국가유전체정보센터의 역할과 비전

국가유전체정보센터 김상수

1. 서 론

2000년 인간게놈프로젝트의 초안 완성이 발표된 이후, 국내외에서 유전체 연구 및 생물정보학에 대한 관심이 고조되었다. 국내에서도 인간유전체기능연구사업단을 비롯하여 대규모 유전체 정보를 생산하는 사업들이 생기게 되었으며, 이들을 체계적으로 활용하여 다가오는 포스트게놈 시대에 능동적으로 대비하기 위해 국가유전체정보센터를 설립하기 위한 범부처적 합의가 있었다. 이에 따라 생명공학분야 중심기관인 한국생명공학연구원(KRIBB)을 주관기관으로 하고 IT 인프라기관인 한국정보기술연구원(KISTI)을 협력기관으로 하여 2001년 10월 과학기술부 지원으로 설립되었다.

2. 비전과 역할

2.1 비 전

국제 경쟁력을 갖춘 국내 생물정보학의 리더로서, 생명과학 및 산업 발전의 견인차

2.2 목 표

가. 국내 유전체 관련 연구주체들 간의 유기적 네트워크를 구축하여 생물정보 기술, 인력양성 및 교류를 촉진

나. 구축된 유전체 통합DB의 공동 활용 기반 구축하여 국내 연구자들에 의한 신속한 특허 획득 및 산업화를 위한 기반 제공

2.3 역 할

2.3.1 국내 유전체 데이터의 공인 등록 기관

최근 국내에서 정부지원으로 생산되는 유전체 정보는 급속히 증가하고 있다. 이들 데이터의 신속한 공개에 의한 공동활용을 위해서는 중앙에 공인된 등록기관이 필요하다. 미국 NCBI의 GenBank, 유럽 EBI의 EMBL, 일본 NIG의 DDBJ 등의 성공적 운영은 유전자 정보의 공유를 촉진하여 생명공학 발전의 밑거름이 되었다. 대부분의 생물학자는 연구의 구상단계에서 이들 데이터베이스를 검색하여 연구계획을 기획을 하는 것이 현재의 일반적인 상황이다. 집합된 유전체 데이터베이스의 중요성은 더 이상 강조할 필요가 없고, 다만 대부분의 국내 연구결과가 국제 DB에 등록되는 상황에서 국내에 독자적인 DB가 필요하냐는 점은 살펴볼 필요가 있다. 더더구나, 국내에서도 원데이터 생산자가 인터넷을 통하여 데이터를 공개하는 것이 보편화 되는 상황에서 말이다. 정부지원으로 생산되는 데이터에 대한 등록의 의무화는 데이터 생산자에 의한 데이터 공개를 촉진할 것이다. 또한 국가유전체DB와의 차별화를 위해서도 양질의 정보를 제공할 것으로 기대된다. 국가유전체정보센터는 개개의 분야에서 데이터 생산자와 질적인 경쟁을 하기는 쉽지 않을 것으로 판단된다. 따라서 데이터 생산자 그룹과의 적절한 상호협력 관계가 절대적이다. 국가유전체DB는 다양한 정보가 모이기 때문에 데이터 통합을 통한 상호 참조 기능에 초점을 맞춰야 할 것이다. NCBI의 Entrez나 EBI의 SRS가 이와 같은 기능을 잘 보여주고 있다. 정부지원 과제의 종료 후에도, 지속적인 데이터 서비스를 보장하는 면도 강조되어야 할 부분이다. 사용자 입장에서는 각 데이터 생산자의 서버를 찾아다니며, 각 서버별로 통일되지 않은 인터페이스를 통해서 검색을 한다는 것도 쉽지 않은 일일 것이다. 국내통합DB의 구축은 국제DB와의 결별을 의미하지는 않는다. 국가유전체DB는 국제기준에 맞는 데이터 검증 시스템을 채용할 계획이다. 그래야 국제DB와 데이터 교류에 문제가 없을 것이

다. 이를 위해, 일본의 DDBJ와의 협력을 통해 데이터 검증 및 서비스의 노하우를 도입할 것이다. 등록 대상은 염기서열, 발현정보, 프로테옴정보, SNP, 네트워크 및 경로 등의 데이터와 정부지원으로 개발된 각종 분석 S/W이다. 등록된 유전체DB는 그 자체로 검색 서비스를 제공하지만, 클러스터링 및 주석 달기 등의 가공 단계를 거친 2차 DB 형태로 검색 서비스를 제공하게 된다.

