국가유전체정보센터(NGIC): 생명정보사회를 대비한 10만 양병

국가유전체정보센터 박종화

1. 서 론

포스트지놈 시대의 genomics, proteomics와 더불어 많은 -omics 분야를 포함하는 생정보학(bioinformatics)은 1960년대 구조생물학과, 1970년대의 최초의 완전 해독된 유전체들에서 시작된 현대 생물학에서 핵심적인 위치를 차지하는 분야이다. 많은 생물학의 분야들이 정보처리학 (informatics)을 필수로 사용한다. 생정보학을 크게 나누면, 여섯 가지이다. 다음이 그것들이다. 유전체학(genomics), 단백체학(proteomics), 발현체학(expressomics), 상호작용체학 (interactomics), 생물구조체학 (Structomics) 및 문헌체학 (textomics). **이들은 서로 유기적으로 연결될 수 있는 그러나 매우 독립적인 연구 단위들이다. 이들을 통합한 인프라 체제를 구축하고, 그것들이 각각의 생물실험실에 실용적으로 이용될 수 있는 것이 생정보학의 가장 이상적인 결과이다.** 그러나, 지난 1990년 중반이후의 급격한 팽창과 기대에도 불구하고, 생정보학은 그 성과가 투자에 비해 세계적으로 부진하다는 것이 일반적 평가이다. 이것은 비전분가들에게도 공평히 나누어 주게 된 연구비, 교육이 안된 인력을 이용한 무리한 프로젝트수행과 이미 지나간 생정보학 문제들을 피한, 새로운 생정보학 문제를 제대로 선정 및 투자하지 못한 때문이다. **NGIC는 유전체정보에서 시작되는 각종 omics 학의 여러 분야를 지원하는 국가적 생명정보 인프라 구축을 통해, 조직적이고, 치밀한 연구와 개발을 위한 생산성 효율계산에 바탕을 둔 프로젝트를 수행하는 것이 목표이다.** 국제적으로는 2007년 동아시아의 핵심 인프라인정을 목표로, 한중일 협동 교육 등을 통한 국제적 교류를 추진하고 있으며, 2010년 세계 3대 생정보처리 및 서비스기관이 되는 것이 목표이다. 결론적으로, 국가유전체센터는 곧 도래될 "개인유전체혁명"이 가져다 줄 기회와 문제를 대비한 국가적 인프라이다.

2. 설립배경 및 목적

국가유전체정보센터(NGIC, National Genome Information Center)는 국내외의 유전체 관련 정보를 체계적으로 수집 집대성하고, 이의 보급 유통을 통해 국내 생정보학 연구지원 및 정보인프라를 구축하여, 국내 생정보학 및 전산생물학 분야의 발전을 위하여 2001년 10월 과학기술부의 지정을 통해 설립되었다.

생정보학(bioinformatics)은 BT와 IT의 융합기술로 생물체의 유전정보와 그들 사이의 복잡한 상호작용 같은 크고, 다양한 생명현상 관련 정보를 컴퓨터를 사용하여 정리 · 분석 · 해석하는 과학 및 공학 분야이며, 차세대 생명공학분야 연구를 위한 필수 연구도구 즉, 연구 인프라로서 매우 중요하게 발전시켜야 할 학문분야이다. 때문에 생정보학 (bioinformatics) 에 대한 기술개발과 연구의 성과가 미래 생명공학분야의 국가적 연구의 성패에 중요한 초석이 된다.

따라서, 국가유전체정보센터는 점차 증대되는 국내 유전체 연구의 체계적이고 통합적인 연구 체제와 인프라를 구축하고, 국내 각종 생물체학들의 연구결과를 통합한 DB의 구축과 활용을 도운다. 그리고 국내 유전체 연구기관 사이의 네트워크를 구축하여(그림 1) 생명공학 연구의 시너지 효과를 도모한다.

한 가지 매우 중요한 것은 NGIC는 핵심 인프라에 바탕을 둔, 자동화된 대량생산체제의 연구개발센터이다. 이것은 일반 생물학내지, 소단위의 생정보학 연구실과는 본질적으로 다르며, 그 지향목표도 다르다.

