바이오 데이터 패턴 분석을 위한 시스템 및 알고리즘 설계

Design of the System and Algorithm for the Pattern Analysis of the Bio-Data

송영옥*, 김성영**, 장덕진*
우송대학교 컴퓨터정보학과*, 경북대학교 전기정보컴퓨터공학부**

Young-Ohk Song(yysong@wsu.ac.kr)*, Sung-Young Kim(sykim@knu.ac.kr)**,
Duk-Jin Chang(djchang@wsu.ac.kr)*

요약

생명과학 분야에서 컴퓨터를 활용할 수 있는 대표적인 예로는 서열화, 서열화 분석, 비교, 진화, 돌연변이 추적, 약 설계를 위한 유사성 비교, 단백질 기능 예측, 그리고 세포 메커니즘과 질병 발생에서의 유전자 역할 예측 등 다양한 분야를 들 수 있다. 생명공학 연구자들에게는 이와 같은 작업을 위한 도구들이 요구되고 있다.

본 논문에서는 바이오 데이터 분석을 위한 기존 시스템의 문제점을 파악하고, 이를 개선할 수 있는 시스템 설계에 초점을 맞추었다. 또한 각각의 분석 작업을 개선할 수 있고 서로 독립적으로 진행되는 기존의 시스템을 통합할 수 있는 통합 분석 시스템을 설계하고자 한다.

■ 중심어 : | 바이오인포메틱스 | 바이오 데이터 | 유전자 데이터베이스 | UML | 마코브 모델 |

Abstract

In the field of biotechnology, computer can play varied roles such as the ordinal analysis, ordianl comparison, nutation tracing, analogy comparison for drug design, estimation of protein function, cell mechanism, and verifying the role of a gene for preventing diseases. Additionally, by constructing database, it can provide an application for the cloning process in other data researches, and be used as a basis for the comparative genetics. For the most of researcher about biotechnology, they need to use the tool that can do all of job above.

This study is focused on looking into problems of existing systems to analysis bio data, and designing an improved analyzing system that can propose a solution. In additional, it has been considered to improve the performance of each constituent, and all the constituents, which have been separately processed, are combind in a single system to get over old problems of the existing system.

■ keyword : | Bioinformatics | Bio-Data | Genom Database | UML | Markov Model |

Ⅰ. 서 론

생명과학 분야에서 컴퓨터를 활용할 수 있는 대표적인 예로는 서열화, 서열화 분석, 비교, 진화, 돌연변이 추적, 약 설계를 위한 유사성 비교, 단백질 기능 예측, 그리고 세포 메커니즘과 질병 발생에서의 유전자 역할

* 본 논문은 한국콘텐츠학회 2010 춘계 종합학술대회 우수논문입니다.
접수번호 : #100715-001
접수일자 : 2010년 07월 15일
심사완료일 : 2010년 08월 09일
교신저자 : 장덕진, e-mail : djchang@wsu.ac.kr

예측 등 다양한 분야를 들 수 있다. 또한 데이터베이스를 구축함으로써 다른 데이터 연구에서 클로닝 작업을 하고자 할 때 가용성을 제공할 뿐만 아니라 비교 유전학을 위한 기반으로 사용될 수 있다[11]. 바이오 데이터 분석의 가장 초기 과정으로 DNA와 단백질 서열에 대한 데이터 정보 검색을 들을 수 있으며, 이와 같은 생물학적 데이터 마이닝 작업을 하기위해 NCBI, EBI, GenomeNet 등에서 제공하는 데이터베이스를 활용하는 각종 도구들이 출시되고 있다. 이와 같은 도구들을 활용하여 바이오 정보검색을 위한 스트링 패턴 검색과 서열이나 구조의 검색, 배열 및 비교를 위한 유사성 검색과 같은 작업을 수행한다[1][2].