2.3.2 유전체 데이터 서비스의 포탈

유전체 연구의 발전에 따라, 전세계적으로 폭넓은 규모의 DB가 다수 제공되고 있다. 많은 국내 생물학 연구자들은 이와 같은 데이터베이스의 활용 및 데이터 분석을 위해 해외의 서비스를 사용하는 실정이다. 국내에서 수집된 DB의 적절한 활용을 위해서는 이와 같은 해외DB와의 상호 비교 및 참조가 필수적이다. 국내DB에서 해외DB로의 링크는 쉽게 달 수 있지만, 해외DB에서 국내DB로의 링크는 제한적일 수밖에 없다. 따라서 국내 생물학자들이 국가유전체DB 서비스를 우선적으로 접속하여도 불편함이 없고 필요한 정보를 쉽게 구할 수 있어야겠다. 따라서 해외DB와는 차별화된 서비스를 폭넓게 제공하도록 구축하여야만 국가유전체DB 구축은 성공할 수 있을 것이다. 특히 원스톱 서비스와 빠른 검색 속도는 필수적이다. 최근 접속자의 급속한 증가로 인해, NCBI의 BLAST서비스 등은 매우 느려진 편이다. 편의성 및 속도 이외에도, 신뢰성과 update가 중요한 이슈이다. 즉, 해외DB를 미러링 하여 제공할 경우, 최신 버전이 제때에 제공되고, down-time이 매우 적은 안정적인 서비스를 제공해야 한다. 유전체, 단백체 등의 다양한 형태의 데이터를 단번의 검색으로 동일한 환경에서 찾을 수 있는 효과적인 검색 시스템이 제공되어야 하는데, 현재 국가유전체정보센터에서는 EBI와 Lion Bioscience에서 개발된 SRS를 기본 시스템으로 하여 통합된 데이터 검색 시스템을 제공하고 있다. 비교적 설치와 운영이 간편하며, 비산업용으로는 무료로 사용할 수 있고, 기본적으로 제공하는 데이터베이스 외에도, 자체 데이터도 통합시킬 수 있기 때문에, 매우 좋은 시스템으로 평가되어 많은 기관에서 사용중에 있다. NCBI나 EBI처럼 데이터를 등록받는 기관도 아니며, 대규모 데이터를 생산하지도 않는 경우에도 훌륭한 검색 시스템을 개발하여 전 세계적으로 활용되는 경쟁력을 확보한 경우가 있다. 이스라엘의 와이즈만연구소에서 개발하여 운영 중인 GeneCards가 대표적인데, 공개된 데이터베이스들을 미러링 하여 가공한 후, 하나의 시스템을 통하여 제공하는 일종의 datawarehouse 기술을 적용한 것이다.

2.3.3 생물정보 핵심 기반기술 및 S/W 개발 및 이에 근거한 데이터마이닝 서비스

분석 알고리즘과 툴이 모두 해외에서 개발된 것이고, 자체적으로 개발한 것이 없다면, 국제 경쟁력의 확보에는 제한적일 수밖에 없고, 포탈서비스의 경쟁력 확보는 쉽지 않을 것이다. 따라서 고유의 경쟁력 있는 알고리즘과 프로그램에 기반한 서비스가 필수적이다. 연구인원의 소규모와 짧은 연구 경험을 고려하여 특정한 분야에 집중할 필요가 있다. 향후, 고급 알고리즘과 DB를 개발할 수 있는 연구인력 육성에 주력할 계획이다. 다양한 데이터베이스를 다양한 분석 툴을 적절히 결합하여 유용한 유전자를 발굴하는 것을 데이터마이닝이라고 한다. 날로 늘어나는 데이터베이스와 새로 개발되는 분석툴을 데이터마이닝에 쉽게 포함시켜 활용할 수 있고, 필요에 따라 분석 및 검색 방법을 자유자재로 변화시킬 수 있는 환경의 제공이 필요하다. 이런 시스템의 개발을 위한 노력이 국제적으로 진행되고 있다(www.biodas.org). 국내의 생물정보 연구인력이 협력하여 참여하면, 국제적으로 초기인 이 분야에서 국제 경쟁력으로 확보할 수 있는 절호의 기회라고 판단된다. 국가유전체정보센터는 국내 연구역량을 규합하는 중재자 역할을 담당할 계획이다.

