2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

전산 클로닝을 위한 Clustered EST 데이터베이스 구축

이진관, 최은선, 류근호
충북대학교 데이터베이스연구실
e-mail:jklee, eschoi, khryu@dblab.chungbuk.ac.kr

Buliding Clustered EST database for In Silico Cloning

Jin Kwan Lee, Eun Sun Choi, Keun Ho Ryu
Database Laboratory, Chung-buk National University

요약

cDNA(complementary DNA)를 복제(cloneing)하여 염기 서열화 한 EST(Expressed Sequence Tag) 데이터는 여러 생물체들의 염기서열 정보들과 비교를 통해 유사점을 찾거나 기능적 부위 검색을 통해 유전자 기능을 추정할 수 있어 기능 유전체 연구에 많이 사용되고 있다. EST 데이터를 식물은 특정 종(Species)별로, 동물의 경우 종의 조직별로 클러스터링 함으로써 아직 알려지지 않은 종의 유전자를 밝혀낼 수 있음은 물론 유전자의 발현에 따른 단백질의 기능도 알아낼 수 있다. 따라서 이 논문에서는 NCBI에서 flatfile 형태로 제공하는 EST 데이터를 분석하여 관계형 데이터베이스로 모델링하고 구축하였다. 또한 EST 데이터의 효율적인 사용을 위하여 데이터를 특정 종의 조직별로 클러스터링하여 제공하는 시스템을 설계하고 구현하였다.

1. 서론

ESTs(Expressed Sequence Tags)는 cDNA (complementary DNA)의 일부분을 나타내는 것으로 새로운 유전자를 밝히고 질병을 일으키는 유전자 집단 구성원을 찾아내는데 이용하며, 아울러 유전자 복제와 유전자 서열을 비교하여 복잡한 유전자 서열 분석 등 다양한 연구에 이용될 수 있다[한윤수00]. 특히, 한 종에서는 알려져 있으나 다른 종에서는 알려지지 않은 유전자나 특정기능을 수행하는 단백질 도메인을 가지는 유전자를 새로 발굴하고자 할 때 매우 유용하다. 이와 같이 EST 염기서열 데이터베이스를 생물정보학적인 방법으로 이용하여 새로운 유전자를 발굴하는 것을 전산 클로닝(in silico cloning)이라고 한다. 이 논문에서는 공개되어 있는 플랫파일 형태의 EST 데이터의 하나인 NCBI의 플랫파일을 전산 클로닝에 이용할 수 있도록 종별로 클러스터링된 정보를 제공할 수 있는 데이터베이스를 설계하고 구현한다.

2장 관련연구에서 EST 유전체 데이터베이스 구축을 위해서 EST 데이터의 특성 및 새로운 유전자를 발굴하는데 EST 데이터를 이용하는 과정에 대해 기술하고, 3장에서는 NCBI에서 제공하는 플랫 파일의 분석과 데이터베이스 스키마의 설계 그리고 클러스터링에 대해 설명한다. 4장에서는 구현 내용과 결과에 대해 알아보고, 5장에서는 결론 및 향후연구에 대해서 논한다.

2. 관련연구

2.1 EST 데이터의 특성

EST는 cDNA 클론 분석을 통해 얻어진 짧은 서열이고, cDNA는 mRNA를 역전사에 의해 이중나선 구조의 DNA로 다시 만든 DNA로 상보적 DNA라고 한다. 세포내의 mRNA가 단백질로 발현되어지는 유전자를 나타내지만 매우 불안정하여 연구에 이용할 수 없기 때문에 cDNA를 이용한다. EST는 cDNA 각 클론의 염기서열을 3' 또는 5' 끝에서 단 한차례 자동화된 기계로 읽어서 서열을 결정한 것으로 실험으로 검증되지 않아 그 염기서열은 다소 부정확할 수 있지만, 서열을 자동화된 기계로 빠르고 값싸게 얻을 수 있다

이 연구는 한국과학재단 특정 기초연구비 (1999-2-303-006-3)의 지원으로 수행되었음.

는 점 때문에 여러 가지 연구에 응용되기 시작하였다 [Brow99].

현재 NCBI의 dbEST에는 총 8,769,068개의 EST가 있다(dbEST release 083101). 전체 EST 데이터 중에서 인간(Homo sapiens)의 EST는 3,760,298개로 약 42.9%, 쥐(Mus musculus + domesticus)의 EST는 2,098,292개로 약 23.9%를 차지하고 있다. 따라서 NCBI에서 제공하는 플랫파일을 토대로 EST 데이터베이스를 구축하게 되면 공개하여 발표되는 모든 EST들의 염기 서열과 기타 정보를 모두 포함하게 된다.

