유전자 예측 프로그램 RepeatMasker 설치와 운용

RepeatMasker Installation Manual

정우근
Chung Woo-Keun
부산대학교 컴퓨터공학과
wkchung@pusan.ac.kr

ABSTRACT

진핵생물의 유전체 서열 중 반복 서열이 가장 많은 영역을 차지하고 있다. Transposon elements를 포함하여 simple repeat region, low complexity 영역이 전체 유전체의 약 70 80% 가량 해당된다. 따라서 반복서열 영역을 우선적으로 선별한 뒤 마스킹 작업을 통해 반복서열 영역에서의 유전자 예측은 예외로 처리한다. 물론, 반복서열 영역내에도 단백질로 코딩되는 부분이 존재 하지만, 극히 일부에 해당하기 때문에 추후에 따로 수행한다. 본 보고서에서는 진핵생물의 유전체 서열 중 가장 많은 영역을 차지 않고 있는 반복 서열의 마스킹 작업을 진행할수 있는 RepeatMasker의 설치 방법과 간단한 실행 방법에 대해 알아보기로 한다. RepeatMasker는 유사성 기반의 검색을 통해 반복서열 데이터 베이스에 존재하는 서열과 비교하여 유전체 내에 존재하는 transposon element와 retrotransposon element, rolling circles를 추출하고, TRF(Tandem Repeat Finder)라는 서브 프로그램에 의해 단순반복 서열을 규명한다.

KEYWORDS RepeatMasker

1 서론

RepeatMasker는 low-complexity sequence 와 interspersed repeats를 포함하는 Genomics 데이터를 규명, 분류 그리고 반복적인 요소들을 Masking 하기 위해 널리 사용되는 툴이다. RepeatMasker는 Nucleotide sequences 에 있는 repetitive elements를 annotation, identify 및 Masking 하기 위해만들어졌다. RepeatMasker는 low-complexity DNA sequences와 interspersed repeats를 포함하는 repetitive elements를 annotation 하고 DNA sequence와 연관되는 데이터 값들을 소문자로 변환하거나 Ns, Xs로 변환한다. 소문자로 변환하는 것은 Soft Masking 이라하고, x로 바꾸는 것은 Hard masking이라고 한다.

최근 RepeatMasker에 대하여 많은 논문들이 비슷한 종류에 뛰어난 성능을 보이는 툴도 선보이고 있다. Bedell2000 [1]는 RepeatMasker 의 성능을 강화하기 위해 만든 것으로 RepeatMasker에서 꼭 필요로 존재하는 Blast 와 같은 프로그램이다. WU-Blast와 MaskerAid의 성능을 분석하였다. 이 논문에서는 MaskerAid의 성능이 정말 뛰어난 것으로 나왔다. Juretic [2]에서는 HMMER 이라는 것과 RepeatMasker에 대한 성능 분석한 부분이 있지만, HMMER Long Sequences에 대한 처리는 불가능 한 것으로 나타났다. Morgulis [3]에서는 RepeatMasker 보다 더 뛰어난 성능을 보이는 WindowMasker가 소개되었으나, 너무나 사용법이 복잡하고 자세한 Manual이 존재하고 있지 않다.

본 보고서에서는 RepeatMasker의 설치 방법과 기본적으로 나타나는 오류해결방법에 대하여 알아보고, 간단하게 실행하는 방법에 대해서 알아보도록하겠다. RepeatMasker를 실행하기 위해서는 repetitive elements consensus sequences를 포함하는 Repeat Library가 필요하다. 현재 사용되고 있는 Repeat Librarys는 Repbase Update 라이브러리이다. 해당되는 라이브러리는 가장 상용적으로 사용되고 있으며, human, rodent(설치류), zebrafish, Drosophila(초파리), Arabidopsis thaliana의 종을 포함하고 있다. 또한 블라스트(Blast)가 필요하다.

기본 국소정렬 검색 도구인 BLAST(the Basic Local Alignment Search Tool, BLAST)는 서열의 유사성을 밝히는 데 가장 많이 사용되는 방법이며, 블라스트 프로그램은 사용자들이 제공한 검색 대상 서열에 대하여 NCBI의 전체 데이터베이스를 대상으로 하여 검색을 수행한다.RepeatMasker에 대하여 간단히 요약하면 다음과 같다.

