웹2.0 기반 DNA서열 분석도구 구현에 대한 연구

김명관, 조충효°
을지대학교
binsum@eulji.ac.kr, jotaja@naver.com°

A Study on Implementation of DNA Sequence Analysis Tool in Web2.0

MyungGwan Kim, Chung-Hyo Jo°
EULJI UNIVERSITY

요 약

최근 컴퓨터를 이용한 유전자 해석 기술이 급속히 발전함에 따라 DNA서열분석도구의 필요성도 늘어나고 있다. 그러나 DNA서열분석에 필요한 데이터베이스는 다양한 형태의 포맷이 제공되어 지고 있고, 유전자 서열 데이터의 처리를 위한 애플리케이션에서도 서로 다른 양식의 포맷이 사용되고 있다. 이로 인해 다른 형태의 포맷이 필요한 경우 별도의 파서를 구현 하는 문제가 발생한다. 이러한 단점을 보안하는 하나의 방법으로 GenBank에서 제공되는 XML파일을 이용한 웹2.0 환경인 RIA(Rich Internet Application)개발방식을 제안한다. RIA개발방식은 XML파서와 XML을 처리할 수 있는 E4X(ECMAScript for XML)와 같은 API를 제공 하여 XML 로 리턴 되는 데이터를 쉽게 처리하여 화면으로 보여준다.

1.서 론

유전체 서열과 기능을 알아내기 위한 인간유전체산업(Human Genome Project, HGP)이 국제협력 사업으로 1990년에 시작되어 1999년에 완료 되었다. 현재 바이러스, 식물, 동물 등 지구상에 존재하는 생물체의 유전체 사업이 전 세계적으로 진행되고 있으며, 1995년 Haemophilus influenzae,RD 의 전체 유전체 서열결정을 시작으로, 현재 생물학의 모든 분야에 대한 서열정보가 대량으로 늘어나고 있다.[1,2]

이에 따라 연구자가 다루는 데이터는 비약적으로 증가되어 데이터를 손으로만 처리한다는 것은 매우 어려운 일이 되었다. 이런 상황에서 바이오 분야의 연구개발의 효율을 높이기 위한 정보기술로서 생명정보학(bioinformatics)이 더욱더 중요해지고 있다.[3]

실제로 생물 정보학의 여러 세부 분야 중에서 서열을 분석하여 유전자 기능을 예측하는 문제를 다루는 여러 가지 소프트웨어 도구가 개발되어 유전체 연구에서 핵심적인 역할을 수행하고 있다.[4]

유전자 서열 분석관련 도구들의 종류로는 Artemis와 AnnHyb 프로그램이 있다. Artemis는 영국의 생어 연구소(Sanger Institute)에서 자바로 개발한 서열분석 프로그램이고, AnnHyb은 1997년에 비주얼 베이직으로 제작된 프로그램이다.[5-7]

기존의 유전자 서열 분석관련 도구들은 리눅스나 유닉스 기반의 프로그램이기 때문에 결과물을 얻어내기 위해서 프로그램을 서버에 설치하고 명령어 사용법과 각종 옵션을 숙지해야만 한다. 최근 웹으로 제공되는 분석도구가 있지만 기존의 웹은 단방향의 성격이 강하다. 사용자는 이미지와 텍스트로 이루어진 웹 페이지를 하이퍼링크를 통해 이동하면서 정보를 제공 받아 대량의 서열을 분석하기위해서는 오래 시간이 걸린다. 따라서 대량의 데이터를 빨리 처리할 수 있고, 사용자가 편리하게 작업을 수행할 수 있는 도구의 개발이 필요한 실정이다.[8]

본 논문에서는 유전자 분석의 필요한 분석도구를 웹 2.0기반 RIA(Rich Internet Application) 개발방식으로 구현한다. 기존의 HTML의 단점을 보완하기 위해 XML 을 사용하여 한 화면에서 필요한 데이터만 받음으로서 페이지 로딩 시간을 줄이고, 대량의 데이터를 빠르게 처리할 수 있었다.[9]

RIA 개발방식을 사용하여 유전자 분석도구를 실제로

구현한 결과 여러 페이지의 이동 없이 본인이 원하는 데이터만을 서버에 요구하고 받아서 처리하였다. 사용자의 편의성 측면에서도 RIA의 그래픽 사용자 인터페이스를 제공함으로서 기능이 개선되었다.

