단백질 서열 정렬을 통한 구조 분류정보 추출

변상희° 김진홍 안건태 이명준
울산대학교
{heeya°, avenue, java2u, mjlee}@mail.ulsan.ac.kr

Extracting Information on Structural Classification through Protein Sequence Alignment

Sang-Hee Byun°, Jin-Hong Kim, Geon-Tae Ahn, Myung-Joon Lee
School of Computer Engineering&Information Technology, University of Ulsan

요 약

인간 지놈 프로젝트가 완료된 이후로 여러 지놈 프로젝트가 수행되었으며 이로 인해 데이터베이스에 수록되는 서열수가 기하급수적으로 증가하고 있다. 최근에는 단순한 서열 분석뿐만 아니라 이미 밝혀진 단백질 정보를 이용하여 새로운 단백질의 기능을 예측하는 연구가 보다 활발히 진행되고 있다. 단백질 기능은 단백질의 삼차구조에 의해 결정된다. 따라서 단백질의 서열을 분석하여 삼차구조를 알아내고 어떤 분류에 속하는지 알아낸다면 단백질의 기능을 예측할 수 있다.

본 논문에서는 단백질 서열 정렬을 통하여 보다 빠르고 효과적으로 단백질 구조 정보를 추출하는 기법에 대하여 기술한다. 개발된 단백질 구조 추출 기법은 Pfam 데이터베이스에서 제공하는 단백질 서열의 샘플링 결과를 기반으로 서열 정렬을 수행하고, 선정된 서열을 대상으로 SCOP 데이터베이스에서 단백질 구조 분류정보(family 및 fold)를 추출함으로써 구조 분류정보 추출 과정의 성능을 향상시키고자 한다.

1. 서 론

최근 다양한 생물에 대한 지놈 프로젝트가 수행되면서 데이터베이스에 수록되는 서열의 수가 기하급수적으로 증가하고 있다. 이에 비하여 각 서열의 구조 및 기능을 실험을 통하여 규명하는 것은 많은 시간과 노력이 필요하여 제한적일 수밖에 없다. 새로운 서열은 이미 알려진 서열과 연관되어 있는 경우가 대부분이다. 따라서 기존의 서열 데이터베이스를 활용하여 새로운 서열을 분석하면 구조 및 기능에 관한 중요한 단서를 얻을 수 있다. 이러한 데이터를 분석하고 처리하는 방법에 대한 연구의 필요성이 증대되고 있다.

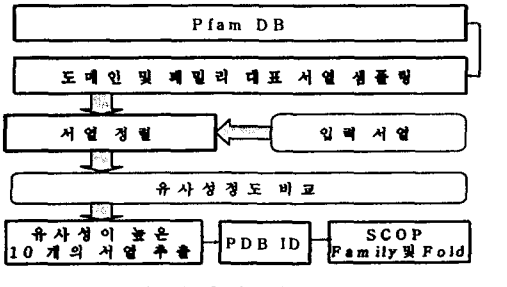
단백질의 기능은 삼차구조에 의해 결정되고, 삼차구조는 일차구조에 의해 결정된다. 따라서 단백질 일차구조, 즉 아미노산 서열을 분석하여 삼차구조에 관한 정보를 알아내고 어떤 분류(family)에 속하는지 알 수 있다면 기능이 밝혀지지 않은 단백질의 기능을 유추할 수 있다[1].

제안된 단백질 구조 분류정보 추출 기법은 Pfam[2], HMM[3], 그리고 SCOP[4][5]을 이용하여 구현되었다. Pfam은 현재 밝혀진 단백질 구조에서 공통으로 많이 나타나는 단백질 도메인(Domains)과 패밀리(Families)에 대한 정보를 제공하고 있다. Pfam에서 제공하는 도메인 및 패밀리 정보는 많은 단백질 서열에서 공통으로 나타나는 보존된 영역(conserved region)을 다중 서열 정렬을 통하여 분석하여 얻어진다. 이러한 단백질 서열과 연관된 단백질 부분 구조에 대한 표현 방법은 HMM(Hidden Markov models)을 이용하고 있다. HMM은 특정 도메인을 효과적으로 표현할 수 있는 도구로써 이용되고 있다. SCOP(Structural Classification of Proteins)은 PDB에 수록된 삼차구조를 체계적으로 분석한 데이터베이스이다. 단백질을 구조의 특성에 따라 총 10개의 class로 분류하고, 각각의 class는 2차 구조의 구성과 *토폴로지(topology)*에 의해 *폴드(fold)*로 나뉜다. 폴드(fold)는 *수퍼패밀리(superfamily)*로, 수퍼패밀리(superfamily)는 *패밀리(family)*로 나뉘고 패밀리(family)는 도메인*(domain)*으로 나뉘게 되어 단백질 구조의 계통 분류학적 분석이 체계적으로 이루진 데이터베이스이다.

