모티프 자원 통합을 이용한 단백질 모티프 예측 시스템 구현

이 범 주[†] · 최 은 선[†] · 류 근 호[††]

요 약

지놈 서열 시퀀싱을 통해 생성되는 원시 데이터에 대한 단백질 기능 및 구조 예측에 사용되는 모티프 데이터베이스들은 원시 데이터들의 폭발적인 성장추세에 맞추어 그 사용빈도가 증가하고 있다. 그러나 이러한 모티프 데이터베이스들은 독자적으로 개발, 발전하여왔고 웹 기반 cross-reference를 이용한 논리적 통합을 추진하여왔기 때문에 이질적인 검색 결과와 복잡한 질의 처리 문제, 중복된 데이터베이스 엔트리 핸들링 문제 등을 갖고 있다. 따라서, 이 논문에서는 이런 문제점들을 개선하기 위하여 물리적인 모티프 자원 통합을 제안하고, 패밀리 기반 단백질 예측 메소드들에 대한 통합 검색 방법을 기술한다. 끝으로 모티프 통합 데이터베이스 구축 및 단백질 모티프 예측 시스템 구현을 통한 결과를 평가한다.

Implementation of Protein Motif Prediction System Using integrated Motif Resources

Bum Ju Lee[†] · Eun Sun Choi[†] · Keun Ho Ryu[††]

ABSTRACT

Motif databases are used in the function and structure prediction of proteins which appear on new and rapid release of raw data from genome sequencing projects. Recently, the frequency of use about these databases increases continuously. However, existing motif databases were developed and extended independently and were integrated mainly by using a web-based cross-reference, thus these databases have a heterogeneous search result problem, a complex query process problem and a duplicate database entry handling problem. Therefore, in this paper, we suppose physical motif resource integration and describe the integrated search method about a family-based protein prediction for solving above these problems. Finally, we estimate our implementation of the motif integration database and prediction system for predicting protein motifs.

키워드 : 모티프(Motif), 모티프 자원 통합(Motif resource integration), 단백질 예측(Protein prediction), 바이오인포메틱스(Bioinformatics)

1. 서 론

생물 정보 데이터베이스 이용 추세에 한 축을 이루고 있는 모티프 데이터베이스는 단백질 아미노산 서열과 3차 구조 정보 사이의 연관 관계를 이용하여 새로이 등장한 단백질의 기능 예측에 사용된다. 모티프는 서열의 부분적인 보존 영역이나 서열 집합이 공유하는 짧은 서열 패턴을 의미하며 분자의 기능을 예측할 수 있는 특정한 서열의 패턴이나 구조적인 특징을 가진다[5, 7, 17]. 지난 10년간 개발된 모티프 데이터베이스들로는 Regular expression, Rule, Profile 데이터 구조를 이용한 PROSITE 데이터베이스[11], 다중 모티프들로 구성된 Fingerprint 데이터 구조를 적용한 PRINTS 데이터베이스[7, 10], HMM(Hidden Markov Models) 데이터 구조를 사용한 Pfam 데이터베이스[8] 등 매우 다양한 모티프 데이터베이스들이 독자적으로 개발, 발전되어져 왔다. 또한 최근에 이르러 이러한 이질적인 데이터 구조로 생성된 여러 모티프 데이터베이스들의 통합을 위해 웹 기반 Cross-reference를 이용한 논리적 통합이 주로 사용되었다[3, 4].

그러나 웹 기반 cross-reference를 이용한 논리적 통합 및 검색 시스템은 엔트리 상호간 데이터 구조를 변경하지 않고 관련된 엔트리간에 유연한 통합을 지원할 수 있는 장점에 비해, 복잡한 질의 처리 문제, Cross-reference된 과도한 엔트리들의 수, 네트웍 과부화 등과 같은 문제점들을 지니고 있다[3]. 기존까지 데이터베이스 검색을 위해서 사용자들은 각 데이터베이스들에 접근하여 중복된 검색 작업을 수행하여야

※ 이 연구는 2002년도 학술진흥재단의 연구비(KRF-2002-072-AM1013)의 지원으로 수행되었음.
† 준 회 원 : 충북대학교 대학원 전자계산학과
†† 종신회원 : 충북대학교 전기전자 및 컴퓨터공학과 교수
논문접수 : 2002년 12월 3일, 심사완료 : 2003년 3월 13일

하고, 검색 결과에 대한 통합된 정보를 얻을 수 없었다[6].

이 논문에서는 위에 기술한 문제들에 대한 해결 방안으로 단백질 모티프들의 Annotation 정보, 3차 구조 정보 및 분류 정보 등을 물리적으로 하나의 통합된 DBMS를 사용하여 저장함으로써 효율적 관리를 위한 기반을 마련하고 기존 모티프 데이터베이스에서 지원하는 검색 메소드들에 대한 장점을 그대로 적용한 새로운 통합 검색 시스템을 제안한다. 이 논문의 핵심적 특징은 다음과 같다.

- PRINTS, Prosite, Pfam 데이터베이스들에서 제공하고 있는 플랫파일을 분석, 분해 및 합병 과정을 통해 중복된 데이터들을 하나의 자원으로 통합한다.
- 통합된 각 엔트리들에 대해 단백질 3차 구조정보를 가지고 있는 PDB 데이터베이스(Protein DataBase)와 단백질 분류 정보를 가지고 있는 SCOP(Structural Classification Of Proteins) 데이터베이스 자원을 통합한다.
- 사용자 편의적 검색과 각 멤버 데이터베이스 검색 프로그램들의 장점을 그대로 살리기 위해 맴버 데이터베이스 검색 프로그램들을 하나의 단백질 모티프 검색 시스템으로 구현한다.

이로써 웹 기반 cross-reference 통합에서 나타나는 복잡한 질의 처리 문제와 중복된 데이터베이스들의 핸들링 문제들에 대한 해결책을 제시하며, 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.

2. 주요 모티프 관련 데이터베이스

단백질 서열 분석 전략에서 표준 툴 역할을 하는 모티프 데이터베이스는 자동화된 분석에 의존하기도 하지만 대부분 전문관리자와 생물학자의 수작업이 데이터베이스 구축에 투입된다. 따라서 모티프 데이터베이스는 GenBank에 비해서 비교적 규모가 매우 작고 단백질 구조나 서열 데이터베이스 차원의 서비스를 지원하지 못하고 있다[19]. 이러한 모티프 데이터베이스에 대한 검색에서 실패했다는 것은 검색 서열에서 찾아낼 만한 패턴이 없다는 것을 의미하지 않는다. 이것은 아직까지 모티프로 밝혀내지 못한 유형의 부분 일수도 있고, 검색한 데이터베이스의 범주 내에 패턴이 포함되지 않았을 수도 있기 때문이다. 주요 모티프 데이터베이스는 다음과 같다.

