단백질 모티프 예측 및 갱신 프로토 타입 구현

노 기 용[†] · 김 원 식[†] · 이 범 주[††] · 이 상 태[†††] · 류 근 호[††††]

요 약

모티프 데이터베이스는 새롭게 등장하는 원시 단백질 서열의 기능 및 구조 예측에 사용된다. 이러한 모티프 데이터베이스들은 원시 단백질 서열의 빠른 성장과 더불어 급속한 이용 증가 추세를 보이고 있으며, 최근에 이르러 모티프 자원 통합에 관한 연구가 진행되고 있다. 그러나 이러한 모티프 데이터베이스들은 각기 개별적인 메소드로 개발되었기 때문에 각기 다른 형식의 검색 결과를 제공한다. 이러한 문제 해결을 위한 데이터베이스 통합에서는 데이터베이스 자동 갱신 문제, 복잡한 질의 처리 문제, 중복된 데이터베이스 엔트리 핸들링 문제, XML 지원 문제 등을 지니고 있다. 이 논문에서는 기존 문제점들을 해결하기 위하여 데이터베이스 자원 통합 방법론을 제안하였고, 통합된 데이터베이스의 주기적 갱신 방안과 XML로의 변환에 관하여 기술하였다. 아울러 구축된 통합 데이터베이스와 사례 데이터베이스를 비교 평가하였다.

Implementation of Prototype for a Protein Motif Prediction and Update

Gi Young Noh[†] · Wuon Shik Kim[†] · Bum Ju Lee[††] · Sang Tae Lee[†††] · Keun Ho Ryu[††††]

ABSTRACT

Motif databases are used in the function and structure prediction of proteins. The frequency of use about these databases increases continuously because of protein sequence data growth. Recently, many researches about motif resource integration are proceeding. However, existing motif databases were developed independently, thus these databases have a heterogeneous search result problem. Database intnegration for this problem resolution has a periodic update problem, a complex query process problem, a duplicate database entry handling problem and XML support problem. Therefore, in this paper, we suppose a database resource integration method for these problem resolution, describe periodically integrated database update method and XML transformation. Finally, we estimate the implementation of our prototype and a case database.

키워드 : 모티프(Motif), 모티프 자원 통합(Motif Resource Integration), 단백질 예측(Protein Prediction),XML, 바이오인포메틱스(Bioinformatics)

1. 서 론

생물학 기반의 BT와 컴퓨터 기반 IT 연구의 결합으로 새롭게 탄생한 생물 정보학(bioinformatics)의 출현은 생물학 연구에 새로운 패러다임을 가져왔다. 대량의 생물학적 정보를 효율적으로 저장, 분석하는 방법에서부터 더 나아가 다양한 생물학적 단위 정보들 간의 유기적 연동 기제를 파악하고 생물학적 의미를 통합적으로 파악하려는 시도와 더불어, 단순한 개개 정보의 통합으로 나타나는 분석 결과 이상의 생물학적 해석을 지향하고 있다[20]. 다양한 생물학적 자원들 중의 하나인 모티프 데이터베이스는 단백질 아미노산 서열과 3차 구조 정보 사이의 연관 관계를 이용하여 새로이 등장한 단백질의 기능 예측에 사용된다[6, 8, 18]. 기존까지 개발된 모티프 데이터베이스들로는 InterPro[3], ProDom[22], BLOCKS[6], PROSITE[12], PRINTS[8, 11], Pfam [9] 데이터베이스 등 매우 다양한 모티프 데이터베이스들이 개별적으로 생성되었다. 이렇게 다양한 데이터베이스들은 사용자들에 대하여 여러번의 중복 검색 문제들과 검색 결과에 대한 이질적 포맷 문제 등을 지니고 있었다[7]. 이러한 문제 해결을 위하여 연구자들은 이질적인 데이터 구조로 생성된 여러 모티프 데이터베이스들의 통합 및 검색을 위해 InterPro, BIOKLEISLI, SRS, PANAL과 같은 웹 기반 Cross-reference를 이용한 접근법이 시도되었다[4-5]. 그러나 이러한 접근법은 데이터 구조를 변경하지 않고 관련된 엔트리 간에 유연한 통합을 지원할 수 있는 장점에 비해,

※ 이 연구는 2003년도 한국과학기술기획평가원의 연구비 지원으로 수행되었음.
† 정 회 원 : 한국표준과학연구원
†† 준 회 원 : 충북대학교 대학원 전자계산학과
††† 정 회 원 : 한국표준과학연구원
†††† 종신회원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
논문접수 : 2003년 10월 24일, 심사완료 : 2004년 2월 2일

정기적 업데이트시 중복 엔트리에 대한 처리문제, 엔트리의 삭제시 링크 처리 문제, 복잡한 질의 처리 문제, 엔트리 통합후의 데이터 표준화 문제, 네트웍 과부화 등과 같은 문제점들을 지니고 있다[4].

우리는 위에 기술한 문제들에 대한 해결 방안을 위하여 기존 논문[2]에서, 멤버 데이터베이스들의 단백질 모티프 자원들에 해당하는 Annotation 정보, 3차 구조 정보[13] 및 분류 정보(SCOP, Structural Classification Of Proteins) 통합 저장과, 통합 검색 메소드를 확장 설계를 제시하였다. 아울러 이 논문에서는 기존의 제안에서 제공치 못했던 정기적 업데이트시 중복 엔트리의 처리문제, 엔트리 삭제시 링크의 처리문제, 엔트리 통합후의 표준화 문제를 해결하기 위하여 다음과 같이 확장 설계하였다. 첫째, 멤버 데이터베이스에서 제공하는 갱신 파일을 이용한 모티프 자원들의 주기적 갱신 모듈들을 설계하여 업데이트시에 발생하는 문제점들을 해결한다. 둘째, 통합 후 데이터 표준화 문제를 해결하기 위하여 통합 데이터를 XML 형식으로 변환할 수 있는 모듈을 구현한다.

