M1-2107-00-0027

EST 및 유전자 발현 프로파일 대량 생산 시스템 개발 (II)

High Throughput Analysis of EST and Gene Expression Profiles

EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석

Genome annotation and finding noble genes using EST clustering

이 화 여 자 대 학 교

과 학 기 술 부

제 출 문

과학기술부 장관 귀하

본 보고서를 "EST 및 유전자 발현 프로필 대량 생산·분석 시스템 개발(II)"과제 (세부과제 "EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석") 의 보고서로 제출합니다.

2005. 1. 17.

주관연구기관명 : 이화여자대학교

주관연구책임자 : 이 상 혁

보고서 초록

<table>
<tr><td>과제관리번호</td><td>M1-2107-00-0027</td><td>해당단계
연구기간</td><td>2002.12.01
-2004.09.30</td><td>단계 구분</td><td>(1단계) / (총1단계)</td></tr>
<tr><td rowspan="2">연구사업명</td><td>중 사 업 명</td><td colspan="4">바이오연구개발사업</td></tr>
<tr><td>세부사업명</td><td colspan="4">생물정보학연구개발사업</td></tr>
<tr><td rowspan="2">연구과제명</td><td>중 과 제 명</td><td colspan="4">EST 및 유전자 발현 프로필 대량 생산·분석 시스템 개발(II)</td></tr>
<tr><td>세부(단위)과제명</td><td colspan="4">EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석</td></tr>
<tr><td>연구책임자</td><td>이 상 혁</td><td>해당단계
참여연구원수</td><td>총 : 10 명
내부 : 7 명
외부 : 3 명</td><td>해당단계
연구비</td><td>정부: 190,000 천원
기업: 천원
계: 190,000 천원</td></tr>
<tr><td>연구기관명 및
소속부서명</td><td colspan="2">이화여자대학교 분자생명과학부</td><td colspan="2">참여기업명</td><td></td></tr>
<tr><td>국제공동연구</td><td colspan="5">상대국명 : 상대국연구기관명 :</td></tr>
<tr><td>위 탁 연 구</td><td colspan="5">연구기관명 : 연구책임자 :</td></tr>
<tr><td colspan="4">요약(연구결과를 중심으로 개조식 500자이내)</td><td>보고서
면수</td><td>71</td></tr>
</table>

유전체 지도의 해석은 기능유전체학의 핵심이며, 향후 생물학 연구의 기반임. 본 과제에서는 유전체 해석을 위한 유전자 구조, 기능, 발현을 예측하는 종합 시스템을 완성하였으며, 그 구체적인 결과는 다음과 같음.

◆ EST clustering을 통한 유전자 예측 알고리즘을 완성
- EST clustering과 transcript assembly 과정을 효율적으로 통합한 ECgene 알고리즘을 개발
- Human, mouse, rat 게놈의 분석 결과를 UCSC genome center를 통하여 제공함.
- ASmodeler 개발: 사용자 서열을 ECgene 알고리즘으로 분석하는 웹 서버 프로그램

◆ 유전자 기능 및 발현 분석 방법 개발
- ECfunction: 서열의 유사성, 도메인/모티프 분석, GO 분석을 통한 종합적인 기능 분석 시스템
- ECexpression: EST에 대한 cDNA library를 분석하여 유전자 발현의 조직, 질병 연관성 등을 예측
- ECprofiler: 특정 조직, 특정 질환에 관련성이 큰 유전자를 게놈 전체를 대상으로 찾는 프로그램

◆ ECgene 유전자 분석 포탈 사이트 개발
○ ECgene, ECfunction, ECexpression, ASmodeler를 종합한 포탈 사이트를 개발
○ ECgene의 결과를 특정 목적에 맞게 가공한 지식기반의 2차 DB 개발
- ChimerDB: 암의 발병, 전이와 관련성이 큰 것으로 알려져 있는 염색체 전좌로부터 발생하는 fusion mRNA와 EST 서열의 데이터베이스
- Antisense DB: 유전자 cluster 중에서 antisense 관계에 있는 서열 데이터베이스 구축

본 연구결과는 생물학, 의학 연구의 기반지식이 될 뿐만 아니라 신약개발과 질병의 예방, 치료 및 예후 예측 등의 의약산업과 밀접한 관련이 있음.

색 인 어 (각 5개 이상)	한 글	유전자 예측, 유전자 발현, 유전자 기능, 게놈 지도 해석, 유용 유전자 발굴
	영 어	gene prediction, gene expression, gene function, genome annotation, candidate gene search system

요 약 문

Ⅰ. 제 목

EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석

Ⅱ. 연구개발의 목적 및 필요성

유전체 지도의 완성은 현대 생물학과 의학에서 가장 획기적인 사건으로 유전체 지도의 해석이 관련분야의 향후 경쟁력을 좌우하는 중요한 요인임. 본 과제의 목적은 유전체 지도의 해석을 위한 종합적인 방법론을 개발하는 것으로, 특히 EST 데이터와 같은 공개된 정보를 이용하여 유전자 구조, 기능 및 발현을 예측하는 시스템을 개발함.

Ⅲ. 연구개발의 내용 및 범위

주요 연구 내용은 (1) EST clustering을 통한 유전자 예측 방법의 개발, (2) 유전자 서열의 분석을 통한 기능 예측, (3) 유전자 예측 결과의 유전체적 분석, (4) library 분석을 이용한 발현 패턴의 예측, (5) 유전체 지도의 분석 결과를 데이터베이스로 구축하고 웹을 통하여 공개하며, (6) 특정 기능과 발현을 지닌 유용 유전자의 발굴 방법을 개발하는 것임.

Ⅳ. 연구개발결과

Human, mouse, rat에 대하여 유전자 구조, 기능, 발현을 예측하는 시스템인 ECgene (gene prediction by EST clustering)을 개발하였음. ECgene은 exon 단위에서의 유전자 변이형을 만드는 alternative splicing의 분석에 강점이 있으며 그 구체적인 내용은 다음과 같음.

- mRNA와 EST의 genomic alignment를 분석하여 서열을 clustering하고, 그래프 이론으로 유전자의 변이형을 예측하는 ECgene 프로그램을 완성하였음. 사용자 서열을 ECgene으로 분석할 수 있는 ASmodeler를 웹서버로 구축하고 Nucleic Acids Research의 Web Server Issue에 발표하였음. 알고리즘 자체에 관한 논문은 2005년 중에 Genome Research에 게재 예정임.
- 예측된 유전자 서열을 분석하여 유전자의 기능과 발현을 예측하는 ECfunction과 ECexpression 시스템을 완성하였음. 그 결과를 데이터베이스로 구축하고 ECgene 웹사이트를 개발하여 국내외에 연구결과를 공개하였고 Nucleic Acids Research의 Database Issue에 게재하였음.
- 기능과 발현 분석을 근거로 게놈에서 특정 기능과 발현 패턴을 갖는 유전자를 검색하는 방법으로 ECprofiler를 개발하였음.
- ECgene cluster의 유전자 모델로부터 antisense DB, ChimerDB와 같은 지식기반의 2차

데이터베이스를 구축하였음.

V. 연구개발결과의 활용계획

ECgene은 게놈 지도가 밝혀진 모든 생물의 유전체 분석에 적용될 수 있으며 EST clustering, 유전자 변이형의 분석, 발현 패턴의 예측 등에서 많은 장점을 지닌 알고리즘임. 이를 다른 생물에 확대·적용하고 시스템 개발과 서비스를 강화하기 위하여 국가유전체정보센터와 협력하여 세계적인 게놈정보 제공 사이트로 개발할 예정임.

특정 조직과 질환에 관련성이 큰 유전자는 신약개발의 타겟, 질병의 진단과 치료에 직접적으로 응용될 수 있음. ECprofiler를 이용하여 다양한 질환에 대한 후보 유전자 목록을 작성하고, 이를 진단용 DNA chip의 디자인과 신약 타겟 유전자의 개발 등에 응용할 계획임.

이 외에도 본 연구결과는 기존의 생물학 연구뿐만 아니라, 유전자 변이형, 약물유전체학, noncoding RNA를 포함한 전사체 연구와 같은 새로운 패러다임의 유전체 생물학 연구의 기반이 될 것으로 예상됨.

SUMMARY

Ⅰ. Title

Genome annotation and finding noble genes using EST clustering

Ⅱ. Background and Aim

Completion of the human genome map is a milestone for research in modern biology and medical sciences. Proper genome annotation is an essential part to accelerate research in related fields. The aim of current project is to develop a pipeline for automatic genome annotation using EST database, which includes gene prediction, functional annotation and gene expression analysis.

Ⅲ. Contents

Major topics of current project are as follows:
(1) Gene prediction using genome-based EST clustering (2) Functional annotation of the resulting clusters, (3) Genome-scale analysis of gene prediction result, (4) Expression analysis of EST clusters using cDNA library information, (5) Construction of database and web site for genome annotation of human, mouse and rat, (6) and Development of profiling method to search for genes with specific function and expression pattern.

Ⅳ. Results

We developed a genome portal site, ECgene (gene prediction by EST clustering), that includes gene prediction, functional annotation, and expression analysis for human, mouse, rat genomes. Its major strength is analysis of alternative splicing using graph theory, and the website provides a suite of tools with many unique feature in analyzing gene expression and function.

- Gene Prediction: ECgene algorithm clusters the genomically aligned mRNA and EST sequences, and we analyze the exon connectivities using graph theory to create assembled sequences. The resulting transcript models include detailed analysis of alternative splicing events. The algorithm is implemented as a web server in ASmodeler, which is published in the Web Server issue of Nucleic Acids Research in 2004. A manuscript on detailed algorithm is in press in Genome Research

currently.

- Functional Annotation and Expression Analysis: Predicted transcriptomes are analyzed in terms of function and expression. Functional annotation, ECfunction, includes domains and motif changes due to alternative splicing. Expression prediction, ECexpression, is based on analysis of cDNA and SAGE libraries. The resulting annotation database is available at http://genome.ewha.ac.kr/ECgene/, and was published in the Database issue of Nucleic Acids Research in 2005.
- ECprofiler searches the ECgene database to find noble genes with desired functions and expression patterns. It can be used to find tissue-specific genes or disease-related genes which are useful for therapeutic drug development.
- The ECgene clusters are further analyzed to construct the databases of antisense transcripts and chimeric sequences. These are just the initial examples of "knowledge-based" secondary databases deduced from the ECgene.

V. Prospects and Future Plans

ECgene can be applied any organism whose genome map is available. Its clustering algorithm is a state-of-the-art in terms of analyzing alternative splicing, and the annotation procedure has many unique features. We plan to expand the number of organisms and develop various tools and the secondary databases in cooperation with the National Genome Information Center of Korea.

Tissue-specific and/or disease-related genes have tremendous values for developing therapeutic drugs and diagnognostic markers. ECprofiler will be used to find various disease-related genes.

Alternative splicing, the major strength of the ECgene database, is an important mechanism of incresing transcriptome diversity in mammalians. We will continue to annotate alternative splicing events, especially its connection to disease, SNP, and pharmacogenomics. Furthermore, regulation by noncoding RNAs is a hot topic these days. ECgene is an ideal system to study noncoding transcriptome, and we plan to develop a database of noncoding RNAs based on the ECgene system.

CONTENTS

목 차

제 1 장 연구개발과제의 개요

1.1 연구개발의 목적

본 과제의 최종 목표는 새로운 방식의 EST clustering을 이용하여 인간 게놈지도에 포함된 미지의 유전자를 예측하고, 그 결과를 유전자의 발현, 기능의 측면에서 종합적으로 분석하는 genome annotation 시스템을 구축하는 것이다. 또한 그 결과를 데이터베이스화하고 다양한 분석 수단을 개발하여 세계적인 수준의 게놈정보를 제공하는 웹 사이트를 구축할 것이다. 또한 유전자의 발현과 기능 분석을 근거로 질병의 진단과 치료 및 신약개발 등의 현대 바이오산업에 유용한 유전자를 발굴하는 시스템을 개발한다. 연구 진행 단계에 따른 세부적인 목표는 다음과 같이 요약할 수 있다.

- mRNA와 EST 서열을 유전체 지도에 mapping하여 clustering하는 genome-based clustering 방법의 개발
- 각 cluster에 포함된 EST의 assembly 과정을 통하여 alternative splicing을 포함한 유전자 모델을 계산하는 알고리즘의 개발
- EST의 조직 발현성과 질병관련성에 대한 데이터베이스의 구축
- 각 cluster에 해당하는 유전자의 기능 분석과 유용 유전자의 발굴
- 각 cluster의 제반 사항을 보여주는 인터넷 사이트 및 인간 유전체 browser의 개발

1.2 연구개발의 필요성

본 연구에서는 인간 유전체 프로젝트의 진행에 따라 대량 생산되고 있는 EST를 분석하여 유용 유전자를 발굴하고 이들의 발현 프로필을 분석하는 시스템을 구축하고자 한다. 21세기 생명공학의 핵심은 유용 유전자 확보로서, 2003년 상반기에 완성되는 인간 유전체 지도의 완성 후 수많은 유전자의 기능을 단시간에 규명하기 위한 치열한 경쟁이 시작될 것으로 예상된다.

인간 유전체 지도의 완성은 단순히 생물학 연구뿐만 아니라 사회 전반에 걸쳐 폭넓은 영향을 미칠 전망이다. 이 유전체 지도의 분석을 통한 유전자의 발굴은 소위 미국 개척시대에 금광 개발의 붐에 비유하여 'golden path'라고 불려질 정도로 전 세계적으로 활발한 연구가 진행되고 있다. 현대 바이오 산업의 핵심은 질병의 원인 규명과 치료에 있다고 할 수 있다. 이에 관련된 연구 분야는 인간 유전체 지도의 분석을 통한 유전자 발굴에 관련된 유전체학(genomics), 이로부터 얻어지는 단백질의 구조, 기능 및 단백질-단백질의 상관관계를 밝히는 단백질정보학(proteomics), 신약개발후보 물질과 이들 단백질간의 상호 작용을 밝히고자 하는 화학정보학(cheminformatics) 등을 들 수 있다. 이를 생물학적 정보의 흐름의 측면에서 분석하면 아래의 그림 1과 같이 나타낼 수 있으며, 그 정보의 시발점이라고 할 수 있는 인간 유전체 지도의 완성은 분자생물학 연구에 한 획을 긋는 중요한 일이라고 할 수 있다. 인간 유전체 지도의 완성은 생명과학

연구에 있어서 유전체 시대(genomic era)의 개막을 의미한다. 가까운 시일 내에 이들 유전자로부터 나오는 단백질을 총체적으로 분석하는 단백질정보학이 중요해 질 것으로 전망되지만, 이는 유전체의 분석을 근거로 하였을 때 더욱 강력한 연구방법이 된다는 점에서 유전체 분석의 중요성은 더욱 커질 것으로 예상된다.

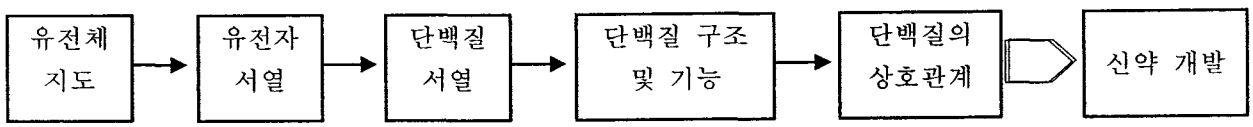


그림 1. 생물학적 정보의 흐름

새로운 유전자의 예측과 유전자 특성의 규명은 현대 바이오산업의 가장 핵심이 되는 부분이다. 특히 유전자의 발현 프로필에 대한 연구는 생명현상의 이해에 필수적인 부분으로 SAGE (Serial Analysis of Gene Expression)와 microarray 등의 다양한 방법으로 연구되고 있다. 또한 유전자와 질병과의 상관관계는 신약 개발을 위하여 필수적인 연구 분야이다.

EST clustering을 통한 유전자 모델링에서 직접적으로 얻어지는 정보의 종류와 응용 분야의 예를 간단히 들면 다음과 같다.

1) 새로운 유전자의 발견 및 유전자 구조의 예측: EST clustering의 가장 기본적인 응용으로 새로운 유전자의 발견을 들 수 있다.

2) Alternative splice variant의 모델링: Alternative splicing에서 특정 exon의 존재 여부는 세포 내에서 일어나는 과정의 on/off 스위치 역할을 하는 경우가 많이 있어서 신약 개발의 주요 목표가 된다. 특히 질병과 관련된 유전자에서 일어나는 alternative splicing은 유전자 칩을 이용하여 발병 여부를 진단할 수 있는 훌륭한 지표가 된다.

3) Assembled sequence에 대한 도메인, 모티프 분석을 통하여 단백질의 기능에 대한 간접적인 정보를 얻을 수 있다.

4) 유전자 발현 프로필의 예측: EST의 출처에 해당하는 조직과 cDNA library를 분석하면 각 cluster에 해당하는 유전자의 신체 조직에 따른 발현성과 질병관련 여부에 대한 간접적인 정보를 얻을 수 있다. 또한 예측된 mRNA 서열에서 SAGE SAGE (serial analysis of gene expression) tag를 뽑아내어 정량적인 발현 패턴을 예측할 수도 있다.

5) SNP의 발굴: EST의 정렬을 분석하면 SNP 중에서도 coding SNP를 발굴할 수 있다.

이와 같이 본 과제의 주된 내용인 EST clustering을 통한 유전자 구조의 예측, 유전자 기능 및 발현의 분석은 기능유전체학 (functional genomics)과 게놈 정보 분석 (genome annotation)의 시발점으로 그림 1에 나타낸 전 분야에 큰 파급효과를 갖는다. 특히 최근의 연구 결과에 의하면 alternative splicing에 의한 유전자 변이가 많은 질병과 관련이 되어 있고, 약물유전체학과도 밀접한 연관이 있는 것을 밝혀졌다. 또한 microarray와 SAGE 실험을 이용한 최근의 연구결과에 의하면 단백질을 만드는 유전자뿐만 아니라 microRNA, antisense RNA, siRNA등과 같이 RNA 상태로 기능을

수행하는 noncoding RNA에 의한 조절 메커니즘이 중요한 작용을 하는 것으로 밝혀지고 있다. 따라서 이와 같은 모든 정보를 포함하고 있는 전사체 (transcriptome)의 분석이 더욱 중요해지고 있다. 본 과제의 유전자 예측 방법은 mRNA와 EST를 이용한 것으로 전사체 연구에 이상적인 유전자 모델링 방법이라고 할 수 있다. 따라서 본 과제의 결과는 향후 alternative splicing에 의한 전사체의 다양성, SNP를 포함한 약물유전체학적인 연구, noncoding RNA에 의한 생체조절 메커니즘의 규명 등에 두루 적용될 수 있는 기반을 제공할 것이다.

1.3 연구개발의 범위

가. EST clustering을 통한 유전자 예측 방법의 개발
(1) mRNA와 EST 서열을 게놈 지도에 mapping한 결과를 분석하여 clustering을 하는 genome-based EST clustering 방법의 개발
(2) EST cluster의 assembly 과정을 통하여 alternative splicing을 고려한 mRNA 서열의 예측 방법 개발

나. 유전자 서열의 분석을 통한 기능 예측
(1) mRNA 서열 분석을 통하여 단백질의 기능에 관련된 도메인, 모티프 분석
(2) Gene ontology database를 이용한 기능 분석 pipeline의 개발

다. 유전자 예측 결과의 유전체적 분석
(1) Alternative splicing에 의하여 기능에 관련된 도메인에 생기는 변화에 대한 연구
(2) 지식 기반의 2차 데이터베이스 구축

라. Library 분석을 이용한 발현 패턴의 예측
(1) EST cluster의 발현 분석을 위한 cDNA library의 분류
(2) mRNA 서열로부터 유추해 낸 SAGE tag을 이용한 유전자 발현 방법의 개발

마. 유전체 지도의 분석결과에 대한 데이터베이스와 웹 사이트 구축
(1) 유전자 예측 및 분석 결과의 데이터베이스 구축
(2) 유전자 구조, 기능, 발현에 관한 정보를 포괄적으로 제공하는 웹사이트 개발

바. 특정 기능과 발현을 지닌 유용 유전자의 발굴 방법을 개발
(1) 장기·조직 또는 질병에서 차등 발현을 하는 유전자의 발굴을 위한 통계기법의 도입
(2) 기능과 발현 패턴을 지닌 유전자 검색 프로그램의 개발 및 검색 결과의 DB화

제 2 장 국내외 기술개발 현황

2.1 국내·외 기술개발 현황

가. EST clustering 알고리즘

(1) 본 과제의 시작 전후에는 UniGene, CAP3, d2_cluster와 같은 대부분의 알고리즘이 mRNA/EST 서열을 자신끼리 비교하는 transcript-based 방법을 사용하고 있음. 2003년 12월부터 NCBI의 UniGene이 본 과제의 알고리즘과 비슷한 genome-based 방법으로 전환하였고 현재 human, mouse, rat과 같이 genome map이 밝혀진 경우에는 이 방법을 사용하고 있음. 이는 genome-based method의 우월성을 보여줌.

(2) 2001년 이후로 alternative splicing의 분석에 관한 무수한 논문이 발표되었음. 대부분의 경우 transcript structure를 모델링하는 것이 최종 결과이고 본 과제와 같이 clone을 clustering하여 모델을 제시하는 것은 NCBI의 AceView 밖에 없음. AceView는 웹을 통하여 결과를 공개하고 있지만 아직도 논문을 발표하지는 않았음.

(3) Genome-based method는 현재 NCBI의 UniGene을 비롯하여 몇 개의 웹사이트가 존재함. 본 과제의 연구결과와 가장 유사한 방식이 NCBI의 AceView이나 최근에야 clustering 결과를 공개하고 있음. EBI에서도 그래프 방법을 이용한 clustering 결과를 개발하였으나 alternative splicing의 분석, clustering 알고리즘에서 본 과제가 개발한 방법이 우월한 것으로 생각됨.

(4) 국내에서는 BLAST를 이용한 간단한 clustering 알고리즘이 개발되었으나 실용화 단계에는 미치지 못한 것으로 판단됨.

나. Genome의 Functional Annotation

(1) 지난 3년간 단백질의 도메인/모티프를 이용한 기능의 예측, Gene Ontology가 부여된 단백질과의 서열 유사성을 이용한 기능 예측, 알려진 유전자와의 발현의 유사성을 이용한 pathway 및 기능의 예측 등의 측면에서 무수한 논문이 발표되었음.

(2) 본 과제의 기능 분석 pipeline은 RefSeq, Ensembl, H-Inv, FANTOM 프로젝트와 같은 대규모 full-length annotation 프로젝트에서 사용하는 방법과 유사함. 그러나 본 과제의 경우에는 alternative splicing에 의한 유전자 구조의 변화에 초점을 맞추어 도메인/모티프의 구체적인 변화를 보기에 적절하게 개발한 점이 장점임.

다. Genome 규모의 유전자 발현의 분석

(1) EST가 얻어진 cDNA library를 분석하여 발현을 예측하고 차등발현을 보이는 유전자를 구하는 방법도 많이 개발되었음. 그러나 대부분의 경우 특정 연구실의 목표를 찾기 위한 1회성 프로젝트임. 본 과제와 같이 계층적 분류 시스템을 만들고 library를 체계적으로 분류한 예는 SANBI에서 개발한 CGP (candidate gene profiler)를 들 수 있으며, 이는 다양한 종류의 조직과 질병에 적용할 수 있음. CGP는 EBI의 EnsMart

의 일부로 포함되어 있음.

(2) SAGE library를 이용한 발현의 예측도 많은 논문이 발표되었음. NCBI의 SAGEmap에서는 UniGene clustering을 근거로 tag를 뽑아서 공개된 library를 검색하는 방식으로 발현을 예측함. 이 방식은 EST 서열의 불확실성, clustering 방식의 결함 그리고 alternative splicing에 의한 variant를 고려하지 않았다는 점에서 많은 문제점이 있음. NCI의 SAGE Genie 방법은 tag extraction 방법에서 조금 더 진보하였고 다양한 보조 프로그램을 개발하여 널리 사용되고 있지만 여전히 clustering 방식의 불완전성에서 기인되는 근본적인 문제점을 해결하지는 못함.

(3) Microarray 데이터를 이용한 발현의 분석은 엄청난 데이터의 양과 함께 수많은 논문이 발표되었음. Stanford 대학의 SMD (Stanford microarray database), NCBI의 GEO (gene expression omnibus), EBI의 ArrayExpress 등이 대표적인 데이터 제공 사이트임. 이들에 대한 메타 분석도 최근 연구의 주요 경향 중 하나임.

라. Alternative splicing의 Annotation

(1) Alternative splicing에 의한 전사체의 다양성에 관련된 논문도 수 십편 이상 발표되었음. UCLA의 Christopher Lee 그룹에서는 UniGene cluster를 분석한 ASAP (alternative splicing annotation project)를 개발하였고, Washington University 그룹에서도 genome-based 방법으로 splice variant를 예측하고 DB화 하였음. 최근에는 EBI의 Tharanaj 그룹을 중심으로 한 국제 협동연구를 통하여 alternative splicing에 대한 지식기반의 데이터 베이스인 ASD (alternative splicing database)를 구축하고 있음. 이 외에도 ASG, ASDB, SpliceNest와 같은 프로그램 등이 있음.

(2) Alternative splicing과 질병, 약물유전체학, 신약개발과의 관련성에 대한 논문도 꾸준히 발표되고 있음. 대부분의 경우 특정 유전자에 대한 것으로 게놈 전체를 대상으로 한 연구는 거의 없는 실정임. 가까운 장래에 이런 방향으로 연구 개발이 진행될 것으로 예상됨.

