제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

모티프 자원 통합 데이터베이스 구축 및 메타엔진 설계

이범주*, 최은선*, 류근호*
*충북대학교 데이터베이스연구실
{bjlee, eschoi, khryu}@dblab.chungbuk.ac.kr

Implementation of integrated motif database and Design of meta engine

Bum Ju Lee*, Eun Sun Choi*, Keun Ho Ryu*
*Database Laboratory, Chungbuk National University

요 약

새롭게 발견되는 단백질의 구조와 기능 예측에 사용되는 모티프는 단백질 원시 데이터가 빠르게 증가함에 따라 그 중요성 역시 나날이 증가하고 있으며 이러한 모티프에 대한 다양한 서열 메소드들과 데이터베이스들이 개발되었다. 그러나, 이러한 모티프 데이터베이스들은 각각 이질적인 데이터 구조를 지니고 독자적으로 개발, 발전되어 왔기 때문에 서로 다른 형태의 검색 결과를 제공한다. 따라서 사용자는 각 데이터베이스에서 사용하는 데이터 구조들에 대한 전반적인 지식을 습득해야 하며, 모티프 데이터베이스들 각각에 대해 중복된 반복 검색 작업들을 수행하여야 한다.

따라서 이 논문에서는 이러한 문제 해결을 위해, 각각의 모티프 데이터베이스들에서 제공하는 자원을 분해하고, 합병하는 과정을 거쳐 하나의 통합된 모티프 데이터베이스를 구축하였고, 기존의 데이터베이스에서 지원하지 못했던 단백질 3차 구조정보, 분류 정보의 지원을 가능케 하였고, 각 멤버 데이터베이스 검색 메소드의 장점을 그대로 적용할 수 있는 메타 엔진을 설계하여 사용자 편의적 검색 환경을 제공한다.

1. 서론

서열 시퀀싱을 통해 빠르게 증가하는 원시데이터들을 대상으로 유사한 서열과 기능 예측에 사용되는 모티프에 대해 지난 10년간 많은 데이터베이스들이 출현하였다[1, 5, 8]. 현재 Prosite, PRINTS, Pfam, ProDom, BLOCKS, SMART 등 각기 다른 데이터 구조를 지닌 데이터베이스들의 등장과[1, 5, 8, 9, 13] 함께, 이러한 이질적인 데이터 구조로 생성된 다양한 데이터베이스 통합을 위해 웹 기반의 cross-reference가 주로 사용되어져 왔다[3, 4]. 그러나 웹 기반 통합은 엔트리 상호간 데이터 구조를 변경하지 않고 관련된 엔트리간 유연한 통합을 지원할 수 있는 장점에 비해, 복잡한 질의를 처리, cross-reference된 엔트리들의 수, 중복된 데이터베이스의 핸들링, 네트웍 과부화 등과 같은 많은 문제점들을 지니고 있다[2]. 뿐만 아니라 데이터베이스 검색시 사용자는 각각의 데이터베이스에서 사용하는 데이터 구조에 대한 전반적인 지식을 습득해야 하며, 각 데이터베이스들에 접근하여 중복된 검색 작업을 수행하여야 하고, 검색 결과에 대한 통합된 정보를 얻을 수 없었다[2, 3].

따라서 이 논문에서는 이러한 문제들에 대한 해결 방안으로 PRINTS, Prosite, Pfam 데이터베이스들에서 제공하고 있는 플랫파일을 분석하여, 분해 및 합병 과정을 통해 중복된 데이터들을 하나의 자원으로 통합하였고, 이렇게 통합된 각 엔트리들에 대해 단백질 3차 구조정보를 가지고 있는 PDB데이터베이스(Protein Database), 단백질 분류 정보를 가지고 있는 SCOP 데이터베이스들을 통합하였다. 또한 사용자 편의적 검색과 각 멤버 데이터베이스 검색 프로그램들의 장점을 그대로 살리기 위해 하나의 검색 메타 엔진을 설계하였다. 이로써 웹 기반

cross-reference 통합에서 나타나는 복잡한 질의 처리와 중복된 데이터베이스들의 핸들링 문제들을 해결하였고, 메타 엔진을 이용하여 검색 결과에 대한 재 조직화를 거쳐 사용자 편의적 통합 검색을 가능케 하였을 뿐만 아니라 기존의 통합 모티프 데이터베이스에서 지원하지 못했던 모티프 3차 구조 정보와 분류 정보 지원이 가능하도록 데이터베이스 기능을 개선하였다.

