인간 cDNA 라이브러리 분석 파이프라인 구축

Construction of Human cDNA Library Analysis Pipeline

정 재 은, 김 대 수
과학기술연합대학원대학교, 한국생명공학연구원

Jung Jaeeun, Kim Dae-Soo
University of Science & Technology,
Korea Research Institute of Bioscience & Biotechnology

요약

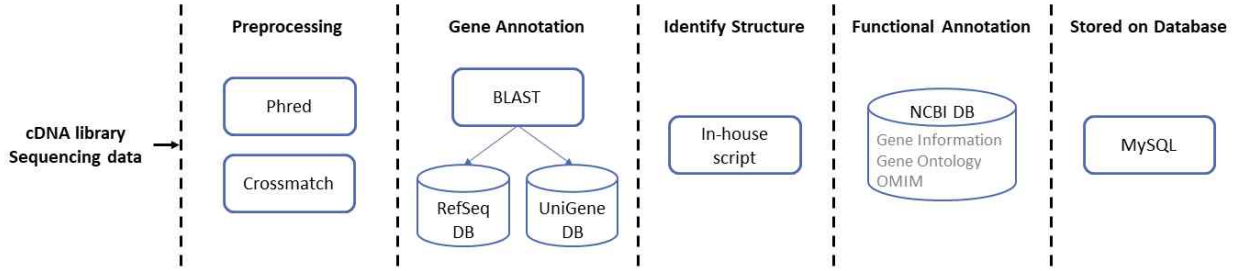
전장 cDNA 클론을 시퀀싱하는 것은 선택적 스플라이싱 형태를 비롯한 정확한 유전자 구조를 정의하는데 유용하게 사용될 수 있으며, 유전자 및 단백질의 생물학적 기능연구에 중요한 자원을 제공한다. 포괄적이며 비 중복적인 cDNA의 생산은 인간 유전체 연구의 중요한 목표이다. 본 연구에서 제공하는 인간 cDNA 라이브러리 분석 파이프라인은 전장 cDNA를 분석하는 자동화 도구로 여러 연구자들에게 활용 될 수 있을 것으로 사료된다.

1. 서론

모든 생물의 유전현상의 핵심은 유전정보의 흐름으로 나타난다. 이 유전정보의 흐름은 기존 DNA에서 새로운 DNA를 생성하는 복제 과정 (Replication), DNA를 주형으로 삼아 RNA를 생성하는 전사 과정 (Transcription), RNA에서 단백질을 생성하는 번역 과정 (Translation)으로 나누어져있다. 복제 과정은 DNA 속에 저장되어 있는 유전정보를 다음 세대로 전달하기 위해 DNA 서열 자체를 복제하는 과정이다. 전사 과정은 DNA 속에 저장되어 있는 유전정보를 RNA로 이동시키는 단계로, 이때 생성된 RNA를 mRNA라고 한다. 전사과정을 통해 mRNA로 옮겨진 유전정보는 번역 과정으로 통해 아미노산으로 번역되고 펩타이드 결합을 이루면서 단백질 형태로 변환된다[1].

한 개체의 모든 세포는 동일한 유전체를 지니고 있지만 특정한 세포, 조직, 기관에 따라 발현되는 유전자의 종류나 발현량은 각기 다르다. 유전자가 발현되려면 mRNA로 전사가 일어나야하기 때문에 일반적으로 세포 내의 mRNA 양은 그 유전자의 발현되는 정도와 비례한다. 각 기관별로 특이적으로 발현되는 유전자와 모든 기관에서 발현되는 유전자가 무엇인지 알 수 있다면 각 기관의 기능 및 특성을 이해하고 나아가 질병의 예방 및 치료에 도움이 될 것이다. 이러한 이유로 각 조직이나 기관에서 발현되는 mRNA를 분리하여 조직 특이적인 cDNA(complementary DNA) 라이브러리를 제작한다. cDNA는 mRNA를 주형으로 역전사 효소와 DNA polymerase에 의해 합성된 DNA를 말하며, 이는 mRNA에 상보적 배열을 가지므로 상보적 DNA라고 불린다. 대량의 cDNA 라이브러리의 분석은 유전자와 단백질의 기능적 유전체 연구에 사용되어 왔고, 따라서, 포괄적인 비 중복 cDNA의 생산은 인간 및 모델 동물 생물체의 유전체 연구에 중요한 목표이다[2].

관련 연구자들이 특정 cDNA 서열을 분석하고자 할 때, 주로 NCBI, Ensembl, DDBJ, UCSC 등 공용 데이터베이스에서 제공하는 웹 기반 검색 프로그램을 사용한다. 그러나 일반적으로 이러한 시스템들은 자원의 한계, 속도 저하 등의 이유로 분석 데이터의 업로드 용량과 수량에 제한을 두고 있으며, 유전자의 기능적 분석을 하기 위해 또 다른 웹 데이터베이스를 검색해야 하는 어려움



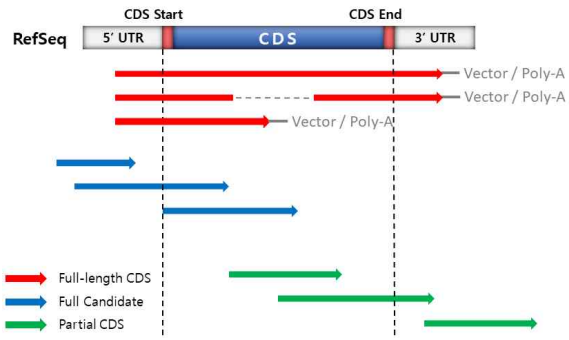
▶▶ 그림 1. 인간 전장 cDNA 라이브러리 분석 과정

이 있다.