2.1 PROSITE

SIB(Swiss Institute of Bioinformatics)에서 운영하고 있는 PROSITE는 지놈 또는 cDNA 서열에서 번역된 단백질의 기능을 식별하여 중요도(significance)가 높은 패턴, 룰, 프로파일들을 생성 및 저장하고 있다. 가중치 메트릭스라고도 불리는 프로파일은 단백질 또는 도메인 발견에 매우 유용하지만, 패턴은 높은 서열 유사성에 대해 작은 지역에 제한적으로 사용되므로 몇몇 패밀리들은 발견하기 어렵다. PROSITE는 이러한 패턴과 프로파일을 이용하여 단백질 패밀리 또는 도메인을 PS_scan, MotifScan, ScanProsite와 같은 신뢰성 있는 툴들을 사용하여 생성한다[11]. 현재 릴리이즈 버전은 17.21 (2002. 9. 21)에서는 1568개의 패턴, 룰, 프로파일/메트릭스들을 저장하고 있다.

2.2 PRINTS

Manchester 대학에서 유지하고 있는 PRINTS 데이터베이스는 PROSITE와 매우 유사하지만 패턴 인식 방법에서 많은 차이점을 나타낸다. 패턴 또는 프로파일을 사용하는 PROSITE 데이터베이스와는 달리 PRINTS 데이터베이스는 하나 이상의 다중 모티프로 구성된 Fingerprint를 사용한다. 모티프는 전체 단백질 서열에 비하면 비교적 짧기 때문에 Fingerprint를 사용하면 더욱 정확한 단백질 서열의 특성을 알아낼 수 있다. 이 Fingerprint는 가중치 부여, 2차 구조정보, 유사성 데이터들을 제공하지 않으며, 오직 빈도 스켄만을 다룬다[7, 10, 15]. 최근에 PRINTS-S라 불리는 관계형 DBMS로 저장소를 확대했으며, 버전 35.0(2002. 7)에서는 1750개의 엔트리들이 저장되어 있다.

2.3 Pfam

Trust Sanger Institute에서 유지하고 있는 Pfam 데이터베이스는 단백질 도메인 패밀리의 정렬 데이터베이스이다. 이 데이터베이스는 Pfam-A와 Pfam-B로 나뉘어져 있는데, Pfam-A는 갭을 허용한 프로필로 설계된 데이터베이스이며 대부분의 단백질 도메인을 망라하고 있다 Pfam-B는 Pfam-A를 만들고 난 나머지 서열들에 클러스터링 기법을 적용하여 자동으로 생성한 구성원으로 이루어져 있다. Pfam-A의 구성원은 신뢰할 수 있는 다중 서열 정렬인 씨앗 서열 정렬 부위로 시작하며 경우에 따라서는 수작업으로 편집하기도 한다. 가장 최근 버전 7.6(2002. 9)에서는 4,463개의 패밀리들을 포함하고 있으며, 플랫파일 형태로 제공하고 있다[8].

2.4 PDB

PDB(Protein DataBank)는 생물학적인 단백질 3차원 고분자 결정 구조를 위한 데이터베이스로서 1971년 부룩헤이번 국립 연구소(BNL)에 의해서 공개되었다. FSSP 데이터베이스(Fold classification based on Structure-Structure alignment of Proteins)는 이러한 PDB 데이터를 원자 구조 비교

프로그램인 Dali를 이용하여 단백질 3차 폴드에 대한 구조 분류 데이터베이스 구축에 사용하며, HSSP(Homology-derived Structures of Proteins) 또한 PDB 데이터를 이용하여 3차 구조 및 1차 구조를 통합한 데이터베이스를 구축하는데 있어서 구조적 유사성에 의한 단백질 서열 패밀리로 그룹화한다. PDB에서 제공하는 플랫파일에서 모티프는 SEQRES 섹션에 서열 정보를 보유하고 있고, ATOM 섹션에서 표준 residue에 대한 모티프 3차 구조 정보를 보유하고 있다.

3. 생물 정보 자원 통합 방법론

현재, 모티프 데이터베이스들은 사용자 편의적 관점에서 중복 접근, 이질적 검색 결과 등의 문제점 들을 내포하고 있다. 따라서 사용자 편의를 위한 모티프 자원 검색을 위해 하나의 자원으로 통합되어져야 한다[20]. 현재 모티프 데이터베이스에 대한 통합 및 검색을 지원하는 방법은 크게 3가지로 나눌 수 있다. 첫째, 웹 기반 cross-reference를 이용한 논리적 통합으로써 대부분의 생물 정보 데이터베이스들이 이러한 구조를 지니고 있다[3]. 다음으로, 여러 생물 데이터베이스들을 물리적으로 하나의 데이터베이스로 통합하는 것이다. 이러한 통합은 어휘 의미, 데이터 표현 등에 대한 문제점을 동반하고 있지만 사용자 측면과 관리적 측면에서 매우 효율적이다[3, 14]. 최근에 이르러 생물 정보데이터베이스들에 대해 이러한 물리적 통합으로 InterPro 데이터베이스가 출현하였다. 마지막으로, 각 생물 정보 데이터베이스에 다중 접근하여 정보를 추출해 오는 가상적인 통합 검색(federated database system)으로써, 이러한 가상 통합 검색으로 SRS[13]와 Entrez[5], PANAL[6] 등의 어플리케이션들이 있다. 생물 정보학에서 통합 데이터베이스 및 검색에 대한 방법론과 그 장단점들을 [1, 3, 4, 22]를 근거로 <표 1>에 기술하였다.