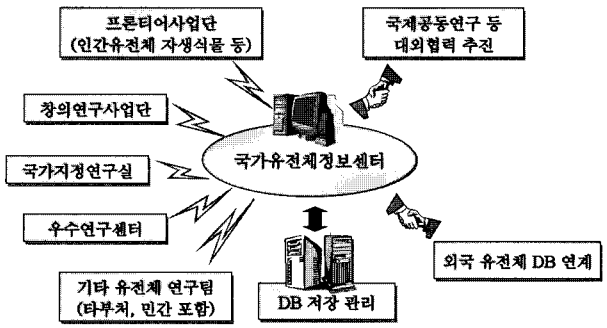

그림 1 국가유전체정보 협력 Network 구성도

이러한 설립취지에 따라서 국가유전체정보센터는 국제 경쟁력을 갖춘 생정보학의 리더로서, 첨단 생명공학과 산업발전의 견인차 역할을 수행하고자 하며, 국내에서 생산된 유전체 관련 정보의 효율적 집대성 시스템을 구축 및 운영할 뿐만 아니라, 국가유전체정보센터에서 연구 개발된 다양한 정보와 도구들을 관련연구자들이 자유로이 이용토록 제공하고자 한다.

3. 주요 추진사업

주요 추진사업의 기본방향은 크게 5가지로 나눌 수 있다.

첫째, 생정보 지원 분야는 국내 유전체관련 정보들의 수집 및 DB구축 그리고 해외 주요 DB의 mirroring을 통한 생정보 Portal Site를 구축하여 국내 연구자들에게 효율적으로 제공하는 유통시스템 개발이고,

둘째, 생정보 개발 분야는 대량 유전자 발굴, 기능예측, 비교 분석 등을 위한 고부가가치 2차 데이터베이스 개발이고

셋째, 생정보 연구 분야는 생정보학과 전산생물학의 연구개발 수행이며

넷째, 연구지원 및 교육 분야는 국내 연구자들에게 생정보학 교육 및 연구수행에 정보학기법을 제공하는 것이다.

다섯째는 위의 4개 사업을 바탕으로 한 국제적인 컨소시엄 참여 및 동아시아 지역의 생정보 데이터 지원 센터화이다.

4. 조직 및 인원

국가유전체정보센터는 생정보핵심개발실(InfoCore)와 생물정보이용 연구개발실(ResearchShell)로 크게 두 가지로 나뉜다. 인포코어분야 내에 여러 개의 능동적인 연구실과 팀 및 사무국 운영관리과가 구성되어 있다. 연구쉘은 첨단 대용량 연구 분야의 개척과 개발에 집중한다.

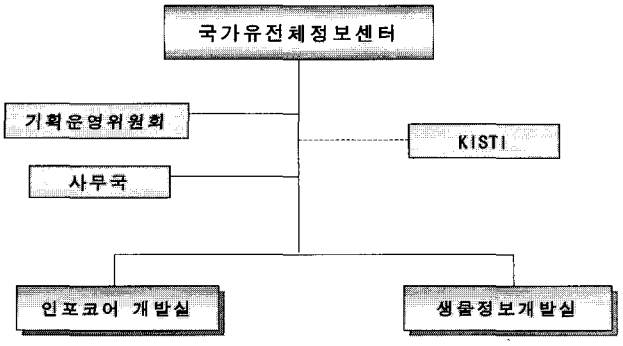

그림 2 국가유전체정보센터 조직도

연구원 구성은 생물물리, 화학공정, 화학, 생물학, 통계학, 전산학 등 다양한 전공분야들로 구성된 박사(Post Doc. 포함) 12명, 석사급 16명, 연수생 등 기타 17명으로 총 45명이 근무하고 있다(그림 2).

4.1 연구실 소개

4.1.1 생정보핵심개발실(InfoCore team)

생물정보핵심개발실은 유전체 분야 데이터베이스 구축 및 관련 연구기관간의 네트워크 구축을 주된 임무로, 첫째, 유전체 분야 연구기관들 간의 네트워크 및 데이터베이스 유통시스템 구축과, 둘째, 인간, 동물, 식물, 미생물의 유전체 종합정보 DB구축 및 공동 활용 기반구축, 셋째, 데이터의 수집, 저장, 조직화, 분석, 시각화하는 전산학적 환경구축, 넷째, System 운영 및 개발에 필요한 전산분야의 기반기술 확립 등의 업무를 수행하고 있다.