2장에서는 InterPro 데이터베이스와 PANAL(an integrated resource for Protein sequence ANALysis) 검색 시스템을 관련 연구로써 기술하였고, 모티프 데이터베이스 통합을 위한 모델링을 3장에서 다루었으며, 4장에서는 메타 엔진 설계에 관하여, 그리고 구현 및 평가와 결론 및 향후연구를 각각 5장, 6장에 기술하였다.

2. 관련 연구

2.1 InterPro 데이터베이스

단백질 패밀리, 도메인, functional site들에 대한 물리적 통합 문서 자원을 목적으로 생성된 InterPro 데이터베이스는 PRINTS, PROSITE, Pfam, ProDom과 같은 시그네쳐 데이터베이스들에 대한 검색 진단 데이터와 문서들을 하나의 집중된 자원으로 통합하였다.

통합 메소드로 parent/child와 contains/found_in을 사용한 이 데이터베이스의 각 엔트리는 functional description, annotation, literature reference를 포함하고 있고, 관련 맴버 데이터베이스에 대한 링크와, SWISS-PROT과 TrEMBL에 대한 매치정보를 제공하고 있다.

이 데이터베이스 버전 5.1(2002. 7월)은 1239개의 도메인, 4280개의 패밀리, 95개의 repeat, 15개의 PTM 사이트들로 구성된 총 5629개의 엔트리를 포함하고 있으며, 웹상에서 엔트리 데이터와 매치 데이터를 XML 형식으로 배포하고 있다[1, 5].

2.2 PANAL 검색 시스템

단백질 서열 분석을 목적으로 제작된 PANAL은 사용자가 여러 개의 모티프 데이터베이스들에 대해 단백질 서열 검색을 동시에 수행하는 것을 목적으로 제작되었다.

BLAST와 FASTA보다 높은 민감도와 신뢰도를 제공하는 패밀리 기반 메소드들을 채택한 이 툴은 각 모티프 데이터베이스들에 대해 사용자가 제시한 E-value cutoff에 따라 단백질 서열 검색을 수행하고 각 데이터베이스에서 제공하는 검색 결과 외에도 그 검색결과들에 대한 요약 정보 창을 사용자에게 제공한다[6].

3. 모티프 데이터베이스 통합 모델링

3.1 플랫파일 분석 및 통합을 위한 메소드

모티프 데이터베이스들은 각각의 고유한 장점(예, PRINTS 데이터베이스는 구별하기 어려운 subfamily relationship 식별에 높은 성능을 발휘하며, Pfam 데이터베이스는 비교적 멀리 연관된 패밀리 맴버를 식별하는데 뛰어남)을 살리기 위해 fingerprint, profile, regular expression, HMMs 등의 데이터 형식으로 각각 서열 패턴 정보들을 나타내고 있고, 플랫파일 상에서 라인당 특별한 문자들로 라인이 포함하고 있는 정보의 의미를 구별한다.

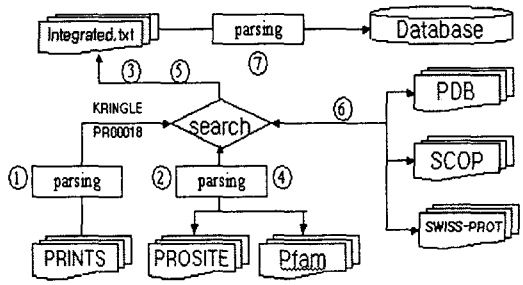

그림 1. 각 플랫파일 및 관련 정보 통합 메소드

우리는 이질적 형식의 모티프 자원을 하나로 통합하기 위해 PRINTS, Pfam, Prosite 데이터베이스에서 제공하는 각각의 플랫파일을 분석하고, 이를 분해 및 합병하였다. 먼저, 하나의 동일한 모티프를 각각의 플랫파일에서 이질적인 형태로 저장하고 있는 것들을 검색하고, 일치하는 엔트리들을 추출하여 비교 분석한 뒤, 단일한 하나의 엔트리로 통합하여 새로운 플랫파일에 저장한다.