우리는 이러한 통합 방법들과 검색 시스템에 대한 분석을 통하여 우리가 진행할 연구에 대한 기준을 채택하였다. 첫째, 사용자 편의적 접근법을 제공하여야 한다. 기존의 독립적 데이터베이스에 대한 반복 접근 검색과 이를 통한 각각의 데이터베이스 검색에 따른 이질적인 검색 결과를 개선하기 위해서는 웹 상의 생물 정보 데이터베이스들을 하나로 통합하여야 한다. 둘째, 물리적인 데이터베이스로 통합되어져야 한다. 웹 기반 cross-reference를 이용한 논리적 통합은 각각의 생물정보 데이터베이스 엔트리들을 수정하지 않는 유연한 통합이 가능하나 위에 기술한 단점 외에도 데이터베이스 업데이트 시 dead link 즉, cross-reference가 끊어지는 문제가 발생하므로 데이터 무결성에 대한 한계점을 가지고 있다. 셋째, 실제적인 통합 검색이 이루어져야 한다. Entrez, SRS와 같은 가상 통합을 이용한 검색은 SQL 질의와 같은 섬세한 부분 검색 및 최적화가 불가능하며, 개별적인 시스템 검색에서 파라미터를 제공하는 세분화된 검색 능력을 상실한다. 넷째, 기존 모티프 자원 통합 데이터베이스인 InterPro에서 제공하지 못한 모티프 3차 구조 정보와 분류 정보를 통합해야 한다. 이것은 "유사한 기능을 가진 모티프는 유사한 구조를 지니고 있다."라는 모티프 기본 명제에 따라 단백질 서열 검색과 구조분석 연구에 상호

〈표 1〉 생물 정보 데이터베이스 통합 방법론의 장·단점 요약

통합 방법론		장 점	단 점	예
물리적 통합	관계형 데이터베이스	• 질의 처리와 최적화가 우수함 • 오류나 불일치에 대한 데이터 무결성을 보장 • 회복, 보안 기능이 뛰어남 • 데이터 조작이 편리함	• 데이터의 표현이 유연하지 못함 • 데이터 출처의 스키마 변경시 어려움이 발생 • 어휘 의미에 대해 쉽게 조화를 이루지 못함	GDB, PfamRDB, PRINTS-S, InterPro 등
	객체 지향 데이터베이스	• 추상 데이터 유형으로 인해 실세계 데이터를 가장 유연하게 표현할 수 있음	• 질의에 있어서 절차적인(Procedural) 경향을 지님 • 관계 대수학, 수학적, 논리적 기초 부족으로 컴퓨터 계산 능력이 떨어짐 • 질의어 최적화가 어려움	AGEDB, EcoCyc 등
웹 기반 cross-reference를 이용한 논리적 통합		• 각 엔트리의 구조를 수정하지 않고 두 엔트리의 관련성을 결정하기 용이함 • 관계형 보다 유연성 있고, 제약 사항이 적은 통합으로 매우 실용적임	• 단방향성 • cross-reference 수의 한계성 및 접근시 불편함 • 복잡한 질의 수행 불가능 • cross-reference된 데이터베이스들은 reference를 포함하지 못하는 경우가 존재 • automatic cross-referencing 문제 • 데이터베이스 업데이트 동기화 문제 • 데이터베이스 Mirroring 문제 • 중복된 데이터베이스 엔트리의 handling • network overload	DBGET, Wiss-Prot 등
가상 통합을 이용한 검색 (Federated database system)		• 한번의 접근으로 다중 자원 검색이 가능 • 각각의 데이터베이스 검색보 다 효율적 • 사용자 편의적 인터페이스	• 데이터의 재조직화가 없음 • 검색 데이터베이스의 구조에 제약을 지님 • 각각의 데이터베이스 스키마 변형 등에 민감 • 각 데이터베이스의 안전성에 의존적	SRS, PANAL, Entrez 등

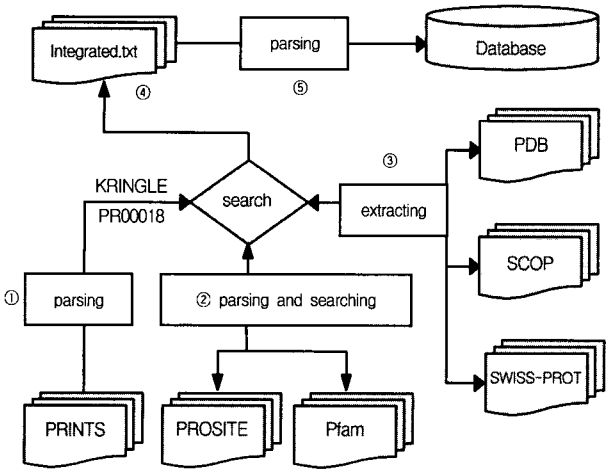
연관성을 지니기 때문이다.

따라서 우리는 이러한 4가지 기준들을 만족하는 모티프 데이터베이스를 관계형 DBMS를 기반으로 구축하였다. 첫째, 두 번째 및 넷째 기준들을 만족시키기 위해 각 모티프 자원들을 하나의 통합 데이터베이스로 재 구축하여 사용자 편의적 측면과 논리적 통합에 따른 문제점 그리고 단백질 3차 구조 지원에 대한 해결책을 제시하였고, 또한 각 통합 데이터베이스에서 사용하는 검색 메소드들을 하나의 예측 시스템으로 통합하였고 옵션 파라미터를 제공하여 세 번째 기준에 대한 해결책을 제시하였다.

4. 모티프 자원 통합을 위한 데이터베이스 설계

4.1 자원 통합 메소드

우리는 기존의 모티프 데이터베이스 통합을 위해 사용된 웹기반 cross-reference의 문제점(즉, 단방향성, 복잡한 질의 처리 불가능 문제, 중복된 데이터베이스 엔트리 handling 등의 문제)를 해결하기 위해 하나의 물리적 통합 데이터베이스에 주목하였다. 따라서 이 논문에서는 이질적 데이터 포멧의 모티프 자원을 하나로 통합하기 위해 PRINTS, Pfam, Prosite 데이터베이스에서 제공하는 각각의 플랫파일을 분석하고, 이를 분해 및 합병하였다. 이러한 과정의 수행은 (그림 1)과 같다.



(그림 1) 모티프 자원 통합 순서

첫째, PRINTS 플랫파일에서 각 ID, Accession number, PROSITE reference, Pfam reference 라인 항목을 파싱한다. 둘째, 파싱한 PROSITE reference와 Pfam reference 항목을 각각 PROSITE, Pfam 플랫파일에서 검색한다. 이때 검색 후 동일한 엔트리가 나타나면 해당 엔트리에서 정보를 추출한다. 셋째, Pfam 플랫파일에 존재하는 PDB와 SCOP에 해당하는 reference 항목을 파싱한 후 그 항목을 이용하여 PDB 플랫파일에서 3차 구조 정보를 추출해 내고 SCOP은 ID만을 추출해 낸다. 넷째, 이렇게 추출된 모든 정보들은 새로운 하나의 플랫파일에 새롭게 저장된다. 마지막으로, 이렇게 생성된 새로운 플랫파일은 다시 파싱과정을 거쳐 데이터베이스에 저장된다. 이러한 과정을 조금 더 자세히 설명하고 구현을 위해 다음과 같은 자원 통합 알고리즘을 설계하였다.