논문의 구성은 제 2장에서 기존의 분석도구 와 RIA(Rich Internet Application)의 특징 및 장점 대해 설명한다. 제 3장에서는 웹2.0기반 RIA 개발방식으로 구현한 DNA서열 분석도구의 구현에 대해 설명하고, 제 4장에서 결론 및 향후 과제를 제시한다.

2. 관련연구

2.1 기존의 DNA 서열 분석도구에 사례

DNA서열 분석 중 실험에서 얻어낸 결과를 해석하기 위하여 이용하는 주석용 도구(annotation tool)는 생물정보학에서 자주 이용된다. 실험에서 얻은 결과에, 이에 관련된 정보를 붙이는 작업을 주석이라 한다.

DNA서열 분석도구에 주석 작업을 할 수 있는 프로그램은 가장 널리 알려진 Artemis 프로그램과 이와 비슷한 기능을 가진 AnnHyb 프로그램이(그림 1) 있다. AnnHyb은 oligonucleotides와 같은 nucleotide sequence나 longer sequence(200,000bp 이하)를 다루는데 도움을 주는 프로그램이다. 이 프로그램에는 format conversion, sequence viewer, sequence editor, oligonucleotides alignment, restriction analysis, pattern searching, retrieval from servers, multi-alignment viewer, consensus determination등의 기능을 쉽게 할 수 있는 도구들이 있다.[10]

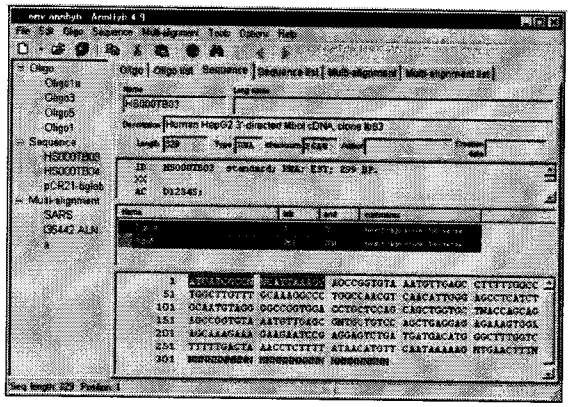

그림 1. AnnHyb 프로그램 서열분석 화면

Artemis는 서열분석기능과 함께 EMBL와 GenBank, GFF 형식을 읽어서 그 내용을 그래픽으로 표현해 준다. (그림 2)는 Artemis의 메뉴와 선택한 서열의 정보요약을 보여준다. 또한 서열을 구성하고 있는 각 염기(A/C/G/T)서열을 그래픽으로 표현 하여 사용자에게 효율적인 서열분석을 할 수 있게 한다. 막대는 색으로 구분하여 막대의 색만으로도 서열의 특징이 무엇인지 구분 지어 놓았다.

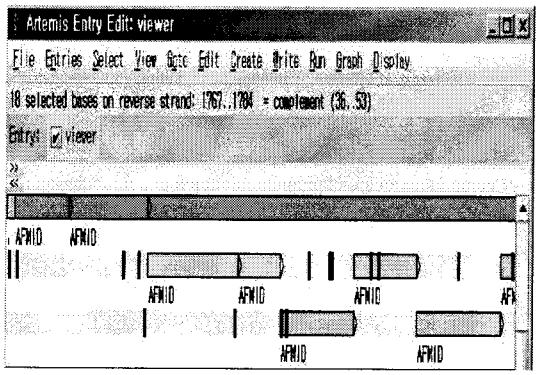

그림 2. 서열분석 프로그램 Artemis의 실행 화면

Artemis의 프로그램은 막대로 표현한 서열들을 염기서열로 표현하는 프레임을 따로 만들어 염기서열을 분석하는 위치를 알 수 있도록 구현 하였다. 마지막 프레임은 막대 색의 정의와 서열의 위치, 주석을 표현한다.(그림 3)

