EST를 이용한 Tissue expression 분석 시스템 구축

생명공학연구소 허철구 · 박경희
이화여자대학교 임소형
배재대학교 신민수
생명공학연구소 고성호

1. 서 론

사람의 유전체는 약 7만개에서 10만개의 유전자로 구성되어 있다는 것은 일반적으로 알려진 사실이다. 사람의 표현형은 이러한 모든 유전자들의 발현에 의해 결정되고, 각각의 기관 및 조직에서 나타나는 생명 현상과 관련된 기능들은 이들에게서 특이적으로 발현되는 유전자의 세트에 의해 조절된다. 지난 십여 년간 분자생물학적 기법의 발달과 Human Genome Project의 성과로 인해 사람의 유전자 중 많은 부분이 밝혀져 왔으며, EST 분석은 그 대표적인 연구 방법이다. EST란, 발현되는 유전자의 일부분을 일컫는 용어로서, 여러 인체 조직과 기타 미생물의 유전자를 찾아내는 연구를 통해 그 유용성이 판명되었다. EST는 임의로 선택된 cDNA 클론의 염기서열을 3' 또는 5' 끝에서 단 한 차례 읽은 것(single-pass)으로, 300-500bp 정도의 유전자 단편이다. 즉, EST는 cDNA 전체의 염기 서열이 아닌 그 일부분을 의미한다. 더욱이, 각 클론에 대해서 자동화된 기계를 단 한 번 사용하여 그 서열을 결정하는 데다가, 생물학적 검증을 거치기 전의 결과물이라는 특성 때문에 다소 부정확하다. 하지만 EST는, 새로운 유전자의 발굴, 질병에 관련된 유전자 집단의 새로운 구성원 발견, 조직간에 차등적으로 발현되는 유전자의 클로닝, 유전체 서열과의 비교를 통한 복잡한 유전체 서열의 분석 등 다양한 연구의 재료로 이용되고 있다. 이런 무작위적 추출의 개념으로 생산된

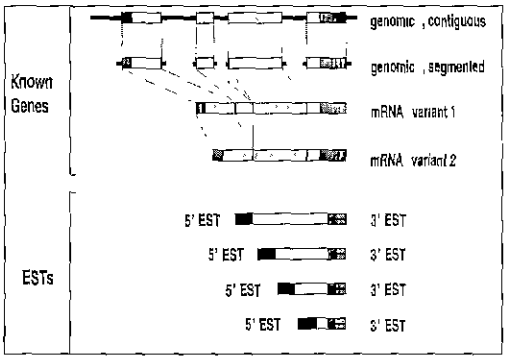

그림 1 EST와 genomics DNA와의 관계
(Ref: *Nature Genetics* 10:369-371; 1995)

데이터를 체계적으로 분석하고, 전체 유전자를 예측할 수 있는 시스템을 갖춘 환경에서 실험 및 연구를 수행하는 것이 필수적인 실정이다. 또, 자료의 양이 적을 때에는 - EST의 개수가 1,000개 미만 - 개인용 컴퓨터를 활용하여 분석하는 것이 가능하나, 이 경우에도 수작업이 약 80%이상 필요하다. 데이터베이스 구축와 상동성 분석 시스템(예 : BLAST 시스템), EST Clustering 시스템, Contig Assembly S/W 등 EST 단편을 분석하기 위한 절차는 매우 복잡하고 지루하며 반복적인 일들을 요구한다. 이미 선진국에서는 오래 전부터 유전체 관련 연구기관을 중심으로 많은 투자를 하는 등 EST 관련 연구와 생물정보 분석 시스템 개발에 상당히 주력해 왔다. 또한, EST 분석 절차에 따른 주요 분석 소프트웨어를 데이터베이스와 연동하여 자동화된 시스

템을 구축함으로써 실험자들에게 효율성과 시간 절감에 많은 효과를 주고 있다 반면 국내에서는, 대용량 EST 분석 시스템을 갖추고 실험을 하는 연구 기관이 현재까지 없으며, 전체적인 시스템을 구축하는 데 고가의 컴퓨터와 DBMS, 전문 인력 확보 등 여러 가지 어려움이 많기 때문에 대학 내 연구실 등에서 짧은 시간 안에 EST 분석 시스템을 구축하지 못하는 어려움이 있다. 따라서 본 연구에서는, 실험자들이 EST 단편을 확보했을 때, 그와 관련되어 이미 밝혀져 있는 유전자 서열과 조직별 유전자 발현(expression) 정도를 일목요연하게 보여주는 동시에, 이 모든 자료들을 쉽게 분석할 수 있는 시스템 구축 방법을 제시하고자 한다.