마. Genome Annotation Portal Site

(1) 이 분야는 세계적으로 가장 경쟁이 치열한 분야로 미국 NCBI의 MapViewer, 유럽의 EBI의 Ensembl, 미국 UCSC genome center에서 제공하는 genome browser가 널리 쓰이고 있음. NCBI와 EBI가 내부의 연구결과를 바탕으로 유전자 정보를 제공한다면, UCSC의 genome center는 자체 연구에 추가적으로 전 세계의 다른 연구진이 제공하는 우수한 결과를 함께 제공하는 장점이 있어 널리 사용되고 있음. 알려진 유전자의 경우 이스라엘의 Weizmann Institute에서 개발한 GeneCards가 있음.

(2) 단백질의 측면에서는 SwissProt과 PIR이 널리 사용되었으며 현재 UniProt으로 통합되는 추세이다. Pathway DB도 현재에는 KEGG가 가장 유명하지만 다양한 종류의 새로운 데이터베이스들이 구축되고 있다.

(3) Alternative splicing에 대한 자세한 정보를 게놈 단위에서 보여주는 곳은 미국 NCBI의 AceView와 유럽 EBI의 ASD가 가장 유명하며 아직도 개발단계에 있음.

2.2 본 과제에서 개발된 기술의 위치

(1) 유전자 모델링 방법

ECgene의 유전자 모델링은 EST clustering과 assembly 과정을 통합한 결과로 기존의 유전자 예측 프로그램의 장점을 취합한 가장 진보된 알고리즘임. EST clustering에서는 예측되는 mRNA의 구조를 그래프 이론으로 풀었다는 측면에서 NCBI의 UniGene보다 우수함. 또한 polyA tail의 분석과 UTR 지역의 연장과 같은 면에서 다른 유전자 예측 프로그램이 비하여 많은 장점을 갖고 있음. 현재 ECgene의 알고리즘은 Genome Research에서 심사중임.

(2) 유전자 기능의 분석

Functional annotation 자체는 InterPro, GO와 같은 공개된 프로그램과 DB를 이용하여 큰 차이가 없음. 그러나 ECgene의 alternative splicing에 대한 모델을 기반으로 ECfunction은 splicing variant의 구조적 차이, 도메인/모티프의 변화 등을 쉽게 볼 수 있도록 만들었다는 점에서 강점이 있음.

(3) 유전자 발현의 분석

EST의 분석은 계층적 분류 시스템을 개발하고 이에 따라 약 8,600개의 human cDNA library, 900개의 mouse cDNA library, 300개의 human SAGE library, 50개의 mouse SAGE library를 수동으로 분류한 것이 큰 장점임. 또한 ECgene clustering을 기반으로 EST를 분석하여 variant에 따른 발현의 차이를 볼 수 있는 유일한 웹 사이트임. SAGE의 경우에는 ECgene의 mRNA 서열에서 SAGE tag를 뽑아내어 계산하였다는 점에서 SAGEmap이나 SAGE Genie에 비하여 큰 차이가 있음. 이렇게 뽑은 SAGE tag는 redundancy가 적고 신뢰도가 훨씬 높음. 이 경우에도 splice variant에 따른 발현의 차이를 볼 수 있다는 점이 큰 장점임. ECgene의 유전자 발현의 분석은 EST와 SAGE를 이용한 것에 관한 한 세계적으로도 가장 발전된 것으로 생각됨.

(4) 유전체 정보 포탈 사이트 개발

NCBI, EBI, UCSC 등의 genome center에서 개발된 genome browser와는 비교할 수 없지만 ECgene 웹사이트는 alternative splicing에 의한 유전자 변이에 관한 한 유전자 예측 방법이 우수하고, 기능 및 발현 분석에서 많은 장점을 지니고 있어 세계적으로도 독특하고 경쟁력이 있는 분석 결과를 제공하고 있다. 그 결과는 2005년 Nucleic Acids Research의 Database Issue에 게재될 예정이다.

제 3 장 연구개발수행 내용 및 결과

3.1 유전자 구조 예측 프로그램 (ECgene & ASmodeler)

가. ECgene 알고리즘의 개요

EST clustering과 transcript assembly 과정을 통합하여서 alternative splicing을 포함한 유전자 모델을 계산하는 방법을 개발하였다. 본 연구에서 개발한 알고리즘은 mRNA와 EST 서열을 genome map에 정렬시켜, 위치의 중복을 통하여 서열을 묶는 genome-based EST clustering 방법이다. 이 방법은 UniGene (Schuler et. al 1996) 과 같이 각 서열들끼리 비교하는 transcript 방법(Schuler et. al 1996; Quackenbush et al. 2001; Christoffels et al. 2001)에 비하여 많은 장점이 있어, 최근 NCBI의 UniGene도 이 방식으로 바뀌었다 [build #162]. 본 연구에서 개발한 방법은 exon-intron의 경계 부분을 공유하는 서열들을 묶었고, exon의 연결 패턴을 그래프 이론으로 분석하여 가능한 모든 경우의 alternative splicing type을 예측한다. 그림 2에 알고리즘의 개요를 정리하였다.

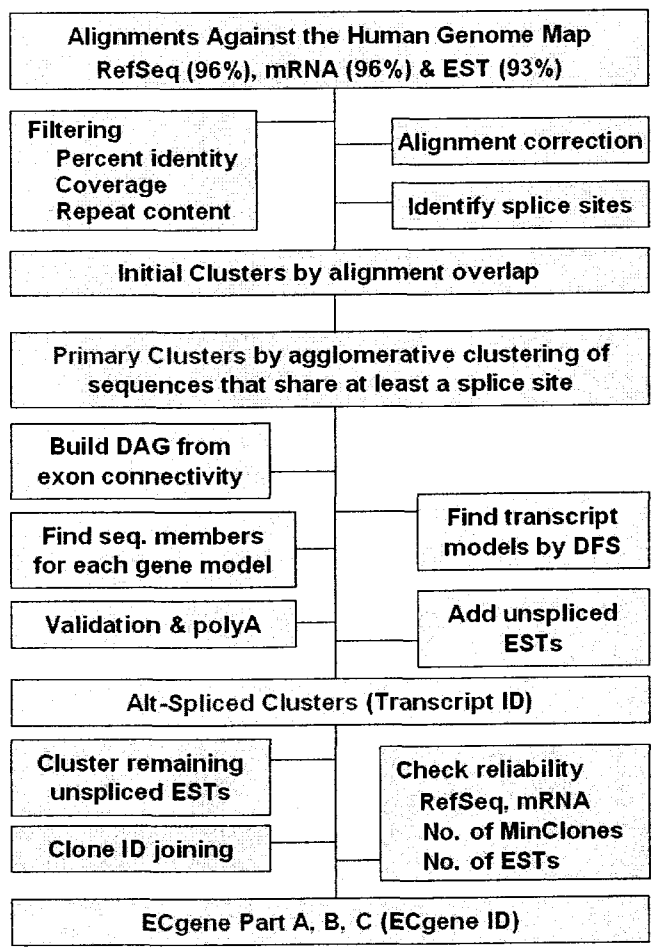


그림 2. ECgene 알고리즘의 개요

알고리즘의 주요 단계는 다음과 같다.

(1) Genomic Alignment
- data sources: dbEST, GenBank, RefSeq DB에서EST, mRNA, RefSeq 서열을 가져온다. 2003년 8월 16일 현재 human sequence는 EST가 5,401,534개, mRNA와 RefSeq는 141,950개의 서열이 있다.
- polyA의 제거: 서열을 genome에 정렬시키기 전에 polyA tail을 제거하기 위하여 EMBOSS (http://www.emboss.org) package의 trimest 프로그램을 사용하여 가능성이 있는 polyA 정보를 추출한다. PolyA 예측 방식은 그 정확도가~ 50% 수준으로 차후 genome alignment를 이용하여 polyA 여부를 재확인할 것이다.
- genomic mapping: PolyA가 제거된 이 서열을 UCSC의 Jim Kent, Ph.D.에 의해 개발된 BLAT(Kent 2002)을 사용하여 게놈지도에 mapping 한다. 사용된 게놈지도는 human hg16, mouse mm3, rat rn3이다. 가장 좋은 hit만을 사용하며, UCSC Genome Track중 RepeatMasker track의 정보를 이용하여 repeat을 제거한 다음 genome에mapping된 서열이 100 bp 이상이 되는 것만EST clustering에 포함시킨다.

(2) Primary Clustering
- initial clustering: genome map에서 각 서열의 시작과 끝을 이용하여 범위가 겹치는 서열을 찾아 묶는다 [initial cluster].
- splice site 검색: 이렇게 묶인 initial cluster를 intron을 포함하고 있는 spliced alignment와 하나의 exon으로 이루어진 unspliced alignment로 나눈 다음, spliced alignment가 canonical intron (GT-AG or GC-AG pair)이 아닌 경우에는 SIM4 (Florea et al. 1998)프로그램의 dynamic programming 방법으로 정렬을 수정, 보완하고, 신뢰도가 떨어지는 EST의 경우에는 ECgene에 포함하지 않는다. 또한 이 과정에서 EMBOSS trimest에의해 polyA로 표시된 서열들의 polyA 여부를 genome alignment를 사용하여 재확인한다. 그 결과, trimest 프로그램에서 찾은 polyA 중 약 50% 정도만이 실제 polyA tail임이 밝혀졌다.
- primary clustering: Initial cluster를 중에서 spliced alignment 만을 사용하여 공통된 splice site의 존재 여부에 따라 묶는다[primary cluster]. Unspliced EST는 아래의 transcript assembly가 끝난 후의 마지막 단계에서 추가하며, 그 결과는 NCBI의 새로운 UniGene 알고리즘과 동등하다.

(3) Graph 이론을 이용한 exon 연결의 분석을 통한 transcript assembly
- search for possible exon paths: 각 primary cluster에 대하여 spliced alignment에 포함된 exon의 연결 패턴을 그래프 형태로 표시하면 그림 3과 같은 DAG(directed acyclic graph)가 된다. 이 그래프에서 시작을 나타내는 exon A 또는 B에서 출발하여 끝나는 부분이 H 또는 I exon까지 도달하는 경로를 DFS (depth-first search) 방법으로 구한다. 이렇게 찾은 각 경로는 한 유전자에서 얻어지는 splice variant에 해당하며 이 경우 그림 4

의 시작부분과 같은 결과를 얻는다.

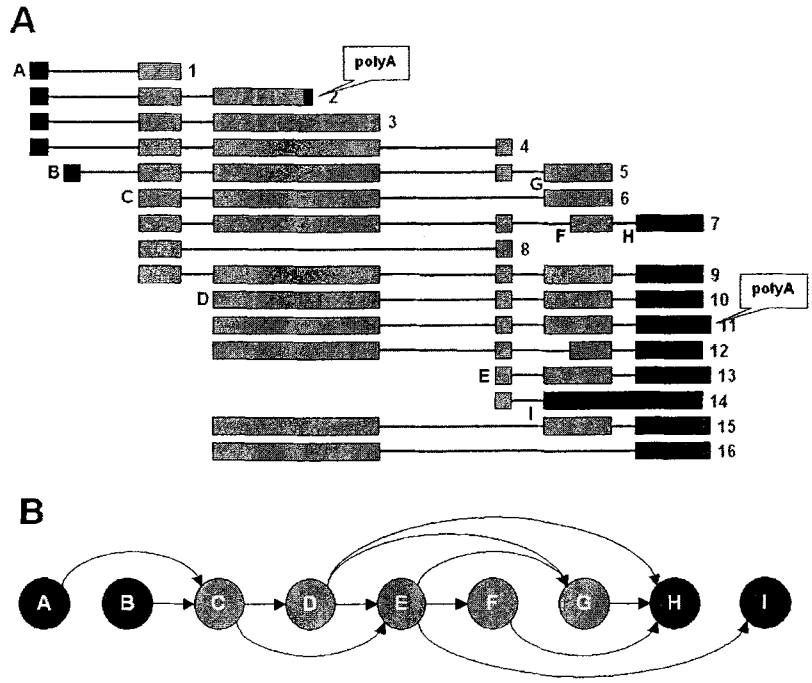


그림 3. Exon의 연결 방식을 그래프로 표현한 모습

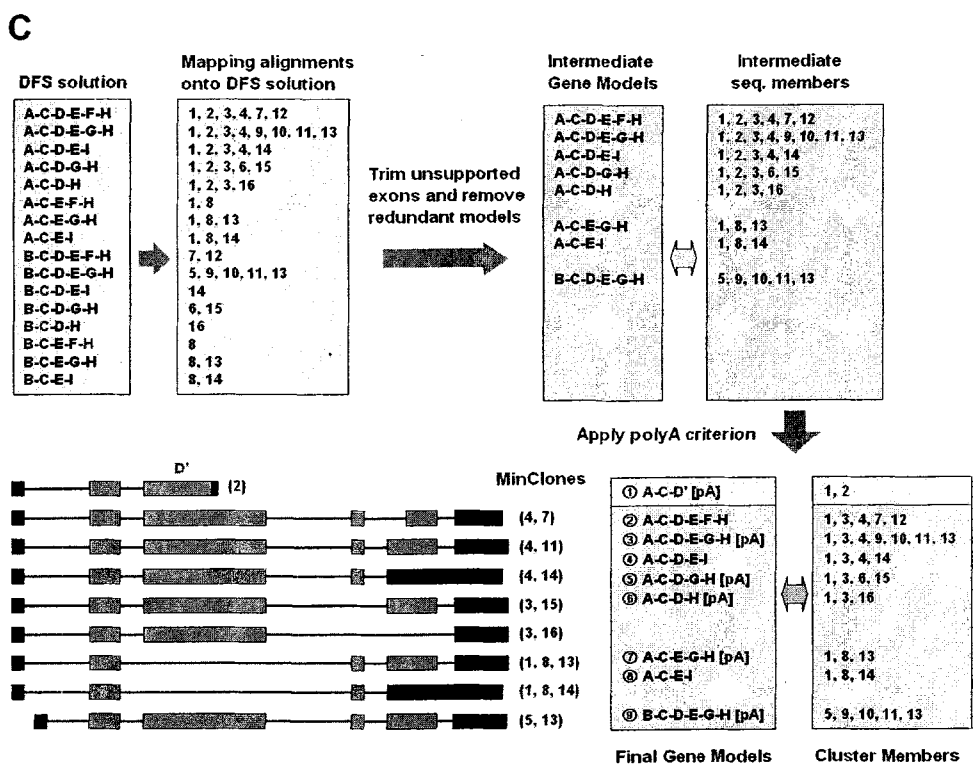


그림 4. DFS 해로부터 각 splicing variant의 구조와 clone 증거를 구하는 알고리즘

- 각 유전자 모델의 확인: 각 alignment를 중복을 허용하면서 exon 경로에 mapping하여 각 gene model과 부합하는 서열을 찾는다. Gene model의 처음부터 끝까지 연결되는 서열 증거가 없는 exon 경로는 유전자 모델에서 제거한다. 이렇게 DFS로 찾아진 gene model도 다음의 몇 가지 재확인 과정을 거친다.
- 유전자 모델의 검증: Splice variant가 아닌 두 개 이상의 mRNA/RefSeq에 걸쳐 있는 EST의 경우에 걸쳐 있는 exon을 제거하거나 혹은 해당 EST를 제거함으로써, 각각의 gene model로 분리한다. 또한 위 과정에서 재확인된 polyA site 정보를 사용하여 alternative polyA가 exon의 중간에 있는 경우에는 해당 gene model을 두 개 이상의 gene model로 분리한다. 그 후 polyA 정보, sense/antisense, EST 5′ & 3′ sequencing 정보를 사용하여 DFS로 찾아진 gene model의 방향을 결정한다.
- Unspliced EST의 추가 및 clone-ID joining: 방향이 결정된 gene model에 위 과정에서 추가되지 않은 unspliced alignment들 중에 genome alignment, polyA 정보, sequencing direction이 gene model과 부합되는 것들만 추가한다. 최종적으로 얻어진 alternative spliced cluster에 존재하는 EST 중에 이웃한 gene과 두 개 이상의Clone ID가 동일할 서열이 존재하는 경우에 1개의 cluster로 묶는다.

나. ECgene 알고리즘의 특징

ECgene 알고리즘은 종래의 transcript 서열끼리 유사성을 분석하여 clustering하는 방법에 비하여 다음과 같은 장점을 갖고 있다.

(1) 유전자 모델과 clustering을 함께 제공: 한 cluster에 포함된 서열은 해당 유전자 모델의 증거 역할을 하고 신뢰도 판단의 척도가 될 뿐만 아니라, 조직 또는 질병 관련성과 같은 유전자 발현에 관한 중요한 정보를 제공한다.

(2) Alternative splicing의 예측: ECgene은 splice site의 공유성을 이용하여 clustering하고 연결 방식을 그래프 이론으로 분석하여 alternative splicing을 예측한다. Alternative splicing은 유전자의 기능, 발현에 큰 영향을 미치는 중요한 요인으로 이를 연구할 수 있는 기반을 제공한다.

(3) polyA tail의 정확한 분석: 많은 경우 mRNA 서열만 보고 판단한 polyA tail의 50% 정도는 게놈지도에 mapping이 되는 가짜 polyA tail이다. ECgene에서는 genomic alignment를 자세히 분석하여 이런 가짜 polyA tail이 존재하지 않는다. PolyA tail의 존재의 그 부분이 실제 전사의 끝부분이라는 강력한 증거가 된다.

(4) Longer UTR: 대부분의 서열 정보와 유전자 예측 프로그램은 단백질 서열 정보가 포함되어 있는 CDS (coding sequence) 부분을 중점적으로 다룬다. ECgene에서는 방향이 일치하는 unspliced EST를 통하여 UTR (untranslated region)을 연장한다. 따라서 그림 5에 나타낸 바와 같이 UTR의 길이가 긴 특성이 있다. 최근 3′ UTR 부분이 mRNA의 안정성

에 중요한 역할을 하며, miRNA와 antisense RNA의 타겟 부분이 되는 것으로 밝혀져 세계적으로 활발한 연구가 진행 중이다. ECgene의 UTR은 이를 위하여 이상적인 모델이다.

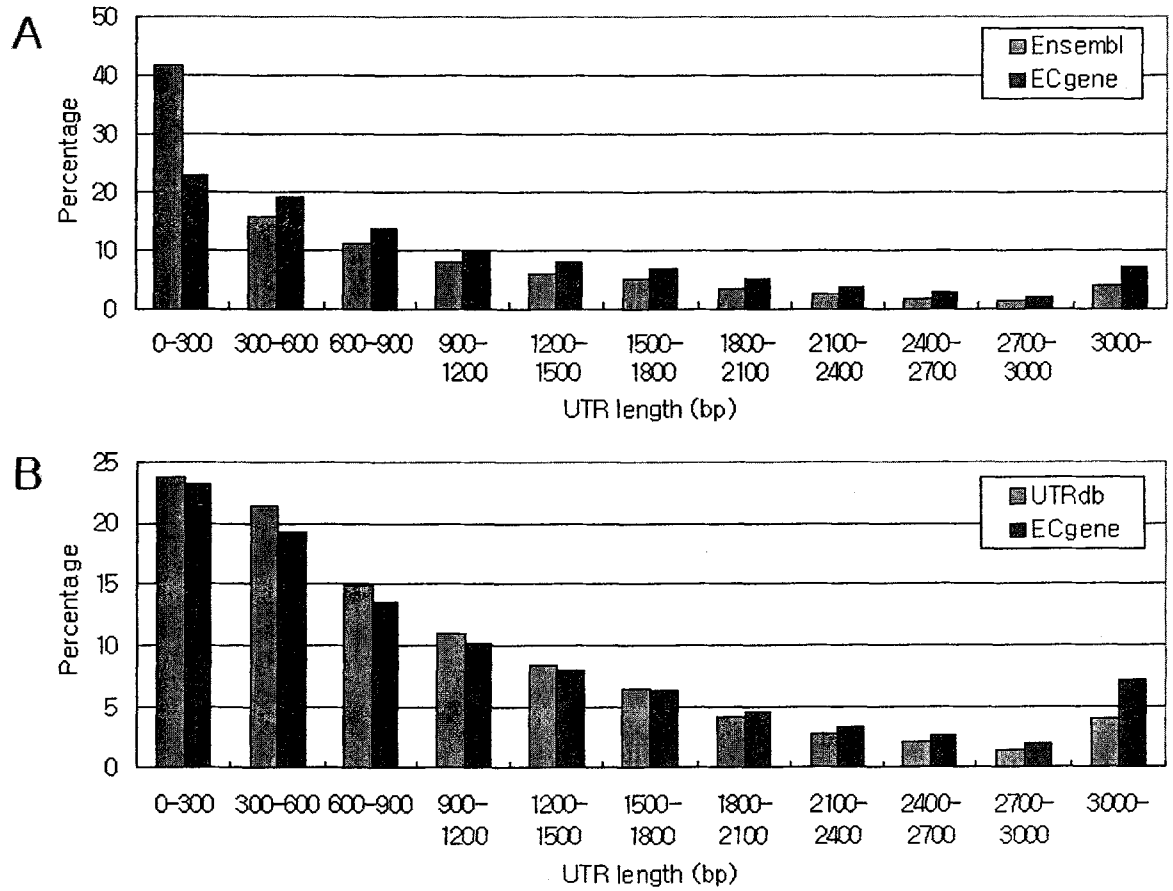


그림 5. ECgene, Ensembl, UTRdb 의 3′ UTR 길이 비교

(5) Neighboring fragments: mRNA와 EST가 유전자의 5′부터 3′ 사이의 모든 exon을 포함하지 않는 경우에는 2개 이상의 cluster로 조각난 결과를 얻게 된다. 그러나 이런 경우에도 ECgene 방법은 이들이 바로 옆에 위치해 있어 전체 모습에 대한 간접적인 정보를 제공한다.

다. ECgene Genome Browser의 개발

유전자 예측 프로그램은 유전자의 구조를 볼 수 있는 수단이 필수적이다. 본 과제에서는 ECgene의 유전자 예측 결과를 UCSC genome browser의 custom track으로 추가하는 방법으로 genome browser를 구현하였다. 이 방법은 UCSC genome browser에 포함된 다양한 종류의 트랙 정보를 쉽게 볼 수 있는 이점이 있다.

ECgene genome browser (http://genome.ewha.ac.kr/ECgene/gbr)의 검색화면은 그림 6의 위부분과 같다. ECgene genome browser에서 genome assembly(hg16, mm3, mm4, rn3)와 version(v1.0, v1.1)을 선택하고 position에 원하는 유전자 이름을 입력 후

show 버튼을 누르면 그림 6와 같은 결과를 볼 수 있다. 제일 위의 파란색으로 나타낸 유전자 모델이 ECgene의 결과이다.

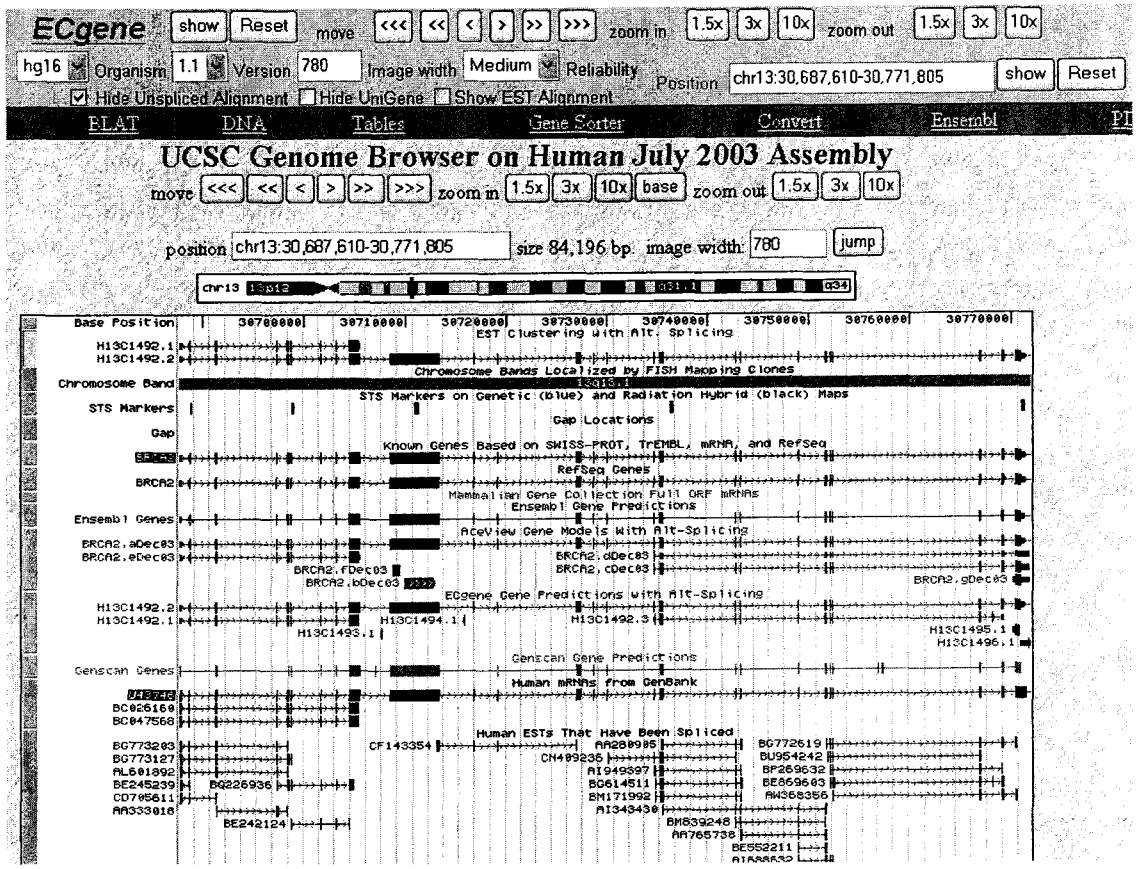


그림 6. ECgene genome browser, http://genome.ewha.ac.kr/ECgene/gbr

ECgene의 유전자 모델은 그 신뢰도를 high, medium, low의 세 종류로 나누었다. 그림 6의 화면에서 reliability를 low로하고 show EST alignment 박스를 체크하면 그림 7와 같이 cluster를 구성하는 mRNA/EST 서열의 정렬을 볼 수 있다. 이는 각 유전자 모델을 검증할 때 유용하다.