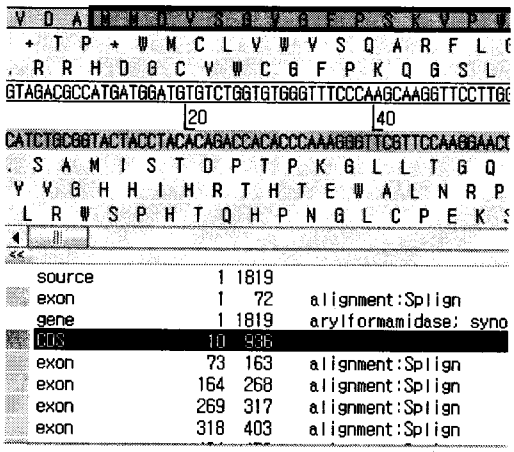

그림 3. Artemis의 염기서열 표현과 주석표현

2.2 RIA(Rich Interne Application)의 특징 및 장점

본 연구에서는 사용자의 편의성과 분석 효율성을 높이

기 위해 유전자 분석과정에서 여러 단계를 거쳐야 했던 기존의 웹의 문제점을 보안한 웹2.0 기반의 RIA개발방식을 제공하고자 한다.

기존의 웹 환경은 페이지단위 기반구성 및 계층적 연결 로 인한 서열분석의 결과 과정이 지연되는 문제점을 RIA의 클라이언트에서 프로세싱 및 프론트 레이어와 데이터 분리를 통해 해결 할 수 있다.(표 1)

표1. RIA의 특징 및 장점

Rich Internet Applications
다양한 디바이스 지원(PDA, 핸드폰, 웹, 디지털 TV)
클라이언트와 서버의 연동
컴포넌트 기반 개발
Vector 기반
HTML 보다 서버 부담 감소
사용자의 편리한 UI(User Interface)

RIA 개발 방식은 페이지의 재 로드 없이 원하는 데이터만 서버에 요구하고 바로 받아서 처리할 수 있다. 또한 실시간으로 서버와 데이터 교환이 가능하고, 다양한 사용자 컨트롤과 그래픽 효과를 개발할 수 있도록 컴포넌트 라이브러리를 제공한다. 이러한 RIA의 장점을 가지고 새로운 분석도구를 구현하면 유전자 분석 효율성을 증진시킬 수 있다.

RIA 개발 방식에는 어도비의 Flex, Ajax, WPF, Lazzlo, Droplet등이 있다. 특히, Flex는 개발 생산성, 애플리케이션 활용, 애플리케이션 성능 향상의 효과를 극대화한 RIA 개발 방식이다. 본 연구에서 RIA 개발 방식에서 가장 널리 알리진 어도비의 Flex2로 DNA서열 분석도구를 구현 하였다.[11]

3. 웹2.0기반 DNA서열 분석도구의 구현

3.1 개발목적

GenBank 파일을 보면서 많은 정보들이 들어있지만, Text 형태로 데이터를 보여주고 있어 한눈에 알아보기 어렵고, 일반 사람들이 보기에는 무슨 내용을 포함하고 있는지 알 수 없는 문제점을 발견하였다. 이러한 문제점을 보완 하기위하여 Text 기반의 서열분석 보다 RIA 환경에서 그래픽 한 서열분석을 보여준다면 서열 분석 효과를 높일 수 있다.

시간과 장소에 구애 받지 않고 손쉽게 정보의 전달과 이용이 가능한 네트워크 기반에 이 서열 분석 프로그램을 적용하면 기존의 서열분석 프로그램을 설치할 필요 없이 웹에 접속만으로 DNA서열 분석을 할 것으로 본다.

3.2 DNA서열 분석도구의 화면구성 원리

HTML 방식보다 역동적인 화면 연출이 가능하며 화면이 바뀔 때마다 페이지를 새로 고쳐야 했던 문제점을 RIA(Rich Internet Application)에서 한 페이지 내에서 모든 정보 이용이 가능하기 때문에 웹 사용성도 비약적으로 높일 수 있는 장점이 있다. (그림 4) 과 (그림 5) 는 HTML으로 구현했을 때와 RIA로 구현한 화면을 비교한 것이다. (그림 4)에서 영화예매 지역을 선택하면 선택한 지역페이지로 화면이 새로 바뀌며 화면이 바뀔 때마다 페이지를 새로 고쳐야 했지만, (그림 5)에서 보여주는 RIA환경은 한 페이지에서 모든 화면을 보여줄 수 있다. 그 결과 대용량 데이터를 다루는 업무의 경우 HTML의 작업에 불편하였던 점을 RIA환경에서 극복할 수 있다.