생물정보핵심개발실은 생명정보의 개발과 연구를 위한 각종 유용 유전자 대량 발굴 및 분석시스템 등 생정보학 핵심기술개발과 등록된 데이터 및 대형 연구 사업에서 얻은 데이터를 해외 공개데이터와 함께 분석한 유용한 고부가가치의 2·3차 데이터베이스 개발을 위해 4개의 팀(Comp-bio팀, Comparative Genomics팀, Biomedical Informatics팀, Proteome Informatics팀)을 구성하여 운영하고 있다. Comp-bio팀은 주로 유용 유전체 데이터의 수학/통계적 분석에 의한 정보와 데이터베이스 개발을 목표로 microarray를 이용한 질병관련 유전자 발현분석, EST data 분석, SNP & HapMap 분석, epigenomics 등의 연구를 수행하고 있으며, Comparative Genomics팀은 비교유전체학을 통하여 신약 및 의료기술 개발에 유용한 정보개발을 주요 목표로 비교유전체학 분석 소프트웨어 개발, Chimpanzee Genome Project, 모델생물체의 서열비교분석 등을 추진하고 있다.

4.1.2 Biomedical Informatics팀

Biomedical Informatics팀은 유전체정보 및 systems biology를 통한 의약용 생정보도구 개발을 주요 목표로 하여, 질병 관련 새로운 유전자의 탐색, shotgun 기술을 이용한 DNA 서열 해독분석, DNA 서열로부터 유전자위치 및 조절영역 탐색, cell/tissue genomics 등을 연구하고 있다. 그리고 Proteome Informatics 팀은 proteome의 구조 예측 및 단백질 상호기능 예측 연구를 주요 목표로 신약개발을 위한 유전자로부터 상호작용하는 단백질의 정보 분석, 단백질의 3차 구조 및 화학물질 간의 상호작용모델링, systems biology 연구를 수행하고 있다.

5. 주요 Service 내역(유전체정보 Portal Site: Bioportal.NET)

국가유전체정보센터의 홈페이지를 이용하여 국내에서 수집된 EST DB 및 인간유전체사업단의 KUGI DB 등을 이용할 수 있으며, 해외 주요 118개 DB의 데이터를 SRS를 통하여 신속하게 최신의 자료를 검색하여 활용할 수 있으며, 생물학자들이 가장 많이 사용하는 도구 및 링크 157종을 일목요연하게 정리하여 정보 분석도구를 쉽게 활용할 수 있다(그림 3).

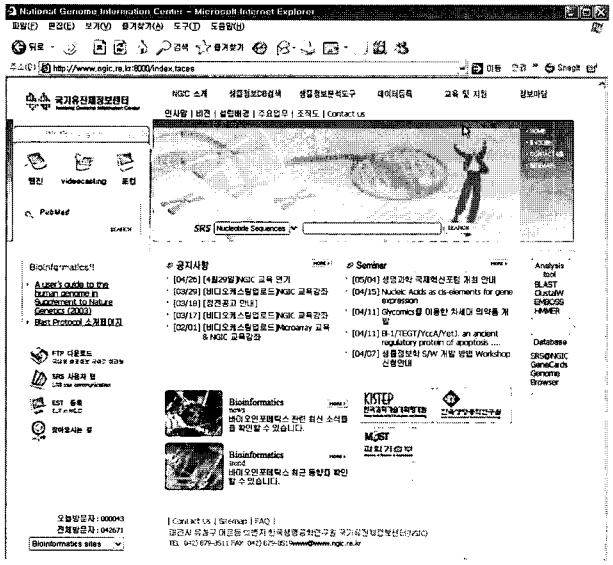

그림 3 국가유전체정보센터 홈페이지

NGIC 서비스의 궁극적 목표는 많은 다양한 생물정보를 각각의 도메인에 가장 특화된 형태로 분리된 상태로 보급하는 것이다. 이를 위해서, 바이오포탈(http://bioportal.net)을 만들었다. 이것은 열리고 자유로운 형태의 인터넷 정보교류형식을 취할 것이다. 이러한 바이오포탈의 모든 하부 도메인은 열린 (편집 가능한) 형태의 웹으로서, 그 예는, "뵤"라는 백과사전식 일반정보보급 서버를 보면 된다. 뵤는 http://bvio.com/에서 현재 보급되고 있으며, 많은 생물학자들의 지원이 필요하다.