1 . Genome 서열에서 Repetitive Sequence를 masking하는 프로그램이다.

2 . RepeatMasker를 사용하면, HumanGenome의 경우 50% 이상이 masking 된다고 합니다.

3 . RepBase라는 Repeat Sequence Database가 필요하며, CrossMatch프로그램을 기반으로 제작되었다.

2 설치

본 단락에서는 RepeatMasker의 기본적인 설치 방법에 대해서 알아보도록 하겠다. RepeatMasker를 사용하려면 기본적으로 필요한 사항은 아래와 같다.

1 . Unix System with perl 5.8.0 or higher installed

2 . Sequence Search Engine(Cross_Match or WU-Blast)

3 . Repeat Database

RepeatMasker를 사용하기 위해서는 위와 같이 3가지 사항이 준비되어 있어야 한다. 먼저 RepeatMasker를 다운받아서 압축을 해제하여 보자. RepeatMasker는 www.repeatmasker.org에서 다운받을 수 있다. 해당 홈페이지에서는 online masking 서비스도 행하고 있다. 하지만 온라인 서비스는 오프라인의 masking 작업과는 다르게 파일 업로드나 텍스트 입력의 양이 제한적이다.

먼저 RepeatMasker를 다운받아서 unix system 기반에 perl 5.8.0 이상 설치되어 있는 컴퓨터에 압축을 해제한다. 본 보고서에서는 Neobio 및 Pearl 컴퓨터에 설치하였다. 그림 1을 살펴보자. 본 그림은 www.repeatmasker.org에서 오프라인용 repeatmasker를 다운받는 페이지이다. 먼저 Installation에 있는 1번 메뉴에 Lastest Released Version을 다운받는다. 다음과 같이 명령어를 실행한다.

```
[root@neobio Desktop]# mkdir RepeatMasker
[root@neobio Desktop]# cd RepeatMasker
[root@neobio RepeatMasker]# gunzip *.tar.gz
[root@neobio RepeatMasker]# tar -xvf RepeatMasker-open-3-2-8.tar
[root@neobio RepeatMasker]# cd RepeatMasker
[root@neobio RepeatMasker]# ls
```

위와 같은 명령어를 실행하고, 맨마지막 명령어를 실행하면 그림 2와 같이 파일이 해제된 모습을 볼 수있다. 이제 RepeatDataBase 파일을 다운받아서 RepeatMasker가 설치되어 있는 Libraries 폴더 밑에 압축을 해제해야한다. RepeatDataBase 파일은 www.girinst.org에 있으나, 홈페이지에 가입을 해야만이 파일을 다운받을 수 있다. 가입은 무료이니 가입을 하기바란다(승인기간이 다소 있으므로 유의하기 바란다).

www.girinst.org에서 그림 3 와 같이 Repbase 항목에 CurrentRelease 서브 메뉴에 보면 RepeatMasker Libraries가 존재한다. 해당 파일을 RepeatMasker 하위 폴더인 Libraries안에 압축을 해제한다. 명령어를 살펴보자.