4.1.1 자원 통합 알고리즘 · 플랫파일 분석기

플랫파일 분석기는 크게 두 가지 알고리즘으로 이루어진다. 첫 번째 알고리즘은 맴버 플랫파일들을 비교 분석한 후 필요한 정보를 추출하여 하나의 플랫파일로 재 생성하는 것이고, 두 번째 알고리즘은 비교 분석에 해당되지 않는, 즉 하나의 플랫파일에서만 존재하는 모티프 엔트리에 대한 정보를 추가시키는 것이다. 이를 위해 첫 번째 알고리즘을 수행한 후 자원이 하나로 통합된 summary.txt 파일이 생성되며, 각 플랫파일에서 하나의 새로운 엔트리로 통합되지 않은 엔트리들의 AC와 ID들을 alone_prosite.txt, alone_pfam.txt에 저장하였다. prints.txt는 이미 첫 번째 알고리즘 수행시 summary.txt 파일에 모두 통합한다. 첫 번째 알고리즘은 다음과 같다.

```
입력 : PROSITE, Pfam, PRINTS 플랫파일들 (prints.dat, prosite.dat,
       pfam.-A.seed), 추출할 라인 표기를 'EXT'로 정의
출력 : summary.txt
01 : prints.dat 파일을 연다.
02 : while (fgets (str, 150, fp1) != NULL)
03 : {
04 :    if (strncmp (str, "EXT : ", 4) == 0)
05 :       각 라인에서 통합에 필요한 정보 추출 후에 summary.txt
           에 write 한다.
06 :    else if (라인중에서 PROSITE 엔트리에 대한 ID와 AC를 파
              싱 한다)
07 :    {
08 :       prosite.dat 파일을 연다.
09 :       while (fgets (str1, 150, fp2) != NULL)
10 :       {
11 :          06 라인에서 파싱한 ID와 AC를 prosite.dat에서 검색해
              서 엔트리를 찾아낸다.
12 :          if (strncmp (str1, "EXT : ", 4) == 0)
13 :             추출한 정보를 summary.txt에 write한다.
14 :          prosite.dat 파일을 닫는다.
15 :       }
16 :    }
17 :    else if (라인중에서 PFAM 엔트리에 대한 ID와 AC를 파싱
              한다)
18 :    {
19 :       pfam-A.seed 파일을 연다.
20 :       while (fgets (str2, 150, fp3) != NULL)
21 :       {
22 :          17 라인에서 파싱한 ID와 AC를 pfam.dat에서 검색해
              서 엔트리를 찾아낸다.
23 :          if (strncmp (str2, "EXT : ", 4) == 0)
24 :             추출한 정보를 summary.txt에 write 한다.
25 :          pfam.dat 파일을 닫는다.
26 :       }
27 :    }
```

```
28 :    else if (라인중에서 INTERPRO 엔트리에 대한 AC를 파싱
              하여 summary.txt에 write한다)
29 :  }
30 :  열린 파일을 닫고 종료
```

두 번째 알고리즘은 alone_prosite.txt를 기준으로 위의 알고리즘에서 통합되지 않은 엔트리들만을 검색하여 summary.txt에 추가한다. 또한 alone_pfam.txt 파일 역시 아래의 두 번째 알고리즘을 수행하여 summary.txt에 추가한다.

```
입력 : alone_prosite.txt, prosite.dat, 추출할 라인 표기를 'EXT'로 정의
출력 : summary.txt
01 : alone_prosite.txt 파일을 연다.
02 : while (fgets (str, 80, fp1) != NULL)
03 : {
04 :       prosite.dat 파일을 연다.
05 :       while (fgets (str1, 80, fp2) != NULL)
06 :       {
07 :          if (strcmp (str, str1) == 0)
08 :          {
09 :             if (strncmp (str, "EXT : ", 4) == 0) {
10 :                추출한 정보를 summary.txt 추가로 write 한다.
11 :                prosite.dat 파일을 닫는다. }
12 :          }
13 :       }
14 :       prosite.dat 파일을 닫는다.
15 : }
16 : 열린 파일을 닫고 종료
```

4.1.2 자원 통합 알고리즘 · 연관정보 추출기

위의 플랫파일 분석기를 통하여 생성된 summary.txt 파일을 기준으로 PDB 플랫파일에서 다음 알고리즘을 이용하여 모티프에 대한 3차 구조정보를 추출한다.

```
입력 : summary.txt, pdb 엔트리들, 추출할 라인 표기를 'EXT'로 정의
출력 : result.txt
01 : summary.txt 파일을 연다.
02 : while (한 라인씩 읽는다.)
03 : {
04 :    읽은 라인을 result.txt에 재 기록한다.
05 :    if (PDB 라인이 존재하면)
06 :       PDB ID를 읽어서 pdb 엔트리들 중에서 찾는다.
07 :       SEQRES 라인과 ATOM 라인을 파싱하여 result.txt에
           write 한다.
08 :       열린 pdb 엔트리 파일을 닫는다.
09 :    end if
10 : }
11 : 열린 파일을 닫고 종료
```

