ABSTRACT

ANALYSIS OF AUF1 TARGETED mRNA SEQUENCES

by
Jiebo Lu

AUF1, an A+U rich element (ARE) binding protein, plays an important role in mRNA decay. To identify the mRNAs that interact with AUF1, mRNA derived from a human cardiac cDNA expression library was purified by AUF1 affinity chromatography and cloned following RT-PCR. 261 sequences were obtained. The sequences were searched against two protein databases and four nucleic acid databases, and the sequence information and database search results were input into a local Microsoft Access database, BLAST-AUF1, by Java applets for parsing document. Analysis of protein information by querying BLAST-AUF1 identified 194 function-known proteins, which were classified into 14 categories. Another 20 clones represented uncharacterized genes encoding hypothetical proteins, which were submitted to a website for analysis of domain and function. The other 47 clones that failed to yield any protein information were located in some chromosomes by aligning their sequences with human genome sequence. The BLAST-AUF1 database provides a data platform for further study of the expression profile of mRNAs that may be involved in AUF1-dependent functional pathways.

ANALYSIS OF AUF1 TARGETED mRNA SEQUENCES

by
Jiebo Lu

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Federated Department of Biology

January 2004

APPROVAL PAGE

ANALYSIS OF AUF1 TARGETED mRNA SEQUENCES

Jiebo Lu

Dr. Gary Brewer, Co-Thesis Advisor — Date
Associate Professor of Molecular Genetics,
Microbiology and Immunology,
University of Medicine & Dentistry of New Jersey

Dr. Michael Recce, Co-Thesis Advisor — Date
Associated Professor of Information Systems, NJIT

Dr. Tamara Gund, Committee Member — Date
Professor of Chemistry and Environmental Science, NJIT

BIOGRAPHICAL SKETCH

Author: Jiebo Lu

Degree: Master of Science

Date: January 2004

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2004
- Master of Science in Medicine,
  Chinese Academy of Preventive Medicine, Beijing, P. R. China, 1996
- Bachelor of Science in Medicine,
  Anhui Medical University, Anhui, P. R. China

Major: Computational Biology

Publications:

Gerald M. Wilson, Jiebo Lu, Kristina Sutphen, Yvelisse Suatez, Smrita Sinha, Brandy Brewer, Eneida C. Villanueva-Feliciano, Riza M. Ysla, Sandy Charles and Gary Brewer, "Phosphorylation of p40AUF1 regulates binding to Arich mRNA-destabilizing elements and protein-induced changes in ribonucleoprotein structure," *Journal of Biological Chemistry,* 278 (35:33039-33048, 2003.

Gerald M. Wilson, Jiebo Lu, Kristina Sutphen, Yue Sun, Yung Huynh, and Gary Brewer, "Regulation of ARE-directed mRNA turnover involving reversible phosphorylation of AUF," *Journal of Biological Chemistry*, 278 (35): 33029-33038.

Jiebo Lu, Chi-Tang Ho, Ghai,Geetha, and Yu K. Chen, "Resveratrol analog 3,4,4',5-tetrahydroxystilbene, differentially induces pro-apoptotic p53/BAX gene expression and inhibits the growth of transformed cells but not their normal counterparts, " *Carcinogenesis*, 22(2):321-328.

Jiebo Lu, Chi-Tang Ho, Geetha Ghai, and Yu K. Chen, "Differential effects of theaflavin monogallates on cell growth, apoptosis, and COX-2 gene expression in cancerous cells versus normal cells," *Cancer Research*, 60:6465-6471, 2001.

Jiebo Lu, Ping Z. Chen, Ping Y. Yan, Knapp S., Harvey Schuger, and Yu K. Chen, "Aminohexanoic hydroxamate is a potent inducer of the differentiation to neuroblastoma cells, " *Cancer Letter*, 160(1):59-66, 2000.

Yu K. Chen, Jiebo Lu, and Alice Y.-C. Liu, "The activation of trans-acting factors in response to hyper- and hypo-osmotic stress in mammalian cells," *Environmental Stressors and Gene Responses (eds. Storey, K.B. and Storey J.)*, Elsevier Science Press, 141-155, 2000.

Jiebo Lu, Hyeon J. Park, Alice Yee-Chang Liu, and Yu K. Chen, " Activation of heat shock factor 1 by hyper-osmotic or hypo-osmotic stress is drastically attenuated in normal human fibroblasts during senescence, " *Journal of Cell Physiology*, 184(2):183-190, 2000.

Jiebo Lu, Jinxiang Huang, and Maobo Ding, "Toxic effects of Carbon disulfide on axonal transport and energy metabolism in peripheral nerves in rats," *Chinese Journal of Industrial Medicine*, 10(4): 198-201, 1997.

Jiebo Lu, Jingxiang Huang, and Shoulin Zhang, " Effects of carbon disulfide on the intracellular free calcium, protein kinase C and cytoskeleton in nerve cells of rats, " *Occupational Health and Emergency Rescue*,14(4):4-6, 1996.

ACKNOWLEDGMENT

I would like to express my deepest appreciation to Dr. Gary Brewer and Dr. Michael Recce, who not only served as my research supervisors, providing valuable and countless resources, insight, and intuition, but also constantly gave me support and encouragement. Special thanks are given to Dr. Tamara Gund for actively participating in my committee. I also wish to thank Mr. Wei Hang, Comrise Technology Inc., for his generous help in computer science.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In eukaryotes, gene expression can be regulated at multiple points, such as transcription, pre-mRNA processing, steady-state level of cytoplasmic mRNA, translation, etc. However, the past two decades of research have demonstrated that mRNA decay is a major control point in gene expression, and regulating mRNA stability is a major posttranscriptional pathway for determining abundance of mRNA in the cytoplasm [1].

A+U-rich elements (AREs) are *cis*-acting determinants of rapid cytoplasmic mRNA turnover located in the 3'-untranslation regions (UTR) of many labile mRNAs, including some encoding oncoproteins, inflammatory mediators, cytokines, and G-protein-coupled receptors [2]. AUF1, an ARE binding factor, has been well-characterized in its ability to direct mRNA decay in *trans*. It has four different isoforms, $p37^{AUF1}$, $p40^{AUF1}$, $p42^{AUF1}$, and $p45^{AUF1}$, based on its apparent molecular weight. In the cytoplasm, $p37^{AUF1}$ and $p40^{AUF1}$ are the dominant isoforms [3].

Many AUF1 targeted mRNAs have been identified. Among those, turnover of β-adrenergic receptor (β-AR) mRNA in human myocardium, which also contains an ARE within its 3'-UTR, appears accelerated accompanied by an increase in intracellular AUF1 concentrations during congestive heart failure (CHF)[4]. Furthermore, recombinant $p37^{AUF1}$ also binds the β-AR 3'-UTR in vitro. β-AR plays a key role in control of heart contractibility, vasodilation, rhythmic beat, and disruption of β-AR activity can lead to CHF [5], [6]. Thus, AUF1 is clearly linked to the β-AR signal transudation pathway, and is possibly involved in regulating the stability of

mRNAs that contribute to CHF. To define the subset of human myocardium mRNAs that interact with AUF1, our laboratory conducted an assay to clone AUF1 targets by screening a human cardiac cDNA library. Totally, 261 cDNA clones were obtained and sequenced.

The objective of this thesis is to analyze these sequences with a series of computational methods. The sequences were searched against selected protein and nucleic acid databases. The data from search results were stored in a designed local database, and then were retrieved for further biological interpretation. The thesis work was focused on creation of local database, application of JAVA applets for data input, and mining of the protein information that the AUF1 targeted mRNAs represent.

CHAPTER 2

METHODS

2.1 Sequence Formatting

All the sequences were re-formatted in Fasta style. Fasta format is a simple and frequently used format for storing, transferring, and viewing of multiple DNA or protein sequences. Basically, the information for each sequence has two parts: part one is a brief description beginning with a ">" sign, and part two is the DNA sequence. For example:

>s1 Temp_71_188538__068

GATTGGGCCGACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATT
CGATTGCTTCAGATCAAGGTGACCCTAGATTGTTCTCTTTCTATATATTCC
TTTGACTTTTTCATCAGATTCTGAAGACTCATTACATAGGGATCTGGGATG
ACTTCTGCCAAAAGTGGTAGATCCTGTTGTTCATTTACTGAGAAGGGACC
ATCAGAAAATAAGAGTTTCTTGTGGGGTTGCTGAAGAGGTCTTTGGAAAA
GCTTCATTTTCTACATGACTAATATTGGAACTATCACATTGCTTCGGAGAA
TTGAATCCTTCTGAATCTCTAGCTAAGTCTATTCCATCAGTTTTACATTGGT
CCTCATTATCCAATGGCAAAATCCCAGCTATCTTATCAAGCTTTGCTGCAG
TAGAGTGTTCCGTATGGCTTGGAAAGCTATTTGGAAATGTAGCAGGAACA
TTCAAGTTTCTGACTTCTGAATTAGAGTAAACTGCTTCCATCTCCCACTGA
TCAAAATCACTGGCATTAGGTGCTTTTCTAACCTGAACATTGTCAAGAATC
TCCTGGACACGAGTCAGTAAAGGCTTTCTTCC.

2.2 Database Search

The sequences were saved as txt. file and submitted to the United Kingdom Human Genome Mapping Project Resource Center (HGMP-RC) as a query. The database

search was performed by the bioinformatics application programs: BLAST. The search options included:

1. Human section.
2. Mask biased regions: This option filtered the query sequence for regions of low compositional complexity. Low complexity regions commonly give spuriously high scores that reflect compositional bias rather than significant position-by-position alignment. Filtering can eliminate these potentially confounding matches (e.g., hits against proline-rich regions or poly-A tails) from the BLAST reports, leaving regions whose BLAST statistics reflect the specificity of their pairwise alignment. Queries searched with the BLASTN program was filtered with the program DUST, which works on nucleotide sequences, and with the program SEG, which works on protein sequences. Low complexity sequence found by a filter program is substituted using the letter "N" in nucleotide sequence (e.g., "NNNNNNNNNNNN") and the letter "X" in protein sequences (e.g., "XXXXXXXXX").
3. Mask repeats: First, a BLAST was carried out to search against a database of human repeat sequences, and any matches were masked out of the query sequences using the program XBLAST. The resulting masked sequence was then used to search against the chosen databases.
4. Matrix: The BLOSUM62 matrix was used because it is a good general purpose scoring matrix.
5. Expect value: 0, as default.

For protein database searches, the cDNA sequence was translated in all 6 reading frames. For nucleic acid database searches, the cDNA sequence was taken as either plus or minus strand. The protein databases were selected as the following:

1. SPTR: a comprehensive protein sequence database that combines the high quality of annotation in SWISS-PROT with the completeness of the weekly updated translation of protein coding sequences from the EMBL nucleotide database. It is composed of three parts:

   A). SWISS-PROT-a manually curated protein sequence database. The SWISS-PROT component of SPTR contains the latest SWISS-PROT release as well as the new or updated entries in SWISSNEW.

   B) TrEMBL-a computer-annotated protein sequence database supplementing SWISS-PROT. It contains translations of all protein coding sequences in the EMBL nucleotide sequence database which are not yet in SWISS-PROT.

   C).TrEMBL-NEW - the weekly update to SP-TrEMBL which contains the protein-coding sequences from EMBLNEW.

2. International Protein Index (IPI): provides cross references to the main databases that describe the human, mouse and rat proteomes: SWISS-PROT, TrEMBL, RefSeq and Ensembl, and maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases. IPI is updated monthly in accordance with the latest data released by the primary data sources.

For nucleic acid database searches, four databases were chosen:

1. Ensembl: provides complete and consistent annotation across the human genome as well as other genomes. Ensembl_cDNAs is one section of the Ensemble database

that contains both known and novel cDNA sequence. These are updated by the Sanger/EBI Ensembl team about every two months.

2. Expressed Sequence Tags (ESTs): The original EMBL database is split up in several taxonomic divisions, each having one or more separate datafiles. HGMP further divided the EMBL_EST database into three sections: human, mouse and others. Thus, Human ESTs becomes a subsection of EMBL's EST sections.
3. UniGene: an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. In addition to sequences of well-characterized genes, hundreds of thousands of novel expressed sequence tag sequences have been included.
4. TIGR Gene Index Project creates organism specific databases aiming to provide an analysis of publicly available EST and gene sequence data to identify transcripts. Human Gene Index (TIGRHGI) is one of sections.

All data from the above database search was downloaded into a local desktop computer as a compacted zip.file. The file was then unzipped with WINZIP® (evaluation version) and the data files were extracted into the corresponding individual folders.

2.3 Setting of Local Database

The local database, called BLAST-AUF1, was built within Microsoft Access (2000 version) to store the data from protein and nucleic acid database searches. Microsoft Access is a relational database used on desktop computers, and it is convenient to manage

small-scale data efficiently. In the BLAST-AUF1database, four objects were created as table, form, query, and report.