이와 같은 기존 분석 도구들의 공통적인 개선점으로는 첫 번째 각각의 분석 작업에 대한 연계성을 배제하고 독립적으로 수행하도록 설계되어있다. 이로 인해 분석 과정에서 불필요한 작업이 반복되고 있다. 두 번째로는 데이터 분석의 연속성을 고려하지 않았다. 기존 대부분의 도구들이 웹 기반으로 제공되고 있는데 사용자 인증과정을 거치지 않는 단순 request와 response만 이루어지기 때문에 일회성 검색 기능을 제공하기 때문에 같은 결과에 대한 결과 값이 다시 요구될 때는 같은 작업을 반복해야 결과 값을 볼 수 있다[4-6].

본 논문에서는 이와 같은 문제점을 개선하여 바이오 데이터 분석 기능에서 주요 기능들을 빠르고 연속성 있게 처리할 수 있는 통합시스템의 필요성을 고려하여 바이오 데이터의 패턴을 분석하고 패턴 분석에 의해 구분된 유용한 바이오 스트링을 기반으로 그 의미를 연구할 수 있는 통합된 시스템에 필요한 전반적인 작업과정을 모델링을 함으로써 시스템 구현에 이용하고자 한다.

이와 같은 시스템 설계를 위한 본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 연구 배경을 알아본다. 이를 바탕으로 본 논문에서 설계하고자 하는 전체 시스템의 구조를 먼저 제시하고 구현에 필요한 부분과 절차를 모델링 도구인 UML을 이용하여 설계하고 서열비교와 유사성 비교를 위해 이용될 알고리즘 제안에 관하여 3장에서 나열한다. 4장에서는 3장에서 설계한 시스템이 구현된다면 어떤 형태로 분석과정이 이루어 질 것인지에 대해 미리 예측해 보고, 본 시스템으로 인한 기대 효과를 논한다. 마지막으로 5장에서는 본 논문의 결론과 향후 계속되어야하는 연구 방향을 제시하고자 한다.

II. 연구 배경

바이오인포메틱스의 분야를 크게 3개의 분야로 나누고 있다. 첫 번째는 방대한 생명과학자료를 분석하기 위한 정보 처리 알고리즘과 통계 이론을 개발하는 분야이고, 두 번째는 다양한 형태의 정보와 분석 이론들을 도구화해서 생물학자가 사용할 수 있도록 구현하는 분야이며 마지막으로 세 번째는 유전자와 단백질의 서열과 구조 및 세포와 개체의 기능에 관련된 방대한 정보를 분석하고 해석하여 생물학적 의미를 찾아내는 분야이다[7][8].

첫 번째와 두 번째의 분야에서 컴퓨터공학적으로 접근할 수 있는 요소들이 주로 존재함으로써 컴퓨터공학자들이 생명공학과의 융합을 찾아볼 수 있는 분야라고 할 수 있다. 이에 비해 세 번째 분야는 생물학자들이 접근하는 분야라고 할 수 있다. 다음 [표 1]에서는 이와 같은 바이오인포메틱스의 연구 분야에 관하여 나열하였다. 많은 부분에서 생명공학의 연구에 컴퓨터공학의 뒷받침이 필요하다는 것을 알 수 있다.

표 1. 바이오인포메틱스의 연구분야

구분	DB	데이터 해석	알고리즘
분자	-분자구조 DB * 염기서열 * 아미노산 서열 * 입체구조(FDB) -분자기능 DB * 핵산 모티프(EFD, Transfac) * 단백질 모티프 * 유전자 주석	-서열/입체구조 해석 * 서열비교 * 입체구조 비교 * 입체구조 예측 - 기능 부위 해석 * 모티프 추출 * 모티프 검색 * 기능 예측	-최적화 알고리즘 * DP(Dynamic Programming) * SA(Simulabed Annealing) * GA(Genetic Algorithm) - 패턴 인식 학습 알고리즘 * ANN(Neural Network) * HMM(Hidden Makov Model) * SVM(support Vector Machine)
Genome (분자의 집합)	-Genome 기능 DB * KEGG OOG * 발현 프로파일 * 유전자 다형	-비교 Genome 해석 -Transcriptum 해석 - Proteome 해석 - 다형 정보 해석	-클러스터링 알고리즘 * 계층적 클러스터 해석 * 코호넨 네트워크