2. EST 분석을 위한 절차

우선 EST 데이터를 입력받아 FASTA 형식의 데이터베이스와 GenBank 형식의 데이터베이스를 구축한다. FASTA 형식의 데이터를 입력받은 다음, 중복된 서열을 제거하기 위하여 ICA Tools 시스템에서 제공하는 EST Clustering S/W를 이용하여 주어진 서열을 대표(parents) 서열과 종속(child) 서열로 구분한다. 그 중에서 대표 서열만 추출하여 다시 데이터베이스에 저장한다. CAP(Contig Assembly Program)을 이용하여, 대표 서열 중 가급적 긴 EST 단편을 찾아내어 미국의 NCBI(National Center for Biotechnology Information)에서 제공하는 단일 유전자 데이터베이스와 비교한다. 동시에 NCBI에서 제공하는 UniGene Database 형식을 발현되는 조직별로 비교하기 위하여 원시 자료를 다시 분류하여야 하는데, 이 과정을 위하여 UniGene 데이터 필드 가운데 EXPRESS 필드에 있는 자료를 조직별로 재분류한다. 기존의 방법을 사용하는 경우, 실험자들이 EST 단편을 가지고 UniGene 데이터에 대해 BLAST를 이용한 상동성 검색을 하면 각 조직별로 정확하게 구분되어 검색 결과가 추출되는 것이 아니라 상동성을 나타내는 서열들이 발현되는 조직에 관계없이 단순히 나열된다. 발현되는 조직과 관련하여 체계적으로 분석이 이루어진 다음에는, 이 EST 단편들을 단백질 서열로 변환시켜 그 기능을 예측한 결과를 실험자들에게 제공하고자 한다. 본 연구에서는, 조직별로 자료를 재분류하여, 실험자가 EST 단편의 서열을 입력하였을 경우 조직별로 발현되는 정도를 일목요연하게 보이고자 하였다. 이 자료를 바탕으로 주어진 유전자를 단백질 서열과 비교, 검색하여 기능을 유추하는 과정에는 SWISS-PROT - 스위스에서 운영중인 단백질 서열 데이터베이스 - 데이터베이스를 활용하였다. 이를 위하여 사용된 분석 도구는 일반적으로 BLAST였으나, 본 연구에서는 보다 정교한 결과를 얻기 위하여 FramePlus를 활용하였다. 이는 EST 데이터 자체가 annotation 에러, sequencing 에러, reading frame 에러, 재정렬 에러, contaminants 등 많은 문제점을 내포하고 있기 때문에 단백질 서열과의 비교 검색 시 이러한 여러 가지 문제점을 고려한 분석 기법이 필요하기 때문이다.