2.3.4 생물정보학 기술보급을 위한 생물학자 대상 교육

지난 수년간에 걸쳐 많은 데이터베이스와 좋은 분석 툴이 등장했으며, 일반 생물학자로서는 이런 발전을 추적하는 것은 불가능에 가까워 보인다. 특히 인간게놈의 완성을 계기로 게놈브라우저가 개발되었으며, 이의 활용은 데이터마이닝의 파라다임을 근본적으로 바꾸는 계기가 되었다. 예를 들면, 종전에는 새로운 염기서열을 발굴하면, 각종 DB에 BLAST 등의 방법을 써서 비교하여 왔으나, 이제는 염기서열을 직접 게놈에 매핑한 후, 그곳에 주석된 유전자 정보들을 참조하면 모든 정보를 단숨에 얻게 된다. 오히려

너무 많은 정보를 제공하기 때문에 정보의 홍수에서 적절한 정보만 추리는 것도 쉽지 않다. 일반 생물학자들로 하여금 이와 같은 새로운 툴과 정보를 활용할 수 있도록 교육하는 것이 절대적으로 필요하며, 이를 국가유전체정보센터에서는 정기적으로 개최할 계획이다.

2.3.5 생물정보학 전문가 양성

생물정보학은 생물학의 전문지식 외에도, 전산학, 통계학, 화학, 물리학 등의 지식을 요구하는 종합학문으로서, 전문가가 부족한 것은 선진국도 마찬가지인 현상이다. 국내에서도 여러 대학에 생물정보학 관련 학과나 학제간 과정이 개설되고 있다. 국가유전체정보센터는 이와 같은 기관과 협력하여 전문가 양성에 노력을 함께 할 것이다. 특히 정보센터에 구축된 DB와 시설을 실습과정에 활용될 수 있도록 할 것이며, 졸업생들을 직업 창출에도 일정한 역할을 할 것이다. 생물정보학의 특성상, 학부에서부터 복수전공을 할 필요가 있다. 따라서 미래를 대비하여 학부생들의 관심을 유도해야 한다.

2.3.6 대형 국책 연구 사업 간의 시너지 창출을 위한 연결고리

국가유전체정보센터의 설립 목적 중의 하나가, 국내 유전체연구 주체 간의 유기적인 네트워크의 구축에 있다. 대형 국책 연구 사업단들과 협력의향서 체결을 통하여, 각 사업단 DB 개발에 참여하고 이들을 통합하면, 각 사업단의 연구결과를 종합하여 새로운 지식을 창출할 수 있는 시너지를 기할 수 있게 된다.

2.3.7 학-연-산 네트워크의 중심체

생물정보 관련 데이터베이스 및 분석 툴들은 일반적으로 비상용은 무료이나, 상용은 매우 고가이다. 또한 일반 생명공학 벤처기업들은 생물정보에 투자할 여력이 부족한 형편이다. 국가유전체정보센터에서는 유용한 정보를 마이닝한 결과를 국내 벤처기업들이 쉽게 활용할 수 있도록 제공할 계획이다. 생물정보 전문벤처기업들과는 고유 알고리즘 개발을 위한 공동연구를 진행하여, 산학연이 공동으로 참여하는 네트워크를 구축하여, 국내 생물정보학 기술 발전을 위한 국내 연구역량을 총동원하는 중재자의 역할을 담당할 것이다.

2.3.8 국제협력을 통한 기술교환 및 인력교류의 창구

날로 발전하는 생물정보 기술의 개발에 보조를 맞추기 위해서는 국제협력이 필수적이다. 일본과는 3년간에 걸쳐 한일과학기술포럼을 통하여 생물정보학 분야 협력이 국가 간에 합의되었다. 이미 한일바이오인포매틱스 교육과정의 실시 등의 성과가 있었다. 이를 계속하여 생물정보학 인재양성을 추구할 것이다. 영국과는 대통령의 방영을 계기로 생물정보학 분야 협력에 관한 협정서가 체결된 상태이다. 영국측 전문가의 방문에 의한 교육 및 공동연구가 계획되고 있다.