그림 7에서 gene model을 펼쳤을 때 나오는 GenBank Accession에는 몇 가지 정보가 포함되어 있다. 맨 앞에 붙은 '#'은 Representative Clone들을 나타낸다. Accession 뒤에 붙은 '[3]', '[5]'는 EST의 read direction을 나타내며, '[m]'은 mRNA, '[R]'은 RefSeq를 나타낸다. 추가로 '[A]'는 해당 sequence가 polyA를 가지고 있다는 의미이다. Track name에 '[10 / 15 / 53]'는 총 해당 유전자에 총 53개의 서열이 묶여 있고, track에 해당하는 gene model에는 15개 서열로 구성되어 있으며 이중에 10개가 spliced alignment를 가지고 있다는 뜻이다. '[Forward]', '[Reverse]'는 유전자의 방향을 나타낸다. 뒷부분에는 reliability와 min. clone의 수를 표시하고 또, Accession에 표시된 것처럼 '[RmA]'로 RefSeq가 존재하는지, mRNA가 존재하는지, polyA를 가지고 있는지를 표시한다. Custom track의 최하단에는 UniGene과 비교를 위한 track이 존재한다. BRCA2에는

UniGene에서 ECgene보다 3개의 서열을 더 가지고 있는 것을 볼 수가 있다.

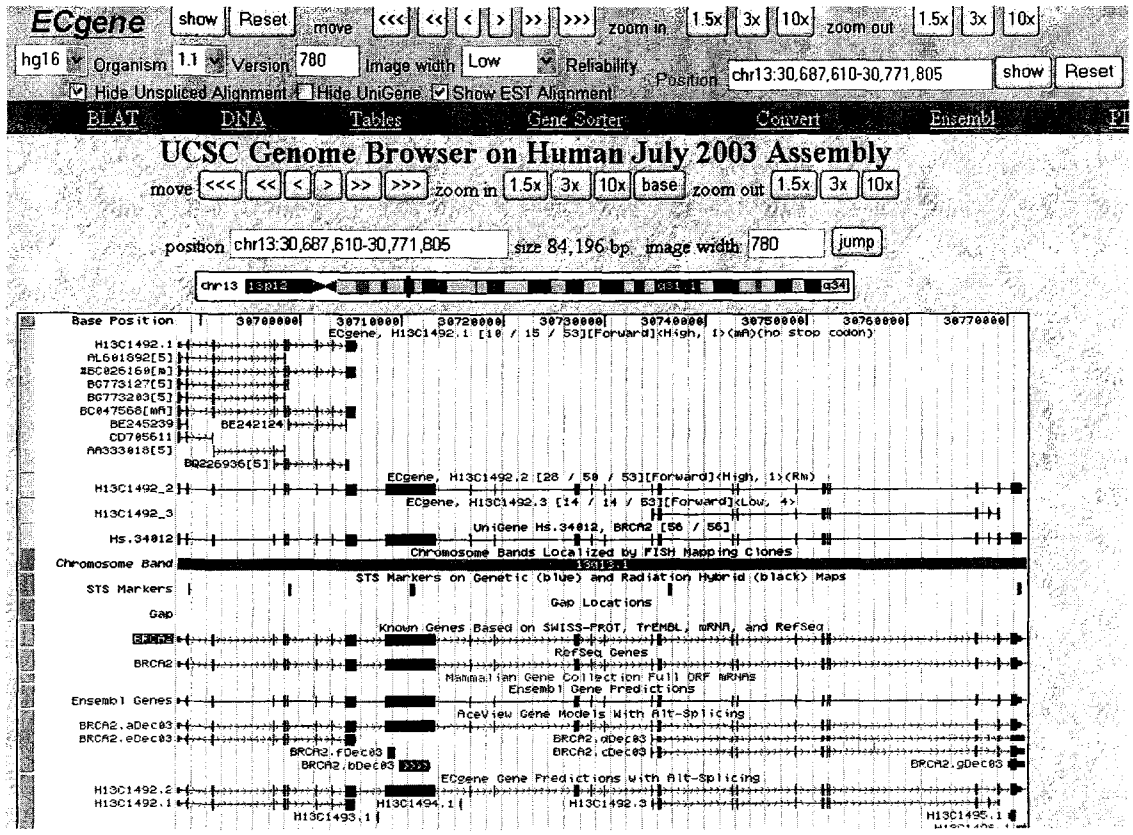


그림 7. Reliability를 low로 하고 Show EST alignment를 체크하였을 때 결과화면

라. 사용자 서열에 대한 유전자 모델링을 위한 ASmodeler의 개발

게놈에 관련된 실험을 하는 연구실에서는 자체 생산된 mRNA/EST 서열을 가지고 있는 경우가 많다. 새로운 유전자 발굴의 시작은 이들 서열과 GenBank의 서열을 이용하여 clustering하는 것으로 시작한다. 본 과제에서는 이 clustering 부분을 ECgene의 방법으로 해결하는 ASmodeler 웹서버를 구축하였다. ASmodeler의 또 다른 장점은 단백질 서열과 gene prediction을 함께 포함할 수 있다는 점이다. 이는 다른 생물에서 나온 단백질을 포함하여 계산함으로써 alternative splicing 패턴이 보존되는지, 또는 다른 연구팀에서 찾은 유전자 구조와 비교하는데 매우 유용하다. ASmodeler는 2004년 8월 Nucleic Acids Research의 Web Server Issue에 게재되었고 현재 human, mouse, rat에 대하여 http://genome.ewha.ac.kr/ASmodeler에서 서비스를 하고 있다. 그림 8은 ASmodeler의 초기화면을 나타내며, 유전자 모델링의 결과는 간단한 표와 ECgene의 genome browser를 이용하여 보여 준다 (화면 생략).

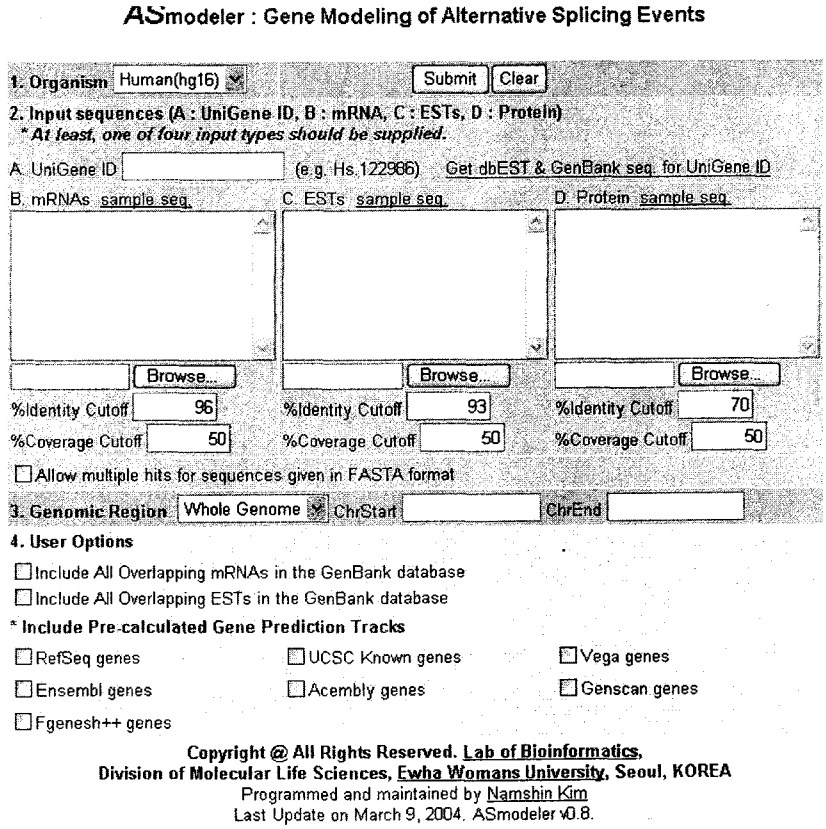


그림 8. ASmodeler의 초기 화면

마. ECgene 알고리즘을 이용한 Human, Mouse, Rat Genome의 분석

이렇게 개발한 ECgene 알고리즘으로 human, mouse, rat genome을 분석하여 alternative splicing을 포함한 유전자 데이터베이스를 구축하였다. 아래의 표 1은 ECgene에서 사용한 입력 데이터이다

Human genome의 경우 아래의 표 2와 같은 결과를 얻었다 (mouse와 rat의 표는 생략). 가장 신뢰도가 높은 part A의 경우 37,497개의 spliced gene과 19,675개의 single-exon gene을 얻었다. Transcript의 수로는 18만개의 transcript 중에서 15.5만개 정도가 multi-exon이었다. Single exon gene의 경우 part A에 포함되려면 mRNA가 있고 cluster의 내의 서열 수가 8개 이상이어야 한다. 이런 조건에서도 많은 transcript들이 있음을 의미하며, 그 중 상당수는 noncoding RNA인 것으로 밝혀졌다. 신뢰도가 낮은 part

Table 1. Summary of Input sequences

	Human			Mouse			Rat		
	RefSeq	mRNA	EST	RefSeq	mRNA	EST	RefSeq	mRNA	EST
Raw data from GenBank	25,975	133,271	5,426,061	40,568	113,526	3,918,650	21,937	11,779	538,134
No. of aligned sequences onto the genome after initial filtering	25,665	118,034	4,836,878	38,137	101,645	3,467,066	20,867	9,996	487,771
No. of sequences after removal of bad alignments[a]	24,895 (96%)	112,933 (84%)	4,408,552 (81%)	37,268 (92%)	100,798 (89%)	3,348,841 (85%)	20,759 (94%)	9,871 (84%)	471,043 (88%)
No. of spliced sequences[b]	22,649 (91%)	86,897 (77%)	2,076,217 (47%)	29,948 (80%)	68,912 (68%)	1,315,511 (39%)	18,289 (88%)	8,404 (85%)	169,604 (36%)

[a]Sequences included in the final clustering of the ECgene. (percentage of aligned sequences)

[b]Input sequences for transcript assembly procedure. (percentage of multi-exon sequences out of all sequences in the ECgene)

C까지 포함하면 31만개의 gene이 나오지만, 이 중 상당수는 하나의 EST만을 포함하는 cluster로 버려야 할 것으로 생각된다. 하지만 49,546개의 spliced gene에서 part C에 속하는 gene이 49,546-43,177=6,369개로 이들은 무시할 수 없다.

Table 2. Summary of ECgene for the human genome

	Part A	Part A+B	Part A+B+C
No. of genes	57,172	82,179	311,252
No. of spliced genes (multi-exon genes)	37,497	43,177	49,546
No. of unspliced genes (single-exon genes)	19,675	39,002	261,706
No. of transcripts	179,810	333,513	658,942
No. of spliced transcripts	154,741	287,934	389,778
No. of protein-coding transcripts	162,645	312,397	558,673
No. of protein-coding transcripts with a polyA tail	82,952	172,888	237,619
No. of non-coding transcripts	17,165	21,116	100,269
No. of non-coding transcripts with a polyA tail	5,082	5,454	5,481
No. of alternatively spliced genes	9,482	14,994	21,266
Percentage of alternatively spliced genes among multi-exon genes	25%	35%	43%
No. of alternative spliced genes with at least one EST-only splice variants	7,524	12,563	18,793
Average number of isoforms for spliced genes[a]	4.1	6.7	7.9

[a]Average number of isoforms per gene for spliced genes = No. of spliced transcripts / No. of spliced genes

각 유전자가 만드는 isoform의 수를 보면 multi-exon gene의 경우 43%의 유전자가 alternative splicing에 의한 variant를 만드는 것으로 보인다. 최근의 실험 데이터에서 얻은 70%-80%를 생각하면 이 숫자는 그리 크지 않은 것으로 보인다. 그리고 spliced gene의 경우 평균 isoform의 수는 4.1-7.9로 비교적 크게 나타났다. 이는 그래프 이론이 여러 곳에서 exon skipping을 보일 때 결과는 각 case의 조합으로 예측하기 때문인 것으로 보인다. 실제로는 이보다 적을 것으로 예상된다.

유전자 중에서 protein coding gene과 noncoding gene의 수를 비교하였다. Part A

의 경우를 보면 총 18만개의 transcript 중에서 16만개 정도가 protein coding transcript 이고 2만개만이 noncoding으로 나타났다. 반면에 모든 경우를 다 포함하면 그 비율은 56만:10만으로 part C에 noncoding RNA가 많이 포함되어 있는 것을 알 수 있다. PolyA tail의 존재는 신뢰도에 따라 달라서 결론을 내기 어렵다.

Alternative splicing을 type별로 통계를 낸 결과는 다음 표와 같다. 전체적으로 human과 mouse는 비슷한 경향을 보인다. Rat의 alternative splicing이 작은 이유는 앞의 표 1과 같이 EST의 수가 human 540만개, mouse 400만개에 비하여 rat은 60만개도 채 되지 않기 때문이 것으로 생각된다. 전체적으로 exon skipping과 splice site variation이 많았으며 intron retention도 예상보다 자주 일어나는 것으로 보인다.

Alternative initiation은 30%의 유전자에서 보이며, alternative termination은 약 70%의 유전자에서 일어나는 현상이다. 이는 전사의 시작과 끝을 조절하는 다양한 메카니즘에 의하여 유전자의 구조가 달라지는 것을 의미하여, 이와 같은 현상은 예상보다 세포 내에서 흔히 일어남을 알 수 있다.

Table 5. Analysis of AS types in the ECgene

Genes	Human	Mouse	Rat
Alternatively spliced genes	21,266	17,706	8,699
with 5' donor splice site variation	10,471 (49%)	7,994 (45%)	2,570 (30%)
with 3' acceptor splice site variation	10,813 (51%)	8,019 (45%)	2,851 (33%)
with exon-skipping event	13,175 (62%)	9,687 (55%)	3,833 (44%)
with intron retention event	2,895 (14%)	1,998 (11%)	212 (2%)
with multiple transcription start sites	6,473 (30%)	5,486 (31%)	1,780 (20%)
with multiple transcription termination sites	15,528 (73%)	12,805 (72%)	4,872 (56%)
with multiple polyadenylation sites	14,835 (70%)	9,123 (52%)	3,524 (41%)

3.2 유전자 기능 예측 프로그램 (ECfunction)

가. 단백질 서열 분석 및 기능 예측 시스템의 개발

ECgene은 EST clustering에 이은 transcript assembly를 통하여 유전자 모델을 제시한다. 또한 대부분의 경우 polyA, intron consensus, sequencing direction 등과 같은 정보로부터 방향이 정해져 있다. 따라서 본 과제에서는 각 유전자의 mRNA 서열을 3-frame translation 시킨 후 가장 긴 ORF를 선택하여 단백질 서열을 구하였다.

각 유전자의 기능 분석은 mRNA 서열 수준과 단백질 서열 수준의 분석을 모두 이용하였다. 각 ECgene의 transcript 중에서 exon의 수와 서열의 길이가 가장 큰 transcript를 선택하여 대표 서열로 정하고 이를 분석한 결과를 유전자의 기능으로 부여하였다. 실제 분석 과정은 다음의 그림 9와 같다.

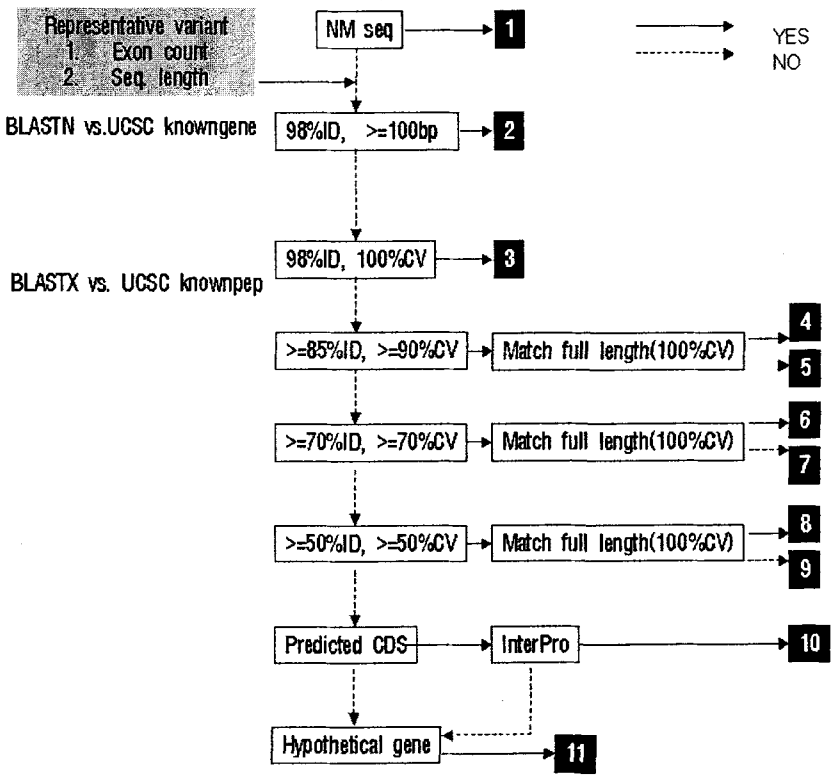


그림 9. 단백질 서열 분석 및 기능 예측 시스템

특정 ECgene cluster가 known gene에 해당하는지의 여부는 RefSeq, UCSC의 Known Genes track의 데이터를 이용하였다. 먼저 RefSeq 서열을 포함하고 있으면 해당 유전자로 할당한다. 그리고 Blastn을 이용하여 mRNA 수준의 유사성으로 단백질을 찾는다. 이렇게 하여 hit이 없는 경우 UCSC의 Known Genes track에 대하여 Blastx로 유사 서열을 찾는다. % identity와 coverage의 기준을 낮추면서 기능을 부여하고, 최종적으로

50% 이하의 % identity 또는 % coverage를 가진 서열은 iprscan(Mulder, et al. 2003) 프로그램으로 특정 모티프의 존재 여부를 확인한다. 이렇게 찾은 서열의 기록 (RefSeq, SwissProt, InterPro)을 참조하여 GO(The Gene Ontology Consortium 2001) 상의 기능을 부여한다.

이 외에도 transmembrane(Krogh, et al. 2001), signal peptide(Krogh, et al. 2001), coiled coil 지역 등의 존재 여부는 단백질의 localization과 기능에 밀접한 연관이 있다. 따라서 이들 성질을 예측하는 공개된 프로그램과 웹 서버를 이용하여 전체 ECgene에서 얻어지는 단백질체에 대한 계산을 수행한 다음 그 결과를 DB화 하였다 .

나. Alternative splicing에 의한 단백질의 domain/motif 변화 분석

모든 ECgene transcript에서 얻어진 단백질에 대하여 Ensembl에서 개발된 InterPro를 이용하여 기능을 결정하는 기본 단위인 도메인과 모티프를 계산하고 그 결과를 DB화 하였다. 특히 2개 이상의 splice variant로 구성된 ECgene의 경우에는 alternative splicing에 의하여 functional domain이 변한 경우가 많으며 이는 다양한 종류의 질병의 직접적인 원인이 된다(Hastings, et al. 2001; Black 2000; Kan et al. 2001; Xu and Lee. 2003; Zhining et al. 2003). 따라서 splice variant에 따라 InterPro 도메인에 생기는 변화를 단백질 motif가 완전히 사라진 형태, 부분적으로 사라진 형태, 모티프의 변화가 전혀 없는 것으로 나누고 모든 유전자 모델에 대한 계산을 수행하였다. 실제 분석 과정은 다음의 그림 10과 같다.

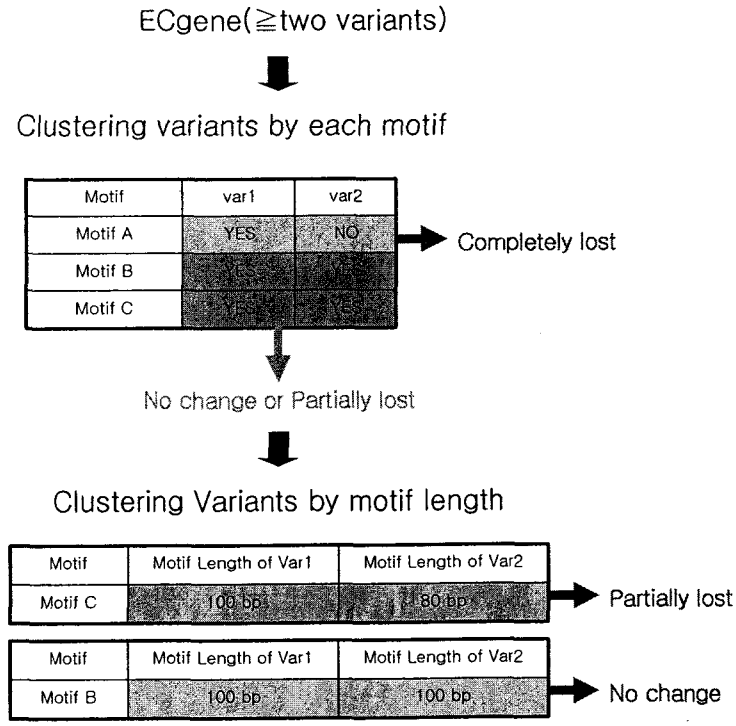


그림 10. Alternative splicing에 의해 변화되는 도메인/모티프 분석 시스템

이와 같은 분석에 의해 얻어진 결과 중 단백질 motif가 완전히 사라진 splice variant들에 대해서 그 원인을 분석하였다. 하나의 ECgene 내에서 variant들 간의 alternative splicing 변화를 비교하기 위하여 모티프의 변화가 그대로 존재하는 variant들 중에서 모티프가 완전히 사라진 variant와 가장 유사한 서열을 가진 isoform을 구하여 비교하였다. 이렇게 구한 similar variant(완전한 모티프를 가진 variant)와 모티프가 완전히 없어진 variant를 비교하였다. 그 결과는 아래 그림 11과 같이 3개의 큰 카테고리로 나뉜다. Category 1은 모티프를 코딩하는 엑손의 skipping이 일어나는 경우, category 2는 모티프를 코딩하는 exon의 alternative splice site 때문에 모티프 코딩하는 일부분이 날아가거나 새로운 엑손이 들어가는 경우이다. category 3은 모티프를 코딩하는 엑손은 그대로 존재하나 frame shift에 의해 CDS부분이 UTR (UnTranslated Region)로 바뀌거나 다른 아미노산 서열이 되는 경우이다.

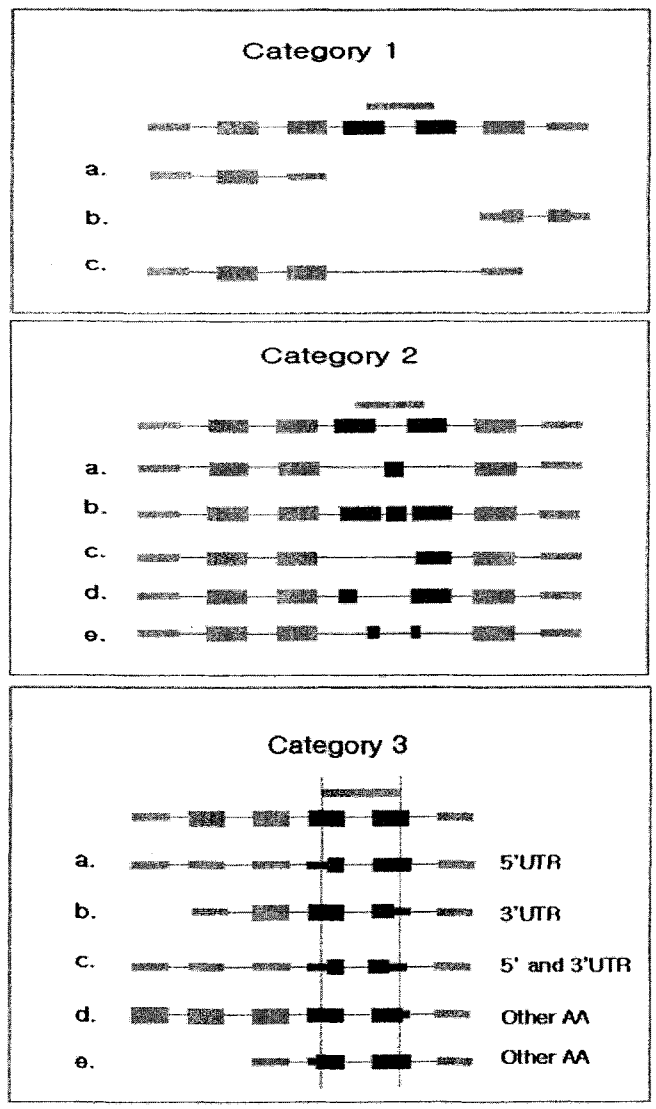
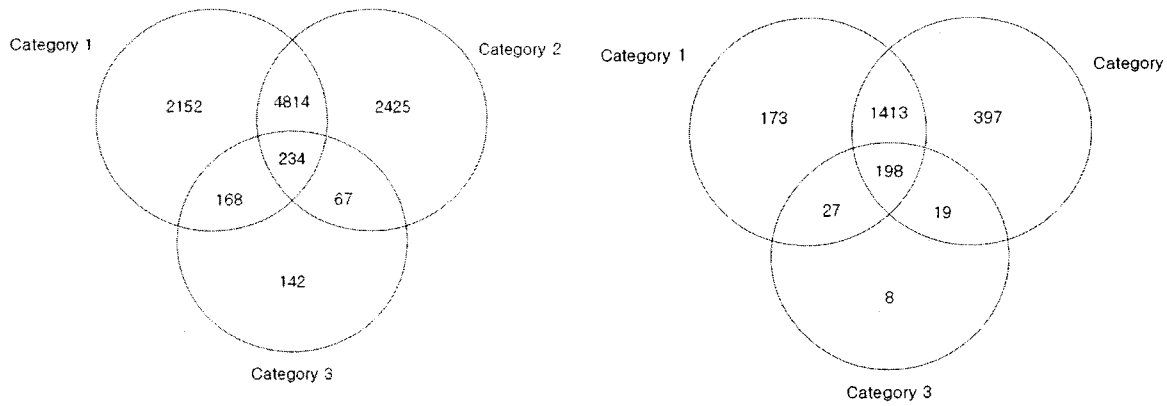


그림 11. 모티프가 완전히 없어진 형태의 분류

ECgene에 대한 계산을 수행한 결과를 그림 12에 나타내었다. Alternative splicing에 의한 단백질 모티프의 상실은 모티프를 코딩하는 엑손의 alternative splice site 선택과 exon skipping이 주 요인이며, 상대적으로 frame shife에 의한 결과는 드물게 나타났다.