따라서, 기존의 통합에서 제공하는 복잡한 질의 처리 지원, 중복된 데이터베이스들의 핸들링, 모티프 3차 구조정보 및 분류정보 지원, 통합 검색 지원 뿐만 아니라 맴버 데이터베이스 갱신에 따른 통합 데이터베이스의 정기적 갱신 지원, 정기적 업데이트시의 중복 엔트리에 대한 처리 지원, 그리고 통합된 모티프 데이터 표준화를 위한 XML 형식의 지원을 가능케 하였다.

이 논문의 구성은 다음과 같다. 2장에서는 현재 주로 사용되어지고 있는 모티프 데이터베이스들과 생물 정보 데이터베이스들에 대한 통합 방법론에 대하여 기술한다. 3장에서는 모티프 자원들에 대한 통합 및 갱신 메소드와 데이터베이스 구현을 위한 시스템 구조에 대하여 기술하고, 4장에서는 단백질 모티프 예측을 위한 검색 메소드에 대하여 다룬다. 5장에서는 프로토 타입 구현과 사례 데이터베이스와의 비교 평가에 대하여 기술하고, 마지막으로 6장에서 결론을 맺는다.

2. 모티프 데이터베이스 및 선행된 생물 정보 자원 통합 방법 검토

이 장에서는 기존의 모티프 데이터베이스들에 대한 개략적인 정보들과 선행 연구된 생물 정보 자원들에 대하여 검토한다.

2.1 모티프 데이터베이스 및 검색 시스템

기존의 모티프 데이터베이스들은 InterPro, ProDom, BLOCKS, PROSITE, Pfam, PRINTS 등 여러 종류의 데이터베이스가 존재하며 데이터베이스 자체적으로 다양한 데이터 생성방식 및 특징들을 지니고 있다. 이러한 모티프 데이터베이스들은 단백질 서열 분석 전략에서 표준 툴 역할을 하며 자동화된 분석에 의존하지만 대부분 전문 관리자와 생물학자의 수작업이 데이터베이스 구축에 투입된다. 이러한 모티프 데이터베이스의 특징 및 버전에 관한 정보는 다음과 같다.

• InterPro 데이터베이스

단백질 패밀리, 도메인, functional site들에 대한 물리적 통합 문서 자원을 목적으로 생성된 InterPro 데이터베이스는 PRINTS, PROSITE, Pfam, ProDom과 같은 시그네처 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. 이러한 InterPro 데이터베이스는 최초의 물리적 통합에 따른 streamline적 검색의 시도와 모티프를 이용한 단백질 서열 검색시 폭 넓은 coverage를 제공하는 장점을 지니고 있으나 모티프의 기능과 구조적 측면에서 아직까지 모티프 3차 구조 정보와 분류정보가 통합되어 있지 않다.

이 데이터베이스 버전 5.1은 1239개의 도메인, 4280개의 패밀리, 95개의 repeat, 15개의 PTM 사이트들로 구성된 총 5629개의 엔트리를 포함하고 있으며, 웹상에서 엔트리 데이터와 매치 데이터를 XML 형식으로 배포하고 있다[3, 6].

• ProDom 데이터베이스

ProDom 데이터베이스의 목적은 유용한 단백질 서열 데이터의 자동화 분석 원리로 상동 도메인들의 패밀리를 분류학적으로 수집하는 것이며, 자동 서열 비교들에 의해서 생성된 SWISS-PROT 데이터베이스로부터 생성된 단백질 도메인 패밀리들을 포함하고 있다. 사용자와 상호 작용이 가능한 그래픽 인터페이스는 도식적인 도메인 배열, 다중 할당들, SWISS-PROT 엔트리들, PROSITE 패턴들, PDB의 3-D 구조들 사이에 손쉬운 항해가 가능하다. 또한 ProDom 데이터의 질을 개선하기 위해, 데이터 생성 절차에서 두 가지 방안을 채택하였다. 첫째, 도메인 경계의 정확도를 개선하기 위해 도메인 패밀리들을 확인하는 몇몇 전문가를 포함시켰으며, 둘째, PSI-BLAST 유사성 검색 알고리즘의 특징을 사용하여 ProDom을 생성하는 데 사용하는 MKDOM의 새로운 버전의 프로그램을 개발하였다[22]. 릴리이즈 ProDom-CG CG47에서 182,217개의 도메인 패밀리를 포함하고 있다. 그러나 ProDom, BLOCKS, PRINTS, Prosite, Pfam 등의 모티프 데이터베이스들 모두 각자 고유한 메소드를 활용하여 데이터베이스 생성 및 성장하여 왔으므로 각기 이질적인 데이

터 형식으로 제작되어 있다. 따라서 기존의 모티프 데이터베이스 통합에 대한 연구들이 진행중이다.