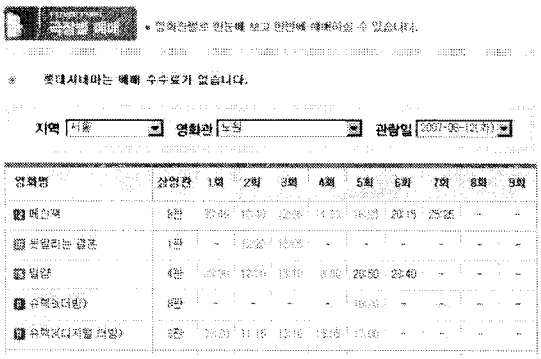

그림 4. HTML로 구현한 영화 예매 화면

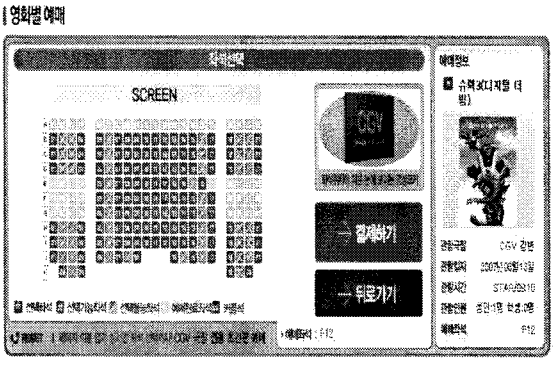

그림 5. RIA로 구현한 영화예매 좌석선택 화면

3.4 GenBank파일의 DNA서열 분석과정

DNA서열은 번역(translation)을 통해 단백질서열로 만들어진다. 번역이란 유전자 부분의 DNA 서열이 RNA 서열로 전사(transcription)된 후 이를 바탕으로 단백질 서열을 만드는 과정이다. 번역과정에서 RNA 서열 자체가 단백질 서열로 변신하지는 않는다. 세포내에서 단백질 서열을 만드는 기능을 가진 호소들이 RNA 서열, 특히 전령 RNA(mRNA) 서열을 읽으면서 이미 세포내에 존재하는 개개의 아미노산 분자들을 순서에 맞게 연결하여 단백질 서열을 만드는 것이다.

(그림 6) 은 RNA서열에서 단백질 서열로 번역하는 알고리즘이다. while 반복문은 i값이 RNA서열의 길이에서 2를 뺀 값보다 작아질 때까지 반복한다. 즉, RNA서열에서 I값이 가리키는 위치로부터 더 이상 세 개의 RNA 분자를 가져올 수 없을 때까지 반복하는 것이다.

```
while(i<len-2){
    codon = seq.substring(i,i+3);
    i+=3;
}
```

그림 6. RNA서열에서 단백질서열로 번역하는 알고리즘

위와 같은 알고리즘을 통해 각각의 코돈이 코딩하는 단백질분자가 생긴다. 본 연구에서는 GenBank의 DNA서열을 가지고 번역한 단백질 서열을 (그림 7) 과 같이 표현하였다. DNA서열에 번역과정을 통한 단백질서열의 표현으로 서열분석을 한 눈에 빠르게 분석할 수 있다.

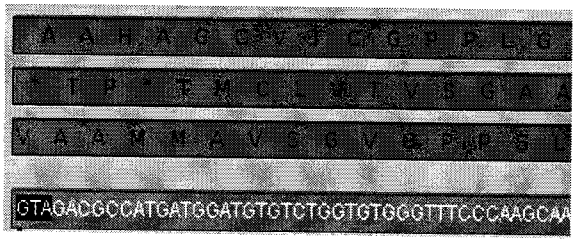

그림 7. DNA서열번역과 단백지서열의 표현

3.4 RIA기반 DNA서열 분석도구의 구현

본 서열 분석 도구는 사용자에게 정보를 알려주는 HOME, 서열의 기본정보, 서열보기 와 같이 3가지 구조로 구현된다. 서열보기는 기존에 Text로 되어있는 데이터를 한눈에 쉽게 분석하기위해 차트와 눈금으로 구현된다.

