제2회 통계유전학 아시아 워크샵 바이오정보학 기초 강좌
2nd Asian Institute in Statistical Genetics and Genomics

김영주
(한국생명공학연구원 의약유전체연구센터 책임연구원)

8월 14 - 18일, 2006
제주대학교 국제교류회관
바이오정보학 기초 (한국어 강의): 8월 14-15일

2003년 인간의 게놈서열이 거의 완결되면서, 이제는 뭔가 인간의 유전 변이를 이용한 질병 유전체 연구가 가능하게 되지 않았을까 하며, 질병원인 유전자 규명을 위한 연구 방법 및 실제적 data 통계분석 방법에 대해 큰 관심을 가지게 되었다. 이와 발 맞추어 2005년도에 "국립보건연구원 유전체센터와 질환군별 유전체 협의회"가 공동으로 주관하여 North Carolina State University의 통계유전학 교육 프로그램을 1주일간 한국으로 옮겨와서 처음 개최하였는데 아주 좋은 평판을 받았다. 다만 작년 수강자들의 의견이, 영어로 접하는 수준 높은 통계유전학으로 일관하기 보다는, 약간의 기초적인 한국어 강좌가 있었으면 좋겠다는 의견이 많아서 올해에는 기초 생물통계학, 질병유전체 분석기법, 바이오정보학 기초를 더불어 개설하였다고 들었다. 올해 제2차 국제 통계유전학 워크샵은 국립보건연구원 유전체센터, 보건의료유전체협의회 및 한국유전체학회가 공동 주관하여 제주대학교 국제교류회관에서 8월 14-18일에 개최하였다. 여기에서는 전체 워크샵 중에서도 한국어 강의로 진행된 바이오정보학 기초 (강의는 이도헌, 김동섭 KAIST 바이오시스템학과 교수와 김영주 생명공학연구원 책임연구원이 맡았다.)에 대하여 얘기를 해볼까 한다.

이도헌 교수의 강의는 분자생물학의 공인된 정설이 무엇인지부터 시작되었다. 살아있는 생물은 생명현상을 위하여, 세포 핵 안에 있는 DNA에서 시작하여 전사체인 RNA가 합성되고, 핵을 탈출한 RNA가 리보솜을 만나 단백질을 만들며, 지질, 탄수화물, 포도당과 같은 여러 대사산물을 만들게 된다. 문제는 인간이 가지고 있는 60~100조 개 각각의 세포 안에 46개의 염색체가 있고, 그 안에 2 m 길이의 DNA가 있으며, 그 안에 30억 개의 DNA 염기들이 있으며, 또 생명현상을 나타내는 유전자가 약 3만여 개 있다는 것이다. 이와 같이 현대의 생물학 연구는 바이오정보학이 없이는 대량의 정보 분석과 깊이 있는 연구가 거의 불가능한 실정임을 강조하였다. 이 바이오정보 기초 워크샵을 통하여 바이오정보학 (bioinformatics)의 기본 개념과 유전체학 (genomics), 전사체학 (transcriptomics), 단백질체학 (proteomics), 대사체학(metabolomics), 시스템 생물학 (systems biology)에 이르는 각 분야별 기본 정보를 정리할 수 있었다.

바이오정보 마이닝 기술을 이용한 마이크로어레이 칩 데이터를 분석하기 위한 강의는 유태우 조교선생이 국내 바이오정보 회사에서 개발한GenPlex소프트웨어를 이용한 컴퓨터 실험을 직접 같이 할 수 있어서 더욱 쉽게 다가올 수 있었다(그림 1). 마이크로어레이 스캐너로부터 얻은 이미지 파일로부터 스팟을 가려내고, 스팟의 밝기에 따라서 오차를 최소화하기 위한 정규화 과정을 거쳐서 여러 가지의 군집분석을 하면