4.1.3 자원 통합 알고리즘 · 통합 파서기

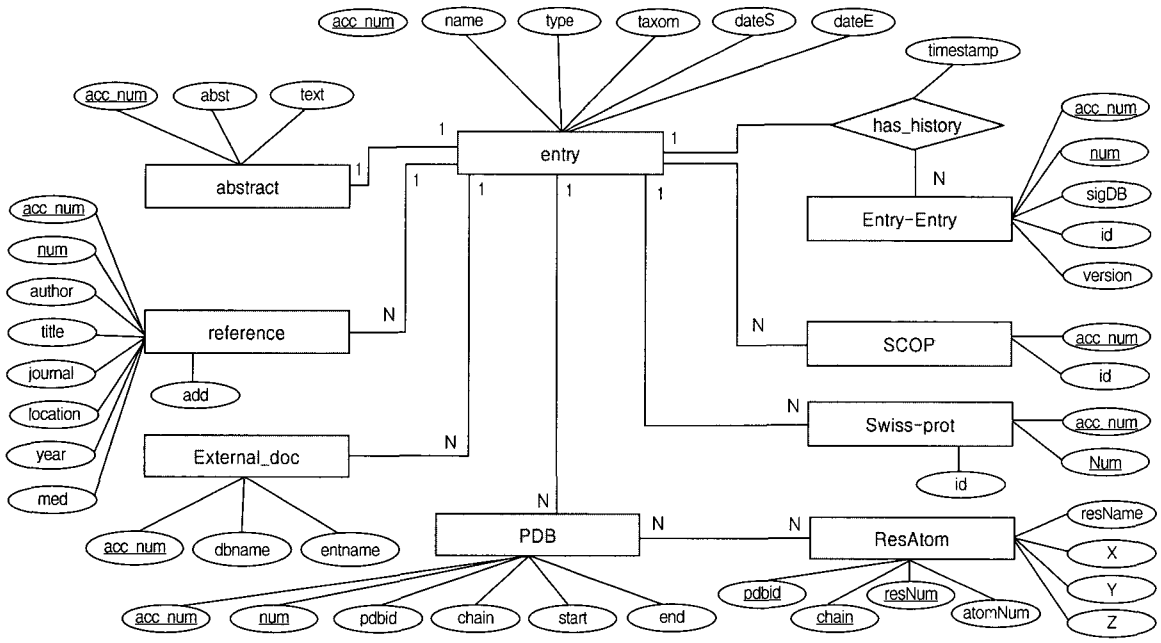
통합 알고리즘은 크게 3단계로 분류되어 진다. 먼저, 플랫파일 분석기는 하나의 동일한 모티프를 각각의 플랫파일에서 이질적인 형태로 저장하고 있는 것들을 검색하고, 일치하는 엔트리들을 추출한 뒤, 하나의 엔트리로 통합하기 위해 각각의 해당 라인에서 통합에 필요한 주요 정보들을 추출해 내어 새로운 플랫파일에 저장한다. 여기서 추출한 주요 정보들은 개체-관계 모델링에 나타내었다. 다음으로 일치하지 않는, 즉 한 데이터베이스 플랫파일에서만 존재하는 엔트리들을 하나의 새로운 엔트리로 재 생성하여 새롭게 생성된 플랫파일에 추가 저장한다. 둘째, 연관 정보 추출기는 통합된 엔트리 각각에 해당하는 모티프 3차 구조 정보를 위해 PDB 데이터베이스에서 제공하는 엔트리 데이터의 Residue 서열들의 시작 위치와 종료 위치 그리고 그 서열들의 Atom 정보에 해당하는 X, Y, Z 구조 정보를 추출하여 새로운 엔트리에 추가하였다. 또한 분류 정보를 위해 SCOP 데이터베이스의 ID와, 샘플 정보를 위한 Swiss-Prot의 ID를 새롭게 생성된 각 모티프 엔트리에 추가하였다. 마지막으로, 아래 기술한 알고리즘으로 구현된 통합 파서기는 이렇게 생성된 통합 플랫파일을 다시 파싱 과정을 거쳐 관계형 데이터베이스에 저장한다.

```
입력 : result.txt
출력 : 데이터베이스 테이블로의 INSERT
01 : Oracle로 connect 한다.
02 : result.txt를 연다.
03 : while (한 라인씩 읽어 들인다)
04 : {
05 :    if (라인 정보 부분을 읽는다)
06 :       라인 정보 부분을 제외한 정보들을 파싱하여 배열에 저장
07 :    if (라인 정보 부분을 읽는다)
08 :       라인 정보 부분을 제외한 정보들을 파싱하여 배열에, 저장
09 :    …
10 :    테이블에 한번에 입력할 정보가 배열에 모두 저장되면
11 :       EXEC SQL INSERT INTO 테이블명(속성명 ,,,)
12 :          VALUES( : 배열명 ,,,)
13 :       …
14 : }
15 : EXEC SQL COMMIT WORK RELEASE Oracle을 Disconnect
     한다.
```

4.2 모티프 자원 통합 데이터베이스 구축을 위한 개체-관계형 다이어 그램

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

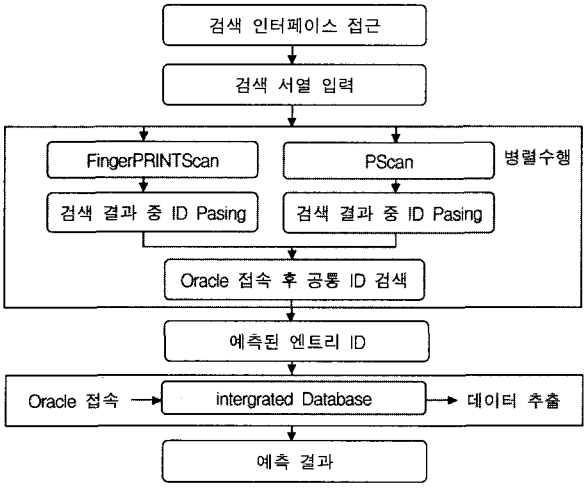
하나의 엔트리는 이름, 타입, taxonomy, abstract, reference, 외부 문서 등에 대한 일반적인 자원들을 위한 엔티티들과, 보다 많은 정보를 지원하기 위해 단백질의 3차 구조정보를 나타내는 PDB와 ResAtom 엔티티, 분류정보를 나타내는 SCOP 엔티티, 샘플정보를 위한 Swiss-prot 엔티티들을 포함하고 있다. 또한, 통합 이전의 각 맴버 데이터베이스의 정보들을 필요로 할 경우와 업데이트를 위해 entry-entry 엔티티(맴버 데이터베이스들이 통합되기 이전 엔트리들의 accession number, 아이디, 버전 정보 등을 보유한 엔티티)와 history 엔티티를 추가하였다.