ECgene　　Pfam 도메인 변화

그림 12. Alternative splicing에 의한 ECgene내 도메인/모티프 변화의 분석

다. ECfunction browser 개발 (http://genome.ewha.ac.kr/ECgene/ECfunction/)

Alternative splicing에 의한 유전자 구조의 변화는 ECgenome browser를 통하여 게놈 지도상에서 볼 수 있으나, 대부분의 유전자는 긴 인트론을 포함하여 막상 엑손의 미세한 변화는 보기 어려운 측면이 있다. 이를 해결하기 위하여 모든 인트론을 짧게 처리한 좌표에서 유전자 구조를 보여주는 ECfunction browser를 개발하였다. 이는 mRNA 좌표이므로 단백질 서열의 좌표에 그대로 비례하고 따라서 도메인/모티프의 존재 범위를 쉽게 표현할 수 있으며, 따라서 아래 그림 13과 같이 각 transcript 별로 alternative splicing에 의한 각 exon의 차이, motif의 변화, CDS의 범위 등을 한 눈에 보여준다.

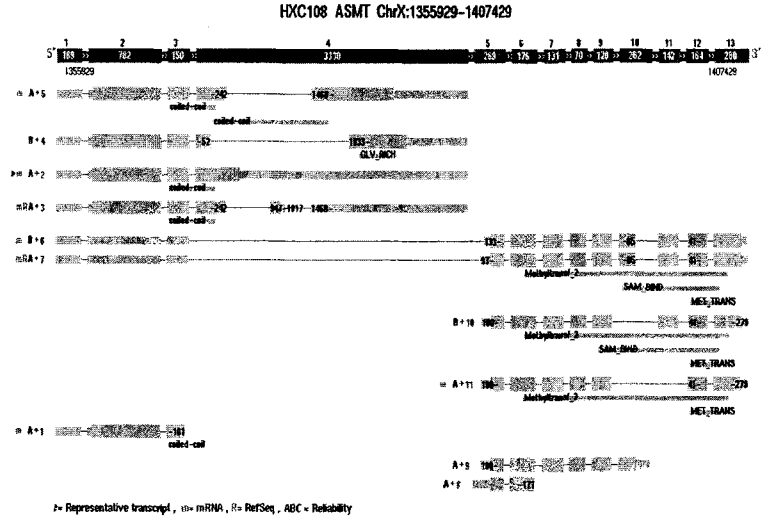


그림 13. ASMT (HXC108) 유전자의 ECfunction 브라우저

라. ASviewer 개발

ECgene과 같이 alternative splicing 정보를 분석한 데이터로서 가장 믿을만한 것이 NCBI의 AceView와 Ensembl gene이 있다. ECfunction 에서는 검색범위가 ECgene에만 국한되었으나 AceView, Ensembl gene, RefSeq 데이터베이스의 alternative splicing 정보로 그 범위를 확장하여 ASviewer (http://genome.ewha.ac.kr/ASviewer) 브라우저를 개발하였다. 이미지를 표현하는 기본 알고리즘은 ECfunction과 동일하다. ASviewer의 장점은 각기 다르게 분석된 alternative splicing variant들을 한 눈에 볼 수 있으며, 또한 RefSeq와 유전자 구조를 비교할 수 있다. motf/domain 정보도 제공되고 있다. 아래의 그림 14는 SRC 유전자에 대하여 ECgene, AceView, Ensembl, RefSeq 데이터베이스에 존재하는 모든 splicing variant의 gene structure를 보여준다.

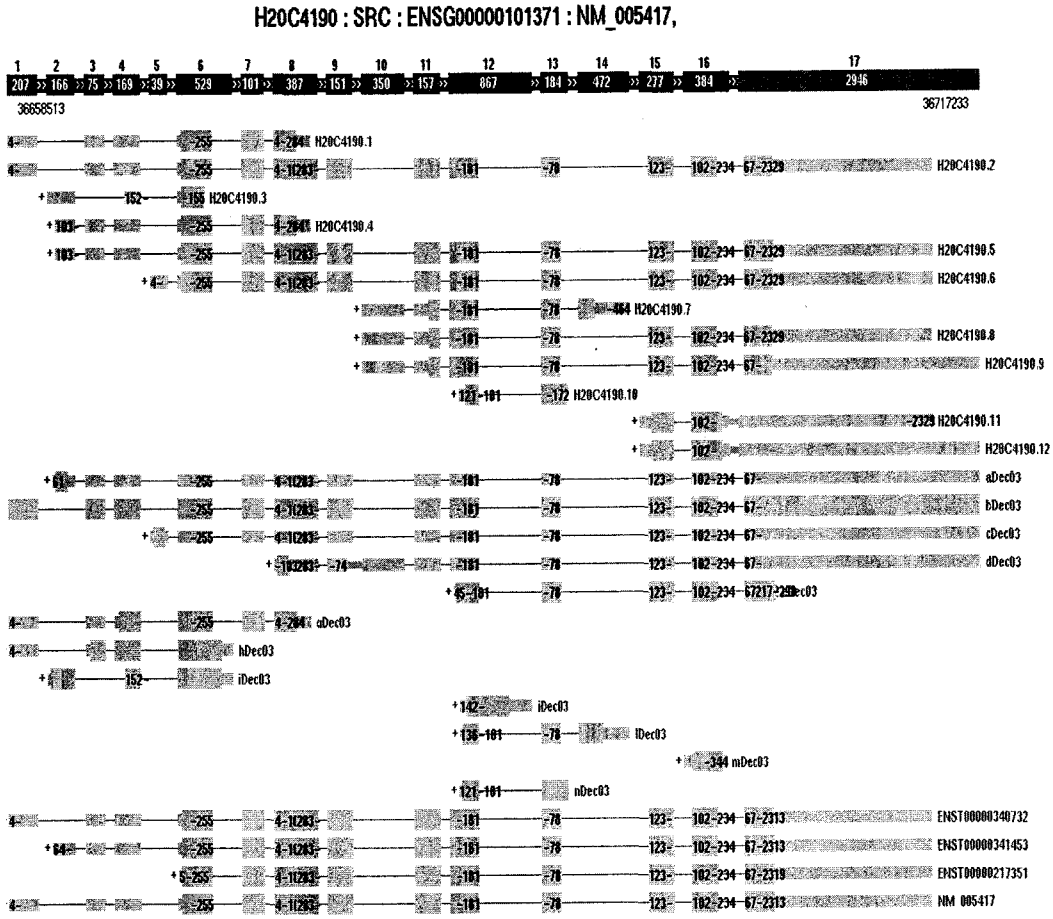


그림 14. SRC 유전자의 ASviewer

3.3 유전자 발현 예측 프로그램 (ECexpression)

가. 유전자 발현 예측을 위한 cDNA 및 SAGE library의 분류

EST cluster 내의 각 서열이 얻어진 조직·병리학적인 정보를 분석하면 해당 유전자의 발현 패턴을 유추할 수 있다. UniGene을 이용하여 많은 연구가 이루어져 왔으나 대부분의 경우 관심 있는 수십 종의 조직 또는 질병에 대하여 분류하는 것으로 그쳤다. EnsMart에서 최근 구현한 EST 발현 profiling 방법은 SANBI(Winston 2003)의 Win Hyde 그룹에서 개발한 계층적 분류를 사용하며, 이는 cDNA libarary를 한 번 분류함으로써 다양한 경우에 적용될 수 있는 이점이 있다. 본 과제에서는 이를 더욱 발전시켜 organ-tissue-cell type, pathology, developmental stage, sex의 네 가지 부문에서 분류체계를 만들고 dbEST 전체 library를 분석하였다. 그림 15는 organ-tissue-cell type의 부문에서 liver가 속한 부분을 보여주고 있다. Organ-tissue-cell type에는 총 781개의 노드가 7 단계로 구성되어 있다.

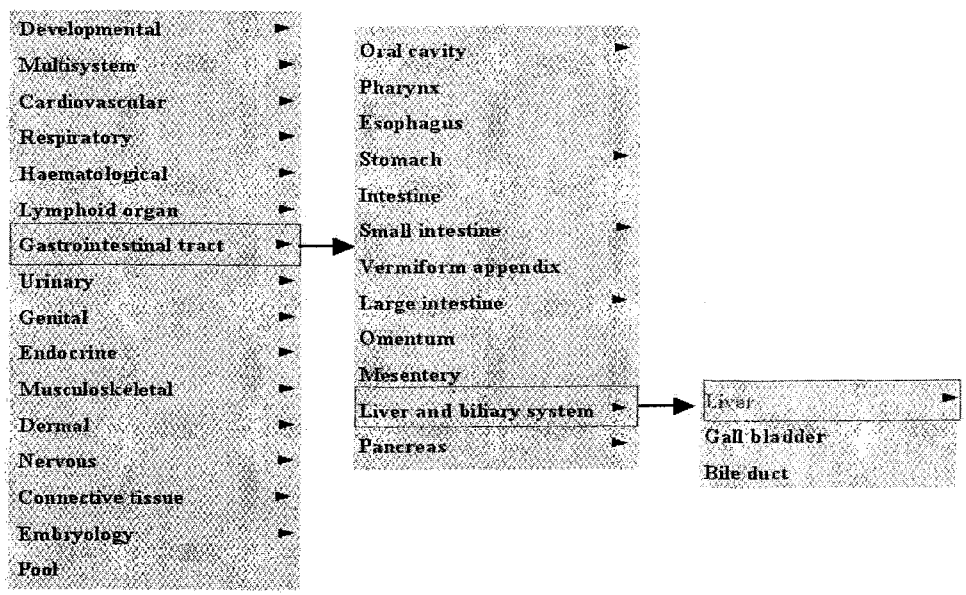


그림 15. Organ-tissue-cell type의 분류체계의 일부

Pathology 분류체계는 121개의 노드가 4 단계로 구성되어 있으며 CGAP (cancer genome anatomy project)의 분류를 참고하여 암의 분류에 강점이 있다. Mouse의 경우에는 발생단계의 데이터가 많은 것을 고려하여 MGI (mouse genome informatics)와 Edinburgh의 mouse atlas의 분류체계를 이용하였다.

이렇게 완성한 계층적 분류체계를 이용하여 NCBI, NCI, SAGE 등의 인터넷 사이트에 공개되어 있는 약 8,600개의 human cDNA library, 350개의 SAGE library, 약 900개의 mouse cDNA library, 100개의 SAGE library를 분류하였다. 모든 분류는 수동으로 이루어져 최대한 구체적인 노드를 할당하였다.

나. 유전자의 발현 정보를 보여주는 웹사이트 구축

EST와 SAGE library를 이용하여 해당 유전자의 발현 정보를 보여주는 웹사이트 ECexpression (http://genome.ewha.ac.kr/ECgene/ECexpression)을 구축하였다. ECexpression은 ECgene의 clustering과 assembly에 기반을 두었기 때문에 기본적으로 splicing variant에 대한 발현 정보를 제공한다. 또한 ECgene clustering의 장점과 transcript-based SAGE tag 방법을 이용하여 현재 공개된 다른 프로그램에 비하여 우수한 결과를 제공한다.

그림 16은 ECexpression의 초기화면이다. 유전자 모델, 예측 방법, 출력 그림의 종류를 선택하고 질의를 보내면 해당 유전자의 전체적인 발현 (EST) 및 isoform 별 발현 정보 (EST & SAGE)를 보여준다(Lash 2000;Riggins 2002). EST의 경우 정량적인 정보를 얻기 위하여 normalized library를 제외시킬 수 있으며, SAGE의 경우에는 tag에 관한 다양한 선택사항을 지원하고 있다.

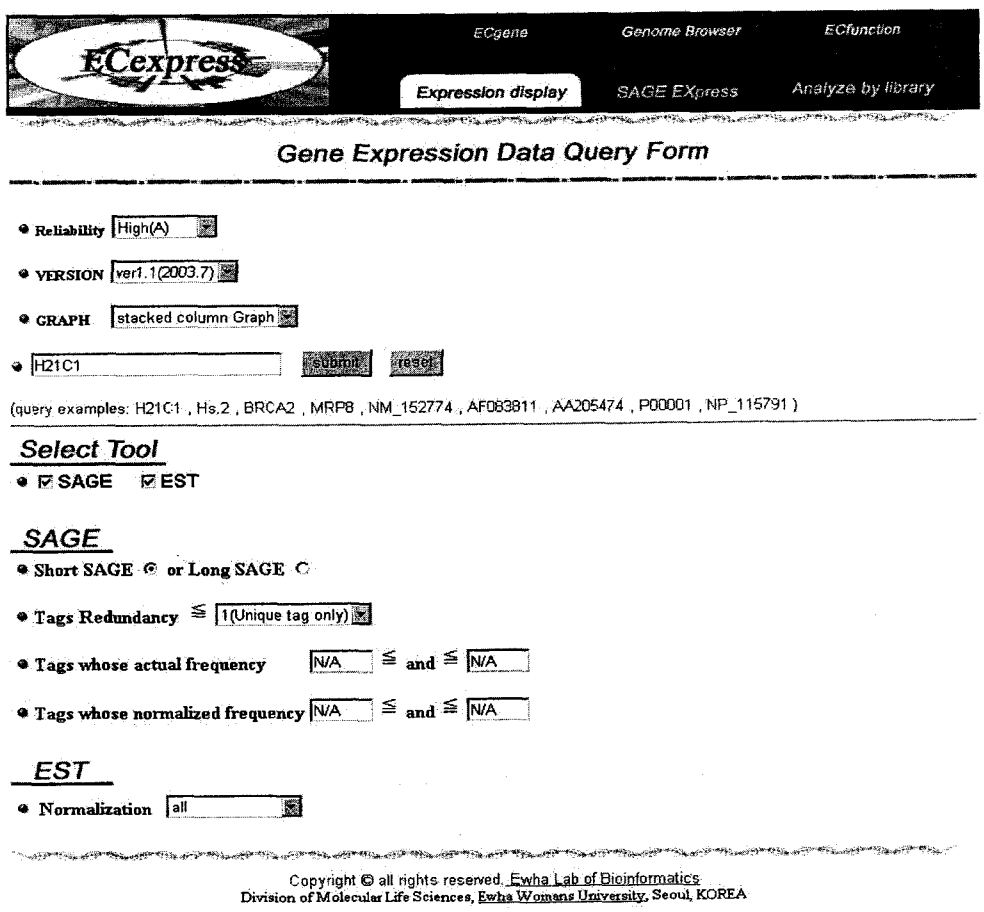
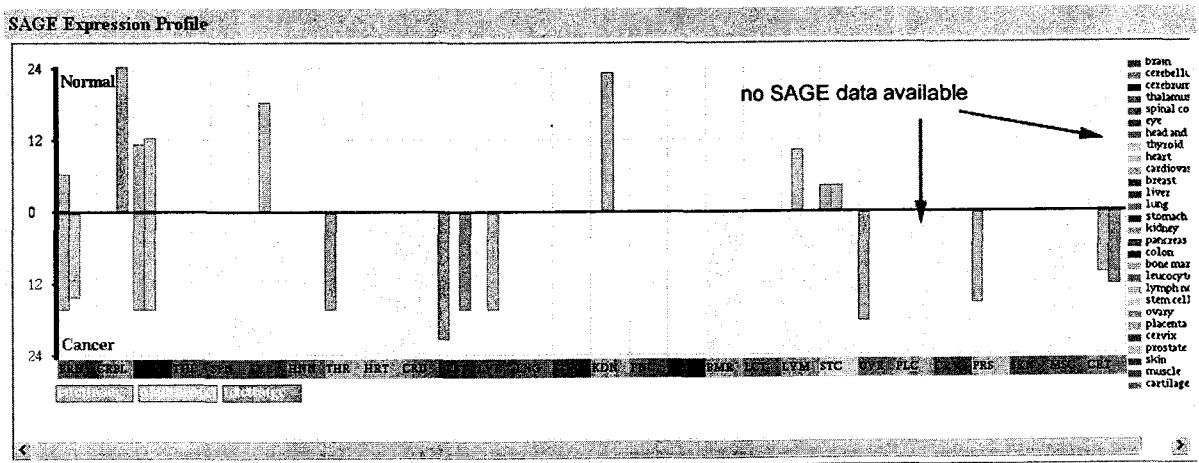


그림 16. ECexpression의 초기 화면

그림 17은 출력 화면 중의 SAGE 부분을 보여주고 있다. RGS17 유전자는 3개의 isoform을 갖고 있으며 각 mRNA 서열에서 3′ 쪽의 SAGE tag을 뽑으면 그림과 같이 세 개의 서로 다른 SAGE tag가 얻어진다. 이들 tag의 빈도를 공개된 SAGE library에서 찾아 그림으로 나타내었다. 막대그래프는 현재 SAGE library 데이터가 있는 28개의 조직에서 정규화된 빈도를 나타낸다. 또한 막대의 위쪽 부분은 정상 조직, 아래쪽 부분은 암조직에서 얻은 library를 의미하여, 한 눈에 각 isoform의 암 관련성 발현을 볼 수 있다. 또한 tag의 중복성을 genome-scale로 조사하여, 같은 tag를 가진 다른 transcript가 있으면 표에 나타내었으며, 각 tag에 대한 더욱 구체적인 정보를 보려면 'Tag Info' 링크를 누르면 된다. 그리고 널리 사용되는 NCBI의 SAGEmap과의 비교·편의를 위하여 링크도 제공하고 있다.



Tag Assignment	ECgene Variant	Links
	1	Tag Info SAGEmap
	3	Tag Info SAGEmap
	2	Tag Info SAGEmap

3 variants in RGS17 gene

Tag Redundancy	ECgene A	ECgene B	ECgene C
	Unique	Unique	Unique
	Unique	Unique	
	H15C9362 ← other genes →	H15C9362	

그림 17. ECexpression에서 SAGE 부분의 출력 화면

그림 18은 특정 tag에 관련된 구체적인 정보를 보여주는 화면이다. 해당 tag를 가지고 있는 ECgene과 UniGene cluster, 그리고 해당 tag가 관찰된 SAGE library 정보를 표로 정리하였다. 정규화된 빈도 (TPM; tag per million counts), 실제 tag의 수, library의 조직, 질병 관련 정보를 보여준다.

SAGEmap, SAGE Genie 등의 널리 알려진 프로그램과 가장 큰 차이점은 ECexpression의 tag 부여 방식이다. 다른 프로그램은 UniGene을 근거로 assembly를 사용하지 않지만, ECexpression에서는 ECgene에서 예측한 mRNA 서열을 근거로 tag를 부여한다. ECgene의 polyA tail 검색, 3′ UTR 부분의 연장 등의 장점을 효과적으로 이용하기 때문에 가장 신뢰도가 높은 tag 부여 방법이라고 할 수 있다. 예를 들면 그림 19

는 TPTE 유전자에 해당하는 cluster를 검색한 결과로, ECgene (H21C1)은 네 개의 isoform이 모두 동일한 tag를 갖지만 UniGene (Hs.122986)을 이용하면 잘못된 tag가 많이 얻어진다.

AAAAAAAAAG

ECgene Result

ECgene	Variants(conf.level)	Name	Symbol	#Total seq.	#RefSeq	#mRNA	#EST	#Total variant	Links
H9C9443	1			1	0	1	0	1	ECgene SAGEmap

UniGene Result

UniGene	Gene description
15159	chemokine-like factor

Lib.Name	TPM	Tag counts	Total tags	Dot Plot	Tissue type	Pathology	Dev.Stage	Sex
SAGE_Prostate_normal_epithelium_CS_confluent	347	25	71897		prostate	normal	child	male
SAGE_Fibroblasts_CL_precrisis	345	3	8681		skin	normal		
SAGE_Prostate_normal_epithelium_CS_senescent	264	19	71717		prostate	normal	child	male
SAGE_Brain_medulloblastoma_B_C609	254	19	74612		brain	neoplasia	toddler	male
SAGE_Breast_carcinoma_B_DCIS-5	208	9	43098		breast	neoplasia	adult	female
SAGE_Breast_adenocarcinoma_CL_SKBR3	184	1	5426		breast	neoplasia	adult	female
SAGE_Fibroblasts_CL_postcrisis	183	4	21833		skin	normal		
SAGE_Brain_astrocytoma_B_H1126	174	3	17178		cerebrum	neoplasia	adult	female
SAGE_Bone_marrow_normal_B_D01	164	6	36577		bone marrow	normal		
SAGE_Prostate_carcinoma_B_pool2	151	10	66034		prostate	neoplasia		male
SAGE_Prostate_carcinoma_CL_LNCaP-C	149	6	40029		prostate	neoplasia	adult	
SAGE_Breast_carcinoma_B_DCIS-4	148	9	60605		breast	neoplasia		female
SAGE_Brain_medulloblastoma_B_98-04-P494	139	6	43068		cerebellum	neoplasia	toddler	male
SAGE_Prostate_carcinoma_CL_PC3_Mock	128	5	38819		prostate	neoplasia		male
SAGE_Liver_cholangiocarcinoma_B_K2D	128	6	46853		liver	neoplasia	adult	male
SAGE_Stomach_carcinoma_B_G189	126	8	63075		stomach	neoplasia		
SAGE_Brain_medulloblastoma_CL_mhh-1	125	6	47858		cerebellum	neoplasia		
SAGE_Retina_Peripheral_normal_B_2	123	13	105312		eye	normal	adult	female
SAGE_Brain_glioblastoma_B_GBM1062	117	7	59762		cerebrum	neoplasia	fetus_16 weeks	male
SAGE_Stomach_carcinoma_B_X43	116	6	51620		stomach	neoplasia	adult	female
SAGE_Ovary_carcinoma_B_OC14	115	2	17298		ovary	neoplasia		female
SAGE_Brain_medulloblastoma_CL_H341	113	5	43920		brain	neoplasia		
SAGE_Brain_medulloblastoma_B_96-04-P019	113	6	52645		cerebellum	neoplasia	toddler	male
SAGE_Brain_medulloblastoma_B_H1413	113	7	61853		brain	neoplasia	toddler	male

그림 18. SAGE tag 'AAAAAAAAAG'에 관련된 상세 정보

H21C1

Mathched UniGene

UniGene Id	desc.
Hs.122986	transmembrane phosphatase with tensin homology

ECgene Result

ecgId	Tag	Frequency
H21C1	TTCATATATC	4/4

UniGene Result

UniGene Id	Tag	Frequency
Hs.122986	TTCATATATC	27/64
Hs.122986	GGGGTTGGGG	6/64
Hs.122986	TTCTTCAATA	5/64
Hs.122986	ATAGTAGAAT	1/64
Hs.122986	ATTCTAACGA	1/64
Hs.122986	ATTGATGATC	4/64
Hs.122986	CCCAGGAAGC	7/64
Hs.122986	GACTTTATAC	1/64
Hs.122986	GATGTTCTTC	8/64
Hs.122986	GCACTGTCCT	1/64
Hs.122986	GGTTTGAAAA	1/64
Hs.122986	TATGTAAATC	1/64
Hs.122986	TTCTTCTTCG	1/64

그림 19. Gene-to-Tag Assignment for the TPTE gene

ECexpression의 EST 발현 정보도 비슷한 형태로 표현된다. 계층적 분류체계를 모두 나타낼 수는 없으므로 대표적인 57개의 조직과 14개의 질병을 선택하여 그림으로 표현하였으며 그 결과는 그림 19와 같다. EST는 normalized library를 포함시키면 정량적인 발현 정보를 얻을 수 없는 단점이 있으나, human에 대하여 공개된 library의 수가 350여개인 SAGE에 비하여 약 8,600개로 압도적으로 많기 때문에 다양한 샘플의 데이터를 포함하고 있다. 사용자가 원하는 경우에는 normalized library를 제외시킬 수 있도록 하였다. 그림 20은 각 ECgene cluster의 전체적인 발현 패턴을 보여주고 있으며, 그 아래 부분에 그림 21에 나타낸 바와 같이 각 isoform 별 EST 서열의 발현 정보를 같은 색으로 표현한 그림이 있다.