그 후 일치하지 않는, 즉 한 데이터베이스 플랫파일에만 존재하는 엔트리들을 하나의 새로운 엔트리로 재 생성하여 새롭게 생성된 플랫파일에 추가 저장한다. 마지막으로, 통합된 엔트리 각각에 해당하는 단백질의 3차 구조 정보로 PDB의 엔트리 데이터를 적용하였다. 이러한 방식으로 분류 정보를 위한 SCOP, 샘플 정보를 위한 SWISS-PROT을 새로운 플랫파일에 추가 저장한다. 이렇게 생성된 플랫파일은 다시 파싱 과정을 거쳐 관계형 데이터베이스에 저장한다. 이러한 과정을 그림 1에서 순서별로 기

술하였고 이 과정을 거쳐 새롭게 생성된 플랫파일을 그림 2에서 나타내었다.

```
AC : 2
DATE : 2002/08/20
sigDB_print : PR00305
NA : 14-3-3 protein zeta signature
sigDB_interpro : IPR000308
sigDB_prosite : 1  PS00796 1433_1: PS00797 1433_2
TAX : Eukaryotes
prosite DOC : PDOC00633
sigDB_pfam : PF00244 14-3-3
TP : Domain
medline : 1  95327195
author : Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken
author : A, Gamblin SJ;
title : Structure of a 14-3-3 protein and implications for coordination of multiple signalling
        pathways.
reference : Nature 1995;376:188-191.
medline : 2  95327196
...
abstract : 1  The 14-3-3 proteins are a family of related proteins found in mammalian
abstract : 2  brain cells, preferentially in neurons, although similar proteins have been
abstract : 3  identified in all eukaryotic species studied to date. They are homodimeric,
abstract : 4  acidic proteins [1] with multiple biological activities: they act as
abstract : 5  protein kinase-dependent activators of tyrosine and tryptophan hydroxy-
...
SCOP : 1  3fxc: fa:
PDB line : 1  1a70 : 8: 83:
PDB line : 2  1roe : 8: 84:
PDB line : 3  2cjn : 8: 84:
PDB line : 4  2cjo : 8: 84:
...
```

그림 2. 자원 통합을 위해 새롭게 생성된 플랫파일

3.2 모티프 데이터베이스의 E-R 다이어 그램

우리는 위의 메소드를 통해 새롭게 생성된 플랫파일을 보다 효율적으로 검색하고 관리하기 위해 관계형 데이터베이스를 구축하였다. 따라서, 각 데이터들의 연관성 분석을 토대로 그림 2와 같은 E-R 다이어 그램을 나타내었다.

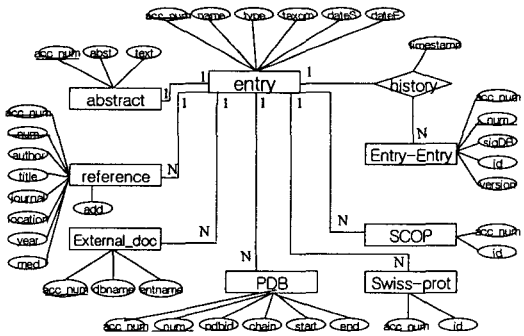

그림 3. 모티프 자원 통합을 위한 E-R 다이어 그램

하나의 엔트리는 축약 아이디, 풀 아이디, 타입, taxonomy, abstract, reference, 외부 문서 등에 대한 일반적인 자원들을 위한 엔티티들과 보다 많은 정보를 지원하기 위한 단백질의 3차 구조정보, 분류정보, 샘플정보의 엔티티들을 포함하고 있다. 또한, 사용자가 통합 이전의 최초 정보들을 필요로 할 경우를 위해 entry-entry 엔티티(맴버 데이터베이스들이 통합되기 이전 엔트리들의 accession number, 아이디, 버전 정보 등을 보유한 엔티티)의 버전 속성과 timestamp를 이용하여 업데이트 시 참조할 수 있도록 하였다.

4. 메타 엔진을 이용한 검색 시스템 설계

각각의 모티프 데이터베이스들은 다음과 같은 데이터 구조를 사용한다.