• BLOCKS 데이터베이스

Fred Hutchinson Cancer Research Center에서 제공하는 BLOCKS 데이터베이스는 단백질의 패밀리 분류를 돕기 위해 시작되었으며, 문서화된 단백질 패밀리들의 block들로 구성되어 있다. BLOCKS 데이터베이스는 씨앗 서열 정렬 부위(seed alignment)를 발견하기 위하여 서열에 있는 3개씩의 아미노산 공간을 세세하게 검색하고 이어, 최대 길이의 정렬 부위를 찾아서 서열을 확장하는 모티프 검출 방법의 조합으로 만들어진다. 이러한 메소드에 대한 입력으로 PROSITE 패밀리, 사용자 서열, 사용자 할당, PRINTS의 자원을 받아들여, 검색을 위해 서열 대 서열, 서열 대 블록, 블록 대 서열, 블록 대 블록을 제공하며, 디스플레이를 위해 logos, trees maps structure를 나타내며 디자인을 위해 PCR primers를 제공한다. 버전 11.0에는 994개의 단백질 패밀리들을 표현하는 4,034개의 block들을 포함하고 있다[6].

• PROSITE 데이터베이스

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, ScanProsite 와 같은 신뢰성 있는 툴들을 사용하여 생성한다[12]. 현재 릴리이즈 버전은 17.21에서는 1,568개의 패턴, 룰, 프로파일/메트릭스들을 저장하고 있다.

• PRINTS 데이터베이스

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[8, 11, 16]. 최근에 PRINTS-S라 불리는 관계형 DBMS로 저장소를 확대했으며, 버전 35.0에서는 1,750개의 엔트리들이 저장되어 있다.

• PANAL 검색 시스템

모티프 데이터베이스들은 사용자 편의적 관점에서 중복 접근, 이질적 검색 결과 등의 문제점 들을 내포하고 있다[21]. 따라서 사용자 편의를 위한 모티프 자원 검색을 위해 하나의 자원으로 통합되어져야 한다. 이러한 명제 하에 제작된 모티프 통합 검색 시스템으로 PANAL(Protein Analysis Application)을 예로 들 수 있다.

Minnesota 대학의 computational Biology Centers에서 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 정보창을 사용자에게 제공한다[7]. 그러나 이러한 검색 시스템은 기존 모티프 데이터베이스들에만 한정된 결과를 제공하므로 모티프 3차 구조 정보 및 분류 정보의 지원을 하지 못하고 있다.

2.2 선행된 생물정보 자원 통합 방법 검토

최근 사용자 편의적 접근을 위한 생물 정보 자원의 통합에 대한 연구가 진행되고 있다. 이것은 사용자들에게 보다 편리한 통합 자원의 제공과, 자원들의 효율적 관리 측면에서 중요한 의미를 가진다. 지금까지 생물 정보 자원에 대한 통합 검색 지원 방법은 다음과 같다.

• 생물정보 자원의 물리적 통합

여러 생물 데이터베이스 자원들을 물리적으로 하나의 데이터베이스로 통합하는 것이다. 이러한 통합은 다시 관계형 데이터베이스와 객체 지향적 데이터베이스로 나뉠 수 있으며, 엔티티 정보의 오류나 불일치에 대한 데이터 무결성을 보장하고, 회복, 보안 기능이 우수하며, 실세계 데이터를 가장 유연하게 표현할 수 있고, 사용자 측면과 관리적 측면에서 효율적인 장점[4, 15]을 지니고 있는 반면에 데이터 표현, 데이터 출처의 스키마 변경시의 수정문제, 어휘 의미 등에 대한 문제점을 지니고 있다. 이러한 통합을 목적으로 제작된 데이터베이스는 관계형 데이터베이스로는 GDB, P-famRDB, PRINTS-S, InterPro 등이 있으며, 객체 지향 데이터베이스로는 AGEDB, EcoCyc 등이 있다.

• 웹 기반 cross-reference를 이용한 논리적 통합

생물 정보 데이터베이스 통합에 있어 현재까지 가장 널

리 사용되는 방법[4]으로서 각각의 엔트리 구조를 수정하지 않고 두 엔트리간의 관련성을 결정하기 쉽고, 물리적 통합보다 유연성이 있으며 제약사항이 적은 실용적 통합이 가능하다는 장점을 지니고 있는 반면에 복잡한 질의 수행이 불가능하고, 각각의 데이터베이스 업데이트시 cross-reference를 끊어지는 문제가 발생하며, cross-reference 수의 한계성 문제가 발생한다. 현재 DBGET, Wiss-Prot 등이 이러한 논리적 통합 방법론을 이용하고 있다.

• 가상 통합을 이용한 검색

각 생물 정보 데이터베이스에 다중 접근하여 정보를 추출해 오는 가상적인 통합 검색(federated database system)으로써, 한번의 접근으로 다중 자원 검색이 가능하고, 사용자 편의적 인터페이스를 지니고 있는 장점이 있으나 검색할 데이터베이스의 구조에 의한 제약사항을 지니고, 각 데이터베이스의 스키마 변형 등에 민감하며, 각 데이터베이스의 안정성에 매우 의존적인 단점들을 지니고 있다. SRS[14], PANAL[7], Entrez[6] 등이 이러한 가상 통합을 이용한 검색 방법을 채택하고 있다.

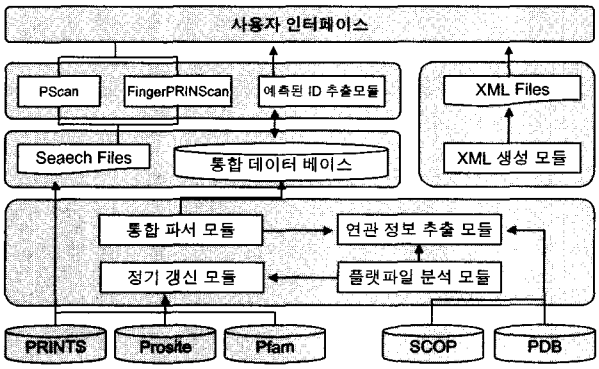
이러한 검색 방법들을 [1, 4, 5, 23]에서 요약하여 [2]에 자세히 기술하였다.