현재까지 일반적으로 이용되고 있는 바이오 데이터 분석 시스템으로는 미국의 NCBI에서 제공되는 데이터베이스를 비롯하여 각종 분석 도구들이 있으며, EBI, GenomeNet등에서 웹 기반 시스템으로 제공되고 있는 분석 시스템을 들 수 있다. 여러 가지 기능을 통합된 형

태로 제공되는 분석 시스템으로 대표적인 것은 NCBI, GeneWeb II등을 들 수 있다. 전 세계적으로 생명공학 연구자들에게 이용되고 있는 대표적인 분석도구들을 다음 [표 2]에 정리 했으며 서비스 되는 사이트 정보를 나열하였다[3][9][10].

표 2. 기존 분석 도구들

기능/기관/도구		종 류	URL
ORF	NCBI	ORF Finder	http://www.ncbi.nlm.nih.gov/gorf/gorf.html http://bioweb.uwlax.edu/GenWeb/Molecular/Seq_Anal/Translation/translation.html
서열유사성 비교	NCBI	BLAST	http://www.ncbi.nlm.nih.gov
	EBI	FASTA	http://www2.ebi.ac.uk/fasta3

또한 다음 [표 3]에서는 대표적인 바이오 데이터베이스 서버들을 소개하고 있으며 일정 시간마다 데이터베이스의 정보가 업데이트되어 생명공학 연구자들에게 공개적으로 제공되고 있다[4-6].

표 3. 바이오 DB 서버

기관	URL	주요검색시스템	주요DB
GenomeNet(동경대)	www.genome.ad.jp	DBGET	KEGG
NCBI	www.ncbi.nlm.nih.gov	Entrez, BLAST	PubMed GenBank
EBI	www.ebi.ac.uk	SRS	EMBL, SWISS-PROT
SIB	www.isb-sib.ch www.expasy.ch	SRS	SWISS-PROT

위의 각각의 표에서 언급한 분석도구나 데이터베이스 서버를 활용할 수 있는 방법들은 아직까지도 바이오 연구자들에게 보편적으로 사용되고 있다[4]. 하지만 이와 같은 도구들은 바이오 데이터의 원하는 연구 결과를 얻기까지는 앞 서 언급된 여러 가지 문제점을 가지므로 여러 가지 반복된 작업들이 행해져야 한다.

본 논문에서는 ORF 검색, 유전자 탐색, 서열 유사성 검색 등의 작업을 고려하여 기존 분석 시스템들의 대표적인 특징 및 개선점들을 조사한 것을 바탕으로 효율적인 바이오 데이터의 분석도구를 구현할 수 있도록 시스템 모델링을 하고자 한다.

본 논문에서는 시스템의 모델링을 위하여 모델링 도구인 UML(Unified Modeling Language)[13][14]을 이용하고 유용하고 정확한 바이오 정보를 찾기 위하여 적절한 알고리즘들을 제시하고자 한다.

Ⅲ. 바이오 분석 시스템 설계

1. 바이오 분석 시스템 구조 설계

바이오데이터 패턴을 이용하여 연구되는 분야들은 다양하며 다음 [그림 1]과 같은 기능들이 기본적으로 구현되고 이용되어야 한다. 기존의 시스템에서 DNA Sequence, RNA Sequence, Protain Sequence의 작업들을 각각 독립된 소프트웨어 시스템 환경에서 이루어 졌다면 본 논문에서 설계하고자 하는 시스템에서는 아래의 각 기능들이 하나의 시스템 환경에서 연계성 있게 이루어 질 수 있도록 설계하고자 한다.