2.2 EST 데이터베이스를 이용한 새로운 유전자의 발굴

그림 1은 EST 데이터베이스에 대한 상동성 검색을 통해 새로운 단편적인 유전자 후보 서열들을 획득하여 이들을 서로 연결시키고, 계속해서 반복적인 DB 검색과 검색으로 얻어지는 EST 서열들의 연결을 반복함으로써 새로운 유전자의 염기서열을 얻을 수 있는 전산 클로닝 과정을 보여준다.

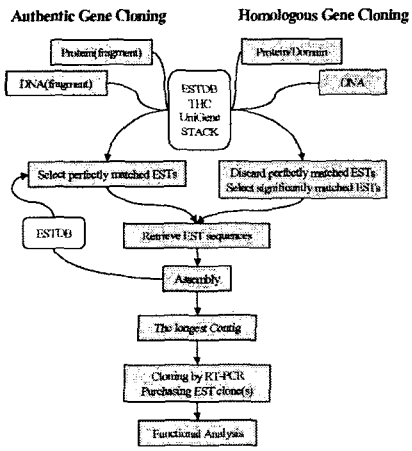

그림 1 전산 클로닝 과정[한윤수00]

이 외에도 서로 진화적으로 가까운 종들간의 상동성의 유전자 또는 특정 기능을 수행하는 단백질 도메인을 가지는 유전자를 새로이 발굴하고자 할 때도 EST 데이터베이스를 이용하면 보다 효과적이다. 그러므로 이러한 종들 사이의 비교 혹은 동일 종 내에서의 비교를 위해서 종별 조직별로 정확한 클러스터링을 제공하는 EST 데이터베이스의 구축이 전산 클로닝을 위해 꼭 필요하다.

3. EST 데이터베이스와 클러스터링의 설계

3.1 NCBI의 EST 플랫파일

플랫파일 하나의 EST 엔트리는 크게 Identification, Library, Citation, Submitter, Map Data와 같은 다섯 가지 정보를 가진다. Identification은 EST 데이터베이스의 가장 중요한 서열데이터와 관련된 정보를 포함하고 Library는 서열에 대한 소스정보를 가지며 생명체와 조직에 대한 정보를 나타낸다. Submitter는 EST서열을 제출한 연구자에 대한 정보로서 EST 소스에 대한 정보를 얻을 수 있고 Citation은 EST서열이 발표된 논문에 대한 정보를 포함한다. 그리고 Map은 유전체내에서의 위치가 확인된 시퀀스의 경우 그 정보를 포함하고 있다.

플랫 파일은 시퀀스를 제출한 연구자에 따라 포함된 정보와 형식이 다르다. 서로 다른 형식을 가진 플랫파일을 분석하여 필수적인 속성을 정의하고 부수적인 많은 속성들은 플랫파일의 형태 그대로 제공받기를 원하는 사용자를 위해 생략없이 모두 설계되었다. 필수적인 속성으로 분류된 속성은 EST서열에 관한 많은 정보 중 ESTdb ID, GenBank Account, Tissue type, Clone Id, Direction of Clone(3′ or 5′), Organism, LibraryId, PolyA tail의 유무, sequence등이다.

따라서 종별, 혹은 종별 조직별로 클러스터링된 데이터를 제공할 때와 사용자가 원하는 데이터를 검색하였을때도 별다른 요청이 없다면 위의 중요 속성들만을 포함하도록 설계한다.

3.2 스키마 설계

EST 데이터베이스는 NCBI의 dbEST에서 제공하는 모든 EST 서열과 이와 관련된 정보를 대상으로 한다. 따라서 데이터베이스 내에 구축된 총 EST 서열은 현재 dbEST가 포함하고 있는 8,769,068개의 엔트리가 된다. 이러한 데이터는 웹과 Unix를 통한 사용자들에게 검색을 통해 모두 제공이 된다. 하지만 클러스터링되어 종별로 파일의 형태로써 제공되는 대상은 동물의 경우 Homo sapiens(human)와 Mus musculus + domesticus(mouse)이며, 식물의 경우는 Arabidopsis thaliana(thale cress), Glycine max(soybean), Medicago truncatula(barrel medic), Lycopersicon esculentum(tomato), Zea mays(maize), Oryza sativa(rice), Triticum aestivum(wheat), Solanum tuberosum(potato)의 8종이다.

데이터베이스로 구축할 추출된 엔티티는 표 1과 같다.

표 1 플랫파일에서 추출된 엔티티들

이름	설명
EST	EST 서열 정보, Identification
Library	cDNA 서열 소스 정보
Submitter	서열 제공자 정보
Citation	서열 게재 논문 정보
MapData	유전체 맵 정보

추출된 엔티티들과 엔티티 사이의 관계를 ER-diagram으로 나타내면 그림 2와 같다.