3. 국가유전체정보센터의 서비스

3.1 유전체데이터 등록 서비스

국가유전체정보센터에서는 국내에서 생산된 다양한 데이터의 등록을 받을 수 있는 시스템을 구축 중이다. 염기서열에는 다양한 종류의 데이터가 있다. 유전자 전장클론의 서열인 mRNA 혹은 cDNA, 그리고 이들의 부분인 EST가 대표적이다. BAC 클론과 같은 게놈서열도 있을 수 있는데, 이들의 경우는 end sequence, 혹은 contig가 있다. 미생물의 경우에는 전체 게놈이나 플라즈미드의 서열이 있다. 국내에서 가장 보편적으로 생산되는 데이터는 EST 염기서열일 것이다. 일반적으로 GenBank나 DDBJ 등에 등록하여 accession code를 부여받고 이를 논문에 적시하여 발표하는 것이 일반적이다. 국가유전체정보센터에서 추진하는 EST 데이터 등록 과정은 그림 1에 나타나 있다. EST 데이터는 하나의 cDNA 라이브러리에서 수천 개의 서열을 얻는 것이 일반적이기 때문에, 이들 서열은 동일한 정보를 중복하여 입력하게 된다. 이를 사용자 편의성 측면에서 도우는 프로그램인 ESTin을 정보센터에서는 무료로 배포하고 있다(그림 2). 이렇게 하여 업로드 된 EST 서열들은 정보센터 서버에서 검증 단계를 거치게 된다. 필수항목의 누락여부 및 포맷, 그리고 간단한 철자 등을 검증한 후, DDBJ에 보내지고, DDBJ에서부터 accession number를 부여받고 이를 사용자에게 전달하는 시스템이다. 이렇게 모아진 EST 데이터는 DB에 따로 모이고, BLAST를 이용하여 서열의 상동성 검사를 할 수 있고, 다양한 키워드 검색도 할 수 있다.

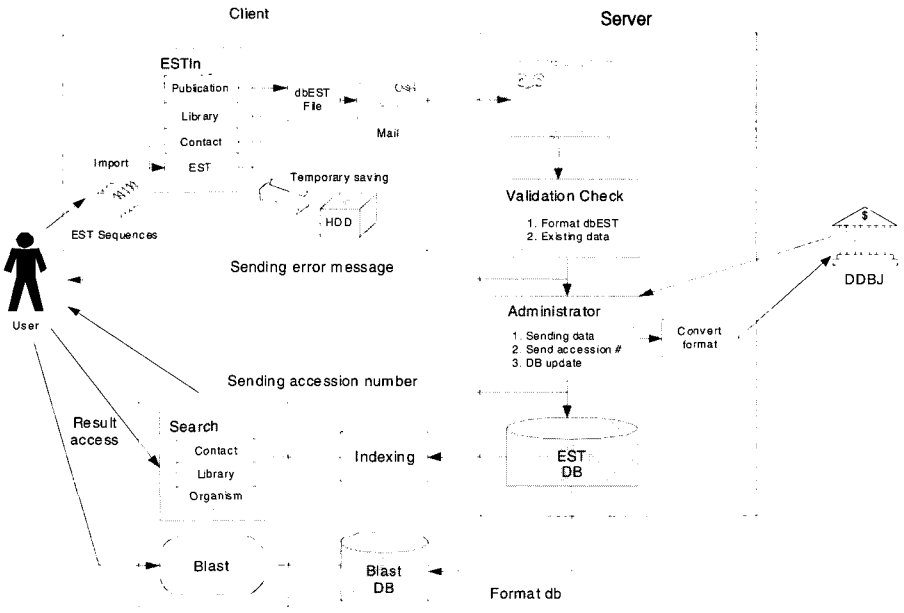


그림 1 EST 데이터의 등록 절차

염기서열 이외에도, SNP, DNA chip 또는 proteomics 관련 데이터의 등록도 추진하고 있으나, 아직 구체적인 데이터 등록 방침이나 관리 지침이 확정되지 않은 상태이다. 프론티어 사업 등에는 정보센터에서 직접 참여하고 있으며 관련 데이터는 이미 데이터베이스화 되어 있기 때문에, 이들 데이터가 공개되는 시점에는 정보센터를 통하여 서비스될 것이다. 일부 과제의 경우에는 정보센터에 등록은 하지만, 일정 기간 비공개를 원할 수 있을 것이다. 논문 등에 데이터가 발표될 때까지 공개를 미루는 것은 가능하다. 일단 등록된 데이터는 등록된 그대로인 1차 DB 형태 외에도, 적절한 주석 달기 과정을 거쳐 2차 DB의 형태로도 일반에 공개된다. 예를 들어 EST 데이터의 경우에는, 중복된 클론을 제거하는 clustering 과정 및 기능 부여, 유사 유전자 확인, 게놈 매핑 등의 단계를 거치게 된다.