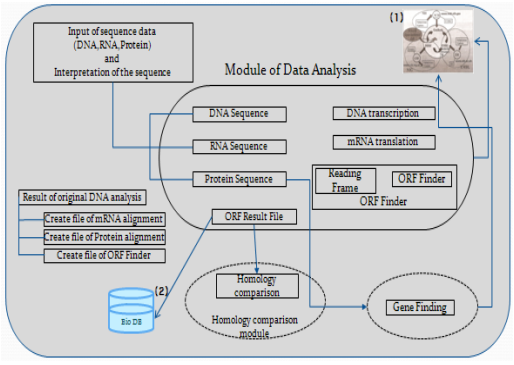


그림 1. 바이오 분석 시스템 전체 구성 요소

위의 시스템에서 가장 먼저 이루어지는 작업인 ORF 축출이란 리딩 프레임을 찾는 것으로 리딩 프레임이란 연구 중인 세포의 DNA에서 실제로 어디에서부터 DNA가 단백질로 번역되는지를 자동으로 판단하여 가능성이 있는 부분을 축출하는 것이다.

[그림 1]의 (1)에 해당하는 데이터베이스는 전 세계 연동된 바이오 DB 서버들이며 일정 시간마다 서로 최대한 중복을 피할 수 있도록 수정되고 있고 (2)는 연구가 진행되는 로컬DB서버로 해당 연구기관에서 연구되는 데이터들만 보관할 수 있는 용도로 이용이 될 것이다.

2. UML 모델링

본 논문에서 설계 모델로 제시한 전체 구조를 조망할 수 있도록 UML 다이어그램으로 모델링하였다. 먼저 다음 [그림 2]에서는 본 시스템에서 제공하고자 하는

주요 요소를 유스케이스 다이어그램으로 표현한다. [그림 2]에서 보는 봐와 같이 본 논문에서 설계하고자하는 전반적인 시스템에는 ORF 축출 기능, DNA 복제 및 DNA 전사과정을 수용할 수 있는 모듈이 포함될 것이며, mRNA 번역과정이 자동화할 모듈이 포함되도록 설계하였다. 이와 같은 각각의 작업은 하나의 소프트웨어 시스템 내에서 이루어지며 NCBI, CIB 와 EBI 등의 바이오 데이터베이스에서의 자료 검색이 필요할 때 간단한 환경 설정을 통하여 자동으로 결과를 얻을 수 있도록 설계한다.

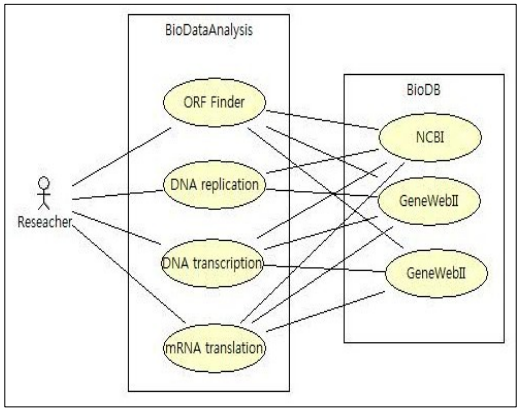


그림 2. 유스케이스 다이어그램

[그림 3]에서는 ORF 축출과정을 예로 전체 시스템 요소 사이의 진행과정을 시퀀스 다이어그램을 통하여 절차를 표현하였다. 이와 같은 과정에서 기존 제공되는 검색 도구들이 이용할 때 1부터 3번까지의 과정을 반복적으로 수행하면서 최종 결과를 얻을 수 있었으나 본 설계과정에서는 ORF 축출된 결과와 DB검색결과들을 저장하고 필요에 따라 계속진행이 가능하도록 하여 무의미한 반복을 배제하고자 한다. 이는 기존 검색도구들이 사용자의 연구 결과와 진행 상황들을 보존해 줄 수 없는 문제점을 극복할 수 있도록 개인 연구 자료의 보존기능을 추가하였기 때문이다.