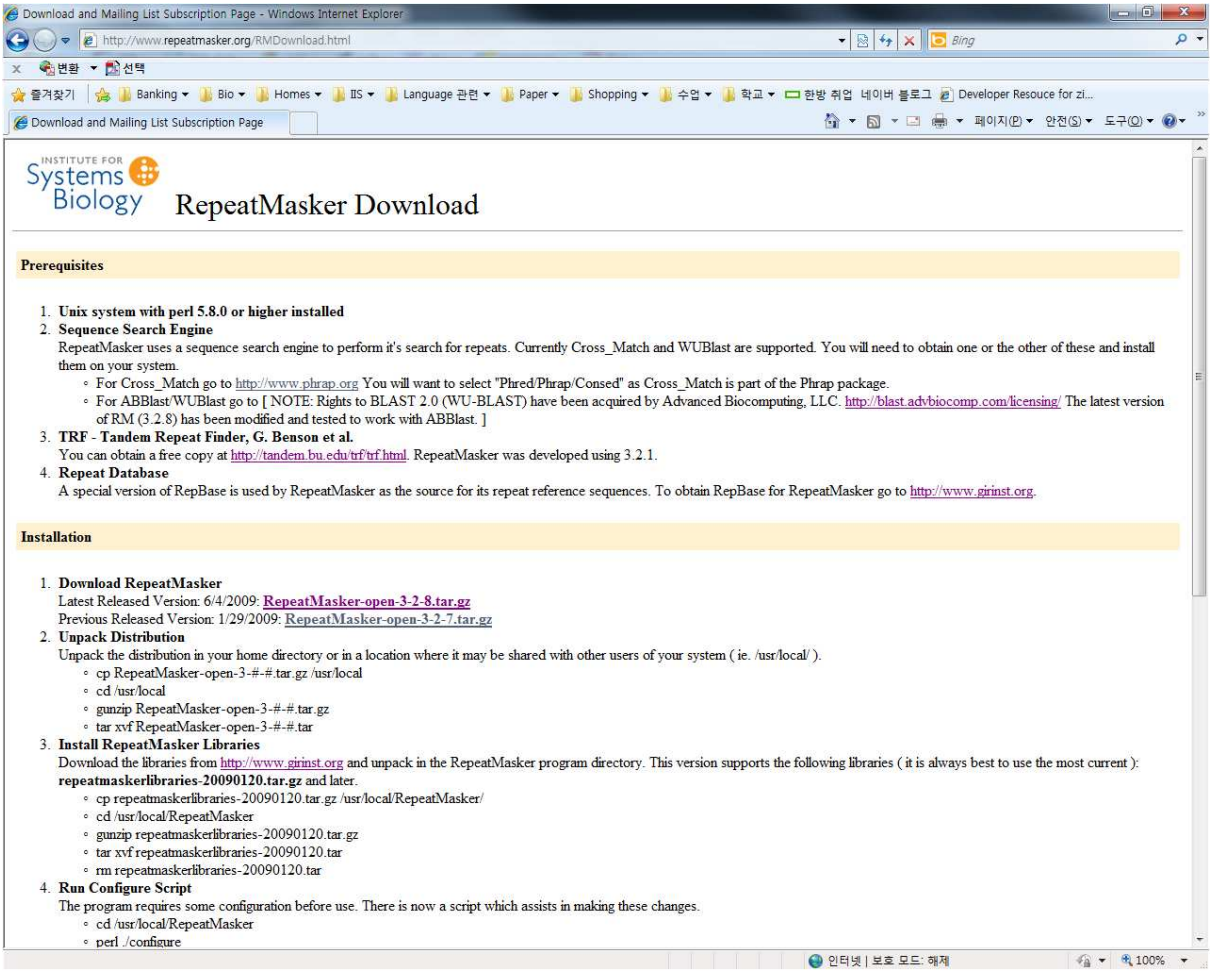

그림 1. www.repeatmasker.org

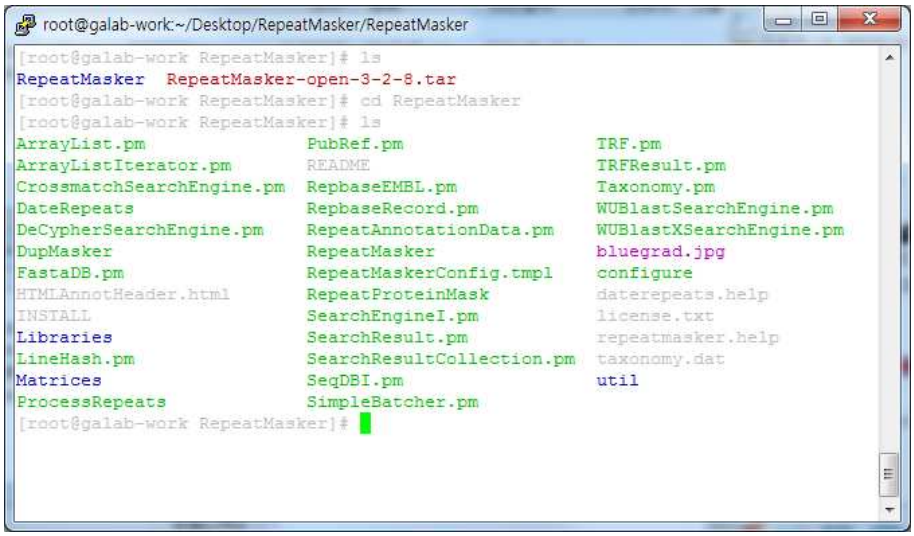

그림 2. www.repeatmasker.org 에서 RepeatMasker 최신버전을 다운 받아 압축을 해제한 snapshot

```
[root@neobio Libraries]# pwd
/root/Desktop/RepeatMasker/RepeatMasker/Libraries
[root@neobio Libraries]# gunzip *.tar.gz
[root@neobio Libraries]# tar -xvf RepeatMasker-open-3-2-8.tar
[root@neobio Libraries]# ls
```

RepeatMasker 하위 폴더인 Libraries 에서 압축을 해제하면 해당 폴더 안에 Libraries 가 또 생긴다. 이 중 첩된 폴더 밑에 RepeatDatabase가 설치되어있다. 이 파일들을 상위 폴더에 전부 옮긴다. Libraries 폴더에 그림 4와 같이 RepeatMaskerLib.embl 파일이 있는 것을 확인한다. 서브 프로그램인 TRF(Tandem Repeat Finder)를 그림 1에 존재하는 링크를 참조하여 다운 받는다. TRF 파일을 저장할 곳은 RepeatMasker 폴더에 저장한다. 저장된 TRF 파일을 아래의 소스를 참조하여 파일을 변경한다.

```
[root@neobio RepeatMasker]# ln -s trf312.linux.exe trf
```

마지막으로 Blast를 설치해야 한다. RepeatMasker에서 사용될 수 있는 Blast는 CrossMatch, WUBlast or ABBlast, Decypher이다. RepeatMasker에서는 Blast가 필수적으로 필요하기 때문에 반드시 설치해야 한다. 본 보고서에서는 Blast로써는 ABBlast를 설치한다. ABBlast는 홈페이지 http://blast.advbiocomp.com/licensing/ 에서 Personal Licensing을 받아서 사용할 수 있다.