우리는 기존의 통합 방법 분석의 검토와 기존 모티프 데이터베이스들에 대한 문제점들의 분석을 통하여 사용자 편의적 접근법 제공, 하나의 물리적 데이터베이스로 통합, 실제적인 통합 검색 가능, 기존 통합 자원들보다 더 많은 자원의 통합, 각각의 데이터베이스 업데이트시 동등한 엔트리의 보유를 위한 갱신 모듈 지원, 통합된 데이터의 재 배포 및 표준화를 위한 XML 형식 지원이 가능한 연구를 진행하였다.

따라서 이러한 6가지 기준들을 만족하기 위한 통합 모티프 데이터베이스 구축과, 구축된 검색 시스템을 이용하여 사용자 편의적 측면과 논리적 통합에 따른 문제점 그리고 단백질 3차 구조 지원에 대한 대안을 제시한 논문 [2]의 확장 설계를 통하여, 맴버 데이터베이스 갱신 정보를 이용하여 통합된 데이터베이스의 업데이트가 가능하도록 각각의 데이터베이스 갱신 메소드들을 구현한다. 또한, 통합된 데이터들의 이질적 형식을 XML 형식으로의 재 변환을 통하여 타 데이터베이스와의 표준화를 가능케 하였다.

3. 모티프 데이터베이스 자원 통합 및 갱신 모듈

(그림 1)은 프로토 타입의 전체 구조도이다. 이 시스템은 크게 6개의 모듈로 구성된다. '정기 갱신 모듈', '플랫파일 분석 모듈', '연관정보 추출모듈'들은 각각의 맴버 데이터베이스에서 제공하는 플랫파일의 엔트리들을 통합 데이터베이스에 삽입하는데 사용된다. 또한 통합 검색 결과를 제공하기 위해 '예측된 ID 추출 모듈'을 사용하고, 통합 검색 결과 또는 통합된 전체 엔트리들을 XML 형식으로 제공하기 위하여 'XML 생성 모듈'을 사용한다. 그리고 FingerPRINTScan과 PScan 검색을 위해 필요한 prints27_0.pval_ blos62 등과 같은 검색 데이터들은 Search Files에 저장한다.



(그림 1) 전체 시스템 구조도

(그림 1)에서 기존의 논문[2]에서 기술한 기능들을 모듈화 과정을 거쳐 플랫파일 분석모듈, 연관정보 추출 모듈, 통합 파서 모듈화를 진행하였고, 특히 각 맴버 데이터베이스에서 제공하는 업데이트 정보를 이용하여 정기 갱신 모듈과 통합된 엔트리 데이터의 표준화 작업 및 재 배포를 위한 XML 형식으로의 변환 모듈로 확장 설계하였다. 정기 갱신 모듈은 각각의 데이터베이스에 대하여 각각 PRINTS 갱신 메소드, Prosite 갱신 메소드, Pfam 갱신 메소드로 구성된다. 또한 통합된 자원 역시 하나의 이질적인 형식으로 인하여 사용자들에 대한 접근시 어려움을 최소화하기 위하여 XML 형식으로 표준화가 가능한 모듈을 설계하였다. 각 모듈에 대한 정보는 다음 장에서 세부적으로 기술한다.

3.1 모티프 데이터베이스 자원의 물리적 통합 모듈

이 절에서는 이질적 데이터 형식으로 제작된 모티프 데이터베이스 자원들에 대한 물리적 통합을 모듈을 설계하였다. 이러한 통합 모듈들은 각기 다른 형식의 모티프 자원들을 하나의 DBMS로 저장하였고, 이러한 과정을 위하여 PRINTS, Pfam, Prosite에서 제공하는 플랫파일을 분석, 및 분해하였다. 이러한 과정을 3개의 모듈화를 통하여 구현하였다.

• 플랫파일 분석 모듈

플랫파일 분석 모듈은 각각의 플랫파일에서 추출 엔트리 정보에 해당하는 라인 정보들을 분석한다. 분석된 라인 정보들은 맴버 플랫파일에서 동등한 엔트리에 대한 정보를 검색한다. 이렇게 검색된 정보들은 결과 플랫파일로 재 생성한다. 이러한 과정을 수행한 후 각기 독자적 맴버 플랫파

일에만 존재하는 엔트리들을 결과 플랫파일에 추가한다. 자세한 알고리즘은 [2]를 참조한다.

• 연관 정보 추출 모듈

하나의 엔트리는 모티프 3차 구조 정보 및 분류 정보에 해당하는 정보들을 보유 할 수도 있고 하지 않을 수도 있다. 이러한 경우를 위하여 연관 정보 추출 모듈은 각기 모티프 엔트리에 해당하는 3차 구조 정보를 추출하고 통합된 플랫파일 엔트리 각각에 해당하는 모티프 3차 구조 정보를 위해 PDB 데이터베이스에서 제공하는 플랫파일에서 엔트리 데이터의 Residue 서열들의 시작 위치와 종료 위치 그리고 그 서열들의 Atom 정보에 해당하는 X, Y, Z 구조 정보를 추출하여 새로운 엔트리에 추가하였다. 분류 정보 또한 해당하는 SCOP 정보들을 추출하여 결과 플랫파일에 저장한다.