(그림 2) 통합 데이터베이스 구축을 위한 E-R 다이어 그램

5. 단백질 모티프 예측을 위한 검색 시스템 설계

서열 검색 메소드로 가장 널리 쓰이고 있는 것은 BLAST와 FASTA이다. 그러나 이러한 메소드들 보다 높은 민감도(Sensitivity)와 신뢰도를 제공하는 패밀리 기반 메소드들인 HMMER, FingerPRINTScan, PScan 메소드는 각각 자신들만의 특별한 장점을 가지고 있다. Pfam은 다소 먼 패밀리 맴버 발견에서 그 성능이 우수하고, FingerPRINTScan은 구별하기 어려운 서브패밀리 관계를 식별하기 위해 디자인되었다[2]. 그러므로, 이러한 상보성을 고려하여 각각의 메소드들을 모두 이용하여 중요한 서열을 스켄하는 것이 최선이다[6]. 따라서 우리는 단백질 예측을 위한 서열 검색 메소드들 중에서 FingerPRINTScan과 PScan을 하나의 검색 시스템 속으로 모듈화 하여 통합 검색이 가능하도록 다자인하였다.



(그림 3) 단백질 예측 수행 순서

이러한 통합 검색 메소드에 사용되는 입력 양식으로, FingerPRINTScan과 PScan은 모두 FASTA 포맷을 사용하며, 검색에 이용되는 데이터베이스는 FingerPRINTScan의 경우 prints27_0.pval_blos62를 사용했으며 PScan의 경우 패키지에서 제공하는 pfscan을 사용한다. 또한 검색 내용으로는 각 검색 메소드의 결과를 각각 새로운 파일에 저장하여 그 파일에서 검색된 엔트리 ID만을 추출한다. 이러한 검색 시스템의 수행과정을 (그림 3)과 같이 나타내었다.

첫째, 사용자는 데이터베이스 인터페이스에 접근하여 단백질 서열과 E-value cutoff 등의 파라미터 정보를 입력한다. 둘째, 사용자가 입력한 정보들은 Sun Solaris상에 설치한 standalone 버전으로 구동되는 FingerPRINTScan, PScan 검색 프로그램들에서 병렬적으로 수행한다. 셋째, 이때 각 메소드의 검색 결과에서 검색된 ID만을 파싱과정을 통하여 추출한다. 이렇게 FingerPRINTScan과 PScan에서 추출된 ID는 크게 4가지의 경우의 수로 나뉠 수 있다. ① 한 개의 검색 메소드에서 1개의 ID가 추출될 경우는 그 검색 메소드에서 추출된 ID를 다음 단계로 넘겨주며, 만일 한 개의 검색 메소드에서만 1개 이상의 ID가 추출될 경우 결과 창에서는 첫 번째 ID에 해당하는 정보를 보여주고 이후에 추가 ID를 보여준다. ② 각각의 검색 메소드에서 ID가 추출되지 않을 경우는 결국 어떠한 예측 결과도 제공할 수 없다. ③ 각각의 검색 메소드에서 1개 이상의 ID가 추출될 경우, 오라클에 저장된 데이터베이스 ENTRY-ENTRY 테이블 중 SigDB 속성과 비교하여 추출된 각 ID를 공통으로 가지고 있는 accnum를 선택하여 선택된 accnum에 해당하

는 정보들을 예측 결과로 나타낸다. 이때 공통 ID가 추출되지 않을 경우, 다시 말해 한 accnum에 해당하는 SigDB 속성에 추출된 ID들이 같이 속해 있지 않을 경우 FingerPRINTScan에서 추출된 ID를 우선시 하여 다음 단계로 넘겨준다. ④ 또한 각각의 검색 메소드에서 1개씩의 ID만이 추출될 경우 ③번과 같이 FingerPRINTScan에서 추출된 ID를 우선시하여 다음 단계로 넘겨준다. 넷째, 세 번째에서 추출된 ID를 이용하여 다시 ENTRY-ENTRY 테이블에서 통합 엔트리 ACCESSION 넘버를 검색하여 그 넘버에 해당하는 정보들을 예측 결과 창에 나타낸다. 따라서 하나의 데이터베이스로 통합된 정보를 이용함에 따라 기존의 데이터베이스들에서 제공하지 못했던 모티프 3차 구조정보, 분류정보 등의 제공이 가능하다. 또한 SQL 문을 직접 기술하여 통합된 자료를 검색할 수 있으므로 기존의 웹 기반 통합 검색에서 나타나는 복잡한 질의 처리문제의 해결과, 모티프 자원 통합시 중복된 엔트리들의 단일화를 통하여 중복된 데이터베이스 엔트리 핸들링 문제를 해결할 수 있다.

6. 구현 및 평가

우리는 각 모티프 데이터베이스에서 제공하는 플랫파일들을 하나로 통합된 새로운 플랫파일로 생성하기 위해 Window 2000 환경에서 C언어를 사용하였으며, 통합된 플랫파일을 오라클 데이터베이스에 삽입하기 위해 ProC언어를 사용하였다. 또한 시스템 기종으로는 Sun사의 Enterprise 250을 사용하였으며, 운영체제로는 Sun Solaris 7(5.7), DBMS로는 Oracle 8i를 이용하였다.

6.1 모티프 자원 통합

모티프 자원 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

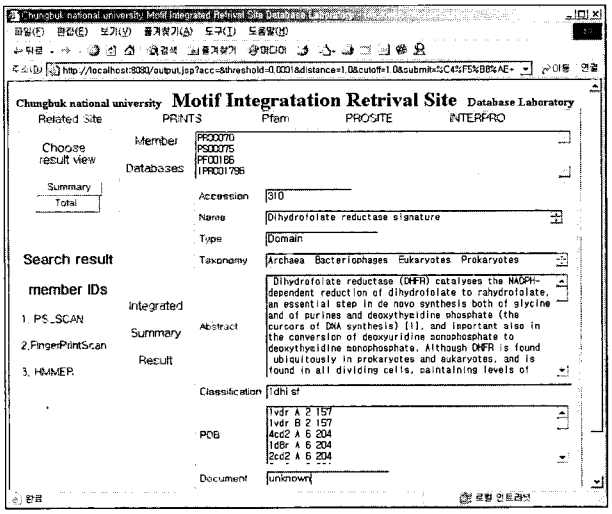
이러한 엔트리들은 플랫파일 분석기, 연관정보 추출기, 통합 파서기를 이용하여 4장에서 기술한 통합과정을 거쳐 5,670개의 새로운 엔트리로 재구성하였다.

6.2 서열 검색 인터페이스

사용자는 이 입력 인터페이스를 통해 검색하고자 하는 단백질 서열을 입력하고, 검색 모듈들을 선택한다. 이때 선택된 검색 모듈에 따른 파라미터 값들을 추가로 입력하여야 한다. 다만 PScan에서는 검색 데이터 구조 Pattern, Rule, Profile에 대해 모두 검색할 것인지 여부를 체크하여야 한다. 정확한 데이터일수록 선택 범위를 적게 하면 정확하고 빠르게 검색 결과를 알 수 있다.