Personal Licensing을 요청하면 요청사항에 기입하였던 E-mail로 ftp 접속할 수 있는 링크를 받을 수 있다. 해당 ftp를 통하여 ABBlast를 받아서 RepeatMasker 폴더에 복사하여 압축을 해제한다. 이제 RepeatMasker를 설치하여 보자. 설치는 RepeatMasker 폴더에서 ./configure 를 실행한다. 설치화면으로 넘어가면 설치환경에서는 RepeatMasker, perl, ABBlast, TRF의 위치를 묻는다. TRF, ABBlast의 위치만 입력하면 된다. 이로써 RepeatMasker의 설치가 끝난다.

3 실행시 문제점 해결방안

앞서 우리는 RepeatMasker 사용할 수 있는 모든 설치를 마쳤다. 본 단락에서는 RepeatMasker 사용방법에 대해서 알아보고, 사용 중에 나타나는 오류에 대해서 알아본다. RepeatMasker의 기본적인 실행 방법은 다음과 같다.

```
RepeatMasker [-options] <seqfiles (s) in fasta format>
```

먼저 RepeatMasker의 option 사항에 대하여 알아본다. 옵션 사항은 다음과 같다.

```
-q Quick search; 5-10\% less sensitive, 2-5 times faster than default
-nolow Do not mask low\_complexity DNA or simple repeats
-div [number] Mask only those repeats < x percent diverged from consensus seq
-species <query species> Specify the species or clade of the input sequence (choose only one!)
```

RepeatMasker에서 사용할 수 있는 옵션은 위와 같이 총 4 가지가 있다. 본 보고서에서는 여기서 -nolow 옵션과 -species 옵션을 사용한다. 옵션을 제외하고 RepeatMasker에서 Masking하기위해서는 fasta format 형태의 파일이 필요하다. 이 파일은 UCSC(http://genome.ucsc.edu/)에서 각 Species 마다 est, mRNA에 대하여 fasta 형식의 파일을 제공하고 있다. fasta 양식은 정의행(definition line)과 서열 문자를 포함하여 다양한 분석 프로그램에 입력용 파일을 뜻하는 것이다. 기본적인 파일의 형태는 다음과 같다.