이러한 열린 형태의 지식보급에 가장 적합한 형태는 Hypertext이며, 위키라고 불리는 프로그램을 사용하여 단백질백과사전 유전자백과사전 조류 백과사전 등 모든 생물학 관련 정보를 대용량으로 저장할 수 있는 바이오피디아 (http://biopedia.org) 라는 조직을 바이오포탈내부에 두는 형태를 띌 것이다.

현재 NGIC의 분석도구의 핵심은 SRS와 바이오인프라(BioInfra)라는 두개의 큰 축으로 표현될 수 있다. SRS (Sequence Retrieval System)는 GenBank 등을 포함한 생물 정보 분야의 주요 118개 DB를 통합하여 제공하고 있으며(세계 6위 규모)[1], XML 형식의 데이터베이스 등 다른 데이터베이스를 통합하기 위한 개발을 진행 중으로, Kugi[2], Kugipathway 등 자체 개발 DB를 SRS에 통합 운영하고 있으며, 자동으로 데이터베이스를 update하고 필요한 index를 재생성하기 위한 script 작성하여 최신의 data를 제공하고 있다. Blast, Fasta, HMMER, Clustal, EMBOSS[3] 등 생정보 분야에 기본적으로 사용되는 157개 해석용 도구가 통합 설치되어 있어 자신이 검색한 유전자 등에 대한 해석을 편리하게 수행할 수 있다(그림 4, 5).

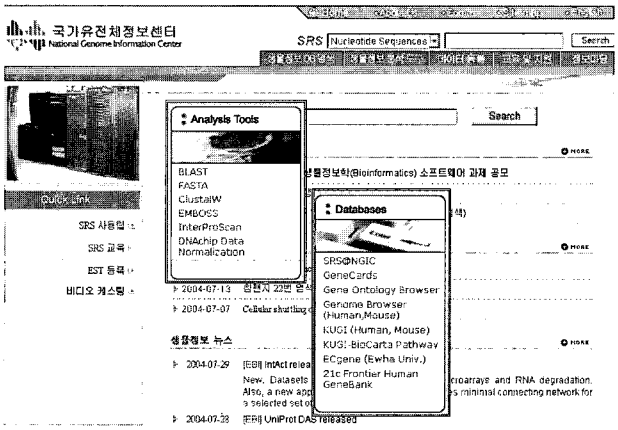

그림 4 국가유전체정보센터 분석도구 및 데이터베이스

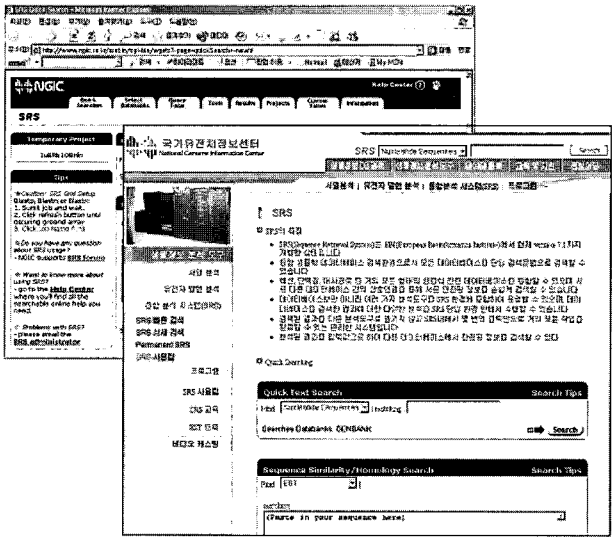

그림 5 국가유전체정보센터 SRS 서비스

다른 하나의 인프라 축인, 바이오인프라는 이미 가공되거나 초기단계의 정보의 보급이 아닌 가공기능을 함께 갖춘 일종의 파이프라인이다. 이것은 연구개발의 핵심 되는 정보를 모두 자동으로 지속적으로 돌리는 작업을 하는 것을 뜻한다. 이 바이오인프라의 핵심을 BioEngine 이라고 부르고 있으며, 이것은 전자동 생정보처리 기계라고 간단히 말할 수 있다.