단백질 서열 검색시 통계적 중요도 나타내는 E-value 값은 사용자의 목적에 따라 다양하게 설정할 수 있다. 단백질 예측시 Remote homology 검색을 위해서는 E-value 값을 좀 더 크게 조정하면 되지만 False positive 배율이 증가하는 것을 감수해야 한다. 또한 유사도가 높은 수준의 검색을 위해서는 E-value 값을 낮게 설정하면 되지만 False negative 비율이 증가하는 것을 감수해야 한다. 이렇게 E-value 값의 설정은 그 설정에 따라 상보적인 관계를 지니고 있다. 우리의 검색 시스템에서는 E-value의 디폴트 값으로 1.0을 적용하였다.

검색 시스템의 예를 위해 우리는 이미 그 기능과 구조가 알려진 서열 Dihydrofolate reductase signature를 예측을 위한 검색 서열로 사용하였다. 또한 검색 조건은 FingerPRINTScan과 PScan을 선택하였고 검색 파라미터들은 위에 기술한 디폴트를 그대로 사용하였다. 예측 결과로 (그림 4)와 같이 통합 정보를 나타냈다. 첫째로, 이러한 Dihydrofolate reductase signature를 포함하고 있는 각 맴버 데이터베이스들의 ID(즉, PR00070, PS00075, PF00186 등)를 알 수 있다. 둘째, 이러한 모티프에 대한 일반적인 정보들인 이름, 타입, 요약 등을 나타낸다. 마지막으로, Dihydrofolate reductase signature에 해당하는 기존에 알려져 있는 3차 구조정보와 분류 정보를 제공한다.



(그림 4) 단백질 예측 결과 출력 인터페이스

6.3 사례 데이터베이스와의 비교 분석

마지막으로 우리는 사례 데이터베이스들과의 비교 분석을 통하여 우리가 구축한 데이터베이스 및 검색 시스템을 평가한다. 비교 데이터베이스 및 검색 시스템 버전은 InterPro Release 5.2(2002.9)와 PRINTS version 35(2002.7)를 기준으로 하였다.

〈표 2〉 통합 방법론에 따른 생물학 데이터베이스 비교

비 교		물리적 통합 데이터베이스		웹 기반 통합 데이터베이스	웹 기반 통합 검색	
		제안한 통합 시스템	InterPro	PRINTS	SRS	PANAL
질의 처리	SQL 질의	가능	가능	가능	불가	불가
	서열 질의	가능	가능	가능	가능	가능
	서열질의 파라미터 지원	가능	불가	가능	불가	가능
통합 자원	총 엔트리 수	5670	5876	1750	-	-
	3차 구조 정보	가능	불가	불가	가능	불가
	모티프 분류 정보	가능	불가	불가	가능	불가
	모티프 서열 지원	불가	불가	불가	불가	불가
브라우징	중복 엔트리 처리	가능	가능	가능	불가	불가
	요약 결과	가능	가능	가능	가능	가능
	그래픽 결과	불가	가능	가능	가능	가능
비 교			EBI	Manchester		Minnesota

<표 2>에서 모티프 분류 정보는 SCOP(Structural Classification Of Proteins)의 정보 지원, 즉 분류 정보 지원 여부에 대한 내용을 제공하는가 하지 않는가에 대해 "가능" 또는 "불가"로 표기하였다. 또한 모티프 서열 지원 역시 기존의 모티프 데이터베이스 검색에서 모티프 시퀀스의 직접적 지원(하이퍼 링크를 제외한) 여부에 대한 내용으로써 기존 데이터베이스에서 제공을 하는지 하지 않는지에 대해 "가능"과 "불가"로 표기하였다. 예를 들어, 모티프 분류 정보에 있어서 우리가 제안한 통합 시스템은 SCOP과 같은 분류(classification) 정보를 하나의 데이터베이스에 포함하고 있으므로 검색 결과에 따라 그 모티프에 해당하는 분류 정보를 제공하는 것이 가능하지만, InterPro, PRINTS 데이터베이스에는 분류정보를 포함하고 있지 않기 때문에 제공할 수 없다. 또한 SRS 검색은 매우 광대한 검색 조건을 제공하고 있기 때문에 SCOP과 같은 분류 정보에 대한 검색을 제공하지만, PANAL과 같이 모티프 검색과 같이 특화된 검색 엔진에서는 모티프 분류 정보를 제공하지 않는다.

<표 2>에서 알 수 있듯이 질의 처리 항목에서 기존 웹 기반 통합 검색보다 물리적인 통합 데이터베이스들이 좀 더 나은 기능을 지원하고 있으며, 통합 자원 측면에서 웹 기반 통합 데이터베이스 보다 물리적 데이터베이스들 자원이 풍부한 것을 알 수 있다. 그러나 브라우징 측면에서는 웹 기반 통합 데이터베이스와 물리적 통합 데이터베이스(InterPro)가 가장 우수한 것으로 평가된다.

InterPro 데이터베이스와 우리가 구축한 데이터베이스를 비교하여 볼 때, 첫째, InterPro는 질의 처리에 사용되는 단백질 예측 시스템에서 총 5가지 메소드들(ScanRegExp, PScan, FingerPRINTScan, HMMER, BlastProdom)을 사용한다. 따라서 우리가 구축한 예측 시스템보다 그 수가 많기 때문에 좀더 포괄적인 검색이 가능하지만 우리가 구축한 검색 메소드와 같이 예측 메소드들에 대한 검색 옵션 파라미터를 지원하지 못한다. 따라서 InterPro 단백질 예측 시스템에서는 사용자 자신이 원하는 특정 E-value 값 등을 기술할 수 없다. 통합 자원 특면에서 볼 때, 총 엔트리수는 InterPro가 206개 더 많은 것을 볼 수 있지만, 우리가 구축한 데이터베이스와 같이 엔트리 3차 구조 정보와 분류 정보를 지원하지 못한다. 마지막으로 브라우징 측면에서 단백질 예측 검색 후 InterPro는 우리가 구축한 검색 시스템에서 제공하지 못하는 그래픽적 형태의 결과 창을 나타낼 수 있다.