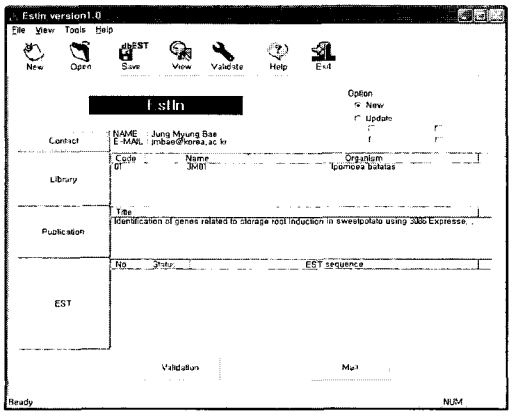


그림 2 국가유전체정보센터에서 무료로 배포하는 EST 등록 프로그램

3.2 유전체 데이터 포탈 서비스

국내에서 생산되어 국가유전체DB에 등록된 데이터 외에도, 해외에서 구할 수 있는 다양한 데이터베이스를 동일한 환경에서 제공하는 것은 매우 중요하다. 정보센터에서는 EMBL에서 개발된 SRS를 이용하여 유전체 정보를 통합하여 제공하고 있다. SRS는 기본적으로 GenBank, Swiss-Prot 같은 서열 정보, InterPro 같은 모티프 정보 등 다양한 정보를 통합하여 동일한 환경에서 제공한다. 여기에 국가유전체DB의 내용도 포함하여 한번에 검색이 가능하도록 할 것이다. SRS를 사용하는 장점 중의 하나는 EMBOSS와 같은 다양한 분석 프로그램과 쉽게 통합할 수 있다는 것이다. SRS로 제공되지 않는 해외데이터는 다양한 분류 체계를 거쳐서 링크를 제공한다. Nucleic Acid Research를 보면, 현재 수백 개의 분자생물학

관련 데이터베이스가 제공되고 있는데, 이들을 사용자의 기호에 맡게 재분류하여 제공한다.

3.3 유전체데이터 분석 응용 프로그램

유전체 데이터 분석에 일반적으로 사용되는 프로그램을 망라하여 제공한다. 최근 NCBI에서 BLAST를 사용하여 서열검색을 하려면, 상당한 시간이 지체되는 것이 일반적이다. 정보센터에서는 병렬서버와 KISTI에 갖춰진 초고속 클러스터 시스템을 이용하여, 빠른 서비스를 제공할 것이다. 특히 사용자수가 폭주할 경우에도 큐시스템을 이용하여 순차적으로 처리할 것이다. ClustalW와 같은 다중서열정렬 프로그램, pfam과 같은 모티프 검색 프로그램 등을 제공할 것이다. 앞서 언급한 EMBOSS와 같은 분자생물학 도구와 함께, 국책과제를 통하여 국내에서 개발된 프로그램들도 일반 사용자에게 제공될 것이다.

3.4 유전체 데이터 마이닝 서비스

날로 증가하고 새로이 개발되는 생물정보학 데이터베이스와 검색 툴을 일반 생물학자가 일일이 파악하여 사용하는 것은 쉽지 않은 일이다. 국가유전체정보센터에서는 주기적으로 교육을 통하여 분석기법을 전파할 것이다. 그러나 모든 사용자가 복잡한 데이터 마이닝을 직접 수행하여 유용한 유전자를 검색하리라고 기대할 수는 없다. 따라서 정보센터에서는 국내에서 수집한 데이터와 해외에서 수집한 데이터를 통합하여 다양한 각도에서 유용한 정보를 직접 마이닝하여 그 결과를 누구나 손쉽게 활용할 수 있도록 제공할 것이다. 예를 들면, 특정 암 세포에서 특이적인 발현 패턴을 보이는 유전자의 리스트, 인간게놈에서 새로 발견된 단백질을 모티프별로 제공하는 것 등이다.

4. 국가유전체 데이터베이스 구성도

국내에서 수집한 데이터와 해외에서 미러링한 다양한 데이터에 대한 통합 데이터베이스 구성도는 다음과 같다. 1차 DB 중에서 염기서열DB는 GenBank 형식으로 EST, mRNA, BAC end, genome 서열을 저장한다. 이들 정보는 clustering과 주석 달기 등을 걸쳐 Korean UniGene Information(KUGI) 시스템을 통해 서비스 된다. KUGI에서 genome browser로 연결할 수도 있고, genome browser에서부터 검색을 시작할 수도 있다. KUGI에서 검색된 유전자에 대해서는 해외 DB로의 링크를 제공할 뿐만 아니라, GeneCards, Stanford Microarray Database(SMD), SRS 같은 미러링된 DB에 연결된다.