5.1 바이오엔진

바이오엔진은 국내 연구자들에게 전 세계의 각종 생물학관련 데이터베이스 및 유용한 생정보학 프로그램이 원활하게 자동으로 제공되도록 할 계획이다. Bioengine의 한 부분은 가장 기초적인 FTP기능이다. 이것은 생정보 데이터는 대부분 용량이 크므로, NGIC FTP 사이트를 이용하면 신속한 다운로드가 가능하다 (그림 6).

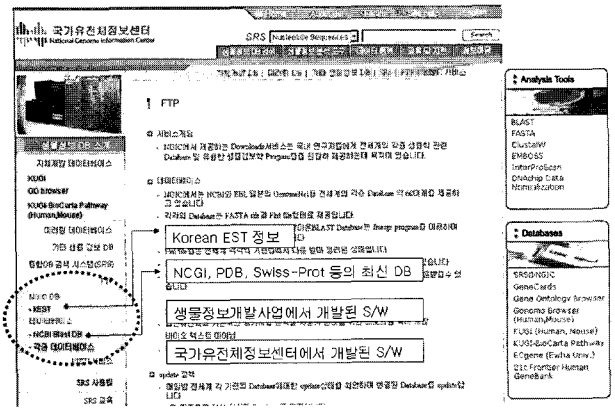

그림 6 국가유전체정보센터 FTP 서비스

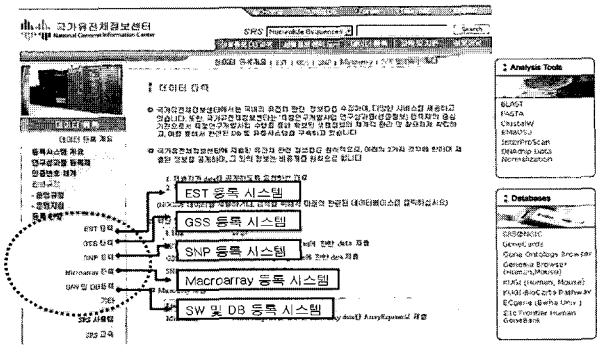

그림 7 국가유전체정보센터 Submission 시스템

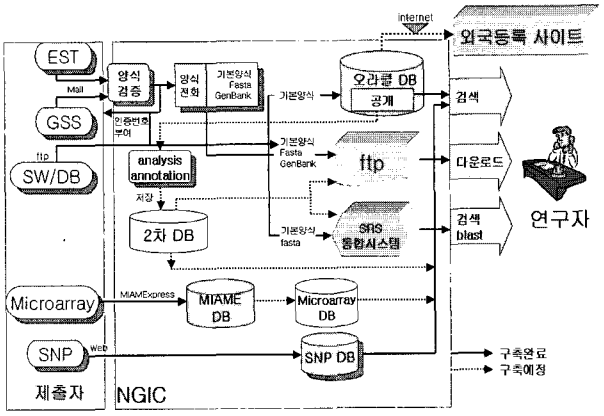

그림 8 등록시스템의 구성도.

자동화된 NGIC의 서비스의 한 기초적 기능으로 현재 만들어지고 있는 것은 자동 등록시스템이다. 현재 국내 생산된 EST 정보를 제출받아 데이터베이스(kEST)화를 완료하였으며, kEST에 대해서는 Web을 통하여 다양한 방법으로 검색할 수 있는 환경을 제공하고 있다(그림 7, 8).