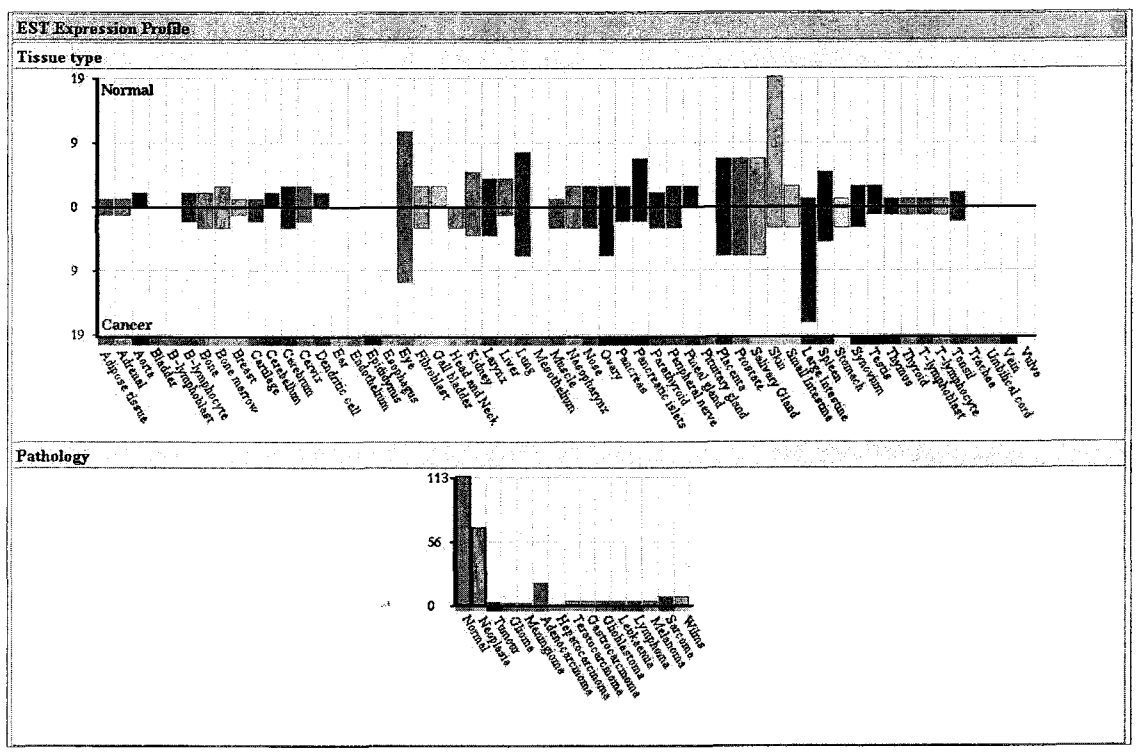


그림 20. Tissue와 질병에 따른 EST 발현 정보

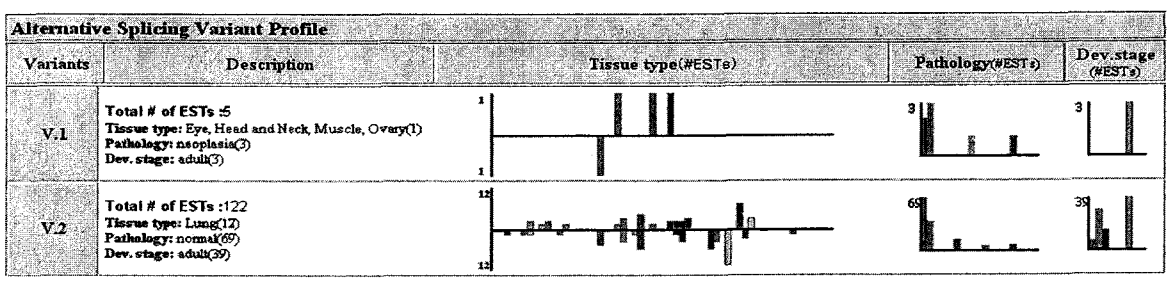


그림 21. EST를 통하여 분석한 각 variant 별 발현 정보

3.4 유용 유전자 검색 프로그램 (ECprofiler)

사용자가 원하는 기능과 발현 양상을 가지고 있는 유전자를 검색하는 것은 유용 유전자 발굴의 시발점으로 질병이 진단과 치료, 신약개발 등에 응용될 수 있는 중요한 기능이다. 특히 현대 유전학의 발달로 positional cloning, association study 등의 방법으로 후보 유전자의 염색체상 위치를 좁혀 놓은 경우가 많기 때문에, 주어진 범위 내에서 원하는 조건을 만족시키는 유전자 검색 프로그램은 활용 가치가 매우 높다.

가. 특정 조직·질병에서 차등발현을 보이는 유전자 검색 프로그램 개발

본 과제에서는 유전자의 발현 분류와 GO 상의 분류를 tree 형태로 표현하고 사용자의 선택에 따라 유전자를 검색하는 ECprofiler를 Java Web Start로 구현하였다. 그림 22는 ECprofiler의 입력 선택 화면이다.

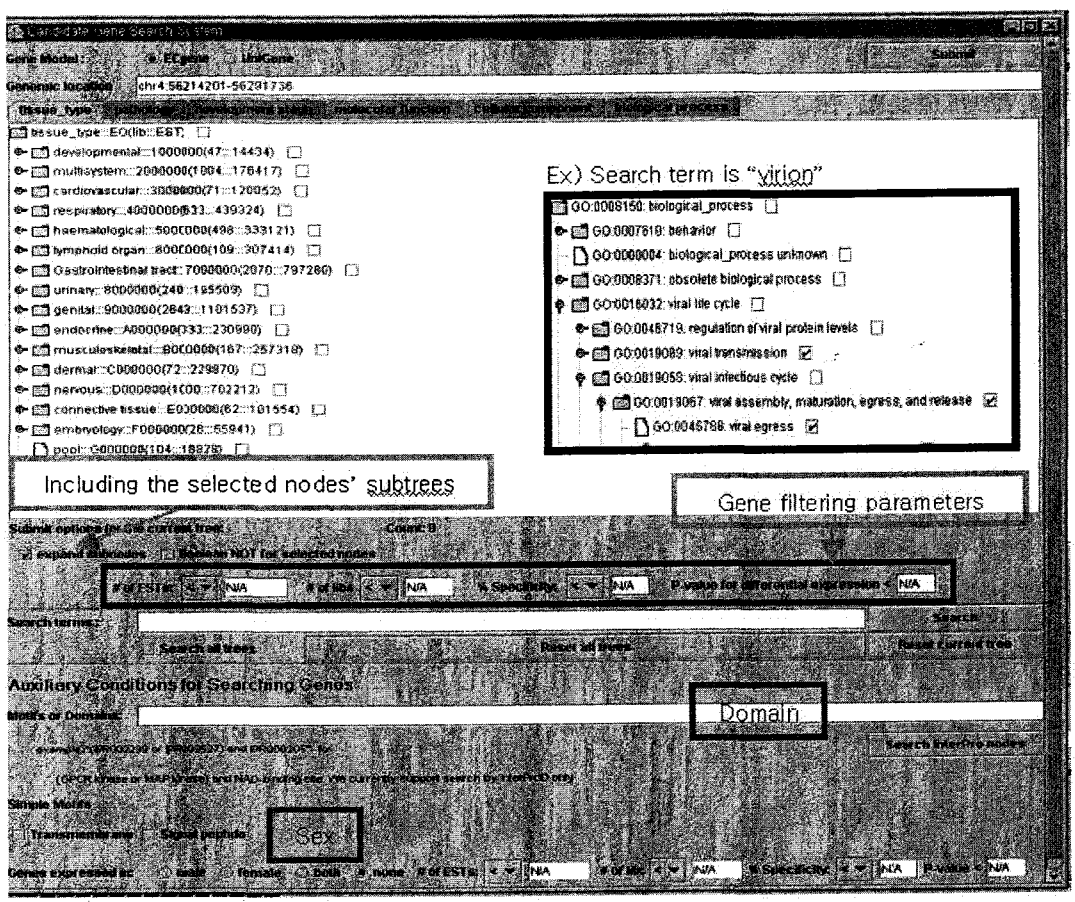
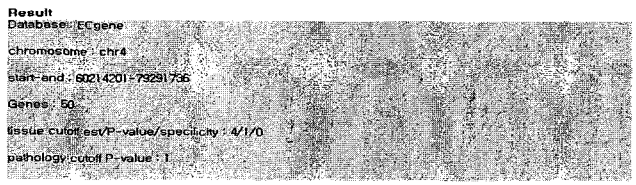


그림 22. ECprofiler의 화면.

Tree 상에서 원하는 노드를 복수로 선택할 수 있으며, tissue-organ-cell type, pathology, developmental stage의 세 부문에서 발현 조건을 선택할 수 있다(Hide 2003). 또한 단어

검색 및 검색 조건의 조합을 지원하여 검색의 편의성을 도모하였다. GO의 tree는 전체 노드 수가 너무 많아 원하는 단어를 포함한 노드를 추출하여 tree 형으로 보여주는 abstracted tree 형으로 구현하였고, 발현에 관련된 EO (expression ontology)는 노드 수가 그리 많지 않기 때문에 모든 노드를 보여준다. 이 외에도 특정 Pfam 도메인을 포함하는 유전자의 검색과 같은 필터 기능도 지원하고 있다. Tree 형태 검색 방법의 가장 큰 장점은 하위 노드를 자동으로 포함하는 검색하는 기능이다. ECprofiler는 자동으로 하위 노드에 할당된 cDNA library도 검색하도록 설계되었다.

이렇게 명시된 검색 조건은 ECgene 서버에 보내져서 발현의 통계 계산, ECgene DB 질의를 통하여 그림 23과 같이 검색 조건에 맞는 ECgene 유전자 목록을 제시한다. 통계계산은 IDEG6 통계 계산 package를 이용하였으며, 결과화면은 조건에 따라서 다른 방식으로 유전자 검색 결과를 볼 수 있도록 하였다. 이러한 유전자 발굴 시스템으로 원하는 조직에 특이 발현되는 유전자나 질병 특이 발현되는 유전자들의 발굴이 가능하게 된다.



ECgene ID	chr	tissue ESTs/Other ESTs	tissue libs	pathology ESTs/other ESTs	pathology libs	nVar	nEST	nmRNA	nRefseq	Symbol	Gene name
H4C6659	4	9/122	6	3/128	2	4	131	3	1	CXCL9	chemokine (C-X-C motif) ligand 9
H4C6162	4	57/469	20	25/501	14	78	526	10	2		similar to hypothetical protein
H4C6552	4	16/152	10	14/154	7	25	168	8	1	ZNF363	zinc finger protein 363
H4C6302	4	73/7842	3	10/7905	1	467	7915	22	1	ALB	albumin
H4C6556	4	11/79	6	3/87	1	9	90	3	1		hypothetical protein MGC30052
H4C6629	4	18/116	8	7/127	3	26	134	4	2	ASAHL	N-acylsphingosine amidohydrolase (acid ceramidase)-like
H4C6938	4	19/320	13	109/230	5	9	339	7	1	MRPL1	mitochondrial ribosomal protein L1
H4C6640	4	8/84	6	3/89	3	4	92	5	1		hypothetical protein FLJ10498
H4C6317	4	6/4	2	1/9	1	1	10	0	0		
H4C6392	4	12/551	4	5/558	1	26	563	6	1	AFP	alpha-fetoprotein
H4C6182	4	7/114	4	8/113	4	3	121	5	1		Mob4A protein
H4C6717	4	15/145	9	5/155	4	12	160	5	1		genethonin 1
H4C6908	4	7/105	4	2/110	2	1	112	0	0		
H4C6456	4	71/56	8	4/123	2	4	127	4	1	CXCL5	chemokine (C-X-C motif) ligand 5
H4C6677	4	16/162	8	10/168	5	50	178	6	1	NUP54	nucleoporin 54kDa
H4C5598	4	7/98	4	3/102	2	10	105	3	1	LPHN3	latrophilin 3
H4C6252	4	7/357	2	6/358	1	30	364	3	1	GC	group-specific component (vitamin D binding protein)
H4C5814	4	5/80	5	2/83	2	8	85	2	1	CENPC1	centromere protein C 1
H4C6366	4	5/23	5	3/25	3	1	28	0	0		
H4C6573	4	37/429	20	25/441	10	59	486	6	1		Ras-GTPase activating protein SH3 domain-binding protein 2
H4C5943	4	19/146	6	4/161	2	1	165	0	1		

그림 23. ECprofiler의 출력화면

나. 특정 조직·질병에서 차등발현을 보이는 유전자 DB 개발

전 항에서 개발된 profiler와 통계 방법은 SAGE library의 분석에도 사용될 수 있다. cDNA library에 있는 57개의 대표적인 조직에 대하여, 해당 조직에서 차등 발현을 보이는 유전자 목록을 Fisher exact test 나 Audic-Claverie test와 같은 통계적인 방법으로 구한다(Romualdi 2003). SAGE 또한 SAGE library에 있는 28개의 대표적인 조직에 대한 계산을 수행할 수 있다. 기존의 차등 발현 연구와 다른 점은 ECgene의 alternative splicing 분석으로 인하여 variant-specific한 발현을 볼 수 있다는 점이다. 아래 표(Su

2004)는 현재 개발된 프로그램을 이용하여 얻은 tissue-specific gene 목록의 일부로, 추가적인 검증과정이 끝나면 데이터베이스를 구축하고 웹을 통하여 공개할 것이다.

Pancreas	PLRP2	PNLIPRP2	H10C13112
pancreas	PLRP1	PNLIPRP1	H10C13104
pancreas	CTRC	NM_007272	H1C2413
pancreas	CPA2	NM_001869	H7C14318
pancreas	FLJ10512	NM_018121	H10C11368
pancreas	ABCD1	NM_000033	HXC8080
liver	SLC22A1	NM_003057	H6C16240
liver	F9	NM_000133	HXC7496
liver	HFL3	NM_005666	H1C22331
liver	SAA4	NM_006512	H11C2697
liver	MASP2	NM_006610	H1C1810
liver	C8A	NM_000562	H1C9323
liver	**SERPINC1**	NM_000488	H1C20410
liver	CRP	NM_000567	H1C18536
liver	ARG1	NM_000045	H6C12869
liver	APOB	NM_000384	H2C1815
liver	SAA4	NM_006512	H11C2697
liver	APCS	NM_001639	H1C18521
placenta	PSG9	NM_002784	H19C7528
placenta	ADAM12	NM_003474	H10C14413
placenta	CYP19	NM_000103	H15C3934
kidney	AQP2	NM_000486	H12C5680
heart	TNNI3	NM_000363	H19C10213
heart	TNNT2	NM_000364	H1C22767

목록에 포함된 유전자의 차등 발현은 ECexpression에서 그래프로 확인할 수 있다. 그 예로 liver에 특이적으로 발현되는 H1C20410 (SERPINC1)의 발현 양상을 SAGE와 EST를 통하여 분석하면 그림 24, 25와 같다.

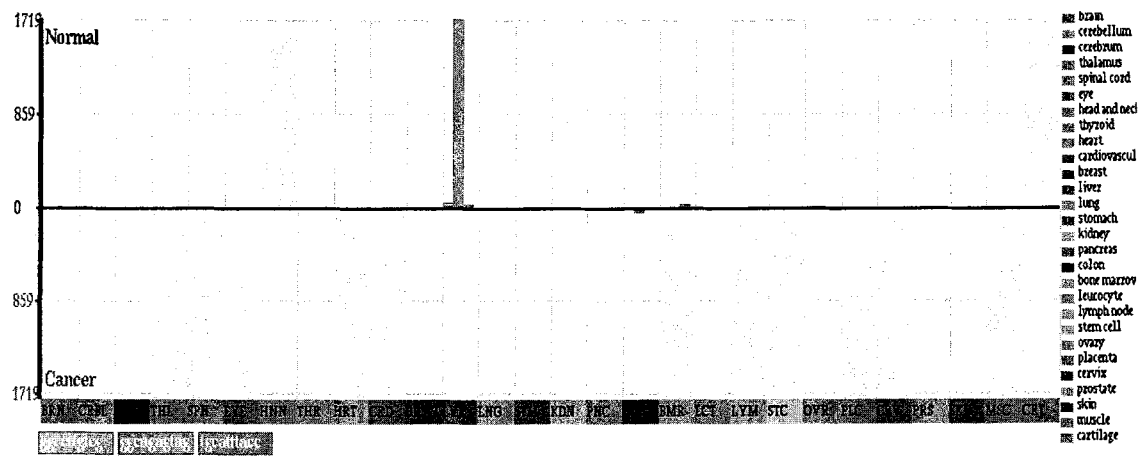


그림 24. 간에서 차등발현을 보이는 SERPINC1 유전자 (SAGE)

H1C20410는 세 종류의 SAGE tag을 가지며 그 중 파란색의 tag가 정상 간 조직에서 압

도적으로 많이 발현되는 것을 볼 수 있다. 다른 tag는 유전자 모델의 신뢰도와 Tag Info를 구체적 살펴보는 등의 자세한 고찰이 필요하다. 그림 25의 EST를 통하여 본 조직 발현별 패턴도 이와 일치하며, 추가적으로 비장 (spleen)에서 발현됨을 알 수 있다. 이와 같이 SAGE는 정량적인 정보를 제공하고, EST는 다수의 library를 반영하는 등의 상호보완적인 역할을 한다.

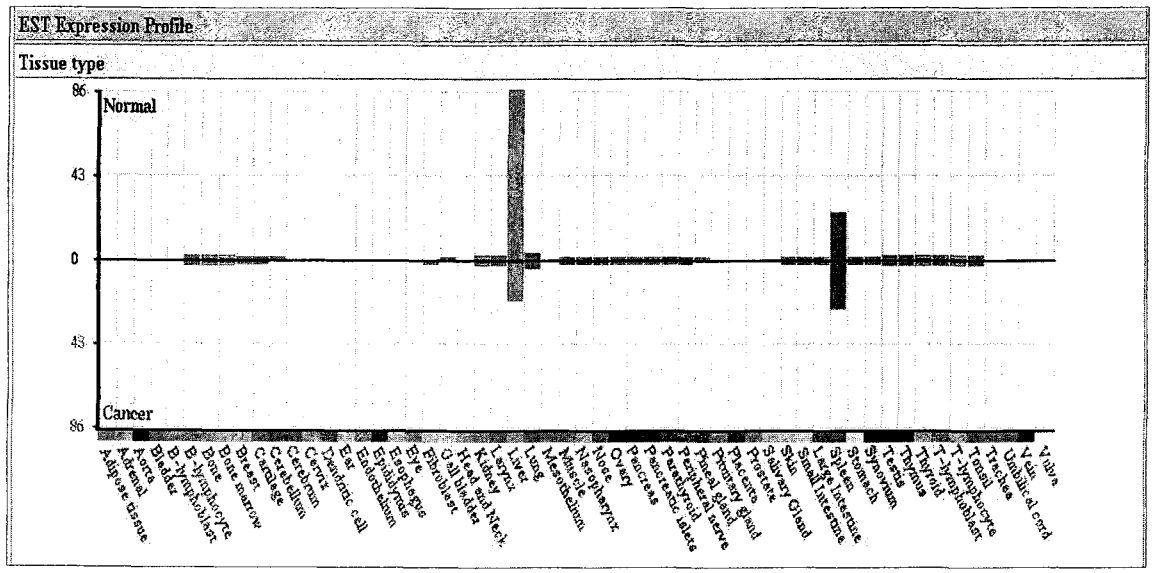


그림 25. 간에서 차등발현을 보이는 SERPINC1 유전자 (EST)

3.5 Genome Portal Site의 구축 (ECgene)

인간 유전자의 70% 이상이 alternative splicing을 보이고 있고(Modrek et al. 2001; Kan et al. 2002; Johnson et al. 2003), 많은 경우 alternative splicing에 의한 유전자 변이는 기능에 직접적인 영향을 미쳐 질병과 밀접한 관련이 있기 때문에(Nuno and Thomas 2003), alternative splicing을 고려한 genome annotation 사이트를 구축하는 것이 중요하다. 이를 위하여 제 2장에서 살펴본 바와 같이 NCBI, EBI를 비롯한 연구그룹들이 다양한 형태의 alternative splicing 데이터베이스(Thanaraj et al. 2003)를 구축하고 있다. 그러나 대부분이 유전자 구조의 모델링에 그쳐서 자세한 기능이나 발현의 분석에 대한 정보는 미미한 실정이다. 본 과제에서는 ECgene의 유전자 모델을 기반으로 ECfunction의 도메인/모티프 분석, ECexpression의 발현 분석을 종합하여 게놈 정보 포탈 사이트를 구축하였다. 특히 alternative splicing에 관련된 정보는 많은 장점을 지니고 있다.

가. ECgene Summary Viewer

그림 26은 ECgene 웹사이트(http://genome.ewha.ac.kr/ECgene/)의 초기 화면의 모습이다. 가장 위 부분에 genome browser, ECfunction, ECexpress, ECprofiler, ASmodeler 등의 사이트에 링크가 있고 그 아래에 검색창이 있다. 기본적으로 human, mouse, rat의 세 종을 지원하고, ECgene의 신뢰도를 선택하여 질의를 한다.

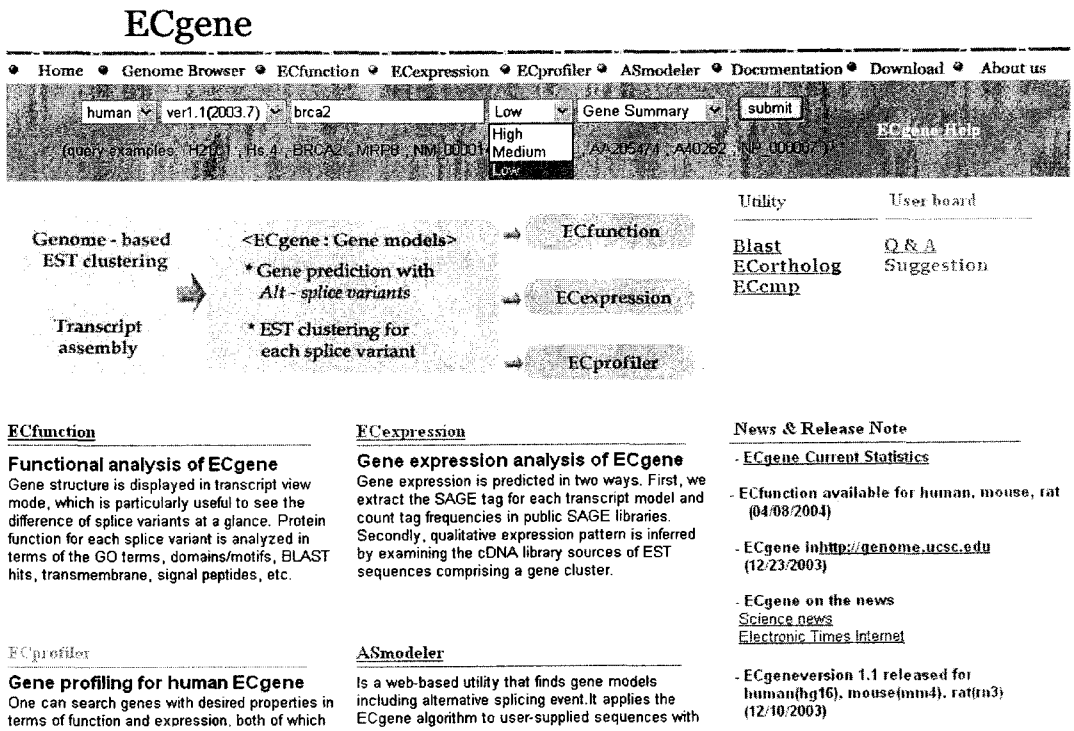


그림 26. ECgene 웹사이트의 초기화면

지원되는 검색 종류는 ECgene ID, HUGO gene symbol, 각 서열의 accession number, UniGene ID 등이 있으며, 이 외의 검색어를 입력하면 그림 27과 같은 에러 메시지와 함께 검색 범위가 보다 넓은 UCSC의 genome browser의 결과를 아래 창에 보여준다.

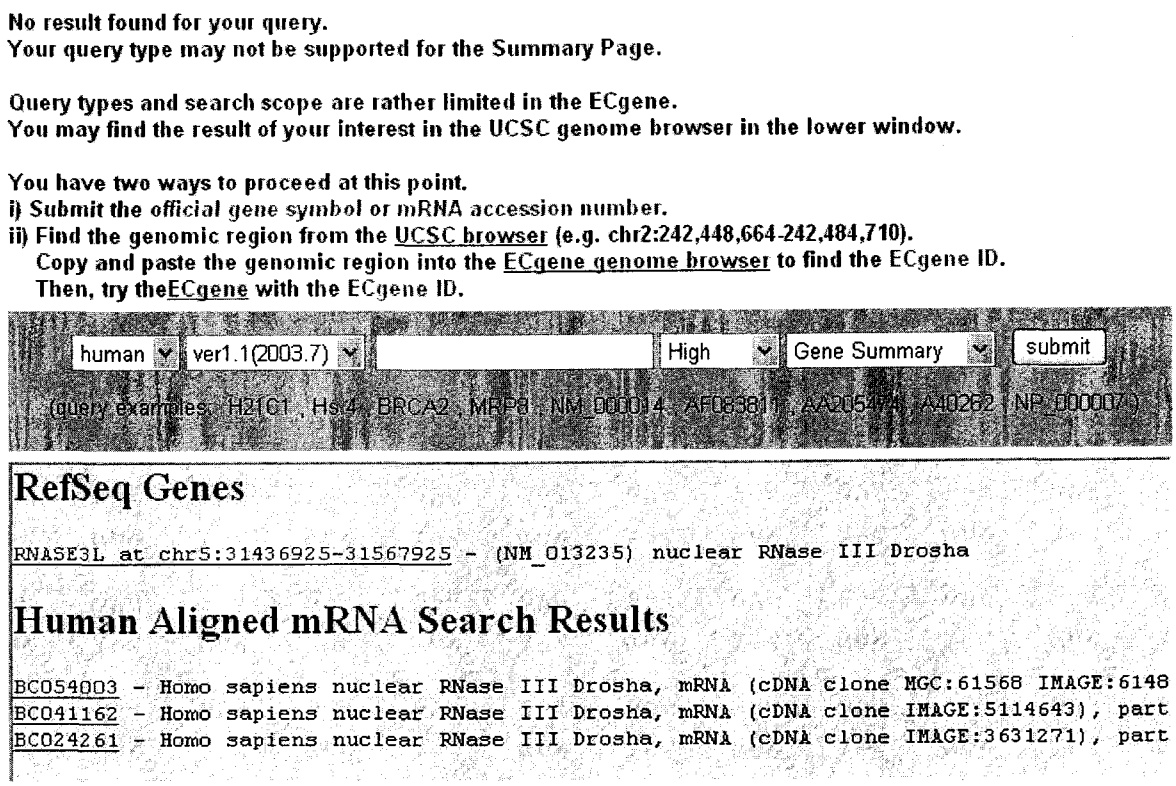


그림 27. ECgene 홈페이지에서 'drosha'로 검색한 결과.