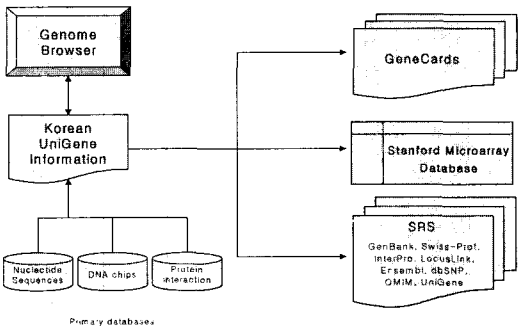


그림 3 국가유전체정보센터의 데이터베이스 구성도 진한색으로 표시된 DB는 해외에서 미러링한 DB들임

5. 국가 e-Bio 시스템(안)

SRS는 많은 데이터베이스를 동일한 환경을 통해서 제공하는 우수한 시스템이지만 몇 가지 문제를 안고 있다. 모든 데이터베이스를 텍스트 파일 형태로 하여 인덱스를 구축하여 검색한다. 유전체 데이터는 다양한 형태로 이뤄져 있는데, 이런 데이터의 특성을 잘 반영하기는 어렵고, 관련된 모든 데이터베이스를 한 곳에 미러링한 후, 인덱스를 만들어줘야 하는 등, 관리에 어려움이 많은 편이다. 한국처럼 초고속 네트워크 등 IT 인프라가 잘 갖춰진 특성을 살려서, 국내의 여러 팀이 각자의 특성에 맡는 데이터베이스를 전담하여 서비스 하고, 이를 검색하는 방법을 표준화한다면, 향후 기하급수적으로 늘어날 것을 대비할 수

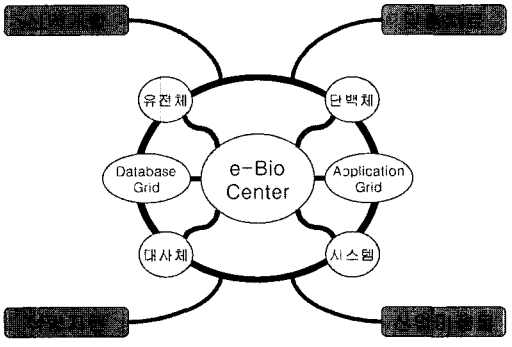
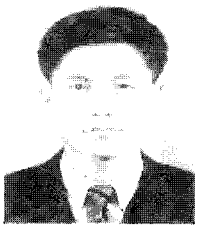


그림 4 국가 e-Bio 시스템 개념도

있는 robust하고 유연한 시스템이 될 것이다. 특히 사용자 인터페이스뿐만 아니라, XML 등에 의한 검색을 제공하여, 프로그램에 의해 자동화된 검색을 지원할 수 있다. 다음 그림은 이와 같은 개념도를 보여주며, 이의 실현을 위해 국내 역량을 총동원할 필요가 있다.

데이터베이스 그리드와 응용프로그램 그리드에는 국내의 연구실이 참여하여 각자 특정 도메인에 대한 전문기술을 개발하고 그 결과를 e-Bio 시스템을 통하여 네트워크화 하고 전국의 연구자에게 공개하는 것이다. 그 응용범위는 보건의료 분야는 물론이고, 식물과 미생물의 응용 분야까지 광범위하게 활용될 수 있다.

김 상 수



1977. 3~1981. 2 서울대학교 화학과 졸업(학사)
1981. 3~1983. 2 서울대학교 물리화학과 졸업(석사)
1983. 3~1986. 12 Iowa State Univ. 물리화학과 졸업(박사)
1986. 12~1988. 12 연구원
1988. 12~1995. 2 LG화학 Biopharmaceutical Design 팀장(선임연구원)
1995. 3~1998. 12 LG화학 Biopharmaceutical Design 팀장(책임연구원)
1999. 1~2000. 2 LG화학 Bioinformatics 팀장(책임연구원)
2000. 3~2003. 1 한국생명공학연구원 인간유전체연구실 책임연구원
2003. 1~현재 한국생명공학연구원 국가유전체정보센터장
E mail : sskimb@kribb.re.kr