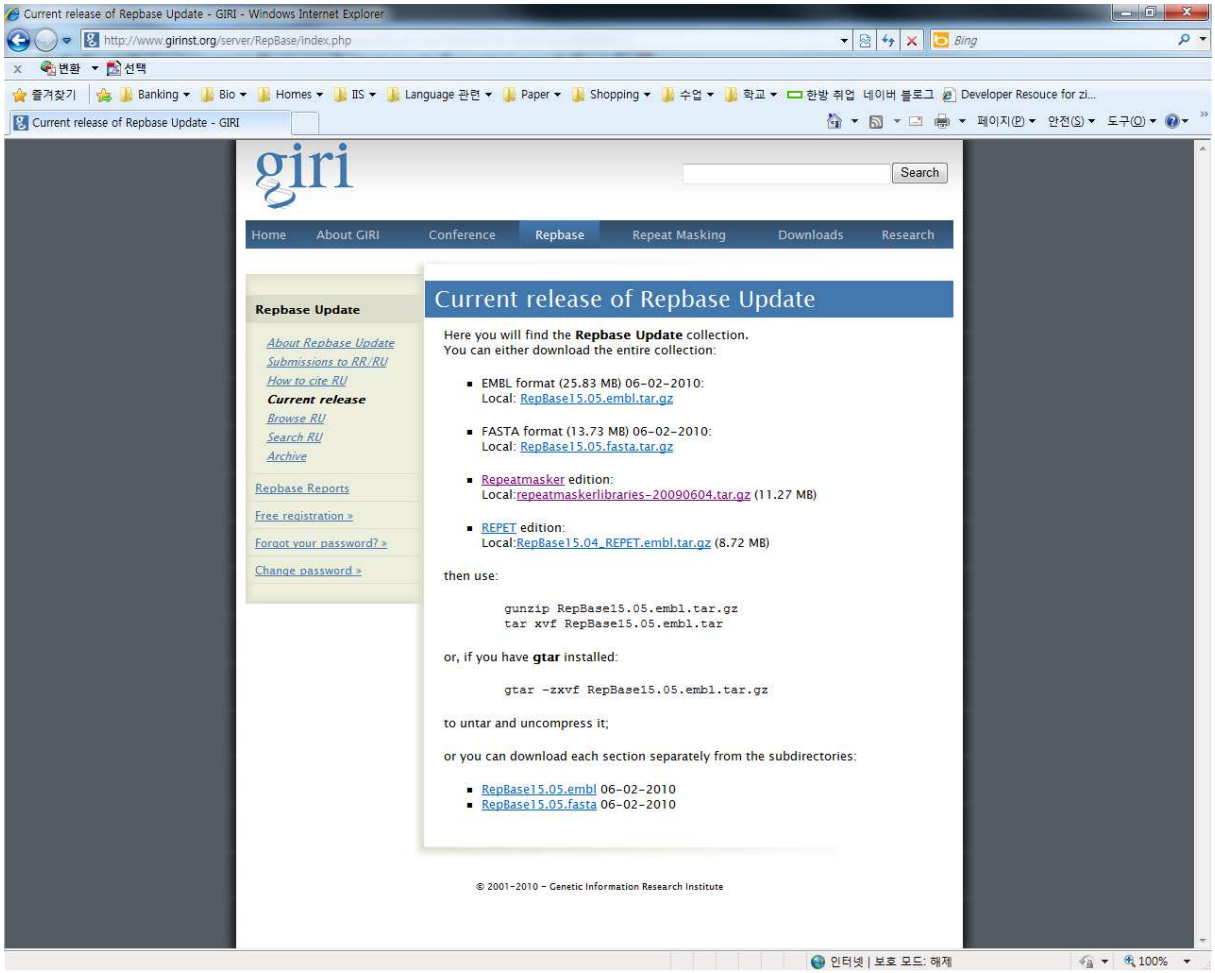

그림 3. www.girinst.org 에서 RepeatMasker Libraries 를 다운받는다.