성공적인 검색의 경우에는 그림 28과 같은 Summary Page를 출력하며 다음과 같은 정보를 포함한다.

- Gene Summary: 알려진 SwissProt 단백질이 있는 경우에는 SwissProt annotation을 parsing하여 clustering 정보와 함께 보여줌. 그리고 mRNA, protein 서열, clustering 결과에 대한 링크를 제공.
- mRNA Structure: ECfucntion에서 그리는 mRNA 좌표에서의 유전자 구조를 보여줌. 제목 줄에 genome browser, motif/domain viewer, alignment viewer 등의 다양한 유전자 분석 프로그램의 링크를 제공함.
- Transcripts Table: 각 splice variant의 성질을 간단한 표로 정리하였음. 신뢰도, cluster를 이루는 서열의 수, exon의 수, polyA tail의 존재 여부, mRNA와 단백질 서열의 길이, UTR 길이 등에 관련된 정보가 있음.
- Functional Annotation: GO, domain/motif, cross-reference를 통한 기능분류의 결과를 요약
- Gene Expression: ECexpression에서 계산한 것과 같이 SAGE와 EST library 분석에 근

거하여 tissue별, 그리고 질병별 발현 양상을 그림으로 나타냄.

- LinkOuts: 가장 널리 사용되는 게놈정보 제공 사이트에 대한 링크를 제공함. 각 링크가 지원하는 질의의 종류에 따라 HUGO symbol, accession number 등의 정보를 구하여 새 창을 띄워서 관련 정보를 보여줌.
- mRNA & protein sequences: 유전자 예측 결과 얻어지는 서열을 FASTA 포맷으로 뿌려주고, 서열을 저장하여 추후 분석할 수 있도록 서열 다운로드 서비스도 제공하고 있다.
- EST clustering result: ECgene cluster를 이루는 서열이 얻어진 cDNA library가 어떤 조직, 질병, 발생단계인지를 하나의 표로 정리하였음.

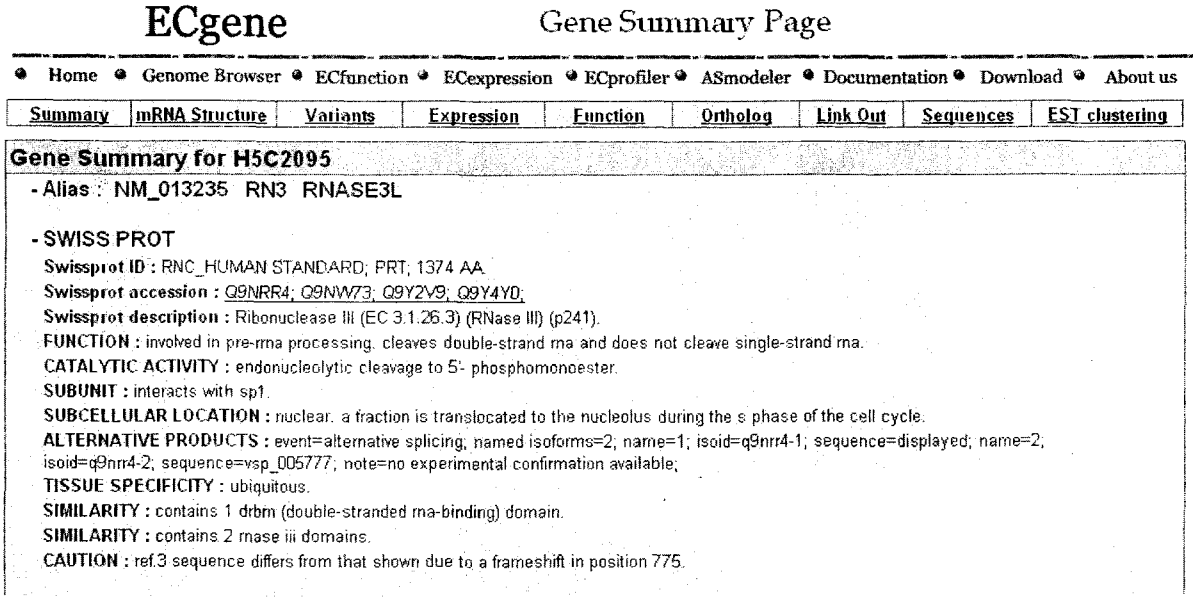


그림 28. ECgene 홈페이지의 위 부분 (Swiss-Prot 내용 부분)

나. ECgene 웹 사이트

다양한 정보로 손쉽게 이동할 수 있도록 페이지 내에 많은 링크를 걸어 놓았음. 그리고Documentation page에 사용 방법을 설명하였으며, Download page에서는 유전자 모델, mRNA 서열, 단백질 서열, clustering 결과 등의 정보를 파일로 download 받을 수 있음.

ECgene 웹 사이트는 국내에서 개발된 Database로는 AngioDB에 이어 두 번째로 2005년 1월에 발표되는 Nucleic Acids Research의 Database issue에 게재될 예정임.

3.6 ECgene 기반의 지식기반 이차DB 구축 (ChimerDB & Antisense)

Alternative splicing을 포함하고, UTR이 길며, polyA tail의 신뢰도가 높은 점과 같은 ECgene의 장점을 살려 다양한 주제에 응용할 수 있다. 현재 완성단계에 있는 두 데이터베이스만 간략하게 살펴보면 다음과 같다.

가. ChimerDB - Database of fusion mRNA and EST sequences in the GenBank

암의 발생과 진행 과정에서 염색체 전좌 (chromosomal translocation)가 흔히 일어나며, 이 경우 두 염색체의 경계면에 놓인 유전자가 섞여서 chimeric gene을 만들 수 있다. 따라서 chimeric sequence에 의한 fusion protein은 치료를 위한 신약 개발의 훌륭한 타겟이 될 수 있다. 본 연구에서는 GenBank의 서열을 genome assembly에 mapping하여 두 곳에 정렬되는 chimeric sequence를 찾았다. 그 결과 20,998개의 chimeric sequence를 발견하였고, 그 중 688개가 mRNA이었다. 그 중 상당수는 cDNA library 제작 과정에서 생기는 cloning artefact로 예상되지만 이를 걸러내면 암치료제 개발의 좋은 타겟을 제공할 것으로 생각된다. 이렇게 얻은 chimer 서열들을 ECgene를 사용하여 다시 clustering하고 데이터베이스를 구축하였다. ChimeDB 웹사이트 (http://genome.ewha.ac.kr/ECgene/ChimerDB)에서 해당 정보를 얻을 수 있으며, UCSC custom track을 사용하여 genomic alignment도 볼 수 있다.

아래 그림은 ChimerDB의 메인 화면이다. 각 ChimerDB의 내용은chromosome 별로 나뉘어 있으며, 원하는 chromosome을 클릭하면 그림 29와 같은 결과 화면을 얻을 수 있다.

ChimerDB : Chimeric Sequence Database for chrY

Select chromosome

chr1 | chr2 | chr3 | chr4 | chr5 | chr6 | chr7 | chr8 | chr9 | chr10 | chr11 | chr12 | chr13 | chr14 | chr15 | chr16 | chr17 | chr18 | chr19 | chr20

Chimer ID	# Chimer mRNA	# Chimer EST	ECgene Summary					
			ECgene ID	# RefSeq	# mRNA	# EST	Gene Symbol	Partner List
R1	0	1	HYC36	0	0	2		
R2	0	1	HYC60	0	0	1		FLJ14888
R3	0	1	HYC68	0	0	1		PACE4
R4	0	2	HYC76	0	0	12		MYH11, EEF1G, FLJ22347
R5	0	2	HYC77	1	11	2644	SLC25A6	MYH11, EEF1G, FLJ22347
R6	0	1	HYC89	1	2	13	FLJ13330	
R7	0	1	HYC96	1	4	51	P2RY8	SPON1
R8	0	1	HYC107	0	2	3		SPON1
R9	0	5	HYC108	2	9	265	DXYS155E, ASMT	RPS2, RNU64, SF1, MMP9, TXN, SET
R10	0	1	HYC109	0	0	1		MMP9
R11	0	3	HYC113	0	0	21		RPS2, RNU64, SF1, SET

그림 29. ChimerDB의 출력화면

그림 29에서처럼 결과 화면에는 각 chimeric sequence들을 ECgene을 사용해서 clustering한 결과를 보여준다. 해당 chromosome에 위치한 chimer 들의 정보 및 partner에 대한 정보도 확인할 수가 있다. 그림 28의 화면에서 EST 5개로 구성된 R9을 누르면

Sequence Information						Chimera Information			Chimera Partner Information			
GenBank Accession	Seq Type	Read Direction	Tissue Source	Seq Length	Seq Alignment	Overlapping ECgene	Gene Symbol	Chromosome	Seq Alignment	Overlapping ECgene	Ge Sym
AW993169	EST		Normal	346	0-95	HYC108, HYC114, HYC115, HYC113	DXYS155E, ASMT	chr9	89-346	H9C01539	SE
BI026039	EST		Neoplasia	404	205-385	HYC109, HYC108	DXYS155E, ASMT	chr20	1-215	H20C5237, H20C5243	MM
AA580963	EST	3		454	127-454	HYC108	DXYS155E, ASMT	chr9	0-98	H9C8843, H9C8844	TX
CB136036	EST	5	Neoplasia	549	0-405	HYC108, HYC113	DXYS155E, ASMT	chr16	406-549	H16C585, H16C589	RPS2;
BI912060	EST	5	Normal	920	399-744	HYC113, HYC108	DXYS155E, ASMT	chr11	18-392	H11C6850	SF

어떤 유전

그림 30. ChimerDB의 출력화면

자 사이에 fusion이 일어났는지를 그림 30과 같이 보여준다. 해당 chimeric EST가 정상(normal)인지 암조직에서 얻어졌는지를 보여주며, 각 EST/mRNA 별로 어떤 유전자와의 fusion을 하고 있으면 정렬에 대한 정보도 한 눈에 알 수 있다.

그림 30에서 ECgene ID를 개별적으로 클릭하면 그림 31과 같이 UCSC custom track을 사용하여 해당 ECgene에 대한 정보 및 여기에 mapping된 chimer 에 대한 정보도 한꺼번에 볼 수 있다.

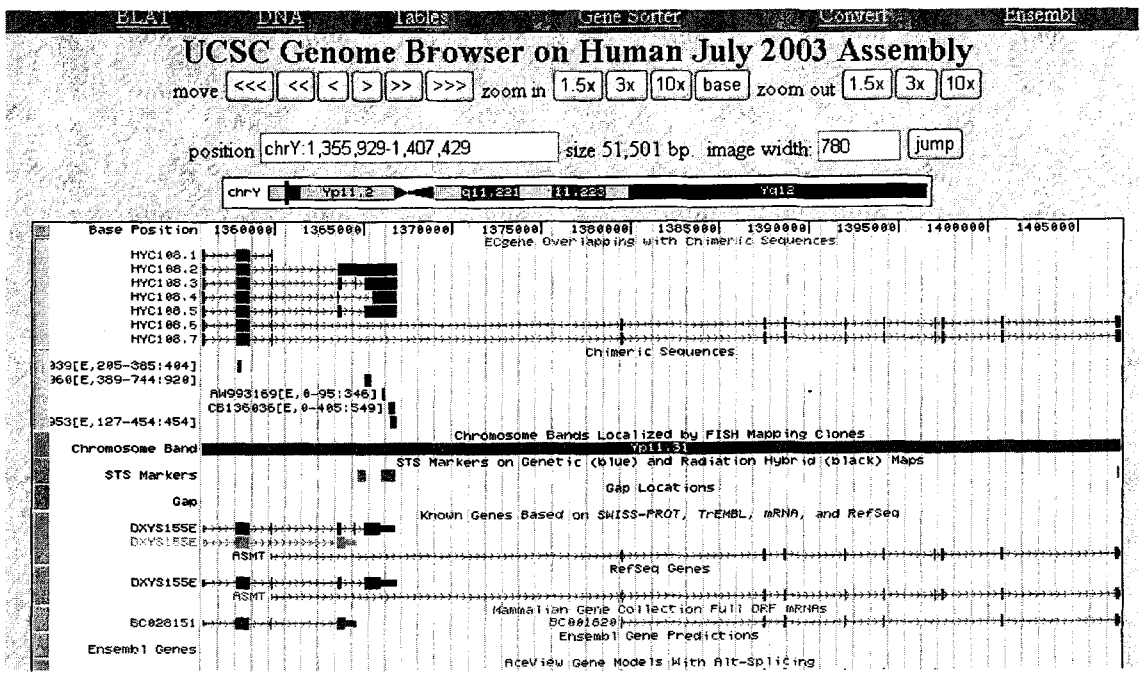


그림 31. ChimerDB의 출력화면

현재 ChimerDB의 업그레이드 버전으로, 사용자 서열이 염색체 전좌에 의한 chimer인지 아니면 cloning artefact에 의하여 얻어진 것인지를 판단할 수 있는 방법을 웹기반으로 구현하고, 그 결과를 데이터베이스로 구축하기 위하여 프로그램을 개발하고 있다.

나. Antisense Database

알려진 유전자의 antisense transcript가 세포내에서 많이 전사된다는 사실이 지난 3년간 여러 편의 논문에서 보고 되고 있다(Chu and Dolnick 2002; Vu et al. 2003; Tufarelli 2003; Shibata and Lee 2004). Human genome 내에는 다양한 형태의 sense-antisense 관계가 존재하며 다른 논문에 많이 보고된 바 있다(Lehner et al. 2002; Shendure et al. 2002; Yelin et al. 2003; Rosok and Sioud 2004; Chen et al. 2004) . 하지만 RNAi에서 많이 사용되는 것과 같이 sense strand에 쏙 들어가는 antisense가 있다면 이에 의한 조절 작용은 더욱 클 것으로 예상된다. ECgene에서는 clustering 과정에서 EST의 방향이 spliced sequence에 의하여 정해지는 방향과 일치하지 않는 서열은 다른 유전자로 분리하였다. 따라서 genomic loci를 공유하고 유전자의 방향이 다른 두 cluster를 뽑으면 자연스럽게 sense-antisense 쌍이 된다.

본 과제에서 찾은 sense-antisense는 다음 그림과 같은 두 종류이다.

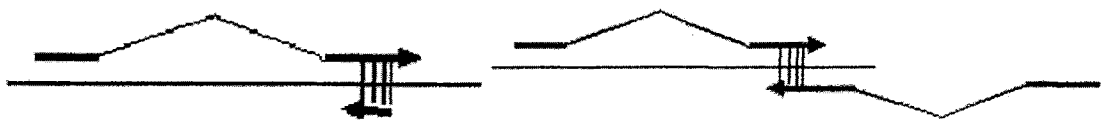


그림 32. Sense-antisense 사이의 관계

그리고 이와 같은 sense-antisense pair가 중요한 조절 기능을 가진다면 진화적으로 다른 종에서 보존되어 있을 가능성이 크다. 따라서 human, mouse, rat의 세 종에서 sense-antisense pairing 관계가 보존되는 것을 선별하여 natural antisense transcript 웹 사이트 (http://genome.ewha.ac.kr/antisense)에 정리하였다.

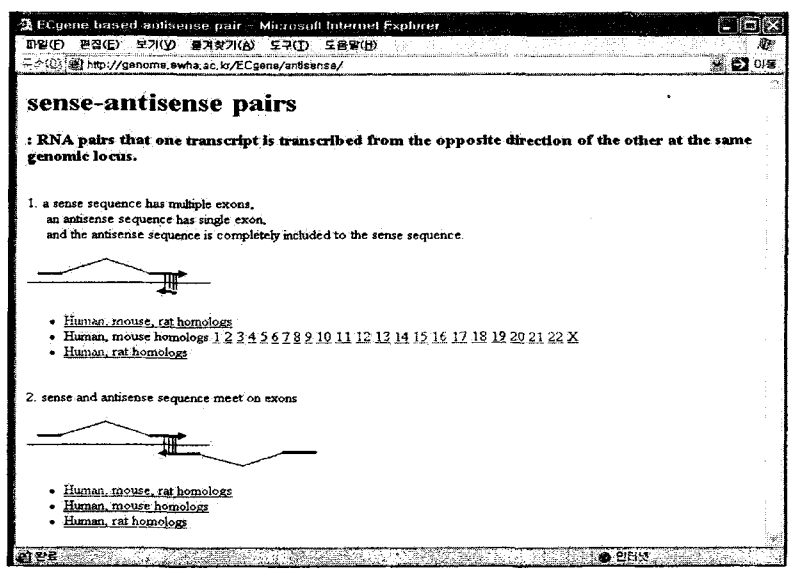
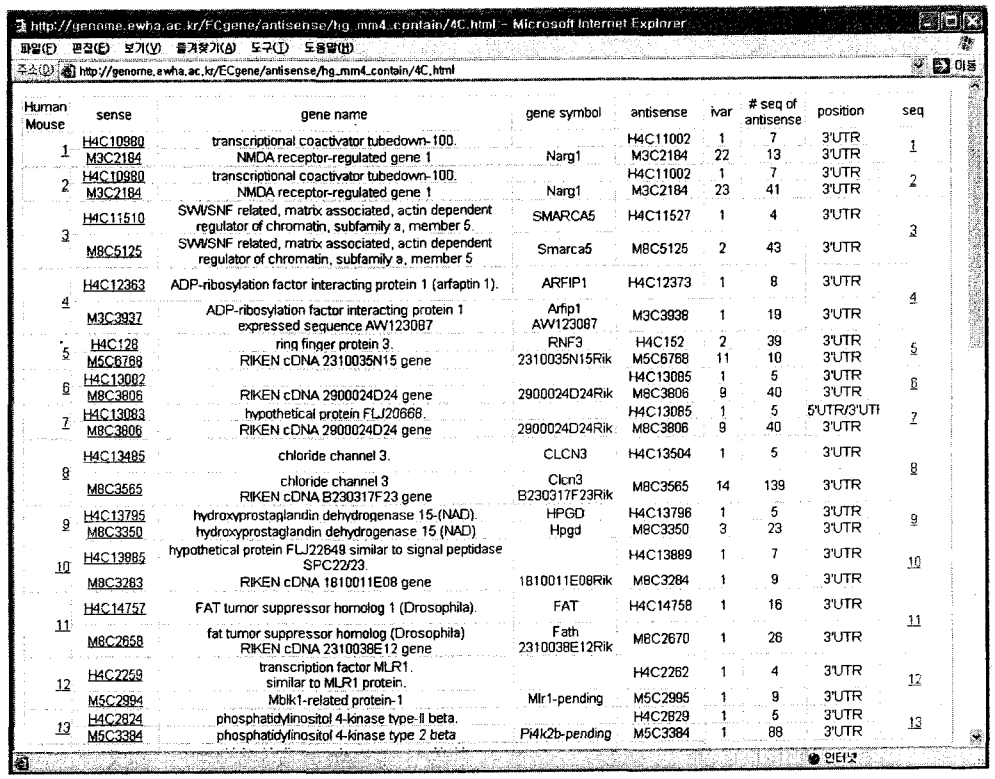


그림 33. Human, mouse, rat에서 보존되는 sense-antisense 관계

진화적으로 보존되는지를 확인하기 위하여, blastp나 blastn으로 human, mouse, rat 사이에서 homolog를 조사하였다. Sense transcript, antisense transcript 모두 homology 관계를 가지고, overlap 되는 부분의 위치나 서열의 homology 관계가 성립되는지 확인하지를 blat program으로 확인하였다. 이렇게 하여 얻은 진화적으로 보존되는 sense-antisense pair를 다음과 같이 보여준다.



http://genome.ewha.ac.kr/FCgene/antisense/hg_mm4_contain/4C.html - Microsoft Internet Explorer

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

주소(D) http://genome.ewha.ac.kr/FCgene/antisense/hg_mm4_contain/4C.html

Human Mouse	sense	gene name	gene symbol	antisense	ivar	# seq of antisense	position	seq
1	H4C10980	transcriptional coactivator tubedown-100.		H4C11002	1	7	3'UTR	1
	M3C2184	NMDA receptor-regulated gene 1	Narg1	M3C2184	22	13	3'UTR	
2	H4C10980	transcriptional coactivator tubedown-100.		H4C11002	1	7	3'UTR	2
	M3C2184	NMDA receptor-regulated gene 1	Narg1	M3C2184	23	41	3'UTR	
3	H4C11510	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5.	SMARCA5	H4C11527	1	4	3'UTR	3
	M8C5125	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5	Smarca5	M8C5125	2	43	3'UTR	
4	H4C12363	ADP-ribosylation factor interacting protein 1 (arfaptin 1).	ARFIP1	H4C12373	1	8	3'UTR	4
	M3C3937	ADP-ribosylation factor interacting protein 1 expressed sequence AW123087	Arfip1 AW123087	M3C3938	1	19	3'UTR	
5	H4C128	ring finger protein 3.	RNF3	H4C152	2	39	3'UTR	5
	M5C6788	RIKEN cDNA 2310035N15 gene	2310035N15Rik	M5C6788	11	10	3'UTR	
6	H4C13002			H4C13085	1	5	3'UTR	6
	M8C3806	RIKEN cDNA 2900024D24 gene	2900024D24Rik	M8C3806	9	40	3'UTR	
7	H4C13083	hypothetical protein FLJ20668.		H4C13085	1	5	5'UTR/3'UTI	7
	M8C3806	RIKEN cDNA 2900024D24 gene	2900024D24Rik	M8C3806	9	40	3'UTR	
8	H4C13485	chloride channel 3.	CLCN3	H4C13504	1	5	3'UTR	8
	M8C3565	chloride channel 3 RIKEN cDNA B230317F23 gene	Clcn3 B230317F23Rik	M8C3565	14	139	3'UTR	
9	H4C13795	hydroxyprostaglandin dehydrogenase 15-(NAD).	HPGD	H4C13796	1	5	3'UTR	9
	M8C3350	hydroxyprostaglandin dehydrogenase 15 (NAD)	Hpgd	M8C3350	3	23	3'UTR	
10	H4C13885	hypothetical protein FLJ22648 similar to signal peptidase SPC22/23.		H4C13889	1	7	3'UTR	10
	M8C3283	RIKEN cDNA 1810011E08 gene	1810011E08Rik	M8C3284	1	9	3'UTR	
11	H4C14757	FAT tumor suppressor homolog 1 (Drosophila).	FAT	H4C14758	1	16	3'UTR	11
	M8C2658	fat tumor suppressor homolog (Drosophila) RIKEN cDNA 2310038E12 gene	Fath 2310038E12Rik	M8C2670	1	26	3'UTR	
12	H4C2259	transcription factor MLR1. similar to MLR1 protein.		H4C2262	1	4	3'UTR	12
	M5C2994	Mblk1-related protein-1	Mlr1-pending	M5C2995	1	9	3'UTR	
13	H4C2824	phosphatidylinositol 4-kinase type-II beta.		H4C2829	1	5	3'UTR	13
	M5C3384	phosphatidylinositol 4-kinase type 2 beta	Pi4k2b-pending	M5C3384	1	88	3'UTR	

인터넷

그림 34. Human, mouse, rat에서 보존되는 sense-antisense 표

예를 들어, human, mouse, rat에서 모두 보존되는 tight junction protein1에서 대해서 세부사항을 본다면, 표에서 H15C818, M7C4439, R1C4327 가 tight junction protein1 (TJP1) 으로 homolog 이고, 이에 대해서 H15C820, H7C4438, R1C4329가 TJP1의 3′ UTR에 포함되는 antisense transcript 이고 서로 homolog 이다.

sense	gene name	gene symbol	antisense	#seq of antisense	position
H15C818	tight junction protein 1 (zona occludens 1)	TJP1	H15C820	5	3' UTR
M7C4439	tight junction protein 1, similar to tight junction protein 1	Tjp1	M7C4438	20	3' UTR
R1C4327	tight junction protein 1		R1C4329	4	3' UTR

또, 이 세 종의 sense transcript와 antisense transcript의 위치 관계가 보존되는지 확인하기 위해 쓴 blat program의 결과를 다음 그림과 같이 UCSC에 custom track으로 확인할

수 있으며, antisense webpage에서 그에 대한 link를 제공하고 있다. 그림에서 보면, 각종의 antisense transcript 가 같은 위치에 있으며 각자의 sense transcript 에 포함되고 있으며, sense transcript 끼리 비슷한 형태를 보임을 알 수 있다.