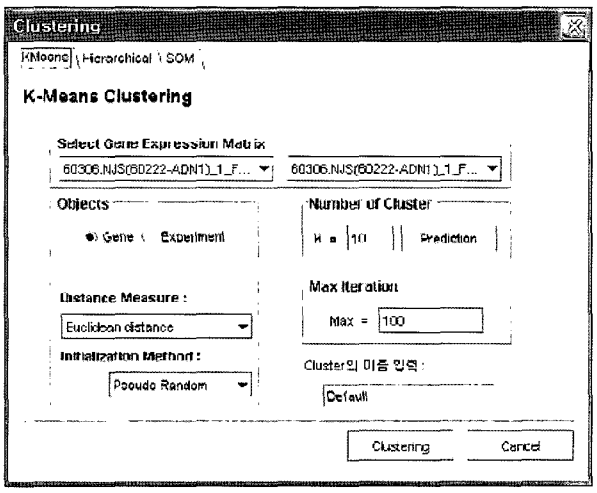

그림 1. 마이크로어레이 분석실습에서 사용한 GenPlex 프로그램

비로소 유의한 정보를 얻을 수 있게 되는 것이다.

이도헌 교수의 강의는 요즘의 뜨거운 연구 분야가 되고 있는 시스템생물학으로 옮겨갔다. 대사물질 경로와 질병의 경로 추정과 같은 연구가 가상적 세포 또는 E-cell 시스템을 이용한 시스템생물학에 의해 더욱 활성화될 수 있으며, 또한 역공학(reverse engineering)에 의하여 더 잘 분석될 수 있다는 것이다. 구성된 생체시스템을 잘 이해하기 위해서는 그 시스템을 구성하고 있는 부품, 구조와 역학을 잘 관찰하면 된다는 것이 역공학이고, 이를 이용하면, 세포공학, 유전자치료, 그리고 신약개발이 가능해진다는 것이다. 추상적이기만 하던 시스템생물학 기술은 나도균 조교선생이 제공한 Cytoscape, Monet, ARACNe 등의 컴퓨터 실험으로 훨씬 손에 닿게 이해를 할 수 있었다(그림 2).

다음 날, 바이오서열분석 및 단백질 구조기능 예측에 관한 강의를 맡은 김동섭 교수는 바이오정보학의 기초라고 할 수 있는 여러 바이오정보 데이터베이스의 소개, 서열비교법, 알고리즘, 컴퓨터 프로그래밍과 이들을 이용한 단백질의 구조와 기능 예측에 대하여 잘 이해할 수 있게 천천히 설명을 이어나갔다. NCBI, ExPASy, EMBL, PDB 등의 데이터베이스 사이트에 직접 인터넷 연결을 해보기도 하였고, DNA와 단백질

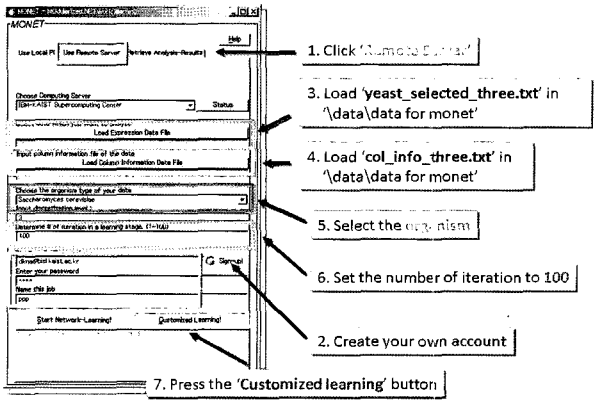

그림 2. 시스템생물학 실습에서 사용한 MONET 프로그램

단계에서 서열정열을 왜 해야 하는지, 그리고 서열정열을 위하여는 전역정열 (global alignment)이 있고, 국부정열 (local alignment)이 있는데, 각각 어떤 알고리즘이 널리 쓰이는지, 장단점에 대하여 자세히 알려주고, 간단한 예를 가져와서 알기 쉽게 설명을 하였다. 서열정렬법의 가장 기본적인 알고리즘인 dynamic programming을 알기 쉽게 설명하였으며, 수강생들에게서 그 동안 궁금해하던 서열정렬 알고리즘을 비로소 이해하게 되었다는 반응이 많음을 느낄 수 있었다. 이어서 Blast와 다중서열정열법에 대하여 설명하고 점진적 정열법인 ClustalW에 대하여 설명을 하였고, 단백질 구조 비교를 하기에 더 알맞은 방법인 Psi-Blast에 대해서도 알고리즘을 설명하여주었다. 바이오정보학의 기초적인 원리와 데이터베이스에 대한 소개를 마친 후, 수업은 단백질의 구조와 기능을 예측하기 위해 필요한 여러 바이오정보학적 도구들에 대한 내용으로 옮겨갔다. 앞에서 소개한 Blast, Psi-Blast, Swiss-model 등을 이용하여 단백질의 구조론 방법을, 구체적인 예를 보여가며 설명했다.

