991 *Introduction to Protein Structure.* Garland, New York

[5] 안건태 윤형석 황의윤 김진홍 이명준 "SCOML과 SCOPBRowser에 대한 연구" 정보처리학회논문지 D2003