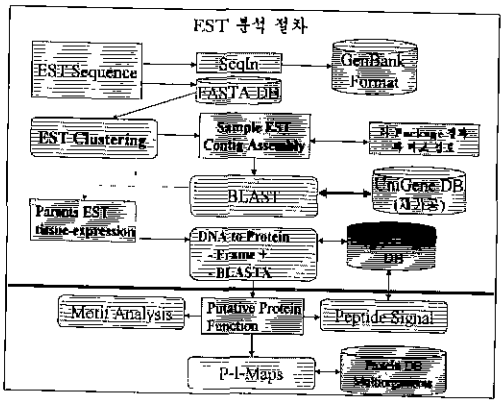

그림 2 EST 분석을 위한 데이터 흐름도

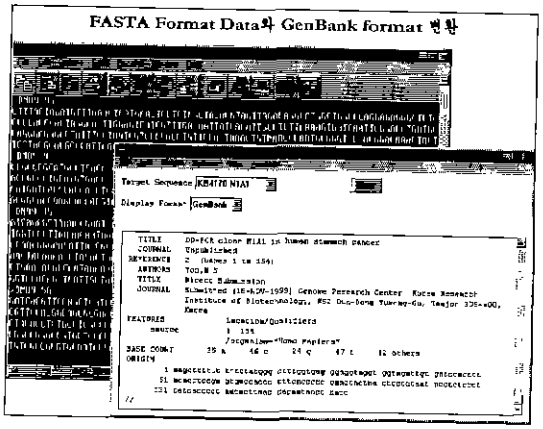

그림 3 FASTA형식의 자료와 GenBank형식으로의 변환 자료

3. 데이터베이스 구축 및 분석

지금까지는 EST 분석을 위한 일반적인 절차에 대하여 간단하게 살펴보았다. 이 장에서는 구체적인 방법론을 살펴보고 그 효율성을 살펴보기로 한다. EST 데이터는 FASTA 형식으로 제공되기 때문에 이를 NCBI에서 제공하는 SeqIn 프로그램을 이용하여 GenBank 형식으로 전환하였다. 즉, EST 데이터를 데이터베이스화 하기 위해서는 FASTA 형식의 데이터베이스와 GenBank 형식의 데이터베이스가 모두 필요하다. FASTA 형식의 데이터는 UniGene 자료와의 상동성을 검색하거나, EST clustering 결과를 재저장할 때, 그리고 clustering 결과를 인터페이스를 통하여 검색할 때 매우 유용하다. 그림 4 (a)에서는 FASTA 형식의 테이블과 GenBank 형식의 데이터베이스 사이의 관계를 보여주고 있으며, EST 단편들을 clustering하기 위한 테이블과의 관계를 설명한 모형, 그리고 UniGene DB를 구축하기 위한 테이블간의 관계를 나타내고 있다. UniGene은 인간, 생쥐(mouse), 쥐(rat)로부터 유래한 GenBank의 염기 서열을 유전자를 기준으로 서로 중복되지 않도록 클러스터화 한

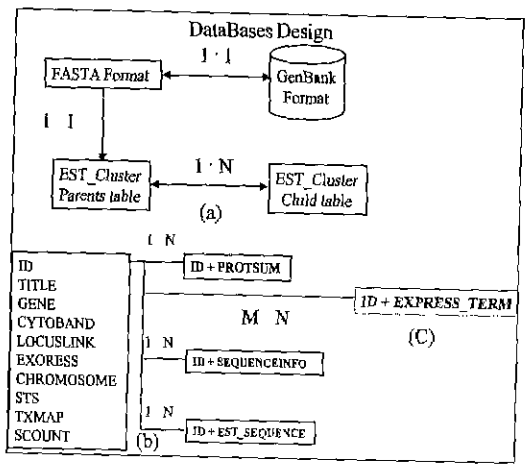

그림 4 (a)의 경우 EST 단편을 저장하기 위한 구조이며, (b)의 경우 UniGene Data를 RDB에 저장하기 위한 테이블 구조이며, (c)의 경우는 UniGene Data를 EST단편과 Tissue expression별로 분석하기 편리하기 위하여 재구성 한 구조이다.