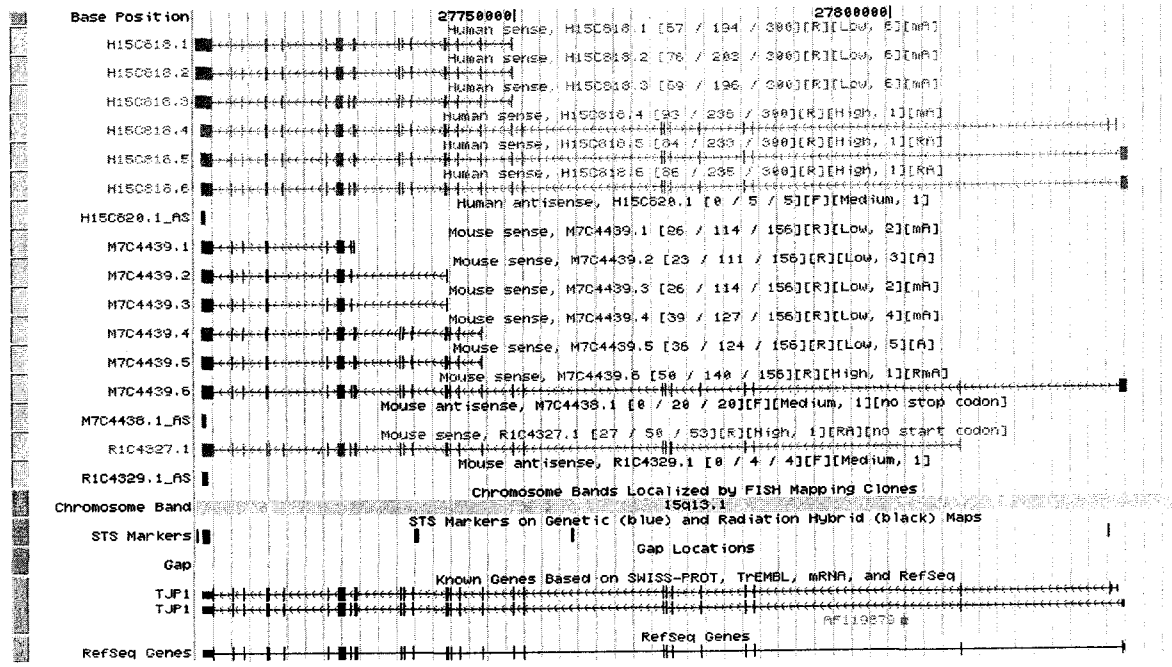


그림 35. Human, mouse, rat에서 보존되는 sense-antisense의 Genomic alignment

아직은 이렇게 구한 sense-antisense pair가 한 세포 내에서 동시에 발현되는지, 상호간에 어떤 조절 역할을 하는지는 많이 밝혀지지 않았다. 하지만 본 과제에서 발굴한 완전히 포함되며 다른 종에서 보존이 되는 antisense를 대상으로 그 기능에 대한 연구가 진행될 것으로 예상된다.

제 4 장 목표달성도 및 관련분야에의 기여도

4.1 목표달성도

가. EST clustering 알고리즘의 완성

(1) ECgene 알고리즘은 genome-based EST clustering과 그래프 이론을 이용한 assembly로 alternative splicing의 분석이 포함되어 있음. 이 외에도 발현의 조절에 중요한 역할을 하는 UTR이 길고, polyA tail의 신뢰도가 높은 장점이 있어, 각종 유전자 예측 프로그램의 장점을 취합한 방법이라고 할 수 있음.

(2) ECgene 알고리즘으로 human, mouse, rat을 분석한 결과는 세계적인 genome center인 UCSC에서 정규 유전자 예측 프로그램으로 선정되어 전 세계를 대상으로 공개되고 있음. 또한 유전자 예측 알고리즘을 생물정보학 관련 분야의 최고 학술지인 Genome Research에 투고하였고 현재 심사 중에 있음. 또한 ECgene 알고리즘으로 사용자의 서열을 분석하는 웹 서비스 프로그램인 ASmodeler를 개발하여 Nucleic Acids Research의 Web Server Issue에 발표하였음.

(3) Alternative splicing을 포함한 유전자 예측을 고려하면 알고리즘의 완성도 측면에서 NCBI의 UniGene보다 우수하다고 할 수 있음. NCBI의 AceView와 EBI의 ASD 프로젝트가 비슷한 수준에 와 있음. 따라서 3년간 개발한 내용으로는 목표를 100% 이상 달성하였음.

나. Genome annotation 방법의 개발

(1) Functional annotation은 세계적인 선도그룹에서 개발된 방법을 그대로 사용하였으나, ECgene의 강점인 alternative splicing에 초점을 맞추어 기능에 대한 분석을 하였음. 이는 질병관련 유전자 변이를 발굴하는데 결정적인 이점으로 작용할 것임.

(2) mRNA 좌표에서 유전자 변이를 보는 프로그램은 ECfunction이 세계에서도 유일함.

(3) 발현의 분석은 EST clustering이 각 variant 별로 되어 있기 때문에 variant-specific expression을 볼 수 있음. 또한 SAGE 분석 방법도 tag를 transcript에서 뽑아내어 발현을 예측하는 방식을 취해 다른 방법보다 신뢰도가 높음. cDNA와 SAGE library를 organ-tissue-cell type, pathology, developmental stage에 따라 계층적으로 분류한 것도 EST 발현 분석에 큰 장점이 됨.

(4) 발현을 보여주는 그래프도 정상과 암조직을 구분하여 차등 발현을 한 눈에 볼 수 있게 하였고, 이도 다른 곳에서는 볼 수 없는 내용임. 따라서 EST와 SAGE를 통하여 유전자의 발현을 in silico로 예측하는 방법에서는 세계적으로도 뛰어난 분석 방법을 개발하였음.

다. Genome 웹 사이트의 개발

(1) ECgene 알고리즘, 기능 및 발현의 분석 방법에서의 강점을 최대한 살리는 방향으로 alternative splicing에 관한 종합적인 정보를 제공하는 웹사이트를 개발하여 2005년 1월의 Nucleic Acids Research의 Database Issue에 발표할 예정임.

(2) Alternative splicing에 강점이 있지만, 일반적인 유전자의 발현 분석도 다른 방법에 비하여 신뢰도가 높음.

(3) 방법론으로는 뛰어나지만 프로그램의 성능은 개선할 여지가 많음. 향후 국가유전체 정보센터와의 공동협력을 통하여 성능을 개선하고 다양한 분석 수단을 개발하여 국가적인 게놈정보 사이트로 발전시켜나갈 예정임.

라. 유용 유전자의 발굴

(1) EST의 발현 분석을 통한 유전자 profiler는 완성하였음. Java web start로 만들어 누구나 쉽게 사용할 수 있으며, 발현과 기능을 tree 형태의 분류체계를 쉽게 navigation 할 수 있는 interface를 개발하였음. EBI의 EnsMart에 비하여 종합적인 profiling 능력은 떨어지지만, 발현의 분석 면에서는 뛰어남. Variant-specific search 기능을 추가하면 isoform까지 검색한다는 측면에서 독특한 profiler가 될 것으로 예상됨. 또한 현재 개발 중인 SAGE 발현의 분석 방법이 완성되면 이를 이용한 profiling 기능도 추가할 예정임.

(2) EST의 분포를 이용한 tissue-specific gene의 목록을 일부 완성하였음. SAGE 데이터는 보다 정량적이고 직접적인 발현 정보를 제공함. 본 과제의 처음 계획에는 포함되지 않았으나 이 부분의 중요성이 크고, 다른 연구그룹의 접근방식에 비하여 우수할 것으로 예상되는 생각이 있어 현재 집중 개발하고 있음. 이 일이 끝나면 EST, SAGE, microarray 데이터를 종합한 발현 분석에 의한 profiling이 가능할 것으로 예상됨. 당초 계획에 비교하면 유용 유전자의 발굴 자체의 진도는 80% 정도로 늦으나 방법론 개발의 측면에서는 목표를 훨씬 초과달성하였음.

(3) ChimerDB, antisense DB와 같은 ECgene에 기반을 둔 새로운 개념의 지식기반 이차 DB를 구축하였음. 그 결과는 기초연구는 물론 신약 개발의 타겟을 발굴에 유용할 것으로 예상됨.

4.2 관련분야에의 기여도

가. ECgene 유전자 예측 결과의 공개

(1) ECgene 알고리즘을 논문으로 발표하고 human, mouse, rat에 대한 계산 결과를 자체 웹사이트와 UCSC genome center를 통하여 공개하였음. Alternative splicing을

genome-scale에서 제대로 모델링하고 지속적으로 업데이트하는 곳은 세계적으로도 2-3 그룹밖에 존재하지 않음.

(2) ECgene 알고리즘을 이용하여 사용자 서열의 분석이 가능하도록 웹 서버 프로그램인 ASmodeler를 개발하였음. 향후 국가유전체로 서비스를 이전하여 보다 많은 연구자들이 사용할 수 있는 방안으로 개발할 예정임.

나 . *ECgene* 게놈정보제공 포탈 사이트 개발

(1) Alternative splicing을 고려한 유전자 구조, 발현, 기능에 대한 종합적인 분석 결과를 제공하는 웹 사이트를 구축하였음. 이는 기초생물학 연구뿐만 아니라 의학·약학 분야에서의 연구에 중요한 정보가 될 것으로 예상됨.

(2) 유전자 구조의 모델링 뿐만 아니라 EST와 SAGE를 통한 유전자 발현의 분석에서도 다른 웹사이트/연구결과에서는 볼 수 없는 독특한 정보를 많이 제공하고 있음. 특히 정상조직과 암조직에서의 유전자 발현의 차이를 웹과 그림을 통하여 쉽게 볼 수 있다는 측면은 관련 연구자들에게 큰 도움이 될 것으로 기대됨.

(3) 현재 Weizmann Institute에서 개발한 GeneCards에서도 ECgene 웹사이트에 대한 링크아웃을 제공하고 있으며, ECgene에 대한 분석을 통한 새로운 가치를 창출하기 위한 노력을 하고 있음. 앞으로 국내외적으로 비슷한 종류의 연구협력이 이루어질 것으로 기대됨.

(4) ECgene 사이트는 대규모 게놈정보를 포함하여 컴퓨터 시스템 자원이 많이 필요하여 국가유전체정보센터로 운영을 이관할 예정임. 공동연구와 협력을 통하여 Human, mouse, rat 뿐만 아니라 다른 모델 생물로 대상을 늘리고, 새로운 종류의 2차 데이터베이스를 구축하며, 다양한 분석 수단을 개발할 것임. 이를 통하여 ECgene을 국제적으로도 경쟁력이 있는 국가적인 게놈정보 제공 사이트로 개발할 것임.

나. 관련 분야의 응용 가능성

(1) 현재의 alternative splicing을 포함한 유전자 모델링 결과는 다음의 분야에 바로 응용할 수 있음. AS의 모델링에서 비교 우위가 있기 때문에 선진국과의 경쟁에서 뒤지지 않는 결과를 얻을 수 있을 것으로 예상됨.

- alternative splicing과 SNP와의 관련을 통한 약물유전체학
- alternative initiation을 일으키는 전사조절 메카니즘 연구
- alternative splicing과 질병 사이의 관계 연구를 통한 질병의 진단 및 치료
- alternative splicing을 고려한 올리고 DNA chip의 개발

이는 기존의 연구팀이 고민하던 문제에 대하여 다른 각도에서의 해결책을 제시할 수 있을 것으로 기대됨.

(2) Tissue-specific 또는 disease-specific한 발현을 보이는 유전자의 목록은 질병의 진단과 치료에 직접 응용될 수 있음. 또한 ChimerDB는 암에 의한 fusion mRNA를 찾아 신약개발의 타겟으로 이용될 수 있을 것임.

(3) ECgene은 EST clustering에서 출발하기 때문에 잘 알려진 유전자뿐만 아니라 noncoding RNA와 같은 아직 알려지지 않은 transcript들이 많이 포함되어 있음. 또한 transcript의 UTR 부분을 연장하였고 polyA tail의 존재를 정확하게 계산하는 등의 많은 장점을 지니고 있음. 이는 3′ UTR을 타겟으로 하는 microRNA의 기능 연구에 직접적인 응용이 가능함. 이와 같은 ECgene 알고리즘의 장점을 이용한 다양한 연구가 가능할 것으로 예상됨.

제 5 장 연구개발결과의 활용계획

5.1 추가연구의 필요성

가. ECgene의 운영·개선 방안에 대한 연구

(1) ECgene의 genome annotation은 알고리즘의 변경이 없어도 GenBank의 mRNA와 EST 서열이 증가함에 따라 그 결과도 바뀌게 된다. 이런 이유로 NCBI의 UniGene도 2-3주에 한 번은 새로운 EST clustering 모델을 제시하는데, ECgene도 genome assembly가 바뀌면 무조건, 그렇지 않더라도 1-2개월에 한 번씩은 새로운 유전자 모델을 만들 필요가 있다. 이 경우 많은 부분을 자동화하여 효율적인 시스템과 데이터의 관리가 필요하다. 또한 현재의 human, mouse, rat genome의 분석 외에도 다른 모델 동물에 대한 annotation을 추가할 것이다. 이는 향후 국가유전체연구센터와의 협력·공동연구를 통하여 해결할 예정이다.

(2) ECgene의 기능과 발현에 대한 분석은 앞으로도 지속적으로 개발할 사항이다. 기능의 경우 pathway에 대한 정보, 논문(문헌)에 대한 정보 등의 추가가 급선무이다. 또한 SAGE에 의한 발현의 분석은 그 방법적인 측면에서 개선할 여지가 많아서 현재 새로운 방법을 개발 중에 있다. 그리고 현재의 발현 분석은 공개된 데이터를 이용하기 때문에 시간이 지나서 데이터가 많이 쌓일수록 더욱 강력한 수단이 된다. 이를 위하여 지속적으로 cDNA와 SAGE library에 대한 분석을 하고 DB화할 필요가 있다. 가능하면 GEO, ArrayExpress와 같은 공개된 microarray 데이터도 추가하면 더욱 좋은 결과가 될 것이다. 이와 같이 앞으로도 꾸준히 새로운 방법을 개발하는 것이 필요하다.

(3) ECgene을 이용한 다양한 종류의 2차 데이터베이스가 구축될 것이다. ChimerDB, Antisense DB, UTRdb 등과 같이 현재 구축되어 있는 정보는 물론 앞으로 추가적으로 분석할 내용(타연구에의 응용 참조)도 ECgene DB의 annotation에 포함시킬 것이다. 이와 같은 지식 기반의 DB는 국내·외 연구자들에게 중요한 정보가 될 것이다.

나. Alternative splicing에 대한 연구 및 타분야에의 응용

(1) Alternative splicing은 유전자의 기능과 발현을 조절하는 중요한 메카니즘으로 질병의 직접적인 원인이 되는 경우가 많이 있는 것으로 밝혀졌다. ECgene에서 유전자를 모델링하고 그 결과를 DB와 웹사이트를 통하여 공개하였으며, 발현과 기능의 변화를 볼 수 있는 기반을 밝혔지만 이는 시작에 불과하다.

(2) Alternative splicing과 질병 관련성의 연구: 본 과제에서 구축한 발현 데이터를 이용하여 특정 질병에 관련된 유전자를 profiling할 수 있다. 이를 연장하여 isoform에 따라 조직 또는 질병에 대한 발현의 달라지는 경우 alternative splicing에 의한 tissue-specific isoform, disease-specific isoform을 발굴할 수 있다. 이와 같은 지식은 신약 개발에 바로 연결시킬 수 있는 새로운 패러다임이 될 것이다. 특히 단일항체를

이용한 암치료 방법으로 최근 각광받고 있는 immunotherapy의 시발점이 될 것으로 기대된다.

(3) Splicing 조절 인자와 메카니즘의 연구: 예상보다 훨씬 많은 유전자가 alternative splicing을 보이는 것으로 밝혀짐에 따라 이를 조절하는 인자를 밝히고 그 메카니즘을 알아내는 것이 중요하다. 이는 antisense 또는 RNAi를 이용하여 splicing을 조절하는 방식으로 질병의 치료에 응용할 수 있다.

(4) Alternative splicing과 SNP, 약물유전체학간의 관련성 연구: 질병과 직접적인 관련이 있는 SNP의 최소 15%는 alternative splicing과 관련이 있는 것으로 밝혀졌다. 따라서 이들 SNP는 단순히 염기 또는 아미노산 한 개를 바꾸는 것이 아니라 exon 단위의 변화를 일으킬 가능성이 크고, 이는 splicing 조절 부위에 나타날 확률이 크다. 이에 대한 연구는 특히 약물유전체학 분야에서 중요한 역할을 할 것으로 기대되며, 신약개발의 필수적인 부분이 될 것이다.

(5) Human gene의 30%는 alternative initiation을, 70%는 alternative termination 현상을 보인다. 이런 현상이 발현과 기능에 어떤 영향을 미치는지에 대한 연구가 필요하다. 또한 alternative initiation을 보이는 경우 전사조절 방법에 대한 연구가 있어야 한다.

(6) Microarray로 분석한 결과를 보면 mRNA와 EST로 예측한 것보다 훨씬 많은 종류의 splice variant가 존재한다. 따라서 microarray 실험에서 사용한 probe를 mapping하여 어떤 isoform이 해당 tissue에서 얼마만큼 발현되고 있는가를 거꾸로 추론할 수 있다. 특히 Affymetrix사의 DNA chip은 유전자 당 10여개의 probe를 사용하기 때문에 AS에 의한 isoform의 차이를 볼 수 있는 가능성이 크다. 따라서 공개된 Affymetrix chip 데이터를 이런 측면에서 분석하는 것도 흥미로운 연구가 될 것이다.

5.2 타연구에의 응용

가. DNA chip의 개발

(1) 현재의 DNA chip은 alternative splicing에 의한 isoform을 제대로 고려하지 않고 디자인되었다. 개인의 genotyping에서 어떤 isoform을 가지고 있는가가 질병에 직접적인 영향을 미치기 때문에 AS를 고려하여 genotyping할 필요가 있다. DNA chip의 probe 디자인을 위해서는 제대로 된 유전자 모델링이 필수적으로 ECgene의 강점을 최대한 이용하면 AS를 고려한 genotyping을 할 수 있는 올리고 DNA chip을 개발할 수 있을 것이다. 현재 Affymetrix사에서는 알려진 모든 유전자의 exon을 고려한 chip을 디자인하여 기초실험을 진행하고 있다.

(2) AS를 고려한 DNA chip은 질병의 진단과 치료에도 중요하다. 어떤 isoform이 특정 질환에서 specific하게 발현한다면 이는 질병의 마커 또는 치료를 위한 타겟으로 사

용될 수 있다.

나. 신약개발의 타겟 유전자 발굴

(1) ECgene의 발현 분석 시스템을 이용하여 tissue-specific, disease-specific gene을 찾아 타겟 유전자로 사용한다.

(2) 마찬가지 방법으로 tissue-specific, disease-specific isoform을 찾아 타겟 유전자로 사용한다.

(3) ChimerDB에서 암과 관련성이 큰 fusion mRNA/EST를 찾아 실험으로 확인하여 타겟 유전자로 이용한다.

다. noncoding RNA에 관한 연구

(1) ECgene은 EST의 clustering에 기반을 두어 단백질을 coding하지 않더라도 EST로 전사만 일어나면 데이터베이스에 포함되어 있다. 따라서 noncoding RNA를 찾기 위한 연구의 시발점이 될 수 있다.

(2) microRNA는 3′ UTR 부분에 결합하여 기능을 발휘하는 것으로 알려져 있다. ECgene은 다른 유전자 모델에 비하여 3′ UTR이 길고, alternative splicing에 의한 isoform이 포함되어 있기 때문에 UTR을 타겟으로 하는 RNA의 binding 연구에 이상적인 모델이다.

5.3 기업화 추진방안

(1) ECgene의 우수성을 감안하면 기업화 가능성도 있지만, NCBI와 EBI의 다양한 데이터베이스나 프로그램과 같이 공개하여 많은 사람들이 사용하는 것이 더욱 중요하다고 생각된다. 따라서 국가유전체정보센터와 협력하여 서비스를 강화하고 다양한 프로그램을 개발하여 더욱 좋은 정보를 제공하기 위하여 노력할 것이다. 따라서 가까운 시일 내에 기업화 추진 계획은 없다.

(2) 다만 국내·외의 생명과학 관련 기업에서 원하는 경우 질병의 진단과 치료, 신약개발과 같은 새로운 가치창출을 위하여 최대한 협력할 것이다.

제 6 장 연구개발과정에서 수집한 해외과학기술정보

본 연구과제의 주제인 유전자 모델링, 기능 및 발현의 분석은 인간게놈지도가 발표됨에 따라 전 세계적으로 가장 치열한 경쟁이 이루어지고 있는 분야 중 하나이다. 따라서 지난 3년간 연구개발을 하는 도중에 수많은 논문이 발표되었고 본 과제와 지향하는 바가 비슷한 프로젝트와 논문도 여러 편 있었다. 이들 논문의 장단점을 분석하여 본 과제의 연구개발에 반영하였고 나름대로의 새로운 아이디어를 접목하였기에 본 과제에서 개발한 ECgene 관련 프로그램과 분석은 세계적으로도 뛰어난 연구 성과라고 자부할 수 있다. 본 과제에 관련된 국내·외 연구현황은 제2장에서 상세히 밝힌 바 있기에 여기서는 그 중 우수한 연구결과를 요약하기로 한다.

현재 우수한 유전자 예측 프로그램은 대부분 UCSC genome center의 genome browser를 통하여 제공되고 있다. 여기에 alternative splicing에 관련된 것을 추가하여 정리하면 다음과 같다.

가. 널리 사용되는 유전자 prediction & annotation 결과

(1) NCBI의 RefSeq와 EBI의 Ensembl: most famous & widely used

(2) NCBI에서 제공하는 GenomeScan: GenScan을 개선한 것으로 가정 정확한 유전자 예측 프로그램. 그 결과는 RefSeq의 XM 서열을 만드는데 사용됨.

(3) UCSC의 KnownGene Set: protein coding gene 중심의 유전자 DB (SwissProt, TrEMBL의 단백질, GenBank의 mRNA 이용)

(4) Fgenesh++: Softberry Inc. 사의 유전자 정보

(5) Acembly: NCBI의 Thierry-Mieg 연구실의 AceView 예측

(6) Twinscan: GenScan과 비슷한 HMM 방법. mouse와 human간에 보존을 gene prediction의 정보로 이용함.

(7) SGP, GeneID: 스페인의 GRIB에서 개발된 ab initio prediction. 신뢰도가 떨어짐.

(8) Alt-splicing: UCSC 그룹에서 개발한 다른 종에서 보존되는 alternative splicing

나. Alternative splicing에 대한 annotation DB

(1) EBI의 ASD (alternative splicing database): Alternative splicing을 annotation하기 위한 유럽 중심의 국제협력 consortium. 시작한지 오래되지는 않으나 많은 연구그룹의 협력을 도모하는 장기 프로젝트로 컴퓨터 프로그램에 의한 예측뿐만 아니라, 실험과 논문으로 확인된 AS에 대한 manual annotation을 추구함.

(2) NCBI의 AceView: 현재 reliability, coverage, annotation 수준에서 ECgene과 함께 가장 뛰어난 DB로 판단됨. AS뿐만 아니라 일반적인 genome annotation을 추구함.

(3) UCLA의 C. Lee 그룹의 ASAP (alternative splicing annotation project): AS 분야 연구의 선구자 중 한 사람으로 alternative splicing에 대한 생물학적 의미에 대한 연구

를 많이 하였음. 유전자 모델링 자체로는 뛰어난 결과는 아님.

(4) 독일 Reich 그룹의 EASED (Extended alternatively spliced EST database): AS 현상을 보이는 EST DB로 구체적인 Gene modeling은 없음.

다. Alternative splicing에 대한 최근의 주요 논문

(1) (Johnson et al. 2003, Science) Exon junction array를 사용하여 alternative splicing을 조사하였음. Human gene 10,000개에 대한 junction probe를 제작하고 52개의 tissue와 cell line에서 각 probe의 발현을 통하여 alternative splicing을 조사한 결과 최소 74%의 유전자가 alternative splicing을 보이는 것으로 판명됨. 이 실험데이터는 GEO에 공개되어 있으며 해당되는 10,000개의 유전자에 대한 alternative splicing의 pattern과 시료별 expression level을 연구하면 흥미로운 결과를 얻을 수 있을 것으로 예상됨.

(2) (Kapranov et al. 2002, Science) & (Kampa et al. 2004, Genome Research) 염색체 21번과 22번에 대한 tiling array를 제작하여 11개의 cell line에서 얻어지는 전사체에 대한 연구를 2002년 수행하였고, 2004년에 새로운 방법으로 분석하였음. 전사체 중에서 31%만이 알려진 유전자이고 추가로 20% 정도가 mRNA 또는 EST 증거가 있음. 나머지 49%가 novel transcript인데 이는 아직도 밝혀지지 않은 RNA가 많음을 의미함. 그 중 상당수는 unknown isoform의 일부이거나 noncoding RNA일 가능성이 큼. 이 tiling array는 염색체 21번과 22번의 전체 지역을 평균 35 bp 간격으로 probing하므로 probe를 genome에 mapping하면 발현 데이터로부터 새로운 alternative splicing candidate을 찾고 발현 level을 유추할 수 있을 것으로 예상됨. 앞의 junction array 데이터가 알려진 RefSeq 10,000개의 알려진 exon-intron 경계의 변화를 볼 수 있다면, 이 데이터는 분석이 좀 더 어렵지만 알려지지 않은 유전자의 변화도 볼 수 있는 이점이 있을 것으로 예상됨.

(3) 다음 일련의 논문들은 alternative splicing과 질병과의 관계에 초점을 둔 review 논문임.