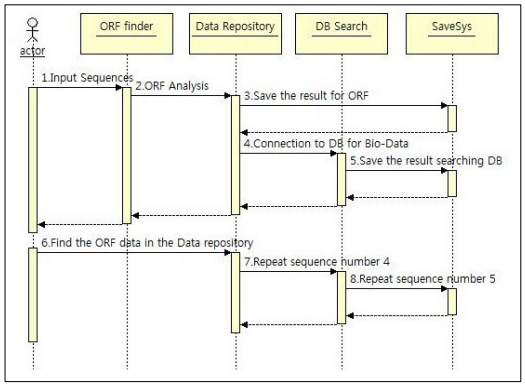


그림 3. 시퀀스 다이어그램

다음 [그림 4]에서는 전체 시스템 구현에 필요한 모듈들의 관계를 클래스 다이어그램으로 표현하였다. 텍스트로 변환하여 연구 자료로 제공되는 DNA 서열정보를 관리할 수 있는 DNASequence 클래스를 기반으로 ORF축출을 담당하는 ORFFinder클래스 및 로컬 데이터베이스와 NCBI, CIB 또는 EBI 등 해당 데이터베이스를 선택적으로 연결하여 결과를 검색할 수 있도록 담당하는 DBControl 클래들의 관계를 볼 수 있다.

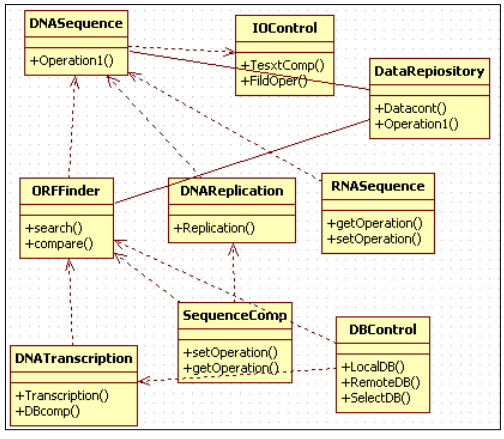


그림 4. 클래스 다이어그램

3. 바이오 데이터 분석을 위한 알고리즘 설계

본 논문에서 설계하고자 하는 전체 과정에서 중요한 부분 중 하나는 DNA 서열에서 정렬시키고 서열 유사성 비교 등을 들을 수 있다. 이와 같은 서열데이터에 대

한 분석 작업을 위하여 본 논문에서는 Markov Model(MM)[12]을 개선시켜 Hidden Markov Model(HMM)[13]을 이용하고자 한다. 다음 식 (1)처럼 먼저 바이오 데이터 분석에 MM을 적용시키기 위해 각 상태 x_i는 이전 상태에 의존적인 서열의 상태를 의미하도록 설정한다. 즉 $x = x_1 x_2 ... x_i .. x_L$와 같이 설정한다.

$$\begin{aligned} P(x) &= P(x_L|x_{l-1},...,x_1)*P(x_{L-1}|x_{L-2},...,x_1)...P(x_1) \\ &= P(x_L|x_{L-1})*P(x_{L-1}|x_{L-2})...P(x_2|x_1)P(x_1) \\ &= P(x_1)\Pi P(x_i|x-i) \quad \text{for } i=2 \text{ to L} \end{aligned} \tag{1}$$

수식에서 $P(x_i|x_{i-1})$은 x_i상태에서 x_{i-1}로 변환하는 변환확률이다.

이와 같은 MM의 기본 알고리즘을 이용한 많은 변형 알고리즘들이 이용되며 그 중 아미노산 서열이 $S = x_1 x_2 ... x_i ... x_n$라고 가정한다면 다음 그림과 같이 기준 서열을 정렬시키고 이를 바탕으로 다음 상태의 가능한 경우를 적용 시킬 수 있다.

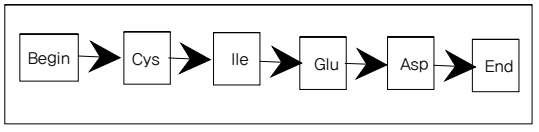


그림 5. HMM의 기준 서열 모델

이와 같은 HMM 알고리즘을 이용하여 바이오 데이터 분석에 활용함으로써 기준 서열을 정렬시키고 이를 바탕으로 다음 상태의 가능성을 예측 할 수 있도록 한다.

Ⅳ. 기대효과

분석 시스템에 포함될 데이터 분석 요소로는 기본적인 작업과정인 DNA전사, mRNA 번역, ORF 검색등과 이러한 작업과정을 통해 얻어진 결과들을 이용하여 데이터베이스로부터 서열 유사성 비교를 하고 전체 유전체로부터 유용한 유전자 탐색 등의 과정이 포함될 것이다.