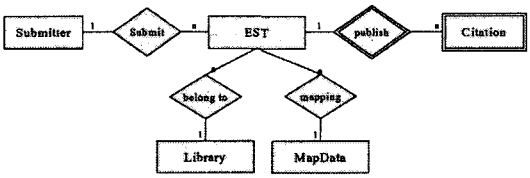

그림 2 EST 데이터베이스 스키마

3.4 클러스터링 설계

EST 플랫파일에 대한 클러스터링은 종별로 분류하는 것을 원칙으로 하되 동물과 식물의 경우 다른 클러스터링 룰을 적용한다.

인간과 쥐와 같은 동물의 경우 종별로 분류하면서 종의 조직별로 클러스터링을 수행하고(표 2), 식물의 경우는 종별로 클러스터링을 수행한다(표 3).

표 2 동물 EST 클러스터링 정보

HomoSapiens & Mouse의 Tissue type별 제공 정보				
Attribute	Type	Null ?	Description	Default
ESTdbID	varchar[]	no	EST id assigned by ESTdb	
GenBankAcc	varchar[]	yes	GenBank accession number	
CloneType	varchar[]	yes	Clone Type	
PolyATail	varchar[]	yes	Y or N to indicate if a polyA tail was or was not found	
TissueType	varchar[]	yes	Tissue Type	

표 3 식물 EST 클러스터링 정보

Arabidopsis 의 식물들				
Attribute	Type	Null ?	Description	Default
ESTdbID	varchar[]	no	EST id assigned by ESTdb	
GenBankAcc	varchar[]	yes	GenBank accession number	
CloneType	varchar[]	yes	Clone Type	
PolyATail	varchar[]	yes	Y or N to indicate if a polyA tail was or was not found	
TissueType	varchar[]	yes	Tissue Type	

4. 시스템 구현 및 결과

4.1 구현 환경

SUN Enterprise 250 머신에서 실행되는 운영체제는 Sun Solaris 7, DBMS는 Oracle 7.3.4를 사용하였다. 플랫파일을 분석하고 클러스터링하는 모듈은 ProC와 C언어를 사용하여 구현하였고, 웹을 통한 검색 시스템은 Perl을 사용하였다.

4.2 시스템 구조

입력 모듈은 입력 프로그램(ESTin)에 한개의 플랫파일을 인자로 받아들인 후, 먼저 파일 분석기를 통해 파일이 삭제할 엔트리를 포함한 파일이라면 그 파일에 속한 엔트리를 데이터베이스로부터 삭제하고 그런 파일이 아니라면 문서 분석기를 통해 문서를 테이블의 속성별로 구분한다. 구분된 속성들은 DB 생성 모듈에서 ProC를 통해 데이터베이스와 접속한 후 갱신되거나 새로 저장된다.

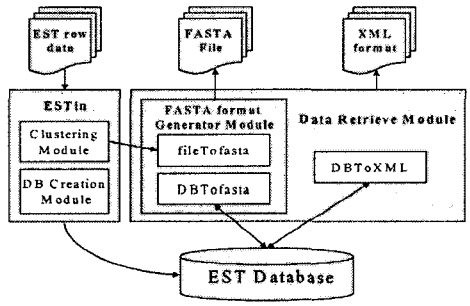

그림 3 시스템 구성도

그리고 플랫파일로부터 별도로 만들어진 EST와 library 정보를 포함하는 파일들은 클러스터링 모듈을 통해 클러스터링 규칙을 적용 각각의 종별, 조직별로 분류된다. 이는 다시 FASTA 포맷 생성 모듈(fileTofasta)에 의해 그림 4와 같은 형태로 저장된다.

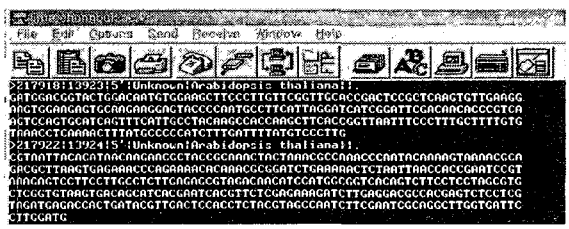

그림 4 클러스터링 결과(식물)

4.3 결과

검색 모듈은 저장된 정보를 웹 검색과 Unix기반 검

색을 지원하도록 구현하였다. 검색은 생명체에 따른 조직의 EST서열 검색을 할 수 있다. EST 데이터베이스에서 출력되는 데이터는 주로 다른 데이터베이스의 입력데이터로 사용되기 때문에 검색의 결과는 FASTA 포맷과 XML 문서 형태로 제공한다.