데이터베이스이다. 각 UniGene 클러스터는 해당 유전자를 대표하는 염기 서열과, 발현되는 조직, 염색체상의 위치 등에 대한 정보를 포함하고 있다. 이는 이미 알려진 유전자뿐만 아니라 EST서열에 대해서도 UniGene 클러스터를 만들어 두고 있기 때문에, 새로운 유전자의 발굴과 유전자 지도 작성 등의 연구에 반드시 참조하여야 하는 자료이다. GenBank 형식의 자료를 데이터베이스화 경우에는 실험자가 원하는 데이터를 검색한 후 상동성 비교를 필요로 한다면, 자료를 복사하여 검색하지 않고 즉시 NCBI의 Advanced BLAST 시스템을 활용할 수 있도록 하였다(그림 5). EST 단편 자료는 짧은 서열로 이루어져 있기 때문에 많은 부분이 중복되어 나타난다. 이러한 데이터에서 EST Clustering시스템을 활용하여 중복된 서열을 제거하고 대표(Parents) 서열과 종속(Child) 서열로 구분한 다음, 대표 서열만 추출하여 데이터베이스에 다시 저장한다(그림 3 (a)는 테이블 관계를 나타내고 그림 6은 실행결과를 나타낸다.). 초기 자료인 T-세포 관련 EST 단편은 1,383건 이였으나 클러스터링 후에는 930건으로 그 수가 줄었다. 이는 약 30%의 자료에 중복이 있었다는 사실을 증명한다. 클러스터링 과정을 거친 EST 서열 중에서 대표 서열만을 추출하여 서열 연결 작업(Contig Assembly)을 실행하였다. 이 때 사용된 S/W는 CAP이다. CAP 프로그램은 두 개 이상의 염기서열을 상호 비교하여 연결시키며, 다수결의 원칙에 따라 염기 서열 결정에 있어서의 실수(error) 등을 교정할 수 있다. 이 프로그램을 이

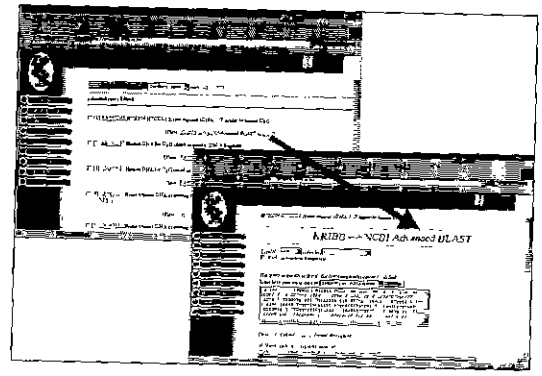

그림 5 EST 서열을 검색한 후 BLAST로 직접 연결 기능 제공

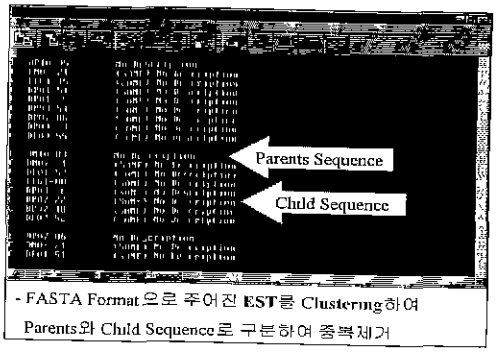

그림 6 Parents서열과 Childs서열 분석 예

용하여 얻어진 염기 서열은 EST 클러스터링 이후의 대표 서열 단편에 비해 어느 정도 길어지게 된다. 이 염기 서열을 가지고 이번에는 blastn을 이용하여 상동성 검색을 수행하였다. Blastx를 사용하여 단백질 관련 기능 분석을 하는 것이 일반적이나, 본 연구에서는 그 방법을 사용하지는 않았다. CAP 프로그램을 수행한 결과를 가지고 상동성 검색만을 거쳐 단백질의 기능을 유추할 수도 있다. 하지만 본 논문에서는, CAP 프로그램에서 나온 긴 EST 유전자를 가지고 UniGene 자료와 비교하여 어느 조직에서 주로 발현되는가를 먼저 살펴본 다음, 해당하는 단백질의 기능을 유추하기 위한 시스템 구축이 목적이므로, 기존의 UniGene 원시 자료(Flat-File)를 적합한 데이터베이스 형태로 재가공 하였다. 그 결과, 그림 4 (c)와 같은 형태로 구성되었다. 또 UniGene 데이터베이스에서 조직별로 유전자가 발현되는 정도를 알아보기 위하여 자료를 12가지 조직에 대해 분류하였고, 그 결과는 그림 7에 나타내었