본 논문에서 설계한 시스템을 바탕으로 이루어질 수 있는 전반적인 과정과 결과는 다음 [그림 6]과 같이 될 것이다. 바이오 데이터 전체 서열정보에서 유용한 ORF 자료가 축출될 것이며 이를 바탕으로 DNA 복제, DNA 전사, mRNA 번역과정이 진행되고 관련 DB서버에 자동 연결되어 가장 유사한 서열정보를 얻게 될 것이다.

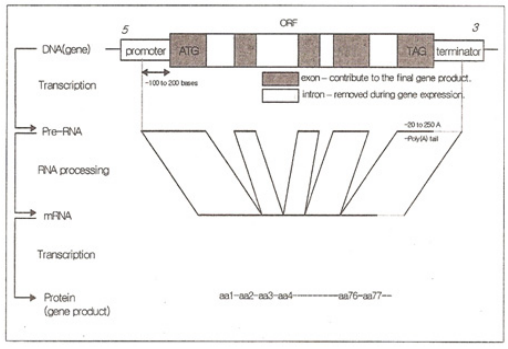


그림 6. DNA에서 단백질 정보 획득 과정

기존 제공된 분석도구들에서 문제시 되었던 분석 작업의 연속성, 수동적인 입력 과정의 과다로 인한 오류 발생률이 증가할 수밖에 없었던 사항에 대해서 본 시스템에서 크게 개선될 것으로 예측한다.

본 논문의 설계를 기반으로 전체 시스템에서 이루어질 작업과정을 샘플 바이오 데이터를 이용하여 조망한 결과를 다음 [그림 7]과 같이 나타낼 수 있다. DNA 서열 데이터는 연구 분야에 따라 다를 수 있으며 연구자들에게 제공될 때 다음과 같은 텍스트로 주어지기 때문에 이를 바탕으로 전체 시스템을 설계하였다.

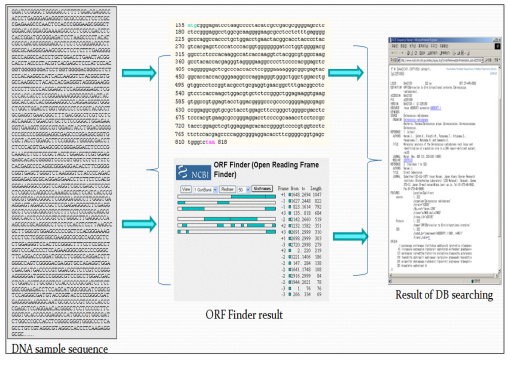


그림 7. 바이오 샘플 데이터의 진행 모형

위 [그림 7]의 진행 모형은 NCBI 사이트에서 제공되는 도구를 이용하여 얻은 결과 값이다. 이는 본 논문에서 설계한 시스템의 실행과정과 결과 값에서는 유사할 수 있으나 실제 이용과정에서는 사용자 편리성 면에서 유리할 것이다.

V. 결론

본 논문에서는 바이오 데이터를 분석하는데 이용되는 하기위한 기존 시스템에게 공통적인 개선점을 찾아보았다. 즉 분석 작업의 연계성 부족으로 인해 분석 과정에서 불필요한 작업이 반복되고 있으며 각각의 기능 분석을 위한 시스템을 제공되는 기관들이 서로 달라 사용자에게 여러 가지 불편함과 불필요한 작업들이 요구되는 문제점들이 존재했다.

본 논문에서는 이와 같은 문제점을 개선하여 바이오 데이터 분석 기능에서 주요 기능들을 빠르고 연속성 있게 처리할 수 있는 통합시스템의 필요성을 고려하여 바이오 데이터의 패턴을 분석하고 패턴 분석에 의해 구분된 유용한 바이오 스트링을 기반으로 그 의미를 연구할 수 있는 통합된 시스템에 필요한 전반적인 작업과정을 모델링을 함으로써 시스템 구현에 이용하고자 한다.