• 통합 파서 모듈

위의 두 단계에서 생성된 결과 플랫파일은 통합 파서 모듈은 다시 파싱 과정을 거쳐 관계형 데이터베이스에 저장한다.

3.2 정기적 업데이트를 위한 갱신 메소드

이 프로토 타입에서는 업데이트를 위해 맴버 데이터베이스들에서 제공하는 각각의 갱신 파일을 이용한다. Pfam 데이터베이스는 새로운 엔트리, 갱신 엔트리, 삭제 엔트리들에 관한 내용을 갱신 파일에 포함하고 있으나 Prosite와 PRINTS 데이터베이스 갱신 파일은 엔트리 삭제에 관한 내용이 없다. 따라서 각각의 데이터베이스마다 갱신 메소드가 다르다.

우선 각 맴버 데이터베이스의 FTP 사이트에서 주기적으로 제공하는 갱신 파일과 최신 데이터 파일들을 다운로드 받는다. 단백질 검색에 필요한 파일들은 Search Files 폴더에 저장한다. 아래는 각 맴버 데이터베이스에서 제공하는 갱신 정보를 이용한 갱신 메소드들이다.

3.2.1 PRINTS 갱신 메소드

1. newpr.lis 파일에서 새로운 엔트리 ID만을 추출
2. 추출된 엔트리 ID를 prints.dat 파일에서 검색
3. 검색된 엔트리에 대한 정보를 플랫파일 분석 모듈, 연관정보 추출 모듈, 통합 파서 모듈을 이용하여 데이터베이스에 저장한다

3.2.2 Prosite 갱신 메소드

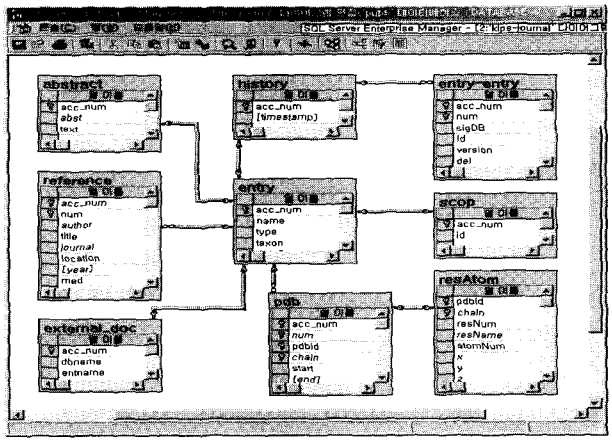
1. prosite.lis 파일을 연다.
2. +, * 인지를 파악한다.
3. +인 경우 엔트리 ID를 추출한다.
   ① 추출한 엔트리 ID를 prosite.dat 파일에서 검색한다.
   ② 검색된 엔트리 정보를 플랫파일 분석 모듈, 연관정보 추출 모듈, 통합 파서 모듈을 이용하여 통합 데이터베이스 안에 삽입한다.
4. *인 경우 엔트리 ID를 추출한다.
   ① 추출된 ID를 데이터베이스내 sigDB테이블 검색을 통하여 acc_num를 추출한다.
   ② 추출된 acc_num에 해당하는 정보 모두를 해당 테이블에서 삭제한다.
   ③ prosite.dat 파일을 열고 추출한 엔트리 ID에 대한 정보를 플랫파일 분석 모듈, 연관정보 추출 모듈, 통합 파서 모듈을 이용하여 통합 데이터베이스 안에 삽입 한다.

3.2.3 Pfam 갱신 메소드

1. diff.gz 파일을 연다.
2. NEW, CHANGE, NOCHANGE, DEAD인지를 파악한다.
3. NEW인 경우 해당 엔트리 이름을 추출한다.
   ① 추출한 엔트리 이름을 Pfam-A.seed 파일에서 검색하여 해당 정보들을 플랫파일 분석 모듈, 연관정보 추출 모듈, 통합 파서 모듈을 이용하여 데이 터베이스 안에 삽입한다.
4. CHANGE인 경우 엔트리 ID를 추출한다.
   ① 추출된 ID를 데이터베이스내 sigDB테이블 검색을 통하여 acc_num를 추출한다.
   ② 추출된 acc_num에 해당하는 정보 모두를 해당 테이블에서 삭제한다.
   ③ Pfam-A.seed 파일을 열고 추출한 엔트리 ID 해당 정보를 플랫파일 분석 모듈, 연관정보 추출 모듈, 통합 파서 모듈을 이용하여 통합 데이터베이스 안에 삽입 한다.
5. DEAD인 경우 엔트리 ID를 추출한다.
   ① 데이터베이스내 sigDB 테이블 검색 후 del 속성에 dead라 기입한다.