7. 결 론

이 논문에서는 각각의 고유한 데이터 형식과 검색 메소드들을 사용하여 성장해온 모티프 데이터베이스들에서 나타나는 이질적인 예측 결과 문제와 웹 기반 cross-reference 통합에 따른 복잡한 질의처리, 중복된 데이터베이스 엔트리 핸들링 문제들을 해결하기 위해 모티프 자원들에 대한 물리적 통합 연구에 대하여 다루었다. 따라서 모티프의 Annotation 정보와 3차 구조 정보 및 분류 정보를 통합하여 하나의 DBMS에 저장하였으므로 데이터의 효율적 관리와 폭발적으로 증가하는 단백질 원시 데이터에 대한 통합 예측 검색을 가능케 하였다. 우리는 이러한 통합 예측 검색을 가능케 하기 위해 아래와 같은 과정을 수행하였다.

- 각 모티프 데이터베이스들에서 제공하는 플랫파일을 분석
- 각 엔트리에 대한 모티프 3차 구조 정보와 분류 정보 통합
- 개체-관계 모델링을 통해 Oracle 8i를 사용하여 통합 데이터베이스 구축
- 각 검색 메소드들을 통합한 검색 시스템 구축과 Web 기반 인터페이스 구현

따라서, 웹 기반 통합에 따른 복잡한 질의 처리 문제, 중복된 데이터베이스들의 핸들링 문제, 기존의 데이터베이스 검색시 사용자가 겪는 이질적 검색환경 및 반복 접근 문제

를 해결하였고 기존의 웹 기반 통합 검색에서 지원하지 못했던 단백질의 3차 구조정보, 분류 정보, 샘플 정보의 지원을 가능케 하였다. 또한 우리가 설계하고 구현한 모티프 예측시스템은 DBMS를 기반으로 하였기 때문에 2차 데이터베이스의 구축 등 데이터의 조작을 수용하였기 때문에 사용자의 편의를 도모하였다.

향후 연구로는, 이 연구에서 제외되었던 다양한 모티프 자원들 즉, BLOCKS, ProDom, SMART, eMOTIF의 모티프 자원 통합과, 분야별로 구축되어진 데이터베이스에 대한 표준화 연구가 필요하다.

참 고 문 헌

[1] 김성진, 이상호, "객체-관계형 데이터베이스 시스템을 위한 새로운 성능 평가 방법론", 정보처리학회논문지, 제7권 제7호, 2000.

[2] 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, pp.37-40, 2001.

[3] M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser, "Proteome Research : New Frontiers in Functional Genomics," Springer-Verlag Berlin Heidelberg, pp.109-175, 1997.

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

[5] David W. Mount, "Bioinformatics : Sequence and Genome Analysis," Cold Spring Harbor Laboratory Press, pp.45-48, 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, pp.1157-1158, 2000.

[7] 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, pp.304-308, 1998.

[8] 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, pp.276-280, 2002.

[9] Jorja G. Henikoff, Steven Henikoff, Shmuel Pietrokovski, "New features of the Block Database servers," Nucleic Acids Research, Vol.27, No.1, pp.226-228, 1999.

[10] 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, pp.417-424, 1997.

[11] 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, pp.235-238, 2002.

[12] 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, pp.235-242, 2000.

[13] Etzold T., Ulyanov A., Argos P., "SRS : information retrieval system for molecular biology data banks," Methods Enzymol, pp.114-128, 1996.

[14] Ramez Elmasri, Shamkant B. Navathe, "Fundamentals of Database Systems," Addison-Wesley, Reading, Massachusetts, 2000.

[15] Philip Scordis, Darren R. Flower, Teresa K. Attwood, "FingerPRINTScan : intellegent searching of the PRINTS motif database," Bioinformatics, Vol.15, No.10, pp.799-806, 1999.

[16] T. K. Attwood, M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, P. Scordis, "PRINTS and PRINTS-S shed light on protein ancestry," Nucleic Acids Research, Vol.30, No.1, pp.239-241, 2002.

[17] Philipp Bucher, Kevin Karplus, Nicolas Moeri, Kay Hofmann, "A Flexible Motif Search Technique Based on Generalized Profiles," Comput. Chem., Vol.20, pp.3-24, 1996.

[18] Doug Brutlag, "Protein Structure & Motifs," Biochemistry 201, Molecular Biology, 2000.

[19] Cynthia Gibas, Per Jambeck, "Developing Bioinformatics Computer Skills," O'REILLY, pp.290-295, 2001.

[20] Attwood, "The Babel of Bioinformatics," Science 290, pp. 471-473, 2000.

[21] Florence Corpet, Florence Servant, Jerome Gouzy and Daniel Kahn, "ProDom and ProDom-CG : tools for protein domain analysis and whole genome comparisons," Nucleic Acids Research, Vol.28, No.1, pp.267-269, 2000.

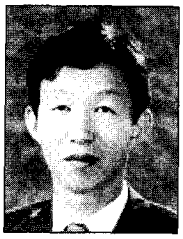
[22] Barbara Eckman, Julia Rice, Bill Swope, "Heterogeneous Data and Algorithm Integration in Bioinformatics," ISMB, 10th International Conference Tutorial, 2002.

이 범 주

e-mail : bjlee@dblab.chungbuk.ac.kr
1997년 충청대학 졸업(공업전문학사)
2001년 서원대학교 졸업(이학학사)
2003년 충북대학교 대학원 전자계산학과 (이학석사)
관심분야 : Bioinformatics, 시공간 데이터베이스, 데이터마이닝, XML 등

최 은 선

e-mail : eschoi@dblab.chungbuk.ac.kr
1993년 충북대학교 전자계산학과(이학사)
1999년 충북대학교 교육대학원 전자계산 교육전공(교육학석사)
2000년 한국생명공학연구원 위촉연구원
2002년 충북대학교 대학원 전자계산학과 박사과정 수료
관심분야 : Bioinformatics, 데이터마이닝, 시공간 데이터베이스, XML 등



류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr
1976년 숭실대학교 전산학과(공학사)
1980년 연세대학교 공학대학원 전산전공 (공학석사)
1988년 연세대학교 대학원 전산전공 (공학박사)
1976년~1986년 육군군수지원사전산실(ROTC 장교), 한국전자통신연구소(연구원), 한국방송통신대 전산학과(조교수) 근무
1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)
1986년~현재 충북대학교 전기전자 및 컴퓨터공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식베이스 시스템, 지식기반 정보검색 시스템, 데이터마이닝, 데이터베이스 보안 및 Bio-Informatics 등