2.3.1 Table

Table is the central point of a database, because all data are stored in tables, and functionality of a database relies on how the tables are designed. In this thesis, six tables were designed to store BLAST data from two protein databases and four nucleic acid databases, and another one was designed to store the cDNA sequence information (**Table 2.1-2.7**). The field in the each table represents the main parameter in the significant alignment with the highest score.

Table 2.1 Design of cDNA_Sequence_Information: table

Field Name	Data Type	Description
RecordNumber	number	number of sequence record
SequenceID (primary key)	number	original number from Sequencing Center
cDNASequence	hyperlink	link to local file for detail cDNA sequence
SequenceLength	number	length of cDNA sequence

Table 2.2 Design of BLAST_in_Human_SPTR: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
SptrIdentifier	text	protein identifier in SPTR
SptrAccessID	text	accession identifier in SPTR
SptrProteinName	text	protein name in SPTR
SptrProteinLength	number	length of subject protein
SptrScore	number	score of alignment
SptrEvalue	text	expect value of alignment
SptrQuerySeqFrame	text	reading frame as query sequence
SptrDetailResult	hyperlink	link to local file for detail blast result

Table 2.3 Design of BLAST_in_IPI: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
IPIAccessID	text	accession identifier in IPI
IPIProteinLocus	text	protein locus in IPI
IPIProteinName	text	protein name in IPI
IPIDetailBlastResult	hyperlink	link to local file for detail blast result

Table 2.4 Design of BLAST_in_Human_EST: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
HEAccessID	text	accession identifier in Human_EST section
HELocus	text	locus in human_EST section
HELength	number	length of subject sequence
HEDetailResult	hyperlink	link to to local file for detail blast result

Table 2.5 Design of BLAST_in_Ensembl_cDNA: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
EncIdentifier	text	identifier in Ensembl
EncDatabaseType	text	core or other database
EncGeneID	text	gene identifier in Ensembl
EncCloneID	text	clone identifier in Ensembl
EncContigID	text	contig identifier in Ensembl
EncChromosome	text	chromosome name
EncBasepair	number	number of contig basepair
EncStatus	text	unknown, known or novel sequence
EncProteinLength	number	length of subject protein
EncDetailBlastResult	hyperlink	link to to local file for detail blast result

Table 2.6 Design of BLAST_in_HSUNIGENE: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
HsugAccessID	text	accession identifier in HSUNIGENE
HsugLocus	text	locus in UNIGEN
HsugLength	number	length of subject sequence
HsugDetailBlastResult	hyperlink	link to to local file for detail blast result

Table 2.7 Design of BLAST_in_TIGRHGI: table

Field Name	Data Type	Description
SequenceID	number	original number from Sequencing Center
TigrhgiAccessID	text	accession identifier in TIGRHGI
TigrhgiLength	number	length of subject sequence
TigrhgiScore	number	score of alignment
TigrhgiEvalue	text	expect value of alignment
TigrhgiDetailResult	hyperlink	link to to local file for detail blast result

2.3.2 Form

Form is the central point of a database. It is used to view, enter, manipulate, and search data. In this thesis, seven forms (**Figure 2.1-2.7**) were created corresponding to the seven tables, and all layouts were designed to be user-friendly.