3.3 갱신 정보가 추가된 모티프 통합 개체-관계형 다이어 그램

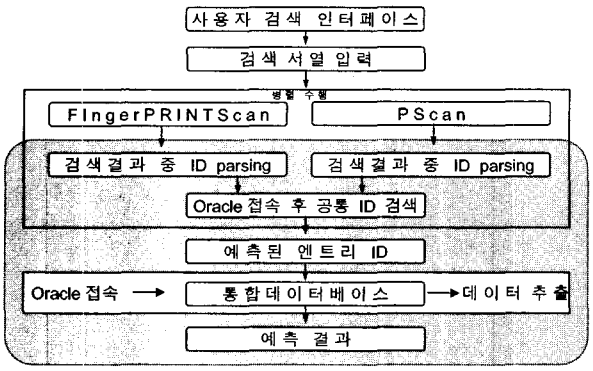
각각의 모듈을 이용한 결과 플랫파일은 통합 파서 모듈을 이용하여 (그림 2)와 같이 설계한 관계형 데이터베이스에 삽입된다. 또한 갱신 메소드를 이용하여 갱신 여부 및 버전을 알 수 있는 속성 del과 version을 이용하여 데이터베이스 관리자는 어떠한 맴버 데이터베이스에서 어떠한 버전까지 업데이트 되었으며, 어떠한 엔트리들이 어떠한 맴버 데이터베이스에서 삭제 되었는지를 알 수 있다. 이러한 개체-관계형 다이어 그램은 데이터 사이의 연관성 정보를 분석하여 entry, abstract, reference, external_doc, history, entry-entry, pab, scop, resAtom 엔티티들로 구성되었다.



(그림 2) 갱신 정보 저장을 위해 확장된 통합 데이터베이스 E-R 다이어 그램

4. 단백질 모티프 통합 예측 검색 모듈 설계

이 장에서는 단백질 모티프 통합 예측을 위한 서열 검색 메소드들 중에서 FingerPRINTScan과 PScan을 사용하였다. 이러한 메소드들은 모두 FASTA 포맷 데이터를 예측에 활용하고 있다. 이러한 검색 모듈의 수행과정을 (그림 3)과 같이 나타내었고 검색 수행 과정 중 예측된 ID 추출 모듈에 해당하는 부분을 점선으로 표시하였다.



(그림 3) 단백질 예측 수행 순서 및 예측된 ID 추출 모듈(점선)

예측 순서는 다음과 같다.

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 9가지의 경우의 수로 나뉠 수 있다. 이 경우의 수에 대한 처리 방법을 아래의 <표 1>에 나타내었다.

예를 들어, <표 2>의 9번과 같이 각각의 검색 메소드에서 1개 이상의 ID가 추출될 경우, 오라클에 저장된 데이터베이스 ENTRY-ENTRY 테이블 중 SigDB 속성과 비교한다. 이때 추출된 각 ID를 공통으로 가지고 있는 accnum를 선택하여 선택된 acc_num에 해당하는 정보들을 예측 결과로 나타낸다. 이때 공통 ID 부재시, 다시 말해 추출된 ID들이 sigDB 속성에서 하나의 acc_num에 속해 있지 않을 경우 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.

5. 구현 및 평가

이 장에서는 이 시스템의 프로토 타입 구현 환경과 인터페이스 및 사례 데이터베이스와의 비교 평가에 대하여 기술한다.

5.1 구현 환경

우리는 각각의 플랫파일들을 통합하기 위해 Window 2000 환경에서 C언어를 사용하였으며, 구축된 데이터베이스에 삽입하기 위해 ProC언어를 사용하였다. 또한 시스템 기종으로는 Sun사의 Enterprise 250을 사용하였으며, 운영체제로는 Sun Solaris 7(5.7), DBMS로는 Oracle 8i를 이용하였다. 그리고 XML 생성 모듈 구현을 위해 JAVA 1.3, JDBC, Graphic User Interface를 위해 JAVA Swing을 사용하였다. 이

〈표 1〉 예측된 엔트리 ID 처리 방안

번 호	FingerPRINTScan의 예측 ID 수	PScan의 예측 ID 수	예측 결과 도출 방법
1	0	0	예측 불가능
2	0	1	PScan 예측 결과를 도출
3	0	N	PScan 예측 결과중 첫 번째 ID 도출
4	1	0	FingerPRINTScan 예측 결과를 도출
5	1	1	DB 검색 후 공통 ID를 예측 결과로 도출 공통 ID 부재시 FingerPRINTScan 예측 결과를 우선시 함
6	1	N	DB 검색 후 공통 ID를 예측 결과로 도출 공통 ID 부재시 FingerPRINTScan 예측 결과를 우선시 함
7	N	0	FingerPRINTScan 예측 결과중 첫 번째 ID 도출
8	N	1	DB 검색 후 공통 ID를 예측 결과로 도출 공통 ID 부재시 FingerPRINTScan 예측 결과를 우선시함
9	N	N	DB 검색 후 공통 ID를 예측 결과로 도출 공통 ID 부재시 FingerPRINTScan 예측 결과를 우선시함

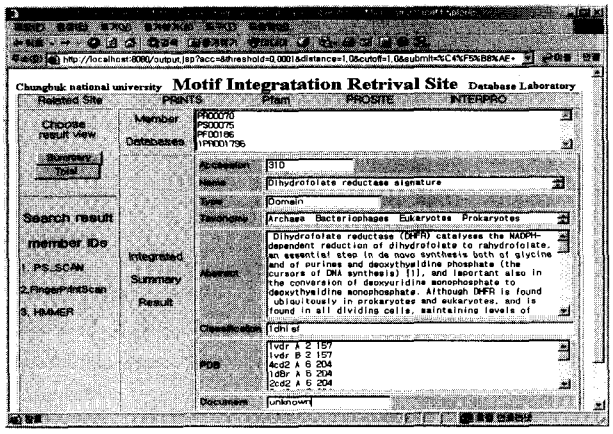
〈표 2〉 확장 설계한 통합 데이터베이스와 기존 데이터베이스와의 비교 평가

비 교		물리적 통합 데이터베이스		웹기반 통합 데이터베이스	웹 기반 통합 검색	
		제안한 통합 시스템	InterPro(Release 5.2)	PRINTS(Version 35)	SRS	PANAL
질의 처리	SQL 질의	○	○	○	×	×
	서열 질의	○	○	○	○	○
	서열질의 파라미터 지원	○	×	○	×	○
통합 자원	엔트리 수	5670	5876	1750	-	-
	3차 구조 정보	○	×	×	△	×
	모티프 분류 정보	○	×	×	○	×
	모티프 서열 지원	×	×	×	×	×
브라우징	중복 엔트리 처리	○	○	○	×	×
	요약 결과	○	○	○	○	○
	그래픽 결과	×	○	○	○	○
	XML 형식 지원	○	○	×	×	×
비교			EBI	Manchester		Minnesota

러한 환경 하에서 각각의 모듈을 거쳐 5,670개의 새로운 엔트리를 구성하였다.