다른 형태를 갖는 유전체 데이터베이스에 EST정보를 제공하는 시스템 사이에서는 데이터 전송을 위해 데이터 교환 표준 형식인 XML을 사용함으로써 이질적인 시스템간의 데이터 교환의 호환성을 높일 수 있다. 뿐만 아니라, 서로 다른 유전체 데이터베이스 시스템에서 XML 문서 형식으로 결과를 전송 받아 하나의 통합된 XML 문서로 데이터를 통합할 수 있다.

```
<!DOCTYPE EST System "Est.dtd">
<ESTInfo>
<EST estid="est00001" type = "EST">
        <cloneID direction=5>IIIIC198</>
        <GeneBankAccountNo>BE735522</>
        <polyAtail>Unknown </>
        <library_info library::IDREF="l1">

<sequence>AATCAGCCTGCAAAAGATAGGAATATTCACAGAGAGTCAGACG
          ACTGACTGGGGCATTAAATTACGA</>
</EST>
<library libraryno::ID=" l1" type="Lib">
        <name>Rat Lambda Zap Express Library </>
        <organism>Rattus norvegicus</>
        <tissue>aorta</>
        <sex>male</>
</library>
<contact contact_no::ID="c1 " type=" con" >
        < name="Sikela JM">
        <tel>3032707097 </>
        <email>tjs@tally.hsc.colorado.edu</>
        <institution>University of Colorado Health Science Center>
</contact>
<publication pubno::ID="p1"  type="pub">
    <title>expressed sequence tags and chromosomal localization of cDNA
clones from a subtracted retinal pigment epithelium library</>
    <author>
        <lastname>L </lastname>
        <firstname>Gieser  </firstname>
    </author>
    <journal name="Genomics ">
    <year>1992</>
    <status>published</>
</publication>
</ESTInfo>
```

그림 4 EST 서열의 XML 문서 형식

5. 결론 및 향후 연구

이 논문에서는 관계형 DBMS를 기반으로 하는 EST 데이터베이스를 구축하기 위해 EST 데이터의 특성과 NCBI에서 제공하는 EST 플랫파일을 분석하여 관계형 스키마를 설계하였고 또한 전산 클로닝에 효과적으로 적용할 수 있도록 클러스터링을 설계하였다. 설계된 스키마를 바탕으로 플랫파일 형태를 테이블로 매핑시키기 위한 ESTin이라는 입력모듈을 작성하였고 클러스터링을 수행하는 모듈을 작성하였다. 저장된 EST데이터의 검색은 UNIX 운영체제와 웹 환경에서 종과 조직별 EST서열을 검색할 수 있으며 검색결과는 FASTA형식과 XML 형태로 제공된다.

또한 특정 조직별로 서열을 클러스터링하여 파일로 직접 제공함으로써 유전자 예측, 더 나아가서 특정 질병을 유발하는 유전자의 발견이나 인간 유전자 기능 추정을 하기 위한 전산 클로닝 작업에 대한 기초로서 이용될 수 있다.

이 논문에서는 XML을 이용함으로써 이질적인 형태의 데이터나 이질적인 시스템간의 호환성을 높일 수 있는 기저를 마련하였는데 향후 이를 이용하여 많은 이질적인 생물학 데이터베이스들과 관련 도구들을 묶어 하나의 통합된 시스템으로 구축할 수 있다. 또한 단백질 데이터베이스를 구축하여 Genomic 레벨에서 수행한 실험결과를 단백질 레벨에서 비교 검증 또한 필요하다.

참고문헌

[Bogu93] Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev. 1993. dbEST: database for "expressed sequence tags". Nat. Gebet. 4: 332-333.

[Adam95] Adams, MD., A.R. Kerlavage, R.D. Fleischman, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, and O. White. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377(Supl. 28):3-174.

[Papa94] Papadopoulos, N. N.C. Nicolaides, Y.F. Wei, S.M. Ruben, K.C. Carter, C.A. Rosen, W.A. Haseltine, R.D. Fleischman, C.M. Fraser, M.D. Adams, et al. 1994. Mutation of a multi homolog in hereditary colon cancer. Science 263: 1625-1629.

[Zhan97] Nan Zhang, Theo Hde, On Modeling Power of Object-Relational Data Models in Technical Applications. ADBIS 318-325,1997

[Guff98] Guffanti, A. and G. Borsani, In Stitu Blasta : a www resource for tissue-specific EST database searches, Trends Genet, 14:518, 1998.

[XML98] W3C, "XML1.0 REC-xml-19980210",1998. http://www.w3.org/TR/REC-xml/

[Brow99] T.A. Brown, Genomes, Bios Scientific publishers Ltd, 1999

[Leto99] Stanley I. Letovsky, Bioinformatics Databases and Systems, Kluwer Academic Publishers, 1999.

[한윤수00] 한윤수, 정재훈, EST를 이용한 유전자 발굴 및 발현 분석, 한국 유전학회, 유전 3월호, 238-258, 2000.

[류근호00] 류근호, 유전체 데이터베이스와 EST 데이터베이스, 연구개발 정보센터, 지식정보인프라 10월호, 48-61, 2000.