이를 해결하기 위해 본 논문에서는 대용량 cDNA 라이브러리 자동화 분석 파이프라인을 제시한다.

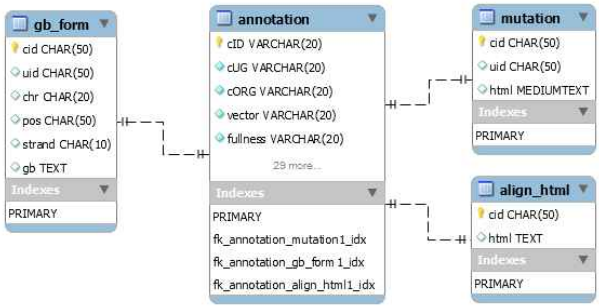
2. cDNA 라이브러리 분석 파이프라인 구축

연구자들이 cDNA 라이브러리 구축 후, 대용량 염기서열 데이터를 쉽게 처리하지 못하는 문제를 해결하기 위해 cDNA 라이브러리 분석 파이프라인을 구축하였다 [그림1]. 분석 파이프라인은 크게 5 부분으로 나누어져 진행되는데, 시퀀싱 데이터의 전처리 단계, 참조 서열과 비교하는 BLAST 단계, 전장 cDNA 구조 확인 단계, 기능 분석 단계, 데이터베이스화 단계이다. 이미 전처리가 된 cDNA 라이브러리는 두 번째 단계부터 진행할 수 있다.



▶▶ 그림 2. cDNA 구조

먼저, cDNA 라이브러리 시퀀싱 데이터의 전처리는 시퀀싱 머신에서 결과로 제공하는 이미지 파일을 컴퓨터로 처리할 수 있는 서열파일로 변환 후, Phred score 20 이상 (시퀀싱 정확도 99%)으로 확실한 서열만 추출하고 cDNA 본연의 서열이 아닌 벡터서열은 제거하는 단계이다. 깨끗한 서열이 얻어지면, 그 다음 단계로 참조 서열 데이터베이스에 대조하여 해당 cDNA가 어떤 유전자인지 확인하게 된다. 여기서 사용되는 데이터베이스는 두 가지인데, 먼저 표준 참조서열 데이터베이스 (RefSeq DB)에서 cDNA의 참조 유전자를 확인하고 첫 번째 데이터베이스에서 결과가 없거나 서열의 유사성이 낮은 cDNA는 유니진 데이터베이스 (UniGene DB)와 비교하여 참조 유전자를 찾아낸다. 세 번째 과정은, 전장 cDNA 구조 여부를 확인하는 단계이다. 각 cDNA는 참조 유전자 서열과 대조하여 Full length CDS, Full Candidate, Partial CDS, Non-CDS, Unknown으로 분류된다[그림 2]. Full length CDS는 coding sequence (CDS)의 개시 코돈을 포함하며 벡터 서열로 시퀀싱이 끝나거나 서열에 poly-A를 포함한다. Full Candidate는 5′ UTR이나 개시 코돈을 포함하고 있지만 벡터 서열이나 poly-A 없이 CDS 중간 영역까지 시퀀싱된 cDNA로 분류하였다. 그리고 참조 유전자의 CDS를 부분적으로 포함하고 있지만 개시코돈이 포함되지 않은 cDNA 들은 모두 Partial CDS로 할당 하였으며, CDS가 없는 유전자로 정의된 cDNA는 Non-CDS, BLAST 결과가 없는 cDNA는 Unknown으로 분류한다. 전장 cDNA를 얻은 후에 참조 유전자 서열과 대조하여 선택적 스플라이싱 구조와 변이 (SNP, INDEL)를 찾고, 시각적으로 볼 수 있는 정렬 파일을 만들어 사용자가 직접 확인 할 수 있도록 하였다. 네 번째 과정은, 참조 유전자가 할당 된 cDNA는 기능 분석을 위해 유전자 정보, Gene Ontology (GO), Cytoband, OMIM 등 추가 주석을 달아주는 단계이다. 마지막으로 모든 분석이 완료되면 cDNA 라이브러리 분석 정보를 데이터베이스화 하여 체계적으로 관리가 가능하도록 하였다[그림3].



▶▶ 그림 3. 분석 파이프라인 데이터베이스 ERD

3. 결론

완전한 서열을 가진 포괄적인 비 중복 전장 cDNA의 생산은 유전자의 exon, intron 구조를 정확하게 결정하는 중요한 자원으로써 가치가 있다. 특히, 선택적 스플라이싱 형태 및 저 발현 유전자가 RNA 시퀀싱의 염기서열 분석에 의해 재구성되기 어려운 경우 더욱 중요한 자원으로 사용될 수 있다. 본 연구에서 제시한 파이프라인은 시퀀싱 raw 데이터를 가지고 전장 cDNA 분석하는 자동화 도구로 여러 연구자들에게 활용될 수 있을 것으로 전망해본다.

▌참 고 문 헌▐

[1] Lu C. et al., "cDNA Libraries Methods and Applications," Methods in Molecular Biology, Vol. 729,

[2] Furuno M. et al., "CDS Annotation in Full-Length cDNA Sequence," Genome Res., Vol. 13, No. 6b, pp.1478-1487, 2003

[3] Sanger F et al., "DNA sequencing with chain-terminating inhibitors.", Proc Natl Acad Sci., Vol. 74, pp.5463-5467, 1977