5.2 단백질 모티프 예측을 위한 결과 인터페이스

(그림 4)는 단백질 모티프 예측을 위한 서열 검색 결과를 나타낸 것이다. 이 검색 모듈에서는 E-value의 디폴트 값으로 1.0을 적용하였다. 예측을 위한 검색 서열로는 Dihydrofolate reductase signature를 사용하였다. 이것은 이미 기능과 구조가 알려져 있다.

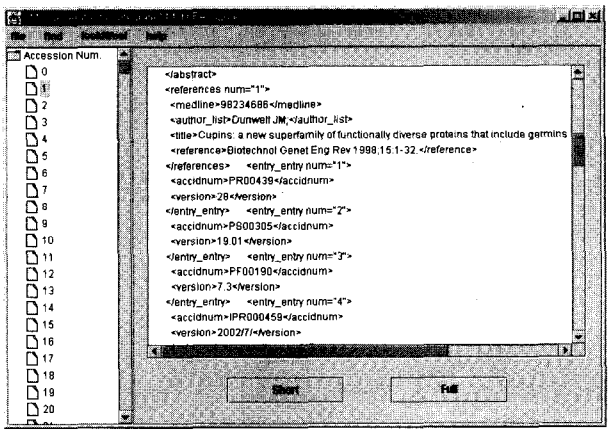


(그림 4) 단백질 모티프 예측 결과 화면

예측 결과로써 검색 서열인 Dihydrofolate reductase signature를 포함하고 있는 각 맴버 데이터베이스들, 이 엔트리에 대한 일반적인 정보들인 이름, 타입, 요약 그리고 Dihydrofolate reductase signature에 해당하는 추가 정보들을 알 수 있다.

5.3 XML 변환기 인터페이스

7.2절에서 통합된 모티프 자원 교환을 위해 그림 XML 생성 모듈을 통하여 (그림 5)와 같이 XML 생성기를 구현하였다. (그림 5)에서 보듯이 XML 생성기의 특징은 크게 두 가지로 나뉜다. 첫째, 사용자가 단백질 모티프 검색 결과를 XML 형식으로 원할 경우 검색 결과 엔티티의 accession number를 이용하여 XML 형식으로 제공할 수 있다. 둘째, 통합된 모티프 데이터들 전체를 하나의 XML 파일로 생성하여 사용자에게 제공한다.



(그림 5) XML 생성기 인터페이스

5.4 기존 모티프 데이터베이스 및 검색 시스템과의 비교 평가

우리는 구현된 모티프 통합 데이터베이스 및 검색 시스템에 대한 비교 평가를 위하여 기존 모티프 데이터베이스인 PRINTS, InterPro와 검색 시스템인 SRS, PANAL에 대

하여 비교 평가하였다. 평가 항목을 크게 세 가지로 나뉘었는데, 질의 처리, 자원 통합, 부라우징으로 평가 하였다. 아래 <표 2>에서 보듯이, 세 가지 항목에 대한 세부 항목들로 평가 항목을 확장하였으며, 각 해당 사항을 "○" 또는 "×"로 표기하였다. 예를 들어, 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. 따라서 모티프 부류 정보 항목에서 SRS는 "○"를 기입하였으며 PANAL은 "×"를 기입하였다. 3차 구조 정보 항목에서 SRS에서는 하이퍼 링크를 이용한 정보를 제공하기 때문에 "△"로 표기 하였다. 이것은 우리가 제안한 시스템과 SRS의 차이점을 보다 명확히 나타내기 위해서이다.

비교 평가를 통하여 질의 처리 항목에서는 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 웹기반 통합 데이터베이스와 물리적 통합 데이터베이스(InterPro)가 가장 우수한 것으로 평가된다. 또한 XML 지원 형식에서는 제안된 통합 시스템과 InterPro에서만 제공되는 것을 알 수 있었다.

6. 결 론

이 논문에서는 그동안 각각 개별적으로 생성 및 성장하여 온 InterPro, ProDom, BLOCKS, PROSITE, PRINTS 등의 모티프 데이터베이스들에 대한 사용자측 접근 문제들과 웹 기반 통합 및 검색에 따른 문제들에 대한 대안책을 제시하였다. 뿐만 아니라 기존 [2]에서 제안치 못했던 정기적 업데이트시 중복 엔트리의 처리문제, 엔트리 삭제시 링크의 처리문제, 엔트리 통합후의 표준화 문제를 해결하기 위하여 확장 설계하였다. 따라서 각각의 모티프 데이터베이스들에서 나타나는 이질적인 형식의 예측 결과 문제와 웹 기반 Cross-reference 통합에 따른 중복된 데이터베이스 엔트리 핸들링 문제 등을 해결하기 위하여 모티프 자원들에 대한 물리적 통합 메소드들과 단백질 모티프 통합 예측 검색 메소드를 제안하였다. 그리고 정기적 업데이트를 위한 갱신 메소드 및 통합된 데이터의 표준화를 위한 XML 형식 변환 모듈을 가능케 하였다. 우리는 이러한 통합 데이터베이스를 구현하기 위하여 아래와 같은 단계들을 진행하였다.