- Bracco and Kearsey 2003, Trends in Biotech., The relevance of alternative RNA splicing to pharmacogenomics.
- Caceres and Kornblihtt 2003, Trends in Genetics, Alternative splicing: multiple control mechanisms and involvement in human disease.
- Faustino and Cooper 2003, Genes & Dev., Pre-mRNA splicing and human disease.
- Garcia-Blanco et al. 2004, Nature Biotech., Alternative splicing in disease and therapy.
- Brinkman 2004, Clinical Biochem., Splice variants as cancer biomarkers.

제 7 장 참고문헌

Black, D.L. 2000. Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. Cell. 103: 367-370.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72,** 291-336.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*, **474,** 83-86.

Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Zhang, R. and Rowley, J.D. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Aicds Res* 32(16): 4812-4820.

Christoffels, A., Gelder, A.V., Greyling, G., Miller, R., Hide, T., and Hide, W. 2001. STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res* **29:**234-238.

Chu, J. and Dolnick, B.J. 2002. Natural antisense (rTSα) RNA induces site-specifc cleavage of thymidylate synthase mRNA. *Biochimica et Biophysica Acta* 1587: 183-193.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8:** 967-974.

Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, **17,** 100-107.

Hannon, G.J. 2002. RNA interference. *Nature* 418: 244-251.

Hastings, M.L. and Krainer, A.R. 2001. Pre-mRNA splicing in the new millennium. Curr.Opin. Cell Biol., 13: 302-309.

Hide, W. 2003. eVOC: A Controlled Vocabulary for Unifying Gene Expression Data. *Genome Res.* 13:1222.1230

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141-2144.

Kan, Z., Oruchka, E.C., Gish, W.R., States,D.J. 2001. Gene Structure Prediction and Alternative Splicign Analysis Using Genomically Aligned ESTs. Genome Res., 11: 889-900.

Kan, Z., States, D. and Gish, W. 2002. Selecting for functional alternative splices in ESTs. *Genome Res*, **12,** 1837-1845.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Krogh, A., Larsson, B., Heijne, G. and Sonnhammer,E.L.L. 2001. Predicting Transmembrane Protein Topology with a Hidden Markov Model : Application to Complete Genome. J. Mol. Biol. 304: 567-580

Lash,E.I. 2000. SAGEmap: A Public Gene Expression Resource. *Genome Res.* 10:1051-1060

Lee, C., Atanelov, L., Modrek, B. and Xing, Y. 2003 ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*, **31,** 101-105.

Lehner, B., Williams, G., Campbell, R.D. and Sanderson, C.M. 2002. Antisense transcripts in the human genome. *Trends Genet* 18: 63-65.

Maniatis, T. and Tasic, B. 2002 Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature,* **418,** 236-243.

Mello, C.C. and Conte, D.J. 2004. Revealing the world of RNA interference. Nature 431: 338-342.

Modrek, B., Resch, A., Grasso, C. and Lee, C. 2001 Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res,* **29,** 2850-2859. Modrek

Mulder, N.L., et al. 2003. The InterPro Database, 2003 brings increased coverage and new features. Nucl. Acids. Res. 31: 315-318.

Pospisil, H., Herrmann, A., Bortfeldt, R.H. and Reich, J.G. 2004. EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res,* **32 Database issue,** D70-74.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* **29:** 159-164.

Riggins, G.J. 2002. An anatomy of normal and malignant gene expression. *PNAS* 99(17):11287-11292

Romualdi, C. 2003. IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. *Physiol Genomics* 12: 159-162

Rosok,Ø and Sioud, M. 2004. Systematic identification of sense-antisense transcripts in mammalian cells. *Nature Biotechnol* 22: 104-108.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E. et al. 1996. A gene map of the human genome. *Science* **274:** 540-546.

Shendure, J. and Church, G.M. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol* 3: research0044.1-research0044.14

Shibata, S. and Lee, J.T. 2004. Tsix transcription- versus RNA-based mechanisms in Xist repression and epigenetic choice. *Current Biol* 14: 1747-175

Su, A.I. et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS* 101(16):6062-6067

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V. and Muilu, J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res,* **32 Database issue,** D64-69.

The Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. Genome Res. 11: 1425-1433.

Tufarelli C., Stanley, J.A.S., Garrick, D., Sharpe, J.A., Ayyub H., Wood, W.G. and Higgs, D.R., 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genet* 34(2): 157-165.

Vu, T.H., Chuyen, N.V., Li, T. and Hoffman, A.R. 2003. Loss of imprinting of IGF2 sense and antisense transcripts in Wilms' Tumor *Cancer Res* 63: 1900-1905.

Xu, Q. and Lee, C. 2003. Discovery of novel splicing and finctional analysis of cancer-specific alternative splicing in human expressed sequences. Nucl. Acids Res., 31: 5635-5643.

Yelin, R., Dahary, D., Sorek, R., Levanon, E., Goldstein, O., Shoshan, A., Diver, A., Biton, S., Tamir, Y., Khosravi,R. et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol* 21: 379-386.

Zhining, W. and et al. 2003. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res. 63: 655-657.

[별지 제16호 서식]

특정연구개발사업 연구결과 활용계획서

사업명	중사업명	바이오연구개발사업		
	세부사업명	생물정보학연구개발사업		
과제명		EST 및 유전자 발현 프로필 대량 생산·분석 시스템 개발(II) EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석		
연구기관		이화여자대학교	연구책임자	이상혁
총연구기간		2002년. 12월. 01일. ~ 2004년. 09월. 30일. (22 개월)		
총 연구비 (단위 : 천원)		정부출연금	민간부담금	합계
		190,000		190,000
기술분야		419 (생물정보학)		
참여기업				
공동연구기관				
위탁연구기관				
연구결과활용 (해당항목에(∨) 표시)	1. 기업화 ( )	2. 기술이전( )	3. 후속연구추진(∨)	4.타사업에 활용(∨)
	5. 선행 및 기초연구(∨)	6. 기타목적활용(교육,연구)(∨)	7. 활용중단(미활용)( )	8. 기타( )

특정연구개발사업 처리규정 제 31조(연구개발결과의 보고) 제 2항에 의거 연구결과 활용계획서를 제출합니다.

첨부 : 1. 연구결과 활용계획서 1부.

2. 기술요약서 1부

2005 년 1월 17일

연구책임자 : 이 상 혁 (인)

연구기관장 : 신 인 령 (직인)

과학기술부장관 귀하

[첨부1]

연구결과 활용계획서

1. 연구목표 및 내용

가. 연구개발의 최종 목표

새로운 방식의 EST clustering 방법을 개발하여 인간 게놈지도에 포함된 미지의 유전자를 예측하고, 그 결과를 유전자의 발현과 기능의 측면에서 종합적으로 분석하는 genome annotation 시스템의 구축

나. 연구의 내용

- mRNA/EST 서열을 이용한 genome-based clustering 방법의 개발
- Alternative splicing에 대한 유전자 모델링
- 유전자의 기능과 발현 분석 시스템의 개발 및 유용 유전자의 발굴
- 유전체 정보의 종합적인 제공을 위한 웹사이트와 프로그램의 개발

2. 연구수행결과 현황(연구종료시점까지)

가. 특허(실용신안) 등 자료목록

발명명칭	특허공고번호 출원(등록)번호	공고일자 출원(등록)일자	발명자 (출원인)	출원국	비고

나. 프로그램 등록목록

프로그램 명칭	등록번호	등록일자	개발자	비고

다. 노하우 내역

(1) Alternative splicing을 포함한 유전자 예측 방법
(2) 유전자의 기능 분석 방법
(3) SAGE와 EST를 이용한 유전자 발현 분석 방법
(4) 게놈 정보 제공 웹 포탈 사이트 개발 및 운영 노하우
(5) 게놈 정보의 효과적인 데이터 마이닝 방법
(6) 원하는 기능과 발현 특성을 지닌 유용 유전자 발굴 방법

라. 발생품 및 시작품 내역

바. 논문게재 및 발표 실적

○논문게재 실적(필요시 별지사용)

국내/외 구분	학술지명	논문제목	게재연월일	권(호)	발행기관	국명	SCI 게재 여부	IF (2003년 기준)	교신저자여부
국외	Nucleic Acids Res	ASmodeler: Gene modeling of alternative splicing from genomic alignment of mRNA, EST, and protein sequence	2004 Jul	32(WS)	Oxford Univ. Press	영국	O	6.575	O
국외	Nucleic Acids Res	ECgene: Genome annotation for alternative splicing	2005 Jan	33(DB)	Oxford Univ. Press	영국	O	6.575	O
국내	Genomics & Informatics	ChimerDB - Database of chimeric sequences in the GenBank	2004 Jun	2(2)	한국유전체학회	한국	X		O
국외	Genome Research	ECgene: Genome-based EST clustering and gene modeling for alternative splicing	accepted		CSHL	미국	O	9.635	O
	계: 4건								

국내/외 구분	학술지명	논문제목	게재연월일	권(호)	발행기관	국명	SCI 게재 여부	IF (2003년 기준)	교신저자여부
국외	Genomics	Prediction of mammalian microRNA targets - Comparative genomics approach with longer 3′ UTR database	심사중		Academic Press		O	3.488	O
		ChimerDB & ChimerSearch - database and inspection tool for chimeric sequences in the GenBank	준비중						
		ECexpress - gene expression analysis using public cDNA and SAGE libraries	준비중						
		Transcript-based Tag Assignment for SAGE Analysis	준비중						
		ECprofiler - A utility program to search for candidate genes	준비중						
		ASviewer - a viewer for functional annotation of alternatively spliced genes	심사중						
		Genomewide analysis of changes in functional domains by alternative splicing	준비중						
	계	5-8건							

※참고사항 (본 과제의 연구성과물로 현재 투고 중, 또는 준비 중인 논문)

○학술회의 발표 실적(필요시 별지사용)

학술회의명	제목	일시	장소국명
2002 ISMB Conference	Top-down EST clustering using the draft human genome map	2002.8.4	Edmonton, Canada
2003 ISMB Conference	Hierarchical calssification of cDNA libraries for gene expression analysis	2003.6.30	Brisbane, Australia
2003 ISMB Conference	ASmodeler: gene modeling of alternative splicing events from genomic alignment of mRNA and ESTs	2003.6.30	Brisbane, Australia
2004 ISMB/ECCB Conference	ECortholog: Finding homologous genes using comparative genomics approach	2004.8.1	Glasgow, UK
The 3rd Japan-Korea Bioinformatics Training Course	Gene modeling and functional annotation for alternative splicing	2004.3.18	Mishima, Japan
2002 KSBI Annual Meeting	Genome-wide detection of splice variants from mRNA and EST alignments against the draft human genome map	2002.11.15	부산, 한국
2002 KSBI Annual Meeting	Expression Ontology DB (EODB): Gene expression using EST database	2002.11.15	부산, 한국
2003 KSBI Annual Meeting	ASmodeler: Gene modeling of alternative splicing frm genomic alignment of mRNA and ESTs	2003.10.31	대전, 한국
2003 KSBI Annual Meeting	Candidate gene search systerm for ECgene and UniGene	2003.10.31	대전, 한국
2003 KSBI Annual Meeting	ECgene: Genome-based EST clustering and gene modeling for alternative splicing	2003.10.31	대전, 한국
2003 KSBI Annual Meeting	In silico detection of cancer-related genes and isoforms using EST information	2003.10.31	대전, 한국
제2회 바이오인포매틱스 포럼	Informatics strategies for finding noble genes	2002.4.30	서울, 한국
제3회 한국유전체학회 국제학술대회	Top-down EST clustering using the draft human genome map	2002.8.23	서울, 한국
제4회 한국유전체학회 국제학술대회	ECgene: Genome-based EST clustering and gene modeling for alternative splicing	2003.9.4	대전, 한국
제5회 한국유전체학회 국제학술대회	ECgene: The genome portal site for alternative splicing	2004.8.18	춘천, 한국
2003 대한생화학분자생물학회 OMICS 연수강좌	EST 분석과 Genomics database 활용	2003.8.30	서울, 한국
제61회 한국생화학분자생물학회 학술대회	Genome browsers - Navigating genomic information	2004.5.28	서울, 한국
2004 생물정보학심포지움	ECgene: Gene modeling for alternative splicing and genome annotation	2004.6.4	서울, 한국
제288회 학연산연구교류회 Statistical Issues in Bioinformatics	유전체 정보 해석과 유용 유전자의 발굴	2004.9.22	서울, 한국
계: 19건			

3. 연구성과

ECgene 유전자 정보 서비스를 한층 원활하게 하고 지원대상 생물체의 수를 늘리기 위하여 현재 국가유전체정보센터에 기술·이전을 추진 중에 있음. 현재 개발이 완료된 부분은 2005년 4월 이내에 국가유전체정보센터에서 서비스를 제공할 것임.

4. 기술이전 및 연구결과 활용계획

가. 당해연도 활용계획(6하원칙에 따라 구체적으로 작성)

- Microarray 실험을 통하여 알려진 tissue(organ)-specific 또는 cancer-specific한 유전자들이 많이 존재함. ECgene에서 개발한 EST와 SAGE를 통한 발현 분석 수단을 array에서 얻은 결과와 비교와 비교할 예정임.
- ECgene의 유전자 모델을 이용하여 human의 microRNA 타겟 유전자를 예측하는 프로그램을 개발할 예정임. MicroRNA는 3′ UTR 부분을 타겟팅하는 것으로 알려져 있는데 ECgene의 경우 overlapping EST를 이용하여 3′ UTR이 다른 유전자 모델보다 길고, alternative splicing에 의한 유전자 구조의 차이를 반영할 수 있는 이점이 있음.

나. 활용방법

(1) 기술 이전

- 현재 국가유전체정보센터와의 협력을 통하여 ECgene 데이터베이스의 내용과 서비스를 한층 강화하는 방안을 추진 중.
- 생물정보 관련 벤처 기업에서 필요한 정보 및 기술의 이전

(2) 타 연구에 활용

- 타 연구에서 축적된 노하우와 체계적으로 정리된 실험 데이터를 ECgene 모델을 적용하여 분석. 다음의 분야에 잘 적용될 수 있음.
- 약물유전체 연구에서 축적된 SNP 데이터의 분석
- 차등발현에 근거한 질병 관련 바이오 마커의 개발

다. 차년도이후 활용계획(6하원칙에 따라 구체적으로 작성)

(1) 국제협력 도모

- 현재 이스라엘의 GeneCards와는 상호 정보제공 및 링크를 통한 협력이 이루어지고 있음. 또한 RGD (rat genome database)와도 유전자 모델의 공유와 정보 교환의 협력을 추진 중임.
- 앞으로 EBI의 ASD 프로젝트와 같은 alternative splicing에 관한 국제 협력 프로젝트에의 참여를 추진.

(2) 연구협력과 공동연구를 통한 ECgene의 개선·발전을 추구

- 현재의 국가유전체정보센터와의 연구협력과 같은 공동연구 방안을 국내·외의

다른 연구팀과도 추진

(3) 기술이전 및 타 연구에 활용

5. 기대효과

향후 활용에 따른 기술적, 사회·경제적 파급효과(정량적 및 정성적으로 전문가입장에서 구체적으로 작성)

- 모든 유전자를 대상으로 다양한 조직, 질환에 대한 발현을 연구하는 것은 많은 연구비가 필요함. 현재의 방법은 microarray를 이용한 발현 프로필 조사가 대세이나, ECgene에서 개발한 EST와 SAGE 데이터를 이용하면 다양한 조직과 질환과의 관련성을 컴퓨터 계산을 통하여 미리 예측할 수 있는 이점이 있음. 앞으로 EST와 SAGE 데이터가 더욱 많이 공개될 것이기 때문에 정확도와 coverage는 점점 좋아질 것임. 공개된 microarray 데이터와의 연동을 할 수 있다면 직접적인 실험을 상당 부분 대체할 수 있을 정도의 정보를 제공할 수 있을 것으로 예상됨. 이와 같은 정보의 거의 모든 생물학, 의약학 연구진에서 필요한 것이므로 예산과 시간의 절약에서 엄청난 국가적인 파급효과를 가질 수 있을 것으로 기대됨.
- 특히 ECgene의 결과를 잘 활용하여 질병관련 바이오 마커를 개발하거나 신약 개발의 타겟을 찾는다면 엄청난 경제적인 효과가 있을 것임. 이는 ECgene의 유전자 예측, 기능, 발현의 분석이 독특한 장점이 많기 때문에 당분간은 경쟁력을 유지할 수 있을 것으로 예상함.
- 게놈 정보를 이용한 연구 분야에서 국가적인 수준의 인프라를 구축하는데 핵심적인 역할을 할 것으로 기대됨.

6. 문제점 및 건의사항(연구성과의 제고를 위한 제도·규정 및 연구관리 등의 개선점을 기재)

건의사항

- 본 과제의 개발 내용은 앞으로도 꾸준한 유지·보수·개발이 필요함. 이와 같이 관련분야에 파급효과가 크고 국제적인 경쟁력을 지니고 있는 연구 결과물이 만들어졌을 때 지속적인 지원을 받을 수 있는 방안이 필요할 것으로 생각됨.

[첨부2]

기술 요약서

■ 기술의 명칭

유전체 지도 분석에 근거한 유전자 구조, 기능, 발현 예측 기술

■ 기술을 도출한 과제현황

<table>
<tr><td>과제관리번호</td><td colspan="4"></td></tr>
<tr><td>과제명</td><td colspan="4">EST clustering을 통한 유용 유전자의 발굴과 인간 유전체의 분석</td></tr>
<tr><td>사업명</td><td colspan="4">바이오연구개발사업</td></tr>
<tr><td>세부사업명</td><td colspan="4">생물정보학연구개발사업</td></tr>
<tr><td>연구기관</td><td colspan="2">이화여자대학교</td><td>기관유형</td><td>대학</td></tr>
<tr><td>참여기관(기업)</td><td colspan="4"></td></tr>
<tr><td>총연구기간</td><td colspan="4">2002.12.01 - 2004.09.30</td></tr>
<tr><td>총연구비</td><td colspan="4">정부(190,000)천원 민간( )천원 합계(190,000)천원</td></tr>
<tr><td rowspan="3">연구책임자 1</td><td>성명</td><td>이 상 혁</td><td>주민번호</td><td></td></tr>
<tr><td>근무기관
부서</td><td>이화여자대학교
분자생명과학부</td><td>E-mail</td><td>sanghyuk@ewha.ac.kr</td></tr>
<tr><td>직위/직급</td><td>부교수</td><td>전화번호</td><td>02-3277-2888</td></tr>
<tr><td rowspan="3">연구책임자 2</td><td>성명</td><td></td><td>주민번호</td><td></td></tr>
<tr><td>근무기관
부서</td><td></td><td>E-mail</td><td></td></tr>
<tr><td>직위/직급</td><td></td><td>전화번호</td><td></td></tr>
<tr><td rowspan="4">실무연락책임자</td><td>성명</td><td>이 상 혁</td><td>소속/부서</td><td>이화여자대학교
분자생명과학부</td></tr>
<tr><td>직위/직급</td><td>부교수</td><td>E-mail</td><td>sanghyuk@ewha.ac.kr</td></tr>
<tr><td>전화번호</td><td>02-3277-2888</td><td>FAX</td><td>02-3277-2384</td></tr>
<tr><td>주소</td><td colspan="3">(120-750) 서울 서대문구 대현동 11-1 이화여자대학교 분자생명과학부</td></tr>
</table>

■ 기술의 주요내용

[기술의 개요]

- 게놈 지도의 분석을 통한 유전자 예측 기술
- 유전자 기능 및 발현 분석 기술
- 차등 발현을 통한 유용 유전자 발굴 기술
- 유전체 정보 제공 웹 포탈 사이트 개발 기술
- 유전체 정보 분석 기술

<기술적 특징>

(1) 유전자 구조 예측 기술
- Genome-based EST clustering과 assembly 과정을 통합
- Alternative splicing에 의한 유전자 구조의 변이 모델 제시
- UTR 부분이 길고 transcript의 끝에 대한 분석이 정확함.

(2) 유전자 기능 및 발현 분석 기술
- cDNA와 SAGE library의 분석을 통한 발현 예측
- 모든 종류의 organ, tissue, pathology에 적용할 수 있음.
- 전체 게놈을 대상으로 차등발현 유전자를 컴퓨터로 검색·발굴할 수 있음.
- 유전자뿐만 아니라 각 isoform에 따른 차이도 분석할 수 있음.

(3) 유전체 정보 분석 및 웹 포탈 사이트 개발 기술
- 유전체 정보의 종합적인 annotation 기술
- ECgene은 유전체 정보를 종합적으로 제공하는 웹 포탈 사이트임.

[용도 · 이용분야]

(1) 유전자 연구에 관련된 생명과학의 제반 분야
- 유전자의 구조, 기능, 발현은 현대 생명과학연구에서 필수 정보임.
- 유전자의 특정 변이를 효과적으로 검증할 수 있는 DNA chip 디자인

(2) 의약학 분야
- 개인 맞춤 치료를 지향하는 약물유전체학
- 조기 진단을 위한 바이오 마커의 개발
- 신약개발 타겟 유전자의 발굴
- 특정 유전자 변이를 타겟으로 하는 RNAi 또는 antisense 방법의 치료제 개발

■ 기술의 분류

[기술분야] (1개만 선택(∨로 표시)하여 주십시오)

☐ 정보산업 ☐ 기계설비 ☐ 소재 ☐ 정밀화학 · 공정 ☑ 생명과학
☐ 원자력 ☐ 자원 ☐ 에너지 ☐ 항공 · 우주 ☐ 해양
☐ 교통 ☐ 보건 · 의료 ☐ 환경 ☐ 기초 · 원천 ☐ 기타

[기술의 활용유형] (1개만 선택(∨로 표시)하여 주십시오)

☑ 신제품개발 ☐ 신공정개발 ☐ 기존제품개선 ☐ 기존공정개선
☐ 기 타 (　　)

[기술의 용도] (복수 선택(∨로 표시)가능합니다)

☐ 기계설비 ☐ 부품소자 ☐ 원료재료 ☑ 소프트웨어
☐ 가공처리기술 ☐ 자동화기술 ☐ 불량률 감소 등 현장애로기술
☐ 제품설계기술 ☐ 공정설계기술 ☐ 기 타 (　　)

■ 산업재산권 보유현황(기술과 관련한)

권리유형	명 칭	국가명	출원단계	일자	등록번호

* '권리유형'란에는 특허, 실용신안, 의장, 컴퓨터프로그램 등을 선택하여 기재
* '출원단계'란에는 출원, 공개, 등록 등을 선택하여 기재

■ 기술이전 조건

이전형태	□ 유상 □무상	최저기술료	천원
이전방식	□ 소유권이전 □ 전용실시권 □ 통상실시권 □ 협의결정 □ 기타( )		
이전 소요기간	년 개월	실용화예상시기	년도
기술이전시 선행요건			

* 기술이전시 선행요건 : 기술이전을 위한 사전준비사항(필수 설비 및 장비, 전문가 확보 등)을 기술

* 실용화예상시기 : 기술을 활용한 대표적인 제품이 최초로 생산이 시작되는 시기를 기재

■ 기술의 개발단계 및 수준

[기술의 완성도] (1개만 선택(∨로 표시)하여 주십시오)

	① **기초, 탐색연구단계** : 특정용도를 위해 필요한 신 지식을 얻거나 기술적 가능성을 탐색하는 단계
	② **응용연구단계** : 기술적 가능성의 실증, 잠재적 실용화 가능성의 입증 등 실험실적 확인 단계
∨	③ **개발연구단계** : Prototype의 제작, Pilot Plant Test 등을 행하는 단계
	④ **기업화 준비단계** : 기업화에 필요한 양산화 기술 및 주변 기술까지도 확보하는 단계
	⑤ **상품화 완료단계**

[기술의 수명주기] (1개만 선택(∨로 표시)하여 주십시오)

	① **기술개념 정립기** : 기술의 잠재적 가능성만 있는 단계
	② **기술실험기** : 기술개발에 성공했으나 아직 실용성, 경제성 등이 확실치 않은 단계
	③ **기술적용 시작기**: 최초의 기술개발국에서만 활용되고 있는 단계
	④ **기술적용 성장기**: 기술개발국 및 일부 선진국에서 활용되고 있는단계
∨	⑤ **기술적용 성숙기**: 선진국사이에서 활발한 기술이전이 일어나며, 기술의 표준화가 되어가는 단계
	⑥ **기술적용 쇠퇴기**: 선진국에서 개도국으로 기술이전이 활발하게 일어나고, 선진국에서는 기술의 가치가 저하되나, 개도국에서는 아직 시장의 가치가 높은 기술

[기술발전 과정상의 기술수준] (1개만 선택(∨로 표시)하여 주십시오)

	① **외국기술의 모방단계** : 이미 외국에서 개발된 기술의 복제, reverse Eng.
	② **외국기술의 소화 · 흡수단계** : 국내시장구조나 특성에 적합하게 적응시킴
	③ **외국기술의 개선 · 개량단계** : 성능이나 기능을 개선시킴
∨	④ **신기술의 혁신 · 발명단계** : 국내 최초로 개발

■ 본 기술과 관련하여 추가로 확보된 기술

기술명	
개발단계	□ 연구개발 계획 □ 연구개발 중 □ 연구개발 완료
기술개요	

[기술을 도출한 과제현황]

과제명				
사업명				
세부사업명				
연구기관			기관유형	
참여기관(기업)				
총연구기간				
총연구비	합계 : ( )백만원 - 정부 : ( )백만원 민간 : ( )백만원			
연구책임자	소속		성명	
	전화번호		E-mail	
연구개발 주요내용				

주 의

1. 이 보고서는 과학기술부에서 시행한 특정연구개발사업의 연구보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 과학기술부에서 시행한 특정연구개발사업의 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 아니됩니다.