본 논문에서는 단백질 서열 정렬을 통하여 보다 빠르고 효과적으로 단백질 구조 정보를 추출하는 기법에 대하여 기술한다. 이를 위하여 Pfam 데이터베이스에서 제공하는 단백질 서열의 도메인 및 패밀리 정보를 대표하는 서열을 추출하는 방법 및 입력서열과 관련된 단백질 구조 분류정보를 SCOP 데이터베이스에서 추출하는 방법을 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템 구조에 대하여 살펴본다. 3장에서는 좀더 자세한 단백질 구조 추출과정에 대해서 기술하며, 4장에서는 구현한 시스템의 결론에 대해서 기술하고자 한다.

2. 시스템 구조



[그림 1] 시스템 구조

[그림 1]은 단백질 구조 분류정보를 추출하는 시스템을 보여주고 있다. 단백질 서열로부터 단백질 구조 분류정보 추출 과정

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

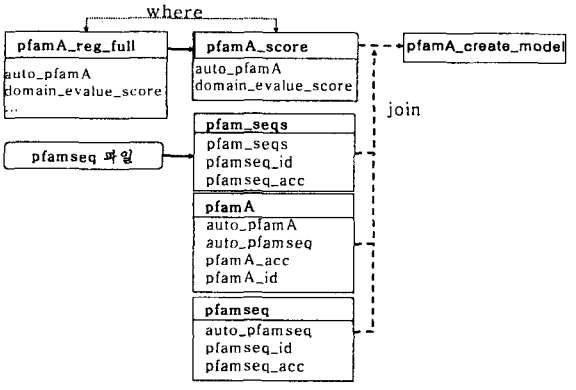
을 크게 3단계로 구분하였다.

첫 번째 단계는 Pfam 데이터베이스에서 제공하는 서열을 샘플링 하여 대표서열을 추출하는 과정이다. 여기서 추출된 대표서열은 Pfam의 각 도메인 및 패밀리 값을 대표하는 서열이다. 두 번째 단계는 샘플링을 통하여 생성된 대표서열 데이터와 입력된 단백질 서열을 정렬하는 과정이다. 이 과정을 통해 얻어진 전체 정렬 결과에서 상동성이 높은 10개의 도메인 및 패밀리 정보를 추출한다. 마지막 단계는 추출된 도메인 및 패밀리의 단백질 구조 분류정보를 추출한다. 두 번째 단계에서 상동성이 높은 10개의 결과를 바탕으로 SCOP 단백질 구조 분류정보를 추출한다.

3. 단백질 분류정보 추출

3.1 Pfam 데이터베이스에서 대표서열 샘플링

2003년 2월 현재 Pfam에서는 5193개의 도메인 및 패밀리의 정보를 서열, HMM 형태의 일반 텍스트와 관계형 데이터베이스로 제공한다[2]. Pfam에서 제공된 도메인의 대표서열을 생성하기 위하여 pfamA_reg_full 테이블의 domain_evalue_score 필드 값을 이용한다. 이 필드는 각 도메인의 HMM과 생성 시에 사용된 서열 사이의 유사성 정도를 나타내는 값을 나타낸다. 각 도메인을 대표하는 서열은 각 도메인을 생성하는 과정에서 사용된 서열 중에서 domain_evalue_score 값이 가장 낮은 값을 가진 서열이다. [그림 2]는 Pfam의 각 도메인을 대표하는 대표서열 데이터베이스를 생성하는 과정에 사용된 테이블을 보여주고 있다.



[그림 2] Pfam 도메인 및 패밀리 대표서열 샘플링에 사용된 테이블

① pfamA_score 테이블 생성

Pfam에서 제공하는 pfamA 테이블은 각 도메인에 대한 정보를 나타내며, pfamA_reg_full 테이블은 각 도메인과 각 도메인을 생성할 때 사용된 서열사이의 관계 정보를 제공하고 있다. 새롭게 생성되는 pfamA_score 테이블은 pfamA에 나타난 도메인을 지정하는 auto_pfamA와 각 도메인과 생성 시에 사용된 서열과의 유사성 정도를 나타내는 domain_evalue_score 값을 저장한다. domain_evalue_score 값이 낮을수록 도메인과 유사성이 많다는 것을 나타낸다.

② pfam_seqs 테이블 생성

Pfam에서 제공하는 pfamseq 파일은 각 도메인을 생성할 때 사용된 모든 서열 정보를 제공한다. Pfam의 각 도메인을 대표하는 서열을 추출하는 중간 과정에 사용되는 pfam_seqs테이블은 pfamseq 파일에 존재하는 모든 실제 서열을 다른 테이블에서 참조할 수 있는 pfamseq_acc, pfamseq_id 필드 값과 함께 제공한다. pfamseq_acc와 pfamseq_id는 각 실제 서열에 부여되는 값으로 주어진 pfamseq 파일에서 추출된다.

③ pfamA_create_model 테이블 생성

pfamA_create_model 테이블은 Pfam에서 제공하는 각 도메인을 대표하는 서열을 가진다. 이 테이블은 ①에서 생성한 pfam_score 테이블에서 제공하는 각 도메인의 대표서열에 해당하는 domain_evalue_score를 이용하여 pfamA_reg_full에서 auto_pfamseq 정보를 얻어낸다. 얻어낸 auto_pfamseq를 이용하여 pfamseq 테이블의 pfamseq_acc 값을 가져온다. 이 값을 이용하여 pfam_seqs 테이블의 실제 서열을 가져온다.

3.2 서열 정렬(Sequence Alignment) 활용

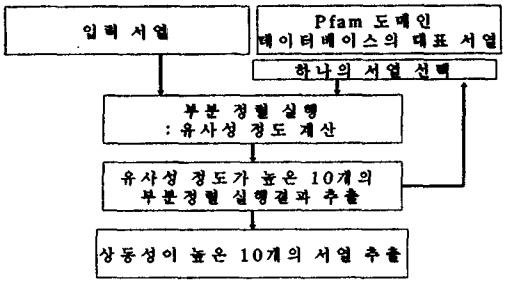
서열 정렬은 입력서열과 상동성(homology)이 높은 서열들을 알아내어 서열의 기능을 유추하거나, 관련 있는 서열들간의 *정량적인 상관관계*나 관련 기능 부위 등을 예측하기 위한 목적으로 이용된다[1].

단백질 서열 정렬 방법은 입력된 두 단백질 서열 사이의 유사성 정도를 측정하는데 사용되고 있다. 본 논문에서는 입력된 단백질 서열과 유사성이 높은 단백질 구조 정보(도메인 및 패밀리)를 대표하는 서열을 찾기 위하여 기존의 부분 정렬 방법을 이용하였다[6]. [표 1]은 Pfam의 대표서열 정보 및 입력서열과 대표서열의 정렬 결과를 저장하는 클래스(class)이다.

[표 1] Pfaminfo 클래스의 멤버필드

멤버 필드	설명
info_pfamA_acc	pfamA 도메인의 accession 값
info_pfamA_id	pfamA 도메인의 id값
info_pfamseq	pfam 도메인의 대표서열
info_align	입력서열과 pfam 도메인의 대표서열을 정렬한 결과
info_pdb	각 pfamA 도메인의 pdb id값들

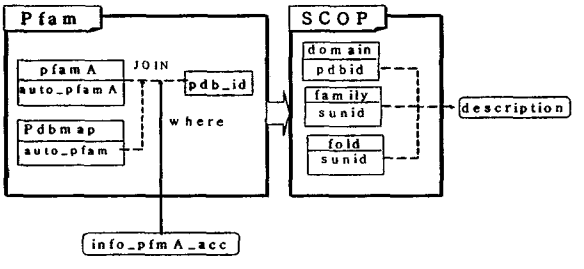
[그림 3]은 입력 서열과 유사성이 가장 높은 10개의 Pfam 도메인을 대표하는 서열을 추출하는 과정을 보여 주고 있다.



[그림 3] 서열 정렬 과정

3.3 SCOP family 및 fold 정보 추출

입력서열과 연관된 SCOP의 단백질 분류정보는 Pfam에서 제공하는 각 도메인 및 패밀리에 대한 정보를 바탕으로 얻어진다. [그림 4]는 Pfam과 SCOP 데이터베이스를 이용하여 입력된 서열과 연관된 단백질 구조 분류정보를 추출하는 과정을 보여주고 있다. 서열 정렬을 통하여 나온 10개의 유사성이 높은 서열의 pfamA_acc 값을 이용하여 pfamA와 pdbmap 테이블을 이용하여 PDB ID를 얻어낸다. pfamA 테이블은 도메인과 연관된 서열에 대한 정보를 기술하며, pdbmap 테이블은 Pfam에서 도메인 및 패밀리와 PDB 데이터베이스에서 제공하는 단백질 사이의 연결 관계를 나타낸다. SCOP 데이터베이스는 PDB 데이터베이스에서 제공하는 모든 단백질에 대한 구조 분류정보(family 및 fold)를 제공한다. 이러한 구조 분류정보는 PDB ID를 통하여 추출될 수 있다[5].



[그림 4] SCOP Family정보 추출

[표 2] 입력서열과 관련된 Pfam 도메인

유사성순위	도메인	PDB ID
1	globin	101m, 102m, 1a00, 1a0u, ...
2	Condensation	없음
3	pp-binding	1acp, 1af8, 1dv5, 1f80, 2af8
4	ALA_synthase	1bs0, 1dj9, 1dje
5	NB-ARC	없음
6	mRNA_cap_C	1ckm, 1ckn, 1cko
7	ACR_tran	없음
8	CheR	1af7, 1bc5
9	CBM_15	1gny
10	Surface_Ag_2	없음

[표 2]는 입력된 1MBA의 서열을 pfam의 도메인을 대표하는 서열 데이터베이스를 대상으로 서열 정렬을 수행하여 유사성이 높은 Pfam 도메인 및 PDB ID를 보여주고 있다. Pfam에서 정의된 도메인 정보에 PDB ID 정보가 없는 경우, SCOP 구조 분류 정보를 추출할 수 없으므로 다른 도메인 정보를 참조로 SCOP 구조 분류 정보를 추출한다.

[표 3]은 pfam 도메인의 PDB ID를 이용하여 SCOP 구조 분류 정보를 추출한 결과를 보여주고 있다.

[표 3] PDB ID를 바탕으로 추출된 SCOP의 패밀리

PDB ID	SCOP 구조 분류 정보	
	Family	Fold
1f7v	Arginyl-tRNA synthetase (ArgRS), N-terminal 'additional' domain	Adenine nucleotide alpha hydrolase-like
1cqx	Ferredoxin reductase FAD-binding domain-like	Globin-like
1bs2	Arginyl-tRNA synthetase (ArgRS), N-terminal 'additional' domain	Adenine nucleotide alpha hydrolase-like
.....	.....	...

4. 결론

본 논문에서는 단백질 서열 정렬을 통하여 보다 빠르고 효과적으로 단백질 구조 정보를 추출하는 기법에 대하여 기술하였다. 개발된 단백질 구조 추출 기법은 Pfam 데이터베이스에서 제공하는 단백질 서열의 샘플링 결과를 기반으로 서열 정렬을 수행하고, 선정된 서열을 대상으로 SCOP 데이터베이스에서 단백질 구조 분류정보(family 및 fold)를 추출함으로써 구조 분류 정보 추출 과정의 성능을 향상시켰다.

5. 참고문헌

[1] Cynthia Gibas and Per Jambeck, "Developing Bioinformatics Computer Skills", 2001

[2] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL, "The Pfam Protein Families Database", Nucleic Acids Research, Vol. 30, No. 1 276-280, 2002

[3] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: Extension and analysis of the basic method", CABIOS 12(2), 95-107, 1996

[4] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540, 1995

[5] 안건태, 윤형석, 황의윤, 이명준, "SCOPML과 SCOP-Browser에 관한 연구", 정보처리학회논문지 D 제10-D권 제1호(2003년 2월)

[6] I Eidhammer, I Jonassen, W R. "Structrue Comparion and Struture Pattern", Reports in informatics, 7, 1999.