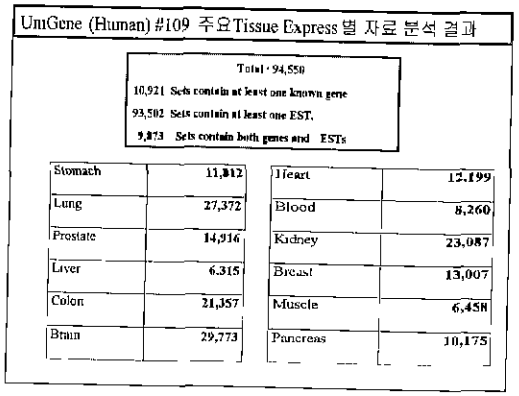

그림 7 UniGene자료를 Tissue별로 재구성

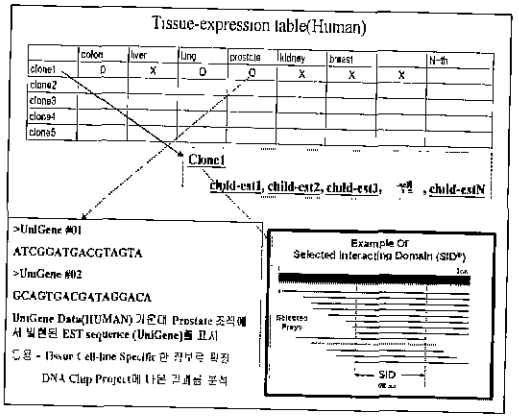

그림 8 EST단편을 가지고 나타나는 Tissue별 빈도 수

다. 즉, 이 자료를 바탕으로 상동성 검색을 함으로써, 실험자가 확보한 EST가 어느 조직에서 가장 많이 발현되는가를 그림 8에서와 같이 쉽게 알아볼 수 있다. EST 클러스터링 데이터를 가지고 기능을 유추하고자 할 경우, 지금까지는 주로 NCBI에서 제공하는 blastx 시스템을 이용하여 단백질과의 상동성 검색을 수행하였으나, 본 논문에서는 이스라엘의 Compugen 회사에 개발된 FramePlus를 활용하였다는 점이 특징적이라 할 수 있다. 이는 EST자료와 단백질 서열비교 시 프레임 이동과, EST 단편의 특징인 서열 에러를 고려한 방법이었다. Blastx 시스템에서 얻어지는 결과와 FramePlus 시스템을 사용하였을 때 얻어지는 결과는 같은 SWISS-PROT 데이터베이스를 사용하더라도 차이가 있다고 알려져 있다. 본 연구에서는 실험자들이 의미 있다고 판단되는 EST 단편을 가지고 조직 발현의 정도를 쉽게 알 수 있도록 하는 데에 중점을 두었다. 또한 EST 서열과 단백질 서열을 비교할 때 BlastX 대신 FramePlus를 사용한 또 다른 근거는, 질이 다소 떨어지는 EST 서열 자료를 처리하더라도 FramePlus가 기존의 알고리즘에 비하여 민감도(sensitivity)가 매우 높게 측정되기 때문이다. 일반적으로 민감도는 다음과 같이 정의한다. $Sn = \frac{TP}{TP + FN}$, TP(원래 positive이면서 positive라고 예측된 것), FN(원래 positive인데 negative라고 예측된 것). 이 값이 클수록 EST와 단백질 서열을 비교한 결과의 신뢰성이 높다. 낮은 민감도를 나타낼 경우에는, 기능 분석

을 위한 실험 횟수가 증가하여 실험자들의 연구 시간이 길어지는 결과를 낳을 수 있다.