- PRINTS, Pfam, Prosite, PDB 등의 데이터베이스들에서 제공하는 플랫파일을 분석
- 각각의 모티프 엔트리에 대한 통합 및 3차 구조 정보와 분류 정보 통합
- 개체-관계 모델링을 이용한 통합 관계형 데이터베이스 구축
- 각 검색 메소드들을 통합한 검색 시스템 프로토 타입 구축과 Web기반 인터페이스 구현
- 통합 데이터베이스 정기적 갱신을 위한 모듈 구현
- 통합 데이터의 XML 형식 변환 모듈 구현

따라서, 기존의 데이터베이스 검색시 사용자가 겪는 이질적 검색환경 및 반복 접근 문제를 해결하였고 기존의 웹 기반 통합 검색에서 지원하지 못했던 단백질의 3차 구조정보, 분류 정보, 샘플 정보의 지원을 가능케 하였다. 또한 정기적 갱신을 통하여 맴버 데이터베이스와 동등한 엔트리들의 보유와, 통합된 엔트리들에 대해 XML 형식으로의 변환을 이용하여 재 배포를 가능케 하였다.

참 고 문 헌

[1] 김성진, 이상호, "객체-관계형 데이터베이스 시스템을 위한 새로운 성능 평가 방법론", 정보처리논문지, 제7권 7호, 2000.

[2] 이범주, 최은선, 류근호, "모티프 자원 통합을 이용한 단백질 모티프 예측 시스템 구현", 정보처리학회논문지D, 제10-D권 제4호, 2003.

[3] 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.

[4] 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.

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

[6] David W. Mount, "Bioinformatics : Sequence and Genome Analysis," Cold Spring Harbor Laboratory Press, pp.45-48, 2001.

[7] Kevin A. T. Silverstein, Alan Kilian, John L. Freeman, James E. Johnson, Ihab A. Awad, Ernest F. Retzel, "PANA

L : an integrated resource for Protein sequence ANALysis," Bioinformatics, Vol.16, pp.1157-1158, 2000.

[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, pp. 304-308, 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, pp.276-280, 2002.

[10] 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.

[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, pp.417-424, 1997.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

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

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

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

[22] 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.

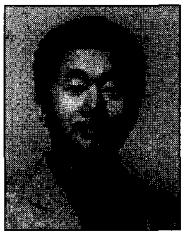
[23] Barbara Eckman, Julia Rice, Bill Swope, "Heterogeneous Data and Algorithm Integration in Bioinformatics," ISMB, 10th International Conference Tutorial, 2002.

[24] Steve Muench, "Building Oracle XML Applications," O' Reilly & Associates, Inc., pp.8-22, 2000.

[25] Steven Holzner, "Inside XML," New Riders, pp.33-42, 2000.

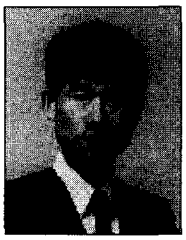
[26] Bill Brogden, Charis Minnick, "JAVA Developer's Guide to E-Commerce with XML and JSP," SYBEX Inc., 2001.

[27] 최명종, 유재우, 최재영, "자바 개발자를 위한 XML", 홍릉과학출판사, 2003.



노 기 용

e-mail : kyno@kriss.re.kr
1981년 충남대학교 물리학과(이학사)
1995년 충남대학교 전산학과(이학석사)
2002년 충북대학교 전산학과(박사수료)
1988년~현재 한국표준과학연구원
관심분야 : DB설계, Image Processing, ATM 등



김 원 식

e-mail : wskim@kriss.re.kr
1979년 아주대학교 전자공학과(공학사)
1984년 고려대학교 물리학과(이학석사)
2004년 연세대학교 의공학과(의공학박사)
1984년~현재 한국표준과학연구원
관심분야 : 생체신호 측정, 처리, 해석 등

이 범 주

e-mail : bjlee@dblab.chungbuk.ac.kr
2001년 서원대학교(이학학사)
2003년 충북대학교 대학원 전자계산학과 (이학석사)
2004년 충북대학교 대학원 전자계산 학과 (박사과정)
관심분야 : 시공간 데이터베이스, 데이터마이닝, XML, Bioinformatics 등

이 상 태

e-mail : stlee@kriss.re.kr

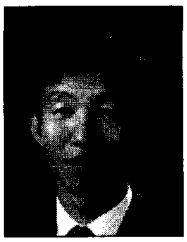
1977년 아주대학교 전자공학과(공학사)

1992년 전북대학교 전자공학과(공학석사)

1998년 전북대학교 전자공학과(공학박사)

1985년~현재 한국표준과학연구원

관심분야 : 지능망, 광대역통신망, 트래픽 제어 등



류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr

1976년 숭실대학교 전산학사(공학사)

1980년 연세대학교 공학대학원 전산전공 (공학석사)

1988년 연세대학교 대학원 전산전공 (공학박사)

1976년~1986년 육군군수지원사전산실(ROTC 장교), 한국전자통신연구원 (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, 에이전트기반 정보검색 시스템, 데이터마이닝, 데이터베이스 보안 및 Bioinformatics 등