본 논문에서 바이오 데이터의 패턴을 효율적으로 분석하고 데이터를 기반으로 데이터베이스를 활용할 수 있는 전반적인 과정을 진행할 수 있는 통합 시스템을 구현하고자 필요한 전체 모듈을 구성요소별로 구분하고 설계하였다. 또한 바이오 데이터 분석에 유용하게 활용될 HMM 알고리즘의 이용을 제안하였다.

본 논문의 설계를 바탕으로 앞으로 보다 구체적으로 각 모듈을 구현하기위한 알고리즘을 작성하고 사용자의 측에서 이용이 용의할 수 있는 인터페이스를 연구함으로써 효율적인 통합 시스템이 구현될 수 있도록 해야 한다.

참 고 문 헌

[1] ETRI(IT-InformationCenter), "bioinformati- cs: Technology & Market Analysis," 2001.

[2] Andreas D.Baxevanis and B. F. Francis Ouellette, Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 2000.

[3] James Tisdall, "Beginning Perl for Bioinformatics," 2001.

[4] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

[5] http://www.genome.ad.jp/dbget/db_growth.html,

[6] http://blast.ncbi.nlm.nih.gov/Blast.cgi

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

[8] Susumu Goto and Kotaro Shiraishi, "Constructing and Annotating Genes Database in KEGG," International journal of Genome Research, 1998(10).

[9] Andreas D. Baxevanis and B. F. Francis Ouleeltte, "bioinformatics : A paractical Guide to the Analysis of Genes and Proteins," WILEY-INTERSCIENCE, 2001.

[10] W. R. Pearson, "Rapid and sensitive Sequence Comparison with FASTP and FASTA," Methods in Enzymology, 183, pp.63-98, 1990.

[11] Des Higging and Willie Taylor, "Bioinformatics: sequence, structure and databanks," Oxford University Press, 2000.

[12] Pachel Karchin and Richard Hughey, "Weighting Hidden Morkov models for maxinum discrimination," Bioinformatics, Vol.14, 9, 1998.

[13] Dan Pilone, Neil Pitman, "UML 2.0 in a Nutshell," OReilly Media, 2005.

[14] He Ke-qing, Jian Hong, He Fei, and Ying Shi, "Extended UML with role modeling," Wuhan University, 2009.

저 자 소 개

송 영 옥(Young-Ohk Song) 종신회원

- 2003년 2월 : 충북대학교 컴퓨터공학과(공학박사)
- 2000년 10월 ~ 2002년 2월 : 대전보건대학 멀티미디어학과 겸임 교수
- 2000년 10월 ~ 2002년 2월 : (주)테라빛테크 책임연구원
- 2002년 3월 ~ 현재 : 우송대학교 컴퓨터정보학과 초빙교수
- 2008년 2월 ~ 현재 : AQUARITAS,INC.USA,CTO

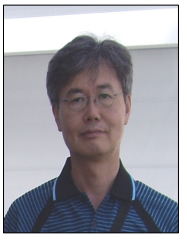
<관심분야> : Embedded System, Software Engineering, IT&BT

김 성 영(Sung-Young Kim) 정회원

- 2010년 2월 : 경북대학교 전자전기컴퓨터공학부 박사 수료
- 2002년 3월 ~ 2005년 2월 : 목원대학교정보통신공학과 겸임교수
- 2005년 3월 ~ 현재 : 상명대학교 컴퓨터정보학과 겸임교수
- 2000년 10월 ~ 현재 : (주)테라빛테크 대표

<관심분야> : Embedded System. Network, USN, IT&BT

장 덕 진(Duk-Jin Chang) 정회원



- 1983년 ~ 1986년 : 미국 텍사스 A&M 대학교 박사수료
- 1984년 ~ 1986년 : 텍사스 교통연구원 연구조원
- 1984년 ~ 1986 : 시스템공학연구소 선임연구원
- 1995년 ~ 현재 : 우송대학교 컴퓨터정보학과 교수

<관심분야> : Software Engineeirng, Project Management, Intelligent Transportation System, IT&BT