4. 결 론

본 연구에서는 EST 단편 서열을 가지고 해당 서열의 유전자 발현과 관계되는 조직의 분포를 알아보기 위한 시스템을 구축하고자 하였다. EST Clustering, Contig assembly, 유전자 상동성 검색, 조직별 발현 정도의 분포 및 단백질 기능 예측에 관한 데이터 처리를 대량으로 할 수 있는 이와 같은 시스템은 국내 EST관련 실험자들에게 좋은 연구 방법으로 활용될 수 있을 것으로 기대된다. 여기서는 인간 유전자 단편을 대상으로 하였지만, 이 시스템은 mouse 데이터는 물론, 유전자 orthologues 상태까지도 검증할 수 있으며, Unigene이 아닌 Rice cDNA, Arabidopsis cDNA 등의 식물 관련 자료로 비교 대상을 확대 할 경우에도 마찬가지로 의미 있는 결과를 보여줄 것이다. 그러므로 본 연구 결과는 유전자 기능 분석(Functional Genomics) 분야에서 앞으로도 계속 사용될 수 있을 것이다. 유전자의 단편인 EST의 양이 적은 경우는 별로 문제가 되지 않으나 1,000여 개 이상을 가지고, 기능을 분석하려 한다면, 대량의 EST를 효과적으로 다룰 수 있는 실험자와 연구 기관은 필수적이다. 유전체 연구가 빠른 속도로 늘어나고 있는 현 시점에서, 정확성과 효율성 및 신뢰성을 동시에 갖춘 EST 분석 시스템은 필수적인 요소임이 분명하며, 이를 좀더 체계적으로 보완하고 여기에 보다 편리한 인터페이스를 추가한다면 post-genome 시대에 적합한 경쟁력을 갖출 수 있을 것이다.

참고문헌

[1] Eran Halperin, Simchon Faigler, and Raveh Gill-More, "FramePlus - Aligning DNA to Protein Sequences", Bioinformatics, Vol 15(11), 867-873, 1999.

[2] Hide, W., Burke, J., Christoffels, A., Miller, R., "A novel approach towards a comprehensive consensus representation of the expression human genome" In Genome Informatics , pp187-196, 1997.

[3] 고성호., "사람 흉선 내 T-세포에서 발현하는 발생분화단계 특이적 ESTs에 관한 genome analysis", 박사학위 논문, 서울대학교, 1999.

[4] Aaronson, J.S., Eckman, B., Blevins, R.A., Borkoski, J.A., Myerson, J., Imran, S., Elliston, K.O., "Toward the development of a gene index to the human genome : an assessment of the nature of high-throughput EST sequence data", Genome Res. 6(9), 829-845, 1996.

[5] Huang, X. "A contig Assembly Program Based on Sensitive Detection of Fragment Overlaps", Genomics 14 : 18 - 25,1992.

[6] Schuler, G.D, "Pieces of the puzzle: expressed sequence tags and the catalog of human genes", J.Mol.Med. 75: 694-698,1997.

[7] Adams,M.D., A.R. Kerlavage,R D. fleschman, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, and O. White. "Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence". Nature 377(Suppl. 28):3-174

[8] http://www.ncbi.nlm.nih.gov/UniGene/index.html

[9] http://genome.kribb.re.kr/seqanal/cap.html

[10] J.D.Parsons, S.Brenner and M.J. Bishop., "Clustering cDNA sequences", CABIOS, Vol 8(5), 461-466, 1992.

허 철 구

1990 충남대학교 계산통계학과 학사
1995 충남대학교 전산학과 석사
1994~현재 생명공학연구소 유전체 연구센터 선임기술원 근무
관심분야 EST 및 genome Analysis, Gene expression Analysis
E-mail:hurice@mail.kribb.re.kr

박 경 희

1996 이화여자대학교 과학교육과 학사
1998 이화여자대학교 생물학과 석사
1999~ 현재 생명공학연구소 위촉연구원
관심분야 Gene expression analysis
E-mail 24ellie@hitel.net

임 소 형

1999 이화여자대학교 화학과 학사
1999.3~현재 이화여자대학교 분자생명과학부 석사 재학 중
관심분야 Protein Motif analysis
E-mail lovesohyung@hanmail.net

신 민 수

1997 경상대학교 컴퓨터과학과 학사
1999~현재 배재대학교 정보통신공학과 석사 과정
관심분야 분산네트워킹, Protein sequence Analysis
E-mail:s_minsu@yahoo.co.kr

고 성 호

1993 서강대학교 생물학과 학사
1995 서울대학교 생물학과 석사
1999 서울대학교 생물학과 박사
1999~현재 생명공학연구소 Post-Doc
관심분야 Genome Analysis and Annotation, Proteins interaction
E-mail gohsh@mail.kribb.re.kr