서열 분석 과정에서 서열의 특성 및 서열이 가진 유전자의 위치 및 기능 등을 분석하여 주석을 작성 하여 다른 연구자들도 알 수 있도록 약속된 형식의 파일로 정리하는데, 그 대표적인 형식이 GenBank 파일 (그림 8) 이다.[12] 서열 기본정보에서 기존에 복잡해 보이는 GenBank파일을 파싱하여 사용자로에게 서열분석의 내용을 쉽고 빠르게 이해하게 한다.(그림 9) 서열의 특징부분을 왼쪽화면에 따로 구현하여 서열이 가진 특징을 빠르게 검사 할 수 있다.

```
LOCUS       NM_001010982              1819 bp
DEFINITION  Homo sapiens arylformamidase (AFM
ACCESSION   NM_001010982 XM_496246
VERSION     NM_001010982.2  GI:142374718
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Cra
            Mammalia; Eutheria; Euarchontogli
            Catarrhini; Hominidae; Homo.
COMMENT     PROVISIONAL REFSEQ: This record h
            NCBI review. The reference sequen
            On Apr 6, 2007 this sequence vers

            Sequence Note: removed 1 bases fr
            to the reference genome assembly.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENT
            1-1819              BX648442.1
FEATURES             Location/Qualifiers
     source          1..1819
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="17"
                     /map="17q25.3"
```

그림 8. GenBank 파일의 원본

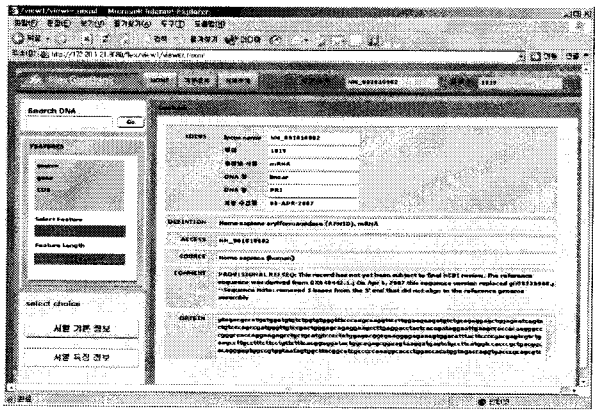

그림 9. RIA환경에서의 GenBank파일을 파싱한 화면

모든 생명현상은 DNA 서열에 들어 있는 정보에 의해 일어나고, 그러한 정보는 서열을 여러 가지 방법으로

조작해야 얻을 수 있다. 서열의 조작 과정에서 효율적인 분류를 위해 (그림 10)에서 보는 바와 같이 서열을 한 줄로 배치하였다. 이러한 방식은 서열 분석을 일괄적으로 확인할 수 있으므로 서열분석의 생산성을 높일 수 있게 된다. 왼쪽화면에는 서열의 특징과 위치를 알 수 있게 하여 정보를 빠르게 얻을 수 있다.

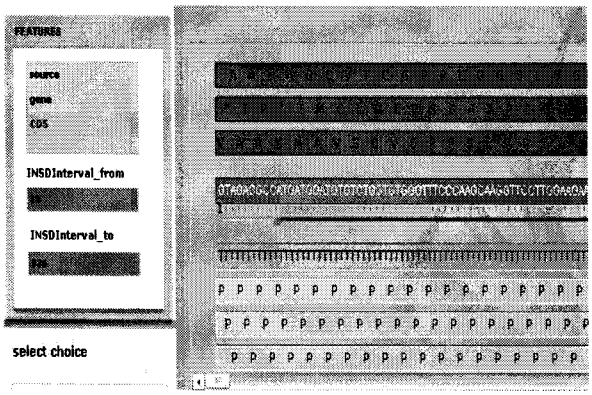

그림 10. 일괄적인 서열분석을 위한 인터페이스

서열분석의 특성에는 서열을 구성하고 있는 각 염기(A/C/G/T)의 비율이 있다. GC-contents는 게놈의 전체 염기서열이나 특징 영역의 부분 서열에서 G와 C가 차지하는 비율을 말한다. RIA환경에서 서열분석 도구는 (그림 11)과 같이 그래프를 이용하여 전체 GC-contents를 한번 에 파악 하고, Text의 표현으로 정확한 수치를 알 수 있다.[13]