현대 유전체학의 흐름은 현재, SNP의 연구에 초점이 맞추어지고 있다. 이것은 SNP가 인간의 가장 원천적인 질문인 우리는 어디에서 왔는가와 같은 것에 답을 줄 수 있기 때문이다. 이러한 유전체연구를 위한 도구로서, NGIC는 유전체 연구기법, 분석 도구 및 DB 개발 보급을 한다. 현재, SNP 정보 분석을 위한 SW 개발 및 데이터베이스 구축서비스를 하고 있으며, 또한 siRNA Design을 위한 알고리즘 개발, 유전자변이 분석 데이터베이스, 비교유전체학 분석을 위한 정보 분석용 SW 등을 개발 보급하고 있다(그림 9, 그림 10). 한국 Hapmap 프로젝트의 과기부 대표로서 NGIC가 참여하고 있으며, Hapmap 프로젝트의 결과를 정보처리적으로 분석 보급하는 임무를 포탈싸이트의 보급을 통해 할 것이다.

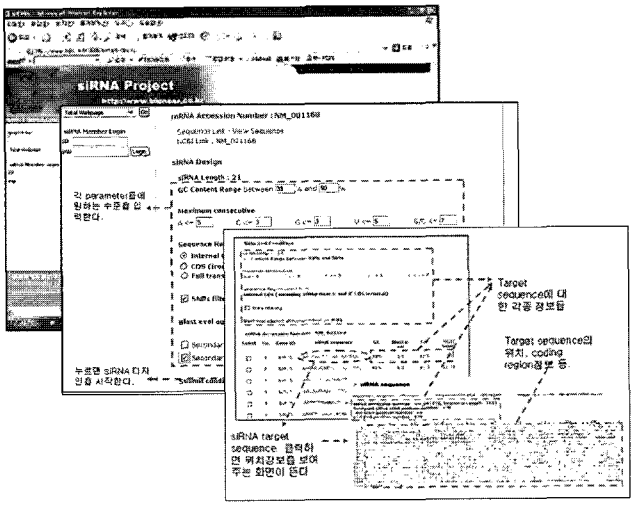

그림 9 국가유전체정보센터 siRNA Design 알고리즘 개발

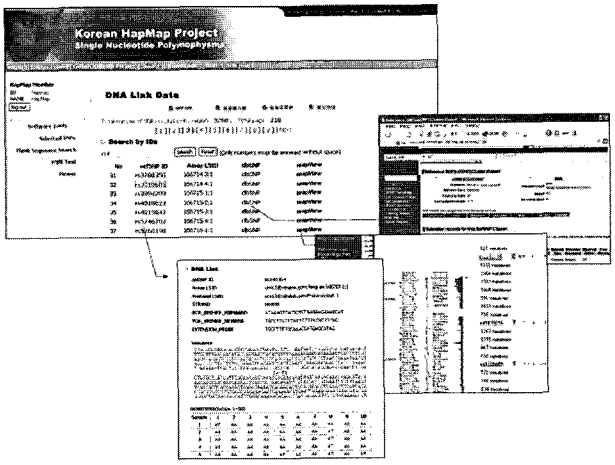

그림 10 국가유전체정보센터 Korean HapMap 데이터베이스 구축

6. 맺음말

생정보학이 미래 생물학의 핵심 역할을 할 것이라는 것은 명백하다. 현재에도 중요 단백질에 대하여 이름, 유전자의 기능, 반응 속도 등의 물리적 상수와 알려진 물리적, 유전적 상호 작용을 computer의 도움 없이 추적하는 것은 불가능하다. **향후 생물학의 연구방향이 묘사적인 생물학으로부터 정량적이며 예상 가능한 생물학으로 변천됨에 따라 새롭고 막대한 양의 data의 발생, 저장, 공유 및 이용을 필요로 할 것이다.**