Figure 2.1 Form of information on cDNA sequence

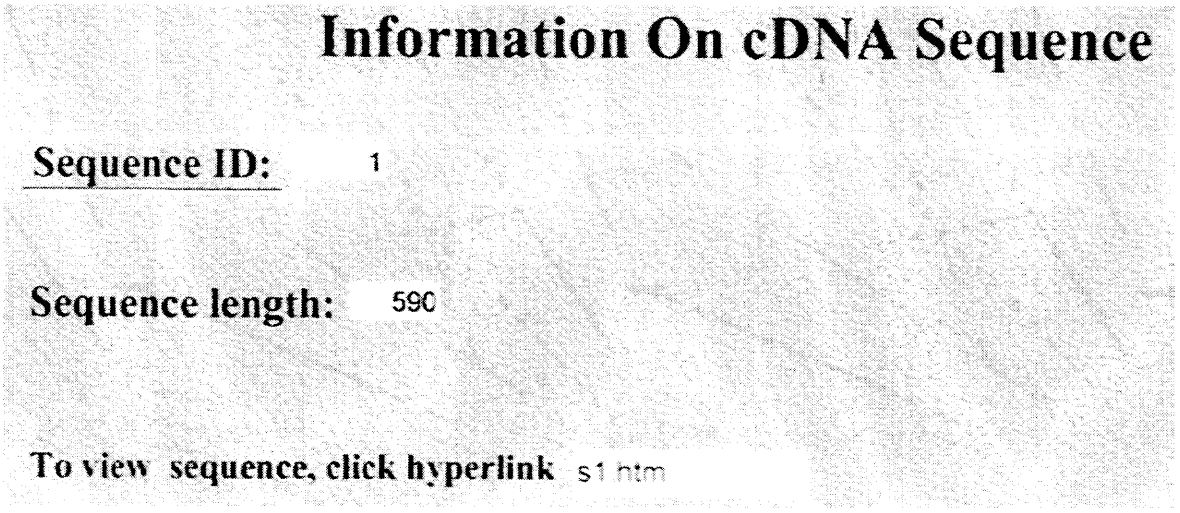


Figure 2.2 Form of BLAST in human section of SwissProt+TrEMBL

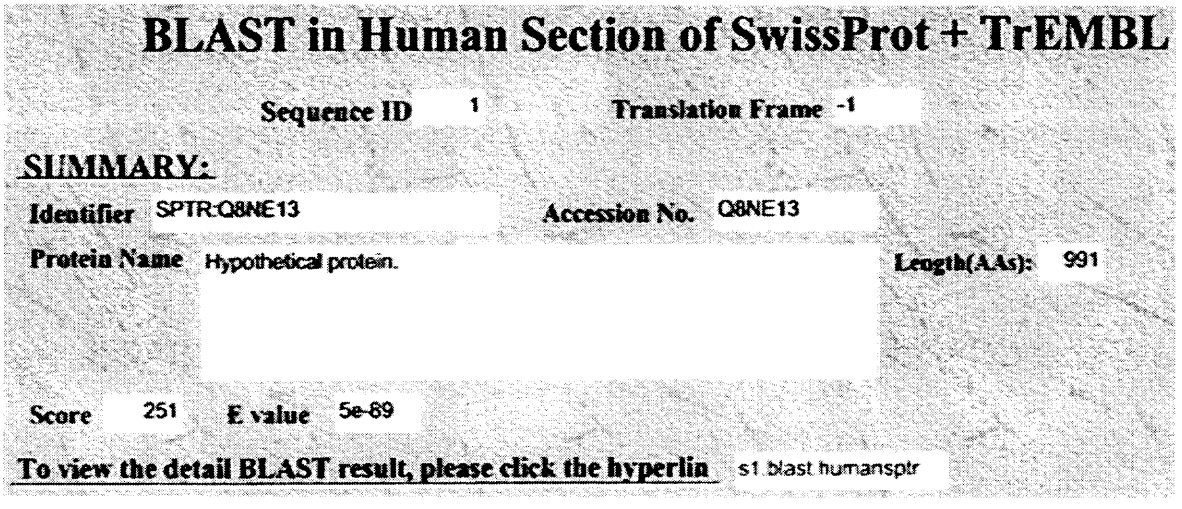


Figure 2.3 Form of BLAST in International Protein Indices

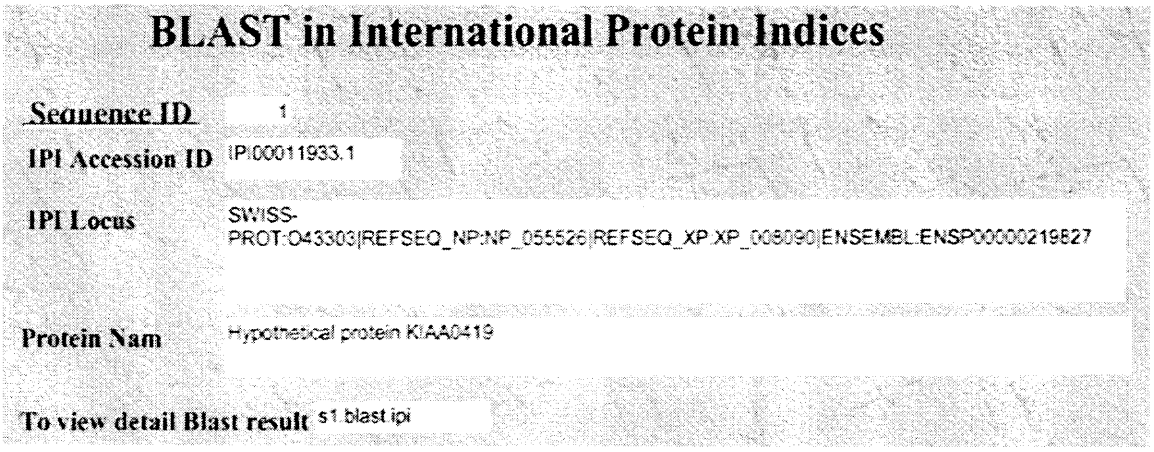


Figure 2.4 Form of BLAST in ENSEMBL cDNA

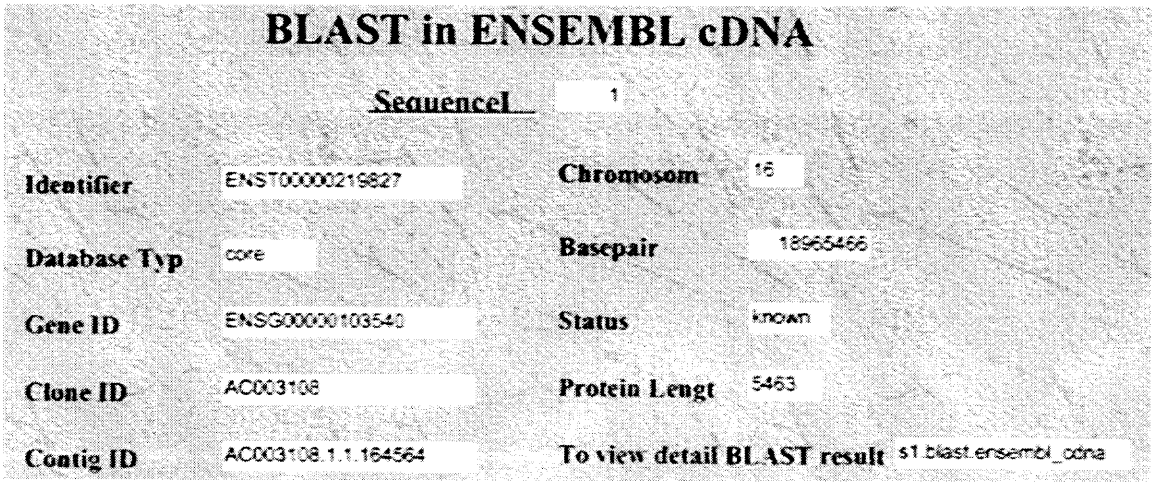


Figure 2.5 Form of BLAST in human EST

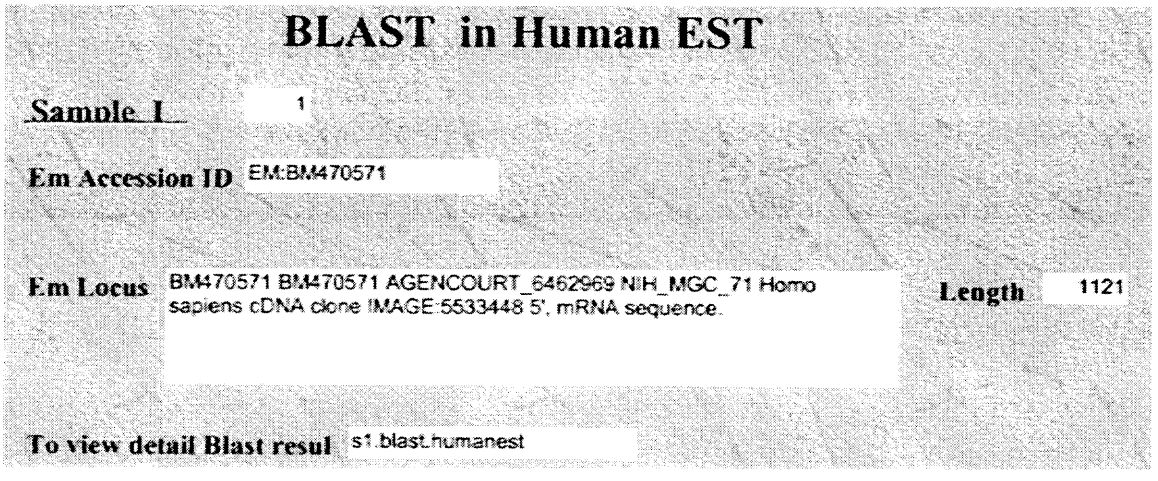


Figure 2.6 Form of BLSAT in HSUNIGENE

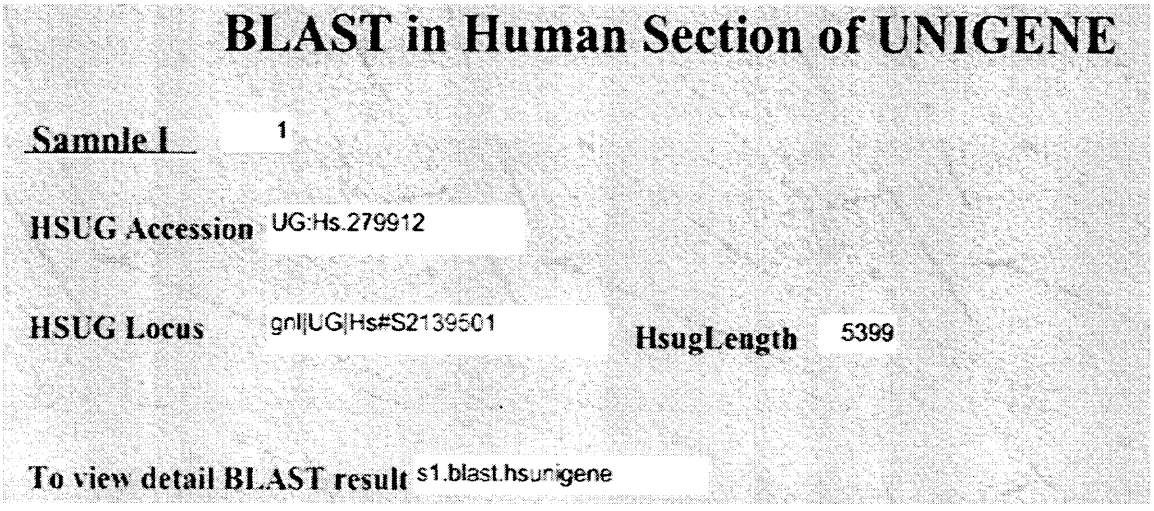
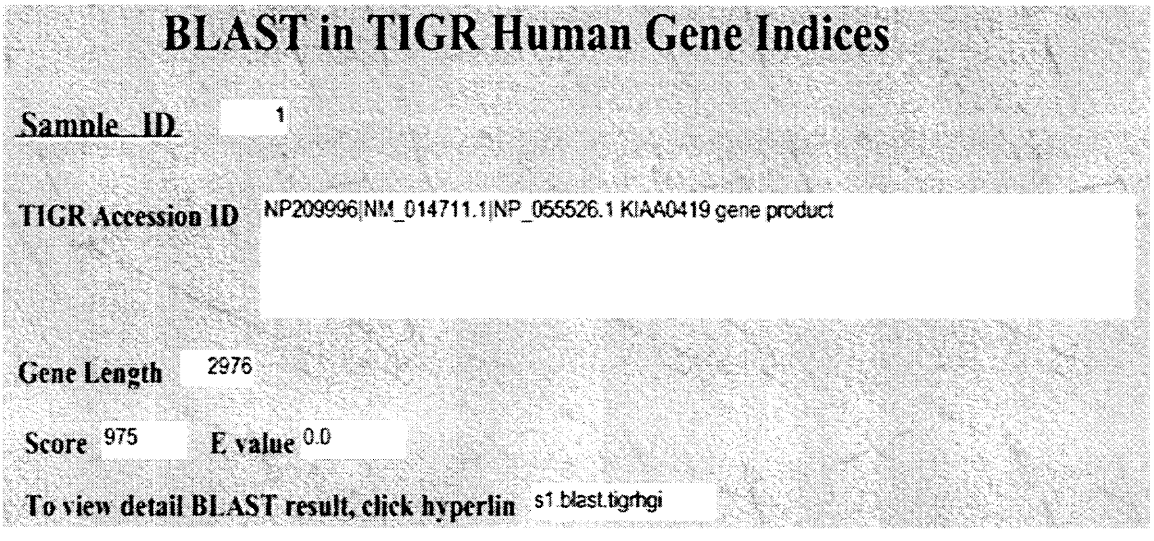


Figure 2.7 Form of BLAST in TIGRHGI



2.3.3 Query

Query was designed to retrieve data from one or more tables by using specified criteria and sorted records in some order. Query was built from Structural Query Language (SQL).

For output of BLAST in SPTR:

```
SQL>
SELECT BLAST_in_Human_SPTR.SequenceID,
       BLAST_in_Human_SPTR.SptrIdentifier,
       BLAST_in_Human_SPTR.SptrProteinName
FROM   BLAST_in_Human_SPTR
ORDER BY BLAST_in_Human_SPTR.SptrProteinName;
```

2.3.4 Report

A report provides an object used to print database records. The reports in this thesis were created to be printer-friendly.

2.4 Java Applet for Data Input

Basically, all applets are a Java document parsing tool. First, it automatically searches the data files under a specified directory in the local computer, then read the file content and filtered result based on a keyword associated with the field name. It then creates a Java Database Connectivity (JDBC) linked to the BLAST-AUF1 database and inputs data into the table. The data from each protein or nucleic acid database search for each cDNA sequence was entered as one record that had the highest score in the BLAST result.

JAVA applets were run on the Windows platform of Java™ 2 SDK, standard edition, version 1.4.1. The Java codes for the data inputs are listed in APPENDIX.

CHAPTER 3

RESULTS

3.1 Data Input

The number of each record was tracked by the field, RecordNumber, in the table of cDNA_ Sequence_ Information. The partial records, from 257 to 261(last five), in each table are shown below as examples (**Table 3.1-3.7**):

Table 3.1 The partial records in cDNA_Sequence_Information: table

RecordNumber	SequenceID	cDNASequence	SequenceLength
257	283	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ cDNA_SEQUENCE\s283	368
258	284	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ cDNA_SEQUENCE\s284	361
259	285	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ cDNA_SEQUENCE\s285	596
260	286	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ cDNA_SEQUENCE\s286	596
261	287	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ cDNA_SEQUENCE\s287	588

Table 3.2 The partial records in BLAST_in_Human_SPTR: table

SequenceID	SptrIdentifier	SptrAccessID	SptrProteinName	SptrProteinLength
283	SPTR:BAC22 594	BAC22594	Peptidylglycine alpha-amidatingmonooxygenase .	973
284	SPTR:AAH15 888	AAH15888	ALDOA protein.	364
285	SPTR:Q9UFK 3	Q9UFK3	Hypothetical protein.	341
286	SPTR:CAH3_ HUMAN	P07451	Carbonic anhydrase III (EC 4.2.1.1)(Carbonate dehydratase III) (CA- III).	259
287	SPTR:BAC22 594	BAC22594	Peptidylglycine alpha-amidatingmonooxygenase .	973

Table 3.2 (Continued)

SptrScore	SptrEvalue	SptrQuerySeqFrame	SptrDetailResult
236	3e-63	-1	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ HUMAN_SPTR\s283
196	3e-51	+1	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_SPTR\s284
359	1e-99	-2	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_SPTR\s285
272	2e-73	-2	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_SPTR\s286
81	7e-16	+3	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_SPTR\s287

Table 3.3 The partial records in BLAST_in_IPI: table

SequenceID	IPIAccessID	IPIProteinLocus	IPIProteinName	IPIDetailBlastResult
283	IPI0001571 4.1	SWISS-PROT:P19021|REFSEQ_ NP:NP_000910|TREMBL :Q13749;O95080|REFSE Q_XP:XP_031121;XP_03 1120|.....	alpha-amidating monooxygenase precursor	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\IPI\s283
284	IPI0002182 0.1	SWISS-PROT:P04075|REFSEQ_ NP:NP_000025|TREMBL :Q9BWD9|REFSEQ_XP: XP_008117;XP_043948; XP_054797;XP_043944; XP_043947;XP_043946; XP_043945	aldolase A	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\IPI\s284
285	IPI0000392 5.1	SWISS-PROT:P11177	Pyruvate dehydrogenase E1component beta subunit, mitochondrial precursor	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\IPI\s285
286	IPI0001167 2.1	SWISS-PROT:P07451|REFSEQ_ NP:NP_005172|REFSEQ _XP:XP_045079;XP_005 207|ENSEMBL:ENSP000 00220705	Carbonic anhydrase III	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\IPI\s286
287	IPI0001571 4.1	SWISS-PROT:P19021|REFSEQ_ NP:NP_000910|TREMBL :Q13749;O95080|REFSE Q_XP:XP_031121.....	alpha-amidating monooxygenase precursor	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\IPI\s287

Table 3.4 The partial records in BLAST_in_Human_EST: table

SequenceID	HEAccessID	HELocus	HELength	HEDetailResult
283	EMU:CD108547	CD108547 CD108547 AGENCOURT_1401671 7 NIH_MGC_179 Homo sapiens cDNA clone IMAGE:30364994 5', mRNA sequence.	848	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_EST\s 283
284	EM:CB113062	CB113062 CB113062 K-EST0154971 L6ChoCK0 Homo sapiens cDNA clone L6ChoCK0-8-G11 5', mRNA sequence.	615	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_EST\s 284
285	EMU:CNSLT0S U3	AL550921 human full-length cDNA 3-PRIME end of clone CS0DI065YD19 of PLACENTA COT 25-NORMALIZED of Homo sapiens (human)	1196	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_EST\s 285
286	EM:BM924263	BM924263 BM924263 AGENCOURT_6630466 NIH_MGC_116 Homo sapiens cDNA clone IMAGE:5760536 5', mRNA sequence.	1108	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_EST\s 286
287	EM:BI430552	BI430552 BI430552 0000243 Human endometrium Homo sapiens cDNA 3' similar to beta-lactamase, mRNA sequence.	659	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HUMAN_EST\s 287

Table 3.5 The partial records in BLAST_in_Ensembl_cDNA: table

SequenceID	EncIdentifier	EncDatabaseType	EncGeneID	EncCloneID	EncContigID
283	ENST000003 25306	core	ENSG000001 45730	AC010250	AC010250.7.1 .157061
284	ENST000003 20381	core	ENSG000001 49925	AC093512	AC093512.2.1 .157481
285	ENST000003 13679	core	ENSG000001 68291	AC116036	AC116036.2.1 .178106
286	ENST000002 85381	core	ENSG000001 64879	AC084734	AC084734.4.1 .181044
287	ENST000003 25306	core	ENSG000001 45730	AC010250	AC010250.7.1 .157061

Table 3.5 (Continued)

EncChromosome	EncBasepair	EncStatus	EncProteinLength	EncDetailBlastResult
5	102632612	known	3685	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ENSEMBL_cDNA\s283
16	30386996	known	2425	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ENSEMBL_cDNA\s284
3	57782043	novel	1560	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ENSEMBL_cDNA\s285
8	86090151	known	2356	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ENSEMBL_cDNA\s286
5	102632612	known	3685	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\ENSEMBL_cDNA\s287

Table 3.6 The partial records in BLAST_in_HSUNIGENE: table

SequenceID	HsugAccessID	HsugLocus	HsugLength	HsugDetailBlastResult
283	UG:Hs.83920	gnl|UG|Hs#S1727286	3960	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HS_UNIGENE\s283
284	UG:Hs.273415	gnl|UG|Hs#S1728284	1464	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HS_UNIGENE\s284
285	UG:Hs.979	gnl|UG|Hs#S1727317	1501	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HS_UNIGENE\s285
286	UG:Hs.82129	gnl|UG|Hs#S1730476	2357	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HS_UNIGENE\s286
287	UG:Hs.446484	gnl|UG|Hs#S2927215	1334	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\HS_UNIGENE\s287

Table 3.7 The partial records in BLAST_in_TIGRHGI: table

SequenceID	TigrhgiAccessID	TigrhgiLength	TigrhgiScore	TigrhgiEvalue	TigrhgiDetailResult
283	NP099080|M 37721.1|AA A36414.1	2925	722	0.0	C:\documents and settings\jeibo lu\.....
284	THC728537 fructose-bisphosphate aldolase A^^aldolase A (AA 1-364).....	2272	704	0.0	C:\documents and settings\jeibo lu\my documents\aufl target.....
285	THC847071 E-1 beta subunit of the pyruvate dehydrogena	1575	1039	0.0	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\TIGRHG I\s285
286	THC795744 carbonic anhydrase III^^carboni c anhydrase III, muscle specific [Homo sapiens]	2377	1053	0.0	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\TIGRHG I\s286
287	THC721240 PRO1708 {Homo sapiens}	1251	650	0.0	C:\documents and settings\jeibo lu\my documents\aufl target\hgmp result\TIGRHG I\s287

3.2 Output of Protein Information

The three fields, SequenceID, SptrIdentifier, SptrProteinName, were queried from BLAST_in_Human_SPTR: table and the output was sorted by protein name (**Table 3.8**). The protein outputs in the field of SptrProteinName were further grouped (**Figure 3.2.1**).