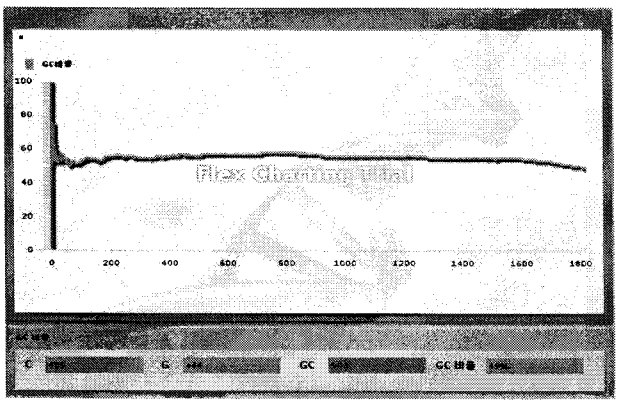

그림 11. GC-contents 인터페이스

4. 결론

본 연구에서는 유전체 사업의 수행에 필수적인 소프트웨어를 웹2.0 기반 RiA로 구현하여 생물학자들의 편리성과 분석의 생산성을 높인 분석도구를 제시하였다.[15]

지금까지 유전체 분석도구는 프로그램을 직접 구하고 설치하여 프로그램에서 요구하는 입력 형태로 데이터를 제공해야 하므로 컴퓨터에 대해 전문적인 지식이 부족한 생물학자 입장에서는 어려운 일이 아닐 수 없다. 또한 기존 웹 프로그램은 페이지 로딩 시간이 많아 대용량 서열 데이터를 분석하기에 많은 시간이 들었다. 이러한 단점을 보완한 RIA환경의 DNA서열 분석도구의 개발은 생물학자들의 편리성과 생산성을 제고할 것으로 보인다.

향후 GenBank 파일에 대하여 실시간으로 서열을 검색하며, 더 나아가 단백질 데이터 뱅크를 이용하여 PDB(Protein Data Bank)를[14] 3D로 표현하는 프로그램을 추가하여 바이오인포맥틱스 연구에 더욱 효율적인 도구로 개발할 계획이다.

참고문헌

[1] Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research Nature , 422:835~847, 2003

[2] 김상수, 유전자 기능 분석을 위한 바이오인포매틱스 기법 소개, 대한간학회지, 제 10권, 제 1호, 11~21, 2004

[3] 니케이 바이오테크, "최첨단 리포트 유전자 비즈니스" , pp304 김영사, 2002

[4] 최범순, 이경희, 권해룡, 조완섭, 이충세, 김영창. 웹 기반 통합 유전체 분석 시스템의 설계 및 구현, Journal of Korea Multimedia Society, 제 7권, 제 3호, pp. 408-417, 2004

[5]Berriman, M., Rutherford, K., Vierwing and annotating squence data with Artemis", Bioinformatice, Vol 4(2), 2003

[6] Rutherford, K., Parkhill,J., Crook, j. et al.,

Aretemis:Seuence visualization and annotation, Bioinformatics, Vol.16(10), pp. 944-945, 2000

[7] URL: http://www.sanger.ac.uk/Software/Artemis

[8] L. Catherine, A Web interface generator for molecular biology program in Unix. Bioinformatics, Vol.7, No.1, pp.73-83, 2001

[9] 옥상훈, 예제로 배우는 Adobe 플렉스2, pp.36-38, 에이콘, 2006

[10] URL: http://bioinformatics.org/annhyb/

[11] 신현삼, 배경희, 박강희, 개발자를 위한 플렉스2 실무테크닉, pp27-43, 성안당, 2007

[12] 제임스 티스달,박현석, "펄로 시작하는 바이오인포맥틱스" ,pp.302-353 한빛미디어,서울,2002.3

[13]박현석, 정철. 자바로 배우는 바이오인포매틱스. pp.72, pp.114 사이텍미디어,2006

[14] 신시아 기버스, 퍼 젬벡, Bioinformatics Computer Skills, pp369-400, O'REILLY,2002

[15] URL: http://jotajan.oranc.co.kr/viewer.swf