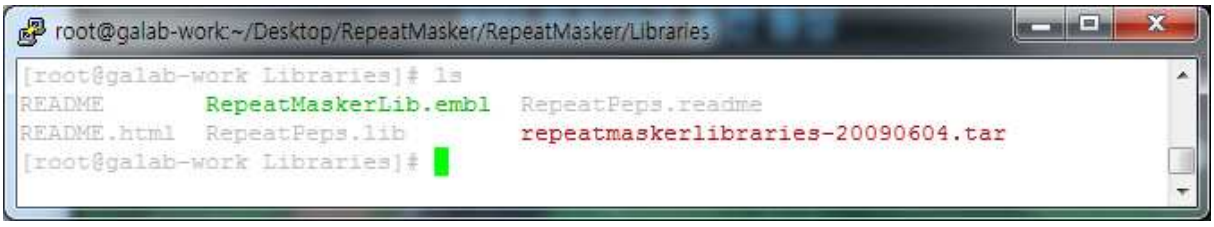

그림 4. RepeatMasker 의 하위 폴더에 RepeatMaskerLib.embl 파일을 확인힌다.

```
>AJ399512 1
tgatctacaaataatgttggataatgccaaaatcaacttaaatgaaaaac
tatctcaactacagacatatgtgatacaatttgatcagtatattaaagat
aattatgatctacatgattttaaactagccgttgctaaaattattgacca
aatcattgaaaaattaaaaattctt
```

예를 들어 fasta 형식의 mRNA 파일을 -nolow 옵션 값으로 masking 하고자 한다면 명령어는 다음과 같다. Species는 Human이라 가정한다.

```
RepeatMasker -nolow -species Human mRNA.fa
```