SNP 해석과 인간유전체 변이를 맡은 김영주 박사는 인간유전학의 목표를 DNA 변이가 인간의 표현형에 어떻게 영향을 미치는지 파악, 이해하고 결국 어떤 DNA 변이가 관련되는지 찾는 것이라고 정의하였다. DNA의 위치를 찾기 위한 많은 마커들이 알려져 있는

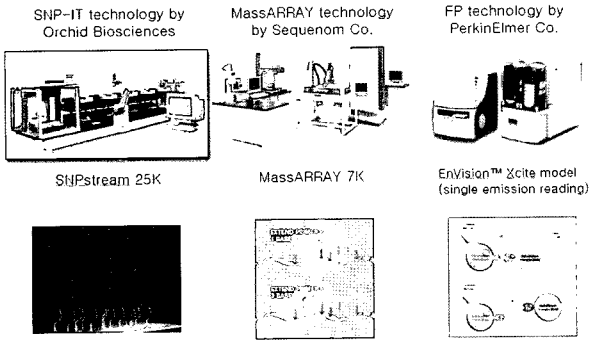

그림 3. 대량 SNP 데이터생산 시스템

데, 그 중에서도 최근 SNP가 각광을 받고 있으며, 그 이유는 유전적으로 좀 더 안정적이고, 약 1,000 bp 당 하나씩의 매우 풍부한 밀도를 가진 마커이며, 단백질 코딩 영역에서도 존재하고, 무엇보다 실험의 자동화를 통한 경제적인 대량 데이터생산이 가능하다는 점이다(그림 3). 단점으로는 SNP 자체가 대립쌍(allele)이 소위성(microsatellite)처럼 다양하지 않고, 질병 등에 적은 영향을 끼친다는 점이다. SNP 데이터로서예를 보여주고, LD(linkage disequilibrium) 지도, 반수체(haplotype), tag SNP 등의 예와 구하는 방법을 보여주었다. 특히 직접 인터넷에 연결하여 FESD(functional element SNPs database), SNP@Domain (단백질 도메인 구조 안에서의 SNP 정보), D2GSNP (인간질병 연관 SNP 정보)를 볼 수 있어 좋았다. 질병 연관 연구는 미국 NIH의 GAIN, 영국의 대규모 당뇨연관 연구, 일본의 30만 코호트 연구 등 많은 프로젝트를 낳고 있으며, 암에 있어서는 향후 5년~10년 안에 정복되지 않을까하는 조심스런 전망도 나오고 있다. 더 나아가 개인맞춤약, 집단맞춤약의 시대가 활짝 열릴 날이 아주 먼 미래가 아니라는 것을 강조하였다.

바이오정보학은 생물과 전산의 융합학문이며, 생물데이터를 데이터베이스화하고, 검색과 및 도구를 만드는 일을 하고, 시스템스와 역공학을 통한 생체시스템을 이해하기 위한 학문이다. 이 워크샵을 통하여 바이오정보학 (bioinformatics)의 기본 개념과 유전체학 (genomics), 전사체학 (transcriptomics), 단백체학 (proteomics), 대사체학 (metabolomics), 시스템 생물학 (systems biology)에 이르는 각 분야별 바이오정보기술과 DNA칩 데이터 분석, 신약개발을 위한 바이오정보 인프라와 개인 맞춤의학을 위한 유전적 다양성 (genetic variation) 분석에 대하여 이해할 수 있었다. 아울러, 유전자 검색, 단백질 구조분석, SNP 해석 등을 위한 대표적인 바이오정보 소프트웨어의 사용기법을 소개받아 바이오정보학에 대한 많은 이해를 가져올 수 있는 기회가 되었다.