모든 개개인의 유전체를 1일 이내에 100만원 이하로 해석이 되는 때가 5년 이내에 도래될 것이다. 이때에 국가적인 인프라와 충분한 수의 생명정보처리 회사가 준비되어 있지 않다면, 이러한 거대한 사회적 변화를 정보처리적으로 재빨리 소화를 할 수가 없게 되며, 국제적 경쟁에서도 뒤지게 된다. 국가유전체센터는 이런 다가올 정보처리의 기회 혹은 위기를 준비할 "10만 양병"을 제때에 양성하는 기관이라 할 수 있다. 따라서 **국가유전체정보센터는 생정보학의 국내 활성화와 유전체정보의 국내 산학연의 원활한 지원을 위하여, 유전체정보를 생산 관리하는 주체들 간의 데이터그리드 구성할 것이다. 그리고, 생정보 분석을 위한 알고리즘, 데이터마이닝, 분석 S/W 등을 연구하는 주체들 간의 응용그리드를 완성하여 국가유전체정보 유통 분석체제 확립에 중심체 역할을 수행하고자 한다.** 또한, 국가적인 인프라 차원에서의 국제협력을 통한 국내 생정보학 연구 분야의 국제화 추진에 이바지하고자 한다.

참고문헌

[1] Baker,P., Brass,A., Bechhofer,S., Goble,C., Paton, N. and Stevens,R :TAMBIS-transparent access to multiple biological information sources. In proceedings of International Conference on Intelligent Systems for Molecular Biology. AAAI Press.

[2] P.Carter,T.Coupaye, D. P. Kreil, and T. Etzold : Analyzing and Using Data from Heterogeneous Textual Databanks with SRS, Kluwer Academic Press(1998).

[3] S.Davidson, C.Overton, V. Tannen, and L.Wong : Biokleisli : A digital library for biomedical researchers, Journal of Digital Libraries (1996).

[4] T.Etzold and P.Argos : SRS an indexing and retrieval tool for flat file data libraries, Appl. Biosci(1993) 49-57.

[5] T.Etzold, A.Ulyanov, and P.Argos : SRS: Information Retrieval System for Molecular Biology Data Banks, Methods in Enzmemolog 226(1996).

[6] J.Hammer, H.Garcia-Molina, J.Cho, R.Aranha, and A.Crespo : Extracting Semistructured Information from the Web, Workshop on Management of Semistructured Data(1997).

[7] A.Y.Levy, A.Rajarman, and J.J.Ordille : Querying Heterogeneous Information Sources Using Source Descriptions , Proc. Of the 22nd Conf. On Very Large Data Bases (VLDB'96).

[8] KUGI: A database and search system for Korean UniGene Information," Jin Ok Yang, Yoonsoo Hahn, Nam-Soon Kim, and Sangsoo Kim, International Women's Conference on BIEN-Technology, Oct., 2003.

[9] Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

[10] Rice P., Longden I, Bleasby A EMBOSS: The European Molecular Biology Open Software Suite 2000, 16:276-277.

[11] Profile Hidden Markov Models. Eddy S Bioinformatics, 14:755-763, 1998.

[12] Pearson WR and Lipman DJ(1988) Improved Tools for Biological Sequence Comparison. PNAS 85:2444- 2448.

[13] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

박 종 화

1990~1994 영국 애버딘대학 생화학(학사)
1994. 10~1997. 6 영국 케임브리지대학 생정보학(석박사)
1997. 8~1997. 12 영국 케임브리지대학 박사후 연구원
1998. 1~1999. 4 미국 하버드 의과대학 박사후 연구원
1999. 4~2000. 12 유럽연합 생물정보학 연구소 박사후 연구원
2001. 1~2003. 4 영국 케임브리지대학 의학연구회 인간영양연구소 그룹리더
2003. 4~2005. 4 한국과학기술원 바이오시스템학과 부교수
2005. 5~현재 한국생명공학연구원 국가유전체정보센터 센터장
2005. 1.~2005.12 한국미생물학회 학술위원
관심 분야 : 생물정보학, 노화 현상을 궁극적으로 규명/교정(생정보적으로, 생물학적으로), 전산학, 프로그래밍, 데이터베이스, 생물의 진화를 포함한 모든 생물학적 문제들, 전산학적 철학문제들, 소프트웨어 공학적 문제들

12th Asia-Pacific Software Engineering Conference (APSEC'05)

- 일 자 : 2005년 12월 15일~17일
- 장 소 : Grand Hotel(타이페이)
- 주 최 : 소프트웨어공학연구회
- 내 용 : 논문발표 등
- 상세안내 : http://selab.csie.ncu.edu.tw/apsec05/