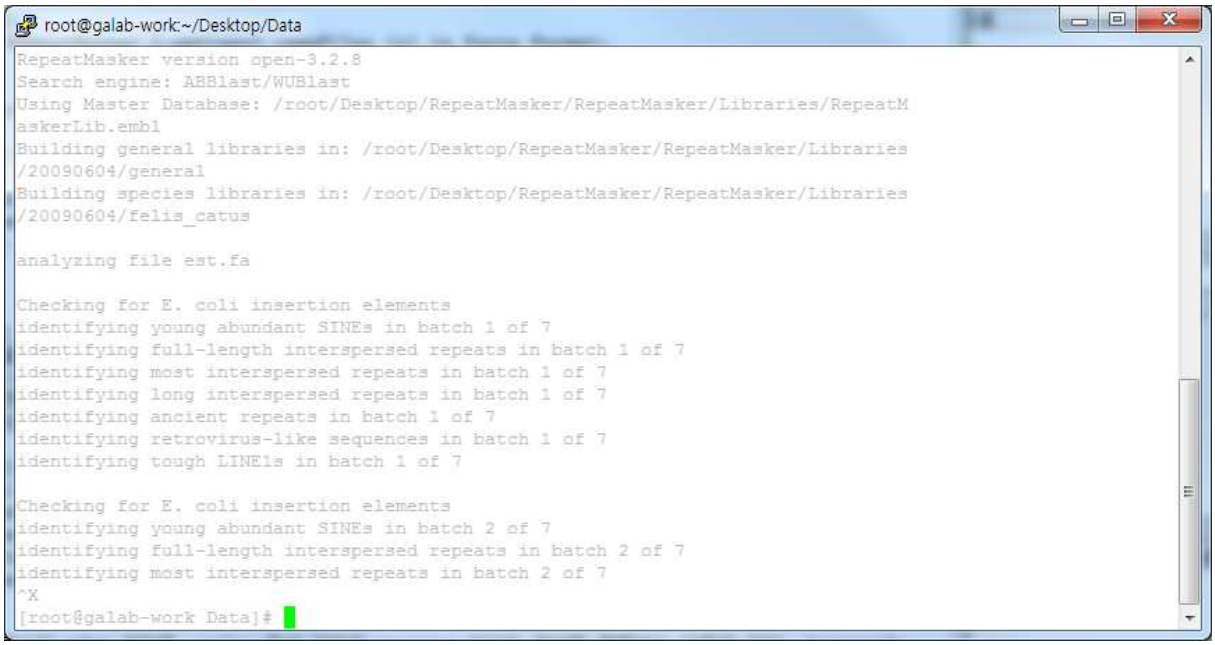

그림 5. RepeatMasker 실행하면 Snapshot. 순차적으로 실행되는 모습을 확인 할 수 있다.

현재까지의 설치를 마치고 위와 같은 명령어를 실행하면, 그림 5과 같이 실행될 것이다. 하지만 예기치 못하게 아래와 같이 오류가 발생할 수도 있다.

```
RepeatMasker mushroom_454LargeContigs.fasta
RepeatMasker version open-3.2.8
Search engine: Crossmatch
Storable binary image v2.7 more recent than I am (v2.6)
at ../../lib/Storable.pm (autosplit into ../../lib/auto/Storable/_retrieve.al) line
328, at /usr/local/genome/repeatMasker/RepeatMasker//Taxonomy.pm line
214
```

위와 같은 소스를 참고해보면 Storable 의 버젼이 낮아서 정상적으로 작동하지 않다는 것을 알 수 있다. 오류를 해결하는 방버은 현재 프롬포트 창에서 cpan을 입력한다. cpan은 리눅스에 사용되는 모듈을 쉽게 다운받을 수 있는 프로그램이다. cpan을 설치하고. cpan> 창에서 install Storable을 입력 한다. 그러면 프로그램 cpan을 통해서 Storable을 설치하고 위와 같은 오류를 해결한 뒤 정상적으로 RepeatMasker가 작동될 것이다. 오류 해결방안을 다시 한 번 살펴 보겠다.

```
[root@neobio ~]# cpan
( cpan> 으로 바뀔때까지 Enter를 입력한다.)
cpan>install Storable
```

4 결론

본 보고서에서는 진행생물의 유전체 서열 중 가장 많은 영역을 차지하고 있는 반복 서열의 Masking 을 수행할 수 있는 RepeatMasker 프로그램의 설치 및 실행 방법 그리고 오류 해결에 대하여 알아보았다. RepeatMasker 의 설치 방법은 의외로 간단하였으나, 국내에 Bioinformatics의 연구가 활발하지 않아 쉬운 설치에도 불구하고, 많은 실패를 겪었다. RepeatMasker의 수행방법은 RepeatMasker Database와 ABBlast를 통해 Masking를 수행한다. 하지만 수행방법이 유연하지 않고 스레드를 사용하지 않아 고사양의 컴퓨터라도 영장류의 est 데이터와 같이 용량이 큰 데이터의 경우 무지 많은 수행시간이 걸릴 것으로 예상된다. est 데이터를 분해하여 여러 대의 컴퓨터를 통해 Masking 하는 방법과 슈퍼 컴퓨터를 이용하는 방법이 있겠지만, 전반적으로 RepeatMasker 자체의 소스 개선이 이루어져야 한다.

참고 문헌

1. J. A. Bedell, I. Korf, and W. Gish, "MaskerAid : a performance enhancement to RepeatMasker," *Bioinformatics*, vol. 16, no. 11, pp. 1040–1041, 2000.
2. N. Juretic, T. E. Bureau, and R. M. Bruskiewich, "Transposable element annotation of the rice genome," *Bioinformatics*, vol. 20, no. 2, pp. 155–160, 2004.
3. A. Morgulis, E. M. Gertz, A. A. Schaffer, and R. Agarwala, "WindowMasker: window-based masker for sequenced genomes," *Bioinformatics*, vol. 22, no. 2, pp. 134–141, 2006.
4. U. G. Bioinformatics, "http://genome.ucsc.edu/."
5. giri(GENETIC INSFORMATION RESEARCH INSTITUTE), "www.girinst.org."
6. repeatmasker, "www.repeatmasker.org."