Table 3.8 The output from BLAST_in_Human_SPTR: table

SequenceID	SptrIdentifier	SptrProteinName
215	SPTR:DHCR_HUMAN	24-dehydrocholesterol reductase precursor(EC 1.3.1.-) (3-beta-hydroxysterol delta-24-reductase)(Seladin-1) (Diminuto/dwarf1 homolog).
118	SPTR:PSDC_HUMAN	26S proteasome non-ATPase regulatory subunit12 (26S proteasome regulatory subunit p55).
30	SPTR:PSDC_HUMAN	26S proteasome non-ATPase regulatory subunit12 (26S proteasome regulatory subunit p55).
231	SPTR:AAH32324	3-hydroxyisobutyrate dehydrogenase,mitochondrial (HIBADH).
99	SPTR:AAH32324	3-hydroxyisobutyrate dehydrogenase,mitochondrial (HIBADH).
127	SPTR:AAH32324	3-hydroxyisobutyrate dehydrogenase,mitochondrial (HIBADH).
222	SPTR:RL31_HUMAN	60S ribosomal protein L31.
155	SPTR:ASAH_HUMAN	Acid ceramidase precursor (EC 3.5.1.23)(Acylsphingosine deacylase) (N-acylsphingosineamidohydrolase) (AC) (Putative 32 KDA heart protein)(PHP32).
162	SPTR:Q8TAQ6	Aconitase 2, mitochondrial.
271	SPTR:Q8TAQ6	Aconitase 2, mitochondrial.
60	SPTR:ARF1_HUMAN	ADP-ribosylation factor 1.
57	SPTR:ARF1_HUMAN	ADP-ribosylation factor 1.
25	SPTR:AAP36057	ADP-ribosylation factor 1.
263	SPTR:AAP36057	ADP-ribosylation factor 1.
100	SPTR:DHAM_HUMAN	Aldehyde dehydrogenase, mitochondrialprecursor (EC 1.2.1.3) (ALDH class 2) (ALDHI) (ALDH-

		E2).
210	SPTR:AAH15888	ALDOA protein.
77	SPTR:AAH15888	ALDOA protein.
211	SPTR:AAH15888	ALDOA protein.
69	SPTR:AAH15888	ALDOA protein.
61	SPTR:AAH15888	ALDOA protein.
284	SPTR:AAH15888	ALDOA protein.
103	SPTR:AAH15888	ALDOA protein.
137	SPTR:AAH15888	ALDOA protein.
163	SPTR:AAH15888	ALDOA protein.
268	SPTR:AAH15888	ALDOA protein.
14	SPTR:AAH15888	ALDOA protein.
273	SPTR:AAH15888	ALDOA protein.
264	SPTR:AAH15888	ALDOA protein.
270	SPTR:AAH15888	ALDOA protein.
188	SPTR:AAH15888	ALDOA protein.
6	SPTR:AAH15888	ALDOA protein.
2	SPTR:ANX2_HUMAN	Annexin II (Lipocortin II) (Calpactin Iheavy chain) (Chromobindin 8) (P36) (Protein I)(Placental anticoagulant protein IV) (PAP-IV).
15	SPTR:AAH52567	ANXA2 protein.
172	SPTR:AAH52567	ANXA2 protein.
94	SPTR:AAH52567	ANXA2 protein.
51	SPTR:AATC_HUMAN	Aspartate aminotransferase, cytoplasmic (EC2.6.1.1) (Transaminase A) (Glutamate oxaloacetatetransaminase-1).
206	SPTR:AAO88883	ATP synthase F0 subunit 6.
86	SPTR:AAP35908	ATP synthase, H+ transporting, mitochondrialF0 complex, subunit F6.
152	SPTR:AAP35792	ATPase, H+ transporting, lysosomal 31kDa, V1subunit E isoform 1.
105	SPTR:ATPB_HUMAN	ATPB_HUMAN P06576 ATP synthase beta chain, mitochondrialprecursor (EC 3.6.3.14).
258	SPTR:ATPB_HUMAN	ATPB_HUMAN P06576 ATP synthase beta chain, mitochondrialprecursor (EC 3.6.3.14).
130	SPTR:Q9H4R9	BA472K17.2 (Collagen type IV alpha 1) (Fragment).
202	SPTR:Q9NTM9	BA483F11.3 (CGI-32 protein).

40	SPTR:Q8IZL7	Basigin long isoform.
119	SPTR:Q9H1D5	Beta-myosin heavy chain.
113	SPTR:Q9H1D5	Beta-myosin heavy chain.
117	SPTR:Q9H1D5	Beta-myosin heavy chain.
214	SPTR:Q9H1D5	Beta-myosin heavy chain.
91	SPTR:Q9H1D5	Beta-myosin heavy chain.
26	SPTR:Q9H1D5	Beta-myosin heavy chain.
78	SPTR:Q9H1D5	Beta-myosin heavy chain.
72	SPTR:Q9H1D5	Beta-myosin heavy chain.
136	SPTR:Q9H1D5	Beta-myosin heavy chain.
46	SPTR:Q9H1D5	Beta-myosin heavy chain.
36	SPTR:Q9H1D5	Beta-myosin heavy chain.
151	SPTR:Q9H1D5	Beta-myosin heavy chain.
207	SPTR:Q9H1D5	Beta-myosin heavy chain.
176	SPTR:Q9H1D5	Beta-myosin heavy chain.
47	SPTR:Q9H1D5	Beta-myosin heavy chain.
20	SPTR:Q9H1D5	Beta-myosin heavy chain.
12	SPTR:Q9H1D5	Beta-myosin heavy chain.
4	SPTR:Q9H1D5	Beta-myosin heavy chain.
286	SPTR:CAH3_HUMAN	Carbonic anhydrase III (EC 4.2.1.1)(Carbonate dehydratase III) (CA- III).
225	SPTR:CAH3_HUMAN	Carbonic anhydrase III (EC 4.2.1.1)(Carbonate dehydratase III) (CA- III).
244	SPTR:CAH3_HUMAN	Carbonic anhydrase III (EC 4.2.1.1)(Carbonate dehydratase III) (CA- III).
282	SPTR:Q9UM53	Cardiac myosin binding protein-C.
157	SPTR:Q8IVC5	Cas-Br-M (Murine) ecotropic retroviraltransforming sequence b.
89	SPTR:CD81_HUMAN	CD81_HUMAN P18582 CD81 antigen (26 kDa cell surface proteinTAPA-1) (Target of the antiproliferative antibody 1).
259	SPTR:CDK4_HUMAN	Cell division protein kinase 4 (EC 2.7.1.-)(Cyclin-dependent kinase 4) (PSK-J3).
139	SPTR:Q9H4N1	Clone CDABP0107 mRNA sequence.
143	SPTR:CO3_HUMAN	Complement C3 precursor [Contains: C3aanaphylatoxin].
59	SPTR:AAP35439	Creatine kinase, muscle.
55	SPTR:AAP35439	Creatine kinase, muscle.
260	SPTR:AAP35439	Creatine kinase, muscle.

140	SPTR:AAP35439	Creatine kinase, muscle.
165	SPTR:KCRS_HUMAN	Creatine kinase, sarcomeric mitochondrialprecursor (EC 2.7.3.2) (S- MtCK) (Mib-CK) (Basic-typemitochondrial creatine kinase).
35	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
68	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
3	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
90	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
120	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
106	SPTR:Q957U9	Cytochrome c oxidase subunit I (EC 1.9.3.1)(Cytochrome c oxidase polypeptide I).
171	SPTR:AAO88880	Cytochrome oxidase subunit I.
182	SPTR:AAO88880	Cytochrome oxidase subunit I.
276	SPTR:AAO88880	Cytochrome oxidase subunit I.
17	SPTR:AAO88880	Cytochrome oxidase subunit I.
169	SPTR:AAO88880	Cytochrome oxidase subunit I.
191	SPTR:AAO88880	Cytochrome oxidase subunit I.
85	SPTR:AAO88880	Cytochrome oxidase subunit I.
29	SPTR:AAO88880	Cytochrome oxidase subunit I.
107	SPTR:AAO88880	Cytochrome oxidase subunit I.
129	SPTR:AAO88880	Cytochrome oxidase subunit I.
178	SPTR:AAO88880	Cytochrome oxidase subunit I.
261	SPTR:AAO88881	Cytochrome oxidase subunit II.
11	SPTR:AAO88881	Cytochrome oxidase subunit II.
27	SPTR:DERM_HUMAN	Dermatopontin precursor (Tyrosine richacidic matrix protein) (TRAMP).
239	SPTR:DERM_HUMAN	Dermatopontin precursor (Tyrosine richacidic matrix protein) (TRAMP).
125	SPTR:DERM_HUMAN	Dermatopontin precursor (Tyrosine richacidic matrix protein) (TRAMP).
141	SPTR:DESM_HUMAN	Desmin.
48	SPTR:DESP_HUMAN	Desmoplakin (DP) (250/210 kDa paraneoplasticpemphigus antigen).

240	SPTR:DYI2_HUMAN	Dynein intermediate chain 2, cytosolic (DHIC-2) (Cytoplasmic dynein intermediate chain 2).
235	SPTR:Q9UHY7	E-1 enzyme.
76	SPTR:EPLI_HUMAN	Epithelial protein lost in neoplasm.
52	SPTR:FGL2_HUMAN	Fibroleukin precursor (Fibrinogen-likeprotein 2) (pT49).
234	SPTR:FGL2_HUMAN	Fibroleukin precursor (Fibrinogen-likeprotein 2) (pT49).
237	SPTR:FGL2_HUMAN	Fibroleukin precursor (Fibrinogen-likeprotein 2) (pT49).
21	SPTR:FGL2_HUMAN	Fibroleukin precursor (Fibrinogen-likeprotein 2) (pT49).
158	SPTR:ALFA_HUMAN	Fructose-bisphosphate aldolase A (EC4.1.2.13) (Muscle-type aldolase) (Lung cancer antigenNY-LU-1).
54	SPTR:ALFA_HUMAN	Fructose-bisphosphate aldolase A (EC4.1.2.13) (Muscle-type aldolase) (Lung cancer antigenNY-LU-1).
209	SPTR:ALFA_HUMAN	Fructose-bisphosphate aldolase A (EC4.1.2.13) (Muscle-type aldolase) (Lung cancer antigenNY-LU-1).
44	SPTR:O95303	Gamma-filamin (Filamin 2).
173	SPTR:Q9NYE5	Gamma-filamin.
104	SPTR:Q16545	GLUCOCEREBROSIDASE precursor (Glucosidase, beta,acid) (INCLUDES glucosylceramidase).
242	SPTR:HSB7_HUMAN	Heat-shock protein, beta-7 (Cardiovascularheat shock protein) (cvHsp).
274	SPTR:Q9UHG4	Heme-regulated initiation factor 2-alpha kinase.
43	SPTR:AAP35586	High-mobility group box 1.
42	SPTR:CAD61872	Human full-length cDNA clone CS0DB001YK19 ofNeuroblastoma of Homo sapiens (human).
149	SPTR:CAD62335	Human full-length cDNA clone CS0DM004YH09 ofFetal liver of Homo sapiens (human).
249	SPTR:CAD62335	Human full-length cDNA clone CS0DM004YH09 ofFetal liver of Homo sapiens (human).
230	SPTR:CAD62335	Human full-length cDNA clone CS0DM004YH09 ofFetal liver of Homo sapiens (human).
243	SPTR:CAD62335	Human full-length cDNA clone

		CS0DM004YH09 ofFetal liver of Homo sapiens (human).
10	SPTR:CAD62335	Human full-length cDNA clone CS0DM004YH09 ofFetal liver of Homo sapiens (human).
49	SPTR:CAD62335	Human full-length cDNA clone CS0DM004YH09 ofFetal liver of Homo sapiens (human).
180	SPTR:Q8IVG9	Humanin.
156	SPTR:Q96E76	Hypothetical protein (EC 2.7.1.40) (Pyruvatekinase) (PK).
201	SPTR:Q9NSL0	Hypothetical protein (Fragment).
71	SPTR:Q9BVX6	Hypothetical protein (Fragment).
248	SPTR:Q9UFG4	Hypothetical protein (Fragment).
167	SPTR:Q9H8L6	Hypothetical protein FLJ13465.
121	SPTR:Q96SW7	Hypothetical protein FLJ14590.
252	SPTR:Q96NC0	Hypothetical protein FLJ31121.
9	SPTR:Q96MD6	Hypothetical protein FLJ32515.
272	SPTR:Q8N9Q1	Hypothetical protein FLJ36757.
199	SPTR:O75160	Hypothetical protein KIAA0672.
70	SPTR:Q9C0G8	Hypothetical protein KIAA1695 (Fragment).
175	SPTR:Q8TF54	Hypothetical protein KIAA1947 (Fragment).
285	SPTR:Q9UFK3	Hypothetical protein.
110	SPTR:O75208	Hypothetical protein.
221	SPTR:Q8WUM6	Hypothetical protein.
1	SPTR:Q8NE13	Hypothetical protein.
66	SPTR:Q96B23	Hypothetical protein.
63	SPTR:Q96DI7	Hypothetical protein.
5	SPTR:O75208	Hypothetical protein.
87	SPTR:IKKB_HUMAN	Inhibitor of nuclear factor kappa B kinasebeta subunit (EC 2.7.1.-) (I-kappa-B-kinase beta)(IkBKB) (IKK-beta) (IKK-B) (I-kappa-B kinase 2) (IKK2)(Nuclear factor NF-kappa-B inhibitor kinase beta)(NFKBIKB).
255	SPTR:AAH01554	Integrin-linked kinase.
277	SPTR:LMA4_HUMAN	Laminin alpha-4 chain precursor.
236	SPTR:LDHB_HUMAN	L-lactate dehydrogenase B chain (EC1.1.1.27) (LDH-B) (LDH heart subunit) (LDH-H).
58	SPTR:AAO15302	MSTP056.
205	SPTR:MYM1_HUMAN	Myomesin 1 (190 kDa titin-associated protein)(190 kDa

		connectin- associated protein).
8	SPTR:MYM1_HUMAN	Myomesin 1 (190 kDa titin-associated protein)(190 kDa connectin- associated protein).
144	SPTR:MYM1_HUMAN	Myomesin 1 (190 kDa titin-associated protein)(190 kDa connectin- associated protein).
166	SPTR:MYM1_HUMAN	Myomesin 1 (190 kDa titin-associated protein)(190 kDa connectin- associated protein).
133	SPTR:AAO47074	Nebulin-related anchoring protein isoform C.
132	SPTR:AAO47073	Nebulin-related anchoring protein isoform S.
50	No Hit	No Hit
53	No Hit	No Hit
193	No Hit	No Hit
194	No Hit	No Hit
150	No Hit	No Hit
247	No Hit	No Hit
251	No Hit	No Hit
146	No Hit	No Hit
67	No Hit	No Hit
145	No Hit	No Hit
196	No Hit	No Hit
75	No Hit	No Hit
147	No Hit	No Hit
262	No Hit	No Hit
183	No Hit	No Hit
7	No Hit	No Hit
184	No Hit	No Hit
168	No Hit	No Hit
186	No Hit	No Hit
23	No Hit	No Hit
41	No Hit	No Hit
266	No Hit	No Hit
45	No Hit	No Hit
31	No Hit	No Hit
32	No Hit	No Hit
37	No Hit	No Hit
64	No Hit	No Hit
39	No Hit	No Hit
197	No Hit	No Hit
189	No Hit	No Hit
24	No Hit	No Hit

93	No Hit	No Hit
223	No Hit	No Hit
216	No Hit	No Hit
80	No Hit	No Hit
228	No Hit	No Hit
97	No Hit	No Hit
111	No Hit	No Hit
112	No Hit	No Hit
226	No Hit	No Hit
96	No Hit	No Hit
232	No Hit	No Hit
123	No Hit	No Hit
82	No Hit	No Hit
142	No Hit	No Hit
138	No Hit	No Hit
83	No Hit	No Hit
238	SPTR:AAH29901	Norrie disease (pseudoglioma).
28	SPTR:Q8NI35	Pals1-associated tight junction protein.
16	SPTR:O43211	Peptidylglycine alpha-amidating monooxygenase(Fragment).
224	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
81	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
56	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
257	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
265	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
65	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
92	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
241	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
275	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
135	SPTR:BAC22594	Peptidylglycine alpha-amidating monooxygenase.
109	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
283	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.

217	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
88	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
213	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
170	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
287	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.
190	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.
148	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.
134	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.
153	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
124	SPTR:BAC22594	Peptidylglycine alpha-amidatingmonooxygenase.
131	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
227	SPTR:AAP36087	Peptidylglycine alpha-amidatingmonooxygenase.
254	SPTR:AAH09601	Peroxiredoxin 3.
114	SPTR:ANKH_HUMAN	Progressive ankylosis protein homolog (ANK).
98	SPTR:Q13345	Protein tyrosine phosphatase epsilon cytoplasmicisoform (Fragment).
34	SPTR:Q9Y6A0	PTH-responsive osteosarcoma D1 protein(Fragment).
62	SPTR:AAH51265	PTMA protein.
95	SPTR:O00375	Putative p150.
250	SPTR:QORL_HUMAN	Quinone oxidoreductase-like 1 (QOH-1)(Zeta-crystallin homolog) (4P11).
278	SPTR:QORL_HUMAN	Quinone oxidoreductase-like 1 (QOH-1)(Zeta-crystallin homolog) (4P11).
229	SPTR:QORL_HUMAN	Quinone oxidoreductase-like 1 (QOH-1)(Zeta-crystallin homolog) (4P11).
279	SPTR:AAP35973	Retinoblastoma binding protein 4.
74	SPTR:SRCH_HUMAN	Sarcoplasmic reticulum histidine-richcalcium-binding protein precursor.

187	SPTR:SRCH_HUMAN	Sarcoplasmic reticulum histidine-richcalcium-binding protein precursor.
267	SPTR:AAH31212	Signal transducer and activator oftranscription 4.
79	SPTR:CBLB_HUMAN	Signal transduction protein CBL-B(SH3-binding protein CBL-B).
269	SPTR:CBLB_HUMAN	Signal transduction protein CBL-B(SH3-binding protein CBL-B).
19	SPTR:AAH12597	Similar to actin, alpha 1, skeletal muscle.
122	SPTR:AAH12597	Similar to actin, alpha 1, skeletal muscle.
246	SPTR:AAH35993	Similar to annexin A1.
73	SPTR:AAH35993	Similar to annexin A1.
159	SPTR:AAH11935	Similar to catechol-O-methyltransferase.
192	SPTR:Q8IV38	Similar to hypothetical protein MGC18754.
253	SPTR:Q8N1E3	Similar to pepsinogen 5, group I (Pepsinogen A).
203	SPTR:Q8IY65	Similar to RIKEN cDNA 0610013D04 gene.
219	SPTR:Q8IY65	Similar to RIKEN cDNA 0610013D04 gene.
198	SPTR:SPSY_HUMAN	Spermine synthase (EC 2.5.1.22) (Spermidineaminopropyltransferase) (SPMSY).
256	SPTR:TLN1_HUMAN	Talin 1.
179	SPTR:Q8WWD0	Triosephosphate isomerase 1 (EC 5.3.1.1) (TIM).
126	SPTR:TRIC_HUMAN	Troponin I, cardiac muscle.
38	SPTR:UCR2_HUMAN	Ubiquinol-cytochrome C reductase complexcore protein 2, mitochondrial precursor (EC 1.10.2.2)(Complex III subunit II).
195	SPTR:VP29_HUMAN	Vacuolar protein sorting 29 (Vesicle proteinsorting 29) (hVPS29) (MDS007) (PEP11) (DC7/DC15).
177	SPTR:AAO84481	Vascular endothelial growth factor and typeI collagen inducible protein.
208	SPTR:VGLN_HUMAN	Vigilin (High density lipoprotein-bindingprotein) (HDL-binding protein).
212	SPTR:Q8IXU7	Vinculin.

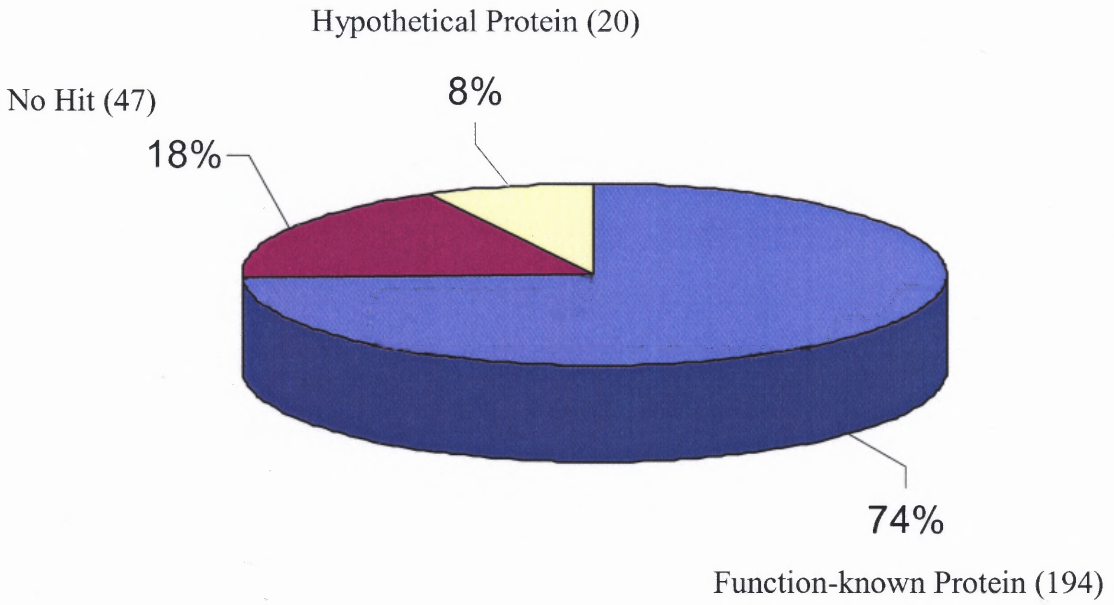


Figure 3.2.1. Group distribution of protein outputs. The number in parenthesis represents the number of cDNA sequences.

3.3 Protein Classification

Based on their functions, the "function-known" proteins were categorized (**Table 3.9**) and the distribution of the categories is shown in **Figure 3.3.1**.

Table 3.9 Function-known protein classification

Category	**Sequence ID**
Enzyme (29)*	**S(102)***
3-beta-hydroxysterol delta-24-reductase	215
3-hydroxyisobutyrate dehydrogenase, mitochondrial	99, 127, 231
Acylsphingosine deacylase	155
Aconitase 2, mitochondrial	162, 271
Aldehyde dehydrogenase-2	100
ALDOA protein	6, 14, 61 ,69, 77, 103, 137, 163,188, 210, 211, 264, 268, 270, 273, 284
Aspartate aminotransferase, cytoplasmic	51
ATP sythase beta chain	105, 258
ATP synthase F0 subunite 6	206
ATP synthase F0 subunite 6, mitochondrial	86
ATP synthase, V1 subunit E isoform 1, lysosomal	152
Carbonate dehydratase III	225, 244, 286
Cholinesterase E1	235
Creatine kinase, muscle	55, 59, 140, 260
Creatine kinase, Basic-type mitochondrial	165
Cytochrome C oxidase subunit I	3, 35, 68, 90, 106, 120
Cytochrome oxidase subunit	17, 29, 85, 107, 129, 169, 171, 178, 182, 191, 276
Cytochrome oxidase subunit II	11, 261
Frutose-bisphosphate aldolase A	54, 158, 209
Glucosidase, beta, acid	104

Table 3.9 (Continued)

L-lactate dehydrogenase B chain, LDH heart subunit	236
Peptidylglycine alpha-amidating monooxygenase	16, 56, 65, 81, 88, 92, 109, 124, 131, 134, 135, 148, 153, 170, 190, 213, 217, 224, 227, 241, 257, 265, 275, 283, 287
Peroxiredoxin 3	254
Protein tyrosine phosphatase epsilon cytoplasmic isoform	98
Pyruvate kinase, M1 isozyme	156
Quinone oxidoreductase-like 1	229, 250, 278
Spermine synthase	198
Triosephosphate isomerase 1	179
Tryptophan--tRNA ligase	10, 49, 149, 230, 243, 249
Substrate of enzyme (1)	**S(1)**
Annexin II	2
Regulator of enzymatic activity(3)	**S(4)**
ADP-ribosylation factor	25, 57, 60, 263
Ubiquinol-cytochrome C reductase complex core protein 2	38
Proteasome (1)	**S(2)**
26S proteasome regulatory subunit 55	30, 118
Ribosomal protein (2)	**S(2)**
60S ribosomal protein L31	222
Humanin	180

Table 3.9 (Continued)

Chromosomal protein (1)	**S(1)**
High-mobility group box 1	43
Membrane / Cytoskeletal protein(17)	**S(43)**
ANXA2 protein	15, 94, 172
Beta-myosin heavy chain	4, 12, 20, 26, 36, 46, 47, 72, 78, 91, 113, 117, 119, 136, 151, 176, 207, 214
Cardiac myosin binding protein-C	282
26kDa cell surface protein TAPA-1	89
Desmin	141
Desmoplakin	48
Dynein intermediate chain 2, cytosolic	240
Elastin microfibril interfacer 3	167
Fibrinogen-like protein 2	21, 52, 234, 237
Gamma-filamin	44, 173
Laminin alpha-4 chain precursor	277
Myomesin 1	8, 144, 166, 205
Nebulin-related anchoring protein isoform C	133
Nebulin-related anchoring protein isoform S	132
Talin 1	256
Troponin I, cardiac muscle	126
Vinculin	212
Matrix / Collagen(2)	**S(4)**
Collagen IV alpha 1	130
Tyrosine richacidic matrix protein	27, 125, 239

Table 3.9 (Continued)

Heat Shock Protein(1)	**S(1)**
Heat-shock protein, beta-7, cardiovascular	242
Immunoglobulin / Immune mediator(4)	**S(4)**
Basigin long isoform	40
C3a anaphylatoxin	143
NF-kappa-B inhibitor kinase beta	87
Progressive ankylosis protein homolog	114
Oncogene / Cell cycle(2)	**S(2)**
Cas-Br-M ecotropic retroviraltransforming sequence b	157
Cyclin-dependent kinase 4	259
Signal transduction(4)	**S(5)**
Heme-regulated initiation factor 2-alpha kinase	274
Integrin-linked kinase	255
Signal transducer and activator of transcription 4	267
Signal transduction protein CBL-B	79, 269
Hormone / growth factor(3)	**S(3)**
PTH-responsive osteosarcoma D1 protein	34
PTMA protein	62
Vascular endothelial growth factor and typel collagen inducible protein	17

Table 3.9 (Continued)

Others(15)	**S(19)**
CGI-32 protein	202
Epithelial protein lost in neoplasm	76
Kidney ankyrin repeat-containing protein	203, 219
Norrie disease protein	238
Putative p150 (zeta-crystallin-like-1)	95
Pals1-associated tight junction protein	28
Retinoblastoma binding protein 4	279
Sarcoplasmic reticulum histidine –rich calcium-binding protein precursor	74, 187
Similar to annexin A1	73, 246
Similar to actin, alpha 1, skeletal muscle	19, 122
Similar to catechol-O-methyltransferase	159
Similar to pepsinogen 5, group I	253
Similar to eukaryotic translation elongation factor 1 alpha 1	58
Vacuolar protein sorting 29	195
Vigilin (HDL-binding protein)	208

*The numbers in the parenthesis following category name and S represent the numbers of member in the category and cDNA sequences, respectively.

3.4 Analysis of Hypothetical Proteins

Besides the named "hypothetical protein" in BLAST_in_Human_SPTR: table, the group of "Hypothetical Protein" here also includes the unknown proteins, whose nucleotide sequence derived from mRNA sequence (Sequence ID 139), or cDNA clone (Sequence ID 42), or similarities (Sequence ID 192) (**Table 3.10**).

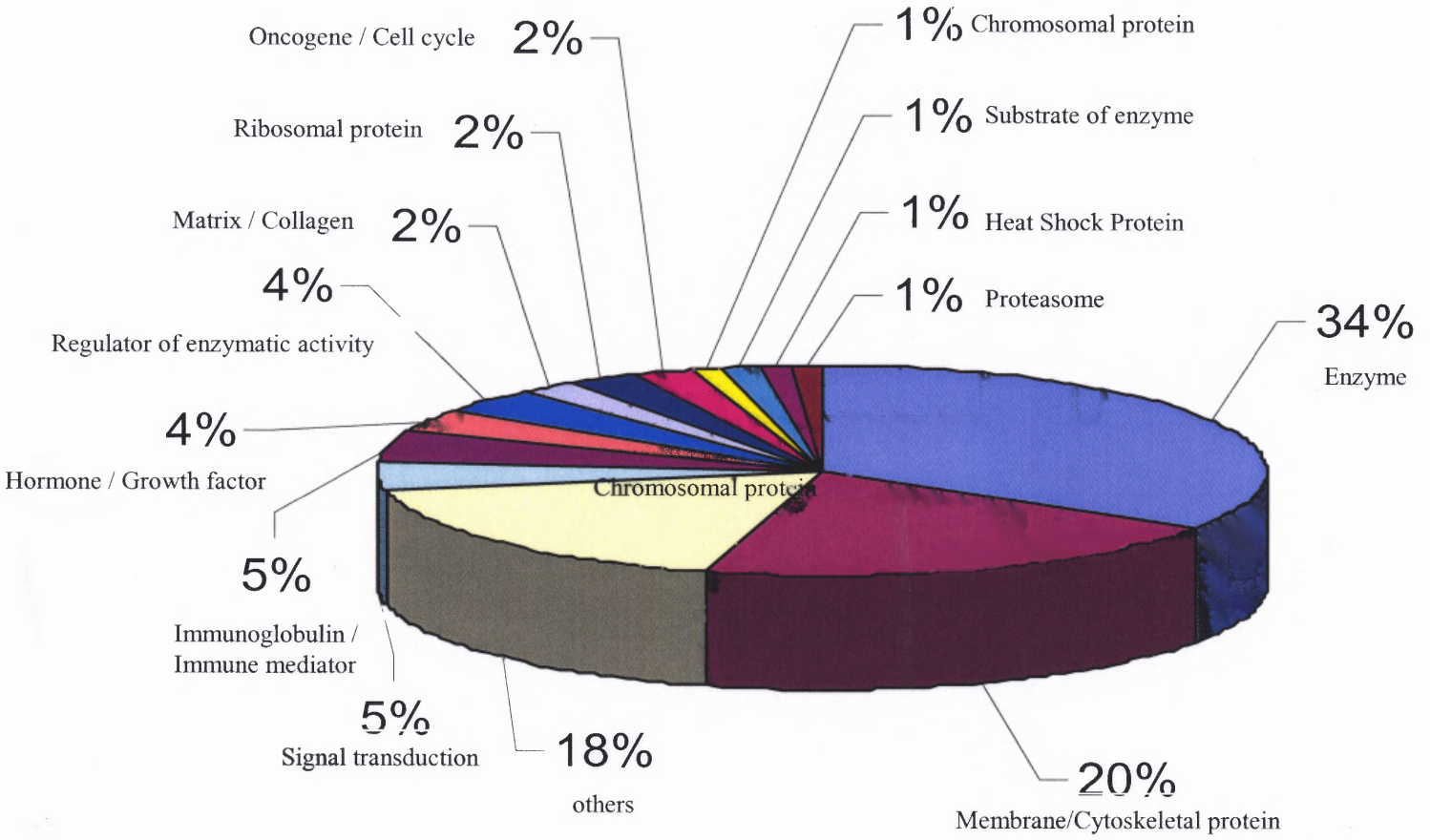


Figure 3.3.1 Distribution of function-known protein categories

Table 3.10 List of hypothetical proteins

Hypothetical Protein*	**Sequence ID**
Clone CDABP0107 mRNA sequence	139
Human full-length cDNA clone CS0DB001YK19 of neuroblastoma	42
Hypothetical protein	1, 5, 9, 63, 66, 70, 71, 110, 121, 175, 199, 201, 221, 248, 252, 272, 285
Similar to hypothetical protein MGC18754	192

* Human full-length cDNA clone CS0DM004YH09 of Fetal liver of Homo sapiens (SPTR: CAD62335) (SequenceID 10, 49, 149, 230, 243, 249), and hypothetical protein (EC 2.7.1.40) (SPTR:Q96E76) (SquenceID 156) and hypothetical protein FLJ13465 (SPTR:Q96E76) (SPTR:Q9H8L6) (SquenceID 167) have been updated as function-known proteins according to the latest release from SPTR.

The hypothetical protein was submitted to the programs, GO, HOVERGEN and IntrPro, for protein function prediction. GO (gene oncology) describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. IntePro is an integrated resource of protein families, domains and functional sites. And HOVERGEN is a homologous vertebrate gene database. The results are listed in **Table 3.11**.

3.5 Analysis of No Hit

The cDNA sequence of "No Hit" in the SPTR database search was submitted for analysis of genomic sequence matching at the public server of genome browser of UCSC (University of California at Santa Cruz) Genome Bioinformatics. The criteria is 70% identity over 100 bp. The results were summarized in the **Table 3.12**.

Table 3.11. Analysis of hypothetical protein

Program	GO	HOVERGEN		InterPro
Sequence ID	Molecular function	No. of AS*	Representative	Domain / functional site
1	structural constituent of ribosome	2		ribosomal protein S2
5		2		
9	---	---		---
42	legumain activity	7	legumain precursor	peptidases C13, Legumain
63		4	FKSG32	tRNA pseudouridine synthase
66		4		hypothetical prenyl group binding site
70	actin-binding	1		actin-binding FH2
71	---	---		---
110		2	HSPC326	
121		8	DNA helicase 1	
139	RNA binding	1		RNA recognition, region 1; Zn-finger, U1-like Zn-finger, C2H2 type;
175		80	Zinc finger protein	Zn-finger, C2H2 type; Zn-finger, U1-like

Table 3.11. (Continued)

Program	GO	HOVERGEN		InterPro
Sequence ID	Molecular function	No. of AS*	Representative	Domain / Functional site
192		2		ankyrin repeat; Zn-finger, MYND type
199		2		RhoGAP; BAR
201	Actin binding	8	leukocyte FORMIN	immunoglobulin/major histocompatibility complex; actin-binding FH2
221	cell adhesion receptor activity	12	integrin beta	integrin beta, C-terminal; von willebrand factor, type A; integrin, beta chain; Plexin / semaphoring integrin
248		5		
252	RNA-binding	1		Zn-finger, C2H2; Zn-finger, U1
272	DNA-binding	1		HMG-1 and HMG-Y DNA-binding domain (A+T-hook)
285		3	2-oxoisovalerate dehydrogenase beta subunit	transketolase, central region; transketolase, C- terminal

* Number of associated protein

Table 3.12. Analysis of No Hit

ID	location on chromosome	conserved region	known gene or gene prediction
7	chr2:162161098-162161529	433bp at 95.20% noncoding	Known Gene: PSMD14, human 26S proteasome-associated pad1 homolog (POH1)
	chr17: 75140810-75148372	266bp at 97.40% noncoding	Ensembl Gene Predictions: ENST00000331934.1 Acembly Gene Predictions With Alt-Splicing: vawfy Genscan Gene Predictions: NT_010641.273
23	chr17:70585910-70586285	382bp at 97.10% UTR	Known Gene: SOX9, transcription factor sox9
24	---	---	---
31	chr1:142583917-142584299	383bp at 100.00% noncoding	Known Gene: AF380581, AG3
	chr1:143634942-143635335	383bp at 95.80% noncoding	Acembly Gene Predictions With Alt-Splicing: smeyloy; Genscan Gene Predictions: NT_034398.5
	chr1:143835398-143835779	383bp at 95.60% noncoding	Acembly Gene Predictions With Alt-Splicing: choyda.a Genscan Gene Predictions: NT_034400.2
	chr1:146056378-146056757	383bp at 94.30% noncoding	Genscan Gene Predictions: NT_034403.1
32	chr21:33715783-33716004	222bp at 88.70% noncoding	Known Gene: IFNGR2, Interferon-gamma receptor beta chain Precursor
37	chr4:77482502-77482962	459bp at 99.80% noncoding	Known Gene: SCARB2, lysosomal sialoglycoprotein
39	chr21:29467690-29468219	528bp at 100.00% UTR	Known Gene: C21orf7
41	chr5:79985417-79985556	149bp at 74.50% noncoding	Known Gene: DHFR, dihydrofolate reductase
45	chr3:198640205-198640674	469bp at 99.40% UTR	Known Gene: AK094447, Hypothetical protein FLJ37128

Table 3.12. (Continued)

ID	location on chromosome	conserved region	known gene or gene prediction
50	chr2:206625093-206625618	526bp at 100.00% noncoding	Acembly Gene Predictions With Alt-Splicing:NRP2.b, neuropilin 2 (UTR) Genscan Gene Predictions: NT_005403.955
53	chr21:33715783-33727798	482bp at 99.60% noncoding	Known Gene: IFNGR2, Interferon-gamma receptor beta chain precursor
64	chr5:43708164-43708417	252bp at 95.60% noncoding	N/A
67	chr2:162161113-162161603	491bp at 99.00% noncoding	Known Gene: PSMD14, Human 26S proteasome-associated pad1 homolog (POH1)
	chr17: 75170798-75170882 75140821-75141075	85bp at 100.00% noncoding 255bp at 99.60% noncoding Total 340bp at 99.7%	Ensembl Gene Predictions: ENST00000331934.1 Genscan Gene Predictions: NT_010641.273
75	chr5:140682957-140683359	220bp at 99.10% UTR	Known Gene: TAF7, Transcription initiation factor TFIID 55 kDa subunit
80	chr15:72726708-72727090	386bp at 98.70% noncoding	Known Gene: SCAMP2, secretory carrier membrane protein 2
82	chr7: 21969801-21970150 21943849-21943967	350bp at 99.70% noncoding 119bp at 100.00% noncoding Total 469bp at 99.8%	Acembly Gene Predictions With Alt-Splicing: GFR, guanine nucleotide exchange factor for Rap1

Table 3.12. (Continued)

ID	location on chromosome	conserved region	known gene or gene prediction
83	---------	----------	----------
93	chr13: 109639348-109639373 109638914-109639347	26bp at 100.00% exon 434bp at 100.00% UTR Total 460bp at 100.0%	Acembly Gene Predictions With Alt-Splicing: COL4A1, collagen, type IV, alpha 1 Genscan Gene Predictions: NT_009952.477
96	---------	----------	----------
97	chr22:24649396-24649503	108bp at 88.90% noncoding	Known Gene: MYO18B, Myosin heavy chain.
111	chr21:33727314-33727801	488bp at 99.4% noncoding	Known Gene IFNGR2, Interferon-gamma receptor beta chain precursor
112	chr2:71636582-71637060	479bp at 99.8% noncoding	Genscan Gene Predictions (NT_022184.1023)
123	chr8: 42000462-42000480 42000481-42000855	19bp at 100.00% exon 375bp at 100.00% noncoding Total 394bp at 100.0%	Known Gene VDAC3, voltage dependent anion channel protein
138	chr13:109638904-109639230	327bp at 99.39% UTR	RefSeq Gene COL4A1, alpha 1 type IV collagen preproprotein
142	chr22: 24647138-24647677	540bp at 99.80% noncoding	RefSeq Gene MYO18B, myosin XVIIIB

Table 3.12. (Continued)

ID	location on chromosome	conserved region	known gene or gene prediction
145	chr1 :31522206-31522603	400bp at 98.30% UTR	K nown Gene: PEF
146	chr7:22690608-22691147	540bp at 99.80% UTR	Ensembl gene prediction: Human gene DRCTNNB1A encoding down-regulated by Ctnnb1, a.
147	chr2:71368510-71368891	382bp at 98.40% UTR	Known Gene: AB032981, Human mRNA for KIAA1155 protein
150	chr2:85505352-85505650	299bp at 100.00% noncoding,	Known Gene: BC028219, Similar to trans-golgi network protein 2
168	chr17:78324011-78324485	475bp at 100.00% noncoding	Acembly Gene Predictions With Alt-Splicing: Human gene Chromo. 1, encoding hypothetical protein MGC10561
183	chr20:24935726-24936207	482bp at 99.80% UTR	Known Gene: ACAS2L, Homo sapiens KIAA1846 protein
184	chr7:130588167-130588596	430bp at 99.10% UTR	Known Gene: PODXL, Homo sapiens podocalyxin-like protein
186	chr5:33799658-33800189	530bp at 99.80% non coding	Known Gene: ADAMTS12
189	chrX:1097847-1098402	555bp at 98.20% noncoding	N/A
193	chr2:161096260-161096745	486bp at 100.00% noncoding	Known Gene: RBMS1, RNA binding motif, single stranded interacting protein 1

Table 3.12. (Continued)

ID	location on chromosome	conserved region	known gene or gene prediction
194	chr9:14089089-14089614	524bp at 99.60% noncoding	Known Gene: NFIB, Human nuclear factor I-B2 (NFIB2)
196	---	---	---
197	chr1:656845-657041	197bp at 97.00% noncoding	Ensembl Gene Predictions: ENST00000332518.1
	chr5:99421060-99421255	194bp at 89.20% noncoding	N/A
	chr2:50773909-50774100	190bp at 86.80% noncoding	Known Gene: NRXN1, Human sapiens KIAA0578 protein
	chr14:30943341-30943519	175bp at 91.40% noncoding	Known Gene: AKAP6, Homo sapiens A-kinase anchor protein (AKAP100)
216	chr6:168748126-168748592	467bp at 100.00% UTR	Known Gene: SMOC2, secreted modular calcium-binding protein 2
223	chr18:47685114-47685576	463bp at 100.00% noncoding	Known Gene: MBD1, protein containing Methyl-CpG binding domain(MBD) 1
226	chr17:41855872-41856349	474bp at 98.90% UTR	Known Gene: BC008286 Acembly Gene Predictions With Alt-Splicing: dual specificity phosphatase 3
228	chr1:199777555-199777873	318bp at 97.50% UTR	Known Gene: FMOD, fibromodulin

Table 3.12. (Continued)

ID	location on chromosome	conserved region	known gene or gene prediction
232	chr15: 58257371-58257412	42bp at 100.00% exon	
	58257413-58257423	11bp at 100.00% UTR	
	58267493-58267766	274bp at 99.30% noncoding	
	58269280-58269310	31bp at 100.00% UTR	
		Total 358bp at 99.5%	Known Gene: ANXA2, annexin A2
247	chr17:34259325-34259851	339bp at 98.80% noncoding	Ensembl Gene Predictions: ENST00000331832.1 Acembly Gene Predictions With Alt-Splicing: feeror; yotoru; matoru (UTR)
251	chr4:77482551-77482952	402bp at 100.00% noncoding	Known Gene: SCARB2, lysosomal sialoglycoprotein
262	---	---	---
266	---	---	---

CHAPTER 4

DISCUSSION

To date, many bioinformatic tools have been developed to assist biologists in organizing and analyzing sequences. Most of the tools are available from web sources or public servers, and the interfaces are well designed to be user-friendly. However, performing these operations on each individual sequence for hundreds or thousands of sequences is a daunting prospect. If the results of these operations are not properly labeled and filed they are difficult to locate and harder still to compare. A database is a much more practical way to store results, and a software "pipeline", computer software that performs a series of operations on each sequence, allows biologists to handle a vast volume data from sequence analysis efficiently.

Microsoft Access, a database management system, was developed as software product to be installed on desktop computers. The database built within Microsoft Access is easily copied and transferred. Therefore, data can be shared with other users or database developers. In this thesis, Java applets for document parsing were applied to maintain correct flow of a data stream into the designed Microsoft Access database, BLAST –AUF1, from more than 1,800 local files containing sequence information and database search results. Correspondingly, various internal objects in BLAST-AUF1, such as table, query, form and report, were created to store necessary information. Consequently, the BLAST-AUF1 database provides a collection of cDNA sequence analysis information organized as to view, search, and retrieval of the correct details in an easy, timely, and effortless manner. These are the features ofthis database resulting from

that the mining of sequence information. This thesis emphasizes efforts to organize protein information that the AUF1 targets encode.

The protein output from BLAST-AUF1 is separated into three main groups as function-known protein (74%), hypothetical protein (8%) and No Hit (18%). The function-known proteins consist of 14 categories involved in diverse cell functions such as metabolism, growth, signal transduction, transcription, immune regulation. A noticeable phenomena is that the combined metabolic enzymes and structural proteins (membrane / cytoskeletal protein and matrix / collagen) occupy 56% of the total protein output, suggesting that AUF1 targets consist mainly of mRNAs that encode fundamental proteins essential for cell survival.

Furthermore, 20 of the 261 cDNAs clones studied represent uncharacterized genes encoding hypothetical proteins, for which there is no experimental evidence of function *in vivo*. The predicted of protein functions are from computational annotation only. The strength of these predictions depends on the quality of the alignment between the associated sequences. In this thesis, three computational programs, GO, IntePro and HOVERGENE, were employed. Sequence ID 9, 71 failed to yield any results from the three programs. Some interesting integrated information concerning domains and functional sites include:

1. RNA recognition, region 1 (Sequence ID 139): also known as the eukaryotic putative RNA-binding region RNP-1 signature, or RNA recognition motif (RRM), is implicated in regulation of alternative splicing and many other posttranscriptional processes.
2. Zn-finger (Sequence ID 139, 175, 192, 252): has the ability to bind to both RNA and DNA and possibly involved in protein interaction.

3. Ankyrin repeat (Sequence ID 221): one of the most common protein-protein interaction motifs in nature, has been found in proteins of diverse function such as transcriptional initiators, cell-cycle regulators, cytoskeletal proteins, ion transporters and signal transducers. Integrins are the major meazoan receptors for cell adhesion to extracellular matrix proteins and, in vertebrates, also play important roles in certain cell-cell adhesions, make transmembrane connections to the cytoskeleton, and activate many intracellular signaling pathways
4. HMG-I and HMG-Y DNA-binding domain (Sequence ID 272): high mobility group (HMG) proteins are a family of relatively low molecular weight, non-histone components in chromatin. It is suggested that these proteins could function in nucleosome phasing and in 3' end processing of mRNA transcripts. They are also involved in transcriptional regulation of genes.
5. RhoGAP domain (Sequence ID 199): a member of the Rho family of small G proteins transduces signals from plasma membrane receptors and controls cell adhesion, motility and shape by actin cytoskeleton formation.
6. Ribosomal protein S2 (Sequence ID 1): has been shown to belong to a family that includes 40S ribosomal subunit 40kDa proteins. Ribosomes are the particles that catalyze mRNA-directed protein synthesis in all organisms.
7. Actin-binding FH2 (Sequence ID 70, 201): FH proteins control rearrangements of the actin cytoskeleton, especially in the context of cytokinesis and cell polarization. Members of this family have been found to interact with Rho-GTPases, profilin and other actin-associated proteins.

Roughly, these hypothetical proteins may be implicated in the regulation of transcription and translation and activation of cytoskeleton and membrane signal transduction.

There are 47 sequences described as "No Hit" in the SPTR database which means no significant alignments can be obtained under the preset conditions for protein database searching. By matching human genome sequences, the positions of 41 No Hit sequences were found on some chromosomes, and another 6 No Hit sequences appeared in no matches. All of the matched sequences are located in noncoding or UTR regions, except that three of them extend to exon with a very short sequence (Sequence ID 93, 123, 232). In addition, 33 of the 41 matched sequences were identified as the partial sequence of known genes. These known genes also have potential multiple roles in the regulation of gene expression, cell metabolism, membrane ion channel, immune response, and so on.

The genome-scale identification of the 261 human cardiac mRNAs associating with AUF1 provides insight into AUF1-dependent potential functional pathways in human myocardial cells. The sequence information stored in the BLSAT-AUF1 database offers a data platform for cDNA microarray design to compare expression patterns of these genes in normal heart tissue with abnormal heart tissue. Such a cDNA microarray is also helpful for identifying the mRNAs that may be misregulated in congestive heart failure.

APPENDIX

JAVA CODES FOR DATA INPUT

The following Java codes were programmed for input of sequence information and database search results into the local database, BLAST-AUF1.

```
//*****************************************************************
  cDNA sequence information input
*****************************************************************//
import java.io.*;
import java.util.*;
import java.sql.*;

public class Sequence extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\cDNA_Sequence_Length";
            new Sequence().processFolder(folderName);
      }

      Connection conn;
      String folder2 = "C:\\documents and settings\\jeibo lu\\my
documents\\auf1 target\\hgmp result\\cDNA_SEQUENCE";


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
                  DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");

            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }
```

```
public void processFile(int i, String fileName) throws Exception{

            String sequenceLength = null;

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));
            String sequenceFile = fName.substring(0,dotIdx)+".htm";
            String cDNASequence = sequenceFile+"#"+
                  folder2+"\\"+sequenceFile+"##";


            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  sequenceLength = line.trim();
                  break;
            }
            fin.close();


            save(id, cDNASequence, sequenceLength);

      }

      void cleanDB() throws Exception{
            Statement stmt = conn.createStatement();
            stmt.execute("delete from cDNA_Sequence_Information");
            stmt.close();
      }

      void save(int i, String cDNASequence, String sequenceLength)
throws Exception{

            String insertValue =
                  "insert into cDNA_Sequence_Information("
                  +"sequenceId, "
                  +"cDNASequence, "
                  +"sequenceLength) values("
                  +i+", "
                  +getSqlStr(cDNASequence)+","
                  +sequenceLength+")";

            System.out.println(insertValue+"\n\n");

            Statement stmt = conn.createStatement();
            stmt.execute(insertValue);
            stmt.close();

      }

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in SPTR database

```
   BLASTX 2.2.3 [May-13-2002]

   Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.(590
   letters)

   Database: humansptr 74,183 sequences; 27,548,462 total letters

   Searching.....................................done

                                                      Score    E
   Sequences producing significant alignments:        (bits) Value

   SPTR:Q8NE13 Q8NE13 Hypothetical protein.            251   5e-89
   SPTR:Y419_HUMAN O43303 Hypothetical protein...      251   5e-89

   >SPTR:Q8NE13 Q8NE13 Hypothetical protein. Length = 991
    Score =  251 bits (641), Expect(2) = 5e-89
    Identities = 123/126 (97%), Positives = 124/126 (97%),
    Frame = -1
*****************************************************************//
import java.io.*;
import java.util.*;
import java.sql.*;

public class Human_sptr extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\HUMAN_SPTR";
            new Human_sptr().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
                  DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");

            cleanDB();
```

```
            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }


      public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            String sptrIdentifier = null;
            String sptrAccessID = null;
            String sptrProteinName = null;
            String sptrProteinLength = null;
            String sptrScore = null;
            String sptrEvalue = null;
            String sptrQuerySeqFrame = null;
            String sptrDetailResult = new File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));



            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  if(sptrQuerySeqFrame!=null) break;

                  if(line.trim().startsWith(">SPTR")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;
                        sptrIdentifier = find2(line, ">SPTR");
                        if(sptrIdentifier!=null)
                              sptrIdentifier = "SPTR:"+sptrIdentifier;
                        sptrAccessID = getToken(line, " ", 1);

                        String t3 = getToken(line, " ", 2);
                        if(t3!=null)
                              sptrProteinName =
line.substring(line.indexOf(t3));

                        while ((line=fin.readLine()) != null){
                              if(line.indexOf("Length")!=-1)
                                    break;
                              sptrProteinName =
sptrProteinName+line.trim();
                        }
                        if(line == null) break;
                  }
```

```
                if(reachFirstLine && line.indexOf("Length")!=-1){
                        if(sptrProteinLength==null)
                                sptrProteinLength = find(line, "Length",
"=");
                }

                if(reachFirstLine && line.indexOf("Score")!=-1){
                        if(sptrScore==null)
                                sptrScore = find(line, "Score", "=");
                        if(sptrEvalue==null)
                                sptrEvalue = find(line, "Expect", "=");
                }

                if(reachFirstLine && line.indexOf("Frame")!=-1){
                        if(sptrQuerySeqFrame==null)
                                sptrQuerySeqFrame = find(line, "Frame",
"=");
                }
            }
            fin.close();

            save(id, sptrIdentifier, sptrAccessID, sptrProteinName,
                        sptrProteinLength, sptrScore, sptrEvalue,
                        sptrQuerySeqFrame, sptrDetailResult);

      }

      void cleanDB() throws Exception{
            Statement stmt = conn.createStatement();
            stmt.execute("delete from Blast_in_Human_SPTR");
            stmt.close();
      }

      void save(int i, String sptrIdentifier, String sptrAccessID,
String sptrProteinName,
            String sptrProteinLength, String sptrScore, String
sptrEvalue,
            String sptrQuerySeqFrame, String sptrDetailResult) throws
Exception{

            String insertValue =
                  "insert into Blast_in_Human_SPTR("
                  +"sequenceId, "
                  +"sptrIdentifier, "
                  +"sptrAccessID, "
                  +"sptrProteinName, "
                  +"sptrProteinLength, "
                  +"sptrScore, "
                  +"sptrEvalue, "
                  +"sptrQuerySeqFrame, "
                  +"sptrDetailResult) values("
                  +i+", "
                  +getSqlStr(sptrIdentifier)+","
                  +getSqlStr(sptrAccessID)+","
                  +getSqlStr(sptrProteinName)+","
                  +sptrProteinLength+","
                  +sptrScore+","
```

```
                +getSqlStr(sptrEvalue)+","
                +getSqlStr(sptrQuerySeqFrame)+","
                +getSqlStr(sptrDetailResult)+")";

        System.out.println(insertValue+"\n\n");

        Statement stmt = conn.createStatement();
        stmt.execute(insertValue);
        stmt.close();

    }

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in IPI database

```
    BLASTX 2.2.3 [May-13-2002]

    Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.
    (590 letters)

    Database: ipi  33,013 sequences; 12,871,200 total letters

    Searching.....................................done

                                                      Score   E
    Sequences producing significant alignments:(bits) Value

    IPI:IPI00011933.1|SWISS-PROT:O43303|REF...  251   3e-89

    >IPI:IPI00011933.1|SWISSPROT:O43303|REFSEQ_NP:NP_055526|REFSE
    Q_XP:XP_008090|ENSEMBL:ENSP00000219827 Hypothetical protein
    KIAA0419 Length = 991

     Score =  251 bits (641), Expect(2) = 3e-89
     Identities = 123/126 (97%), Positives = 124/126 (97%)
     Frame = -1
 *****************************************************************//
import java.io.*;
import java.util.*;
import java.sql.*;

public class Ipi extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\IPI";
            new Ipi().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
                  DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");
```

```
            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }


public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            boolean reachLength = false;
            String ipiAccessID = null;
            String ipiProteinLocus = null;
            String ipiProteinName = null;
            String ipiDetailBlastResult = new
File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));

            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            String combinedLine = null;
            while ((line=fin.readLine()) != null){
                  if(line.trim().startsWith(">IPI:")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;

                        combinedLine = line.substring(5).trim();
                        while ((line=fin.readLine()) != null){
                              if(line.indexOf("Length")!=-1){
                                    reachLength = true;
                                    break;
                              }
                              combinedLine += line.trim();
                        }
                        if(line == null || reachLength) break;
                  }
            }

            if(combinedLine!=null){
                  combinedLine = combinedLine.trim();
                  ipiAccessID = getToken(combinedLine, " |", 0);
                  int idx1 = combinedLine.indexOf("|");
                  int idx2 = combinedLine.indexOf(" ");
                  if(idx1!=-1 && idx2!=-1){
                        ipiProteinLocus =
combinedLine.substring(idx1+1, idx2);
                        ipiProteinName =
combinedLine.substring(idx2+1);
```

```
                }
            }
            fin.close();

            save(id, ipiAccessID, ipiProteinLocus, ipiProteinName,
ipiDetailBlastResult);

      }

      void cleanDB() throws Exception{
            Statement stmt = conn.createStatement();
            stmt.execute("delete from Blast_in_IPI");
            stmt.close();
      }

      void save(int i, String ipiAccessID, String ipiProteinLocus,
            String ipiProteinName, String ipiDetailBlastResult) throws
Exception{

            String insertValue =
                  "insert into Blast_in_IPI("
                  +"SequenceID, "
                  +"ipiAccessID, "
                  +"ipiProteinLocus, "
                  +"ipiProteinName, "
                  +"ipiDetailBlastResult) values("
                  +i+", "
                  +getSqlStr(ipiAccessID)+","
                  +getSqlStr(ipiProteinLocus)+","
                  +getSqlStr(ipiProteinName)+","
                  +getSqlStr(ipiDetailBlastResult)+")";

            System.out.println(insertValue+"\n\n");

            Statement stmt = conn.createStatement();
            stmt.execute(insertValue);
            stmt.close();

      }

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in Human_EST database

```
    BLASTN 2.2.3 [May-13-2002]

    Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.
             (590 letters)

    Database: humanest1; humanest2  5,360,316 sequences; 2,905,283,263
    total letters

    Searching.....................................done

    Sequences producing significant alignments:     Score    E
                                                    (bits) Value

    EM:BM470571 BM470571 AGENCOURT_6462969 NIH_...   967   0.0
    EM:BQ423812 BQ423812 AGENCOURT_7918651 NIH_...   924   0.0
    EM:BM462895 BM462895 AGENCOURT_6427522 NIH_...   833   0.0
    EM:HSZZ17176 AA312040 EST182740 Jurkat T-cc...   484   e-134
    EM:BQ213066 BQ213066 AGENCOURT_7594527 NIH_...   438   e-120

    >EM:BM470571 BM470571 AGENCOURT_6462969 NIH_MGC_71 Homo sapiens
    cDNA clone IMAGE:5533448 5', mRNA sequence.
    Length = 1121

     Score =  967 bits (488), Expect = 0.0
     Identities = 511/516 (99%), Gaps = 2/516 (0%)
     Strand = Plus / Minus
*****************************************************************//
import java.io.*;
import java.util.*;
import java.sql.*;

public class Human_est extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\HUMAN_EST";
            new Human_est().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
```

```
            conn =
                  DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");

            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }



      public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            String emAccessID = null;
            String emLocus = null;
            String emLength = null;
            String emDetailResult = new File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));



            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  if(emLength!=null) break;

                  if(line.trim().startsWith(">EM")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;
                        emAccessID = find2(line, ">EM");
                        if(emAccessID!=null)
                              emAccessID = "EM:"+emAccessID;

                        String t2 = getToken(line, " ", 1);
                        if(t2!=null){
                              emLocus =
line.substring(line.indexOf(t2));
                              while ((line=fin.readLine()) != null){
                                    if(line.indexOf("Length")!=-1)
break;
                                    emLocus = emLocus+" "+line.trim();
                              }
                              if(line == null) break;
                        }
                  }
```

```
                if(reachFirstLine && line.indexOf("Length")!=-1){
                        if(emLength==null)
                                emLength = find(line, "Length", "=");
                }
        }
        fin.close();

        save(id, emAccessID, emLocus, emLength, emDetailResult);

    }

    void cleanDB() throws Exception{
        Statement stmt = conn.createStatement();
        stmt.execute("delete from Blast_in_HumanEST");
        stmt.close();
    }

    void save(int i, String emAccessID, String emLocus,
        String emLength, String emDetailResult) throws Exception{

        String insertValue =
                "insert into Blast_in_HumanEST("
                +"sampleID, "
                +"emAccessID, "
                +"emLocus, "
                +"emLength, "
                +"emDetailResult) values("
                +i+", "
                +getSqlStr(emAccessID)+","
                +getSqlStr(emLocus)+","
                +emLength+","
                +getSqlStr(emDetailResult)+")";

        System.out.println(insertValue+"\n\n");

        Statement stmt = conn.createStatement();
        stmt.execute(insertValue);
        stmt.close();

    }

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in ENSEMBL_cDNA database

```
   BLASTN 2.2.3 [May-13-2002]

   Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.
            (590 letters)

   Database: ensembl_cdna 37,347 sequences; 87,467,415 total letters

   Searching.....................................done
                                                 Score    E
   Sequences producing significant alignments: (bits) Value

   ENC:ENST00000219827 Database:core Gene...     975   0.0

   >ENC:ENST00000219827 Database:core Gene:ENSG00000103540
   Clone:AC003108 Contig:AC003108.1.1.164564 Chr:16 Basepair:18965466
   Status:known Length = 5463

    Score =  975 bits (492), Expect = 0.0
    Identities = 512/516 (99%), Gaps = 2/516 (0%)
    Strand = Plus / Minus
   ******************************************************************
```

```
import java.io.*;
import java.util.*;
import java.sql.*;

public class Cdna extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\aufl target\\hgmp result\\ENSEMBL_cDNA";
            new Cdna().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
                  DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
aufl.mdb", "", "");
```

```
            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }


      public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            String encIdentifier = null;
            String encDatabaseType = null;
            String encGeneId = null;
            String encCloneId = null;
            String encContigId = null;
            String encChromosome = null;
            String encBasePair = null;
            String encStatus = null;
            String encProteinLength = null;
            String encDetailBlastResult = new
File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));



            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  if(encProteinLength!=null) break;

                  if(line.trim().startsWith(">ENC")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;
                        encIdentifier = find2(line, ">ENC");
                        encDatabaseType = find2(line, "Database");
                        encGeneId = find2(line, "Gene");
                        encCloneId = find2(line, "Clone");
                  }

                  if(reachFirstLine && line.indexOf("Contig")!=-1){
                        if(encContigId==null)
                              encContigId = find2(line, "Contig");
                        if(encChromosome==null)
                              encChromosome = find2(line, "Chr");
                        if(encBasePair==null)
                              encBasePair = find2(line, "Basepair");
                  }
```

```
                if(reachFirstLine && line.indexOf("Status")!=-1){
                      if(encStatus==null)
                            encStatus = find2(line, "Status");
                }

                if(reachFirstLine && line.indexOf("Length")!=-1){
                      if(encProteinLength==null)
                            encProteinLength = find(line, "Length",
"=");
                }
          }
          fin.close();

          save(id, encIdentifier, encDatabaseType, encGeneId,
                      encCloneId, encContigId, encChromosome,
                      encBasePair, encStatus, encProteinLength,
                      encDetailBlastResult);

    }

    void cleanDB() throws Exception{
          Statement stmt = conn.createStatement();
          stmt.execute("delete from Blast_in_Ensembl_cDNA");
          stmt.close();
    }

    void save(int i, String encIdentifier, String encDatabaseType,
String encGeneId,
          String encCloneId, String encContigId, String
encChromosome,
          String encBasePair, String encStatus, String
encProteinLength,
          String encDetailBlastResult) throws Exception{

          String insertValue =
                "insert into Blast_in_Ensembl_cDNA("
                +"sequenceId, "
                +"encIdentifier, "
                +"encDatabaseType, "
                +"encGeneId, "
                +"encCloneId, "
                +"encContigId, "
                +"encChromosome, "
                +"encBasePair, "
                +"encStatus, "
                +"encProteinLength, "
                +"encDetailBlastResult) values("
                +i+", "
                +getSqlStr(encIdentifier)+","
                +getSqlStr(encDatabaseType)+","
                +getSqlStr(encGeneId)+","
                +getSqlStr(encCloneId)+","
                +getSqlStr(encContigId)+","
                +getSqlStr(encChromosome)+","
                +encBasePair+","
                +getSqlStr(encStatus)+","
```

```
                +getSqlStr(encProteinLength)+","
                +getSqlStr(encDetailBlastResult)+")";

            System.out.println(insertValue+"\n\n");

            Statement stmt = conn.createStatement();
            stmt.execute(insertValue);
            stmt.close();

        }

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in HSUNIGENE database

```
   BLASTN 2.2.3 [May-13-2002]

   Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.
            (590 letters)

   Database: hsunigene 111,064 sequences; 121,581