표 1. 맴버 데이터베이스의 데이터 구조 및 검색 프로그램

데이터베이스	데이터 구조	검색 프로그램
Prosite	profile, regular expression, rule	PS_scan
PRINTS	fingerprint	FingerPRINTScan
Pfam	HMMs	HMMER

이러한 검색 프로그램들은 그 데이터 구조 고유의 특성에 따른 장점과 단점을 가지고 있다.

따라서 우리는 기존의 검색 프로그램의 장점을 그대로 살리고 검색 결과에 대한 재조직화 및 통합 결과를 제공하기 위해 메타 엔진을 설계하였다. 이 엔진은 다음과 같은 순서로 수행된다.

첫째, 사용자는 데이터베이스 인터페이스에 접근하여 검색 서열, E-value cutoff 등의 정보를 입력한다. 둘째, 사용자가 입력한 정보들은 HMMER, FingerPRINTScan 등의 전용 검색 프로그램들에서 병렬적으로 수행한다. 셋째, 각 검색 프로그램들에서 수행된 결과 값을 메타 엔진을 이용하여 하나의 통합된 결과 창에 나타낸다. 이로써 기존에 사용자가 각각의 데이터베이스에 반복적으로 접근해야 하는 문제, 중복된 정보 입력 문제, 검색 결과에 대한 통합 결과 문제 등을 해결 할 수 있다. 이러한 메타엔진 구조와 전체 데이터베이스 아키텍처를 그림 4에서 나타내었다.

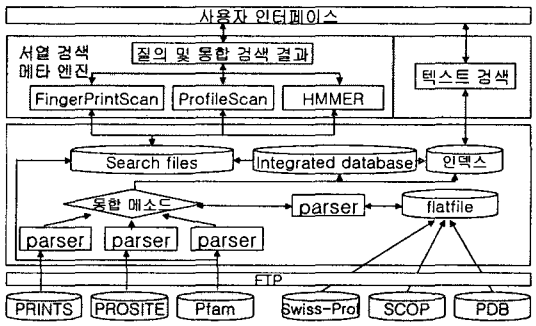

그림 4. 통합 데이터베이스 및 검색 시스템 아키텍처

5. 구현 및 평가

이 논문에서 모티프 자원 통합 메소드 및 통합 관계형 데이터베이스 구축을 위해 C와 pro*C를 사용하였고 DBMS는 Oracle8을 기반으로 하였다. 그리고 통합에 이용한 맴버 데이터베이스들 즉, Prosite, Pfam, PRINTS의 엔트리들은 다음과 같다.

① PRINTS에서 제공하는 1,410개의 fingerprint들
② Prosite에서 제공하는 1,510개에 해당하는 rule, regular expression, profile들
③ Pfam-A.seed에서 제공하는 3,849개의 엔트리들

이러한 엔트리들을 저장하고 있는 플랫파일을 분해, 통합과정을 거쳐 5,670개의 새로운 엔트리로 재구성하였다.

이렇게 생성된 데이터베이스에서 accession number 99번 annexin에 해당하는 정보를 추출하기 위해 SQL문을 통한 검색 결과를 그림 5와 같이 나타냈다. sig 컬럼은 PRINTS, Pfam, Prosite, InterPro 맴버 데이터베이스에서의 통합 이전 accession number와 ID를 나타낸다.

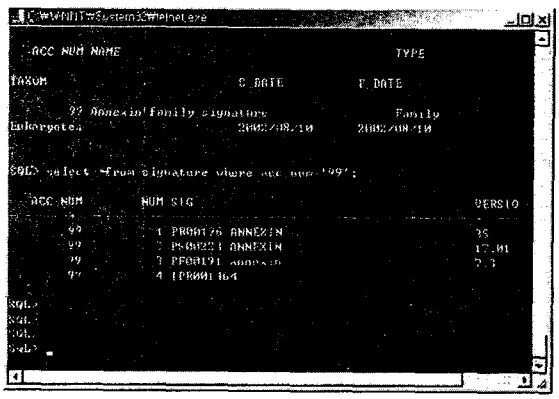

그림 5. annexin 엔트리 검색 결과 창

6. 결론 및 향후 연구

우리는 빠르게 등장하는 단백질 서열들에 대한 기능 및 구조 예측에 사용되고 있는 모티프 데이터베이스들을 하나의 자원으로 통합하였고, 통합된 자원의 효율적 관리를 위해 새로운 관계형 데이터베이스를 설계 및 구축하였다. 뿐만 아니라 맴버 데이터베이스에서 사용되고 있는 검색 프로그램들에 대한 장점을 그대로 제공하면서 검색 결과들에 대한 통합 결과를 제공할 수 있는 메타 엔진을 설계하였다.

따라서, 웹 기반 통합에 따른 복잡한 질의 처리 문제, 중복된 데이터베이스들의 핸들링 문제, 기존의 데이터베이스 검색시 사용자가 겪는 이질적 검색환경 및 반복 접근 문제를 해결하였고 더 많은 정보를 지원하기 위해 단백질의 3차 구조정보, 분류 정보, 샘플 정보의 지원을 가능케 하였다. 향후 연구로는 검색 시스템의 구현 및 더 많은 모티프 자원 통합을 위한 연구가 진행 중이다.

참고문헌

[1] R. Apweiler, T.K. Attwood, A.Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D.R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H.Hermjakob, N. Hulo, L.Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J.A. Sigrist and E.M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nuleic Acids Research, Vol. 29, No.1, page 37-40, Oxford University Press, 2001.

[2] M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser, "Proteome Research: New Frontiers in Functional Genomics", Springer-Verlag Berlin Heidelberg, page 125-127, 1997.

[3] M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser, "Proteome Research: New Frontiers in Functional Genomics", Springer-Verlag Berlin Heidelberg, page 150-175, 1997.

[4] Minoru Kanehisa, "Post-Genome Informatics", Oxford university press, page 35-47, 2000.

[5] David W. Mount, "Bioinformatics : Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press, page 429-431, 2001.

[6] Kevin A. T. Silverstein, Alan Kilian, John L. Freeman, James E. Johnson, Ihab A. Awad, Ernest F. Retzel, "PANAL: an integrated resource for Protein sequence ANALysis", Bioinformatics, Vol. 16, p1157-1158, 2000.

[7] M.R. Willcins, K.L. Williams, R.D.Appel, D.F. Hochstrasser, "Proteome Research: New Frontiers in Functional Genomics", Springer, p109-114, 1997.

[8] T.K. Attwood, M.E. Beck, D.R. Flower, P. Scordis, N. Selley, "The PRINTS protein fingerprint database in its fifth year", Nucleic Acids Research, Vol.26, No.1 p304-308, 1998.

[9] Alex Bateman, Evan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R. Eddy, Sam Griffiths-Jones, Kevin L. Howe, Mhairi Marshall, Erik L.L. Sonnhammer, "The Pfam Protein Families Database", Nucleic Acids Research, Vol.30, No.1 p276-280, 2002.

[10] Jorja G. Henikoff, Steven Henikoff, Shmuel Pietrokovski, "New features of the Block Database servers", Nucleic Acids Research, Vol.27, No.1 p226-228, 1999.

[11] T. K. Attwood, H. Aviison, M. E. Beck, M. Bewley, A. J. Bleasby, F. Brewster, P. Cooper, K. Degtyarenko, A. J. Geddes, D. R. Flower, M. P. Kelly, S. Lott, K. M. Measures, D. J. Parry-Smith, D. N. Perkins, P. Scordis, D. Scott, C. Worledge, "The PRINTS Database of Protein Fingerprints: A Novel Information Resource for Computational Molecular Biology", J. Chem. Inf. Comput. Sci. 37, p417-424, 1997.

[12] Wolfgang Fleischmann, Steffen Moller, Alain Gateau, Rolf Apweiler, "A novel method for automatic functional annotation of proteins", Bioinformatics, vol 15, p228-233, 1999.

[13] Laurent Falquet, Marco Pagni, Philipp Bucher, Nicolas Hulo, Christian J. A. Sigrist, Kay Hofmann, Amos Bairoch, "The PROSITE database, its status in 2002", Nucleic Acids Research, Vol.30, p235-238, 2002.

[14] Loredana Lo Conte, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia and Alexey G. Murzin, "SCOP database in 2002: refinements accommodate structural genomics", Nucleic Acids Research, Vol.30, p264-267, 2002.

[15] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T.N.Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, "The Proten Data Bank", Nucleic Acids Research, Vol. 18, p235-242, 2000.

[13] Bairoch, A., Apweiler, R., "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", Nucleic Acids Res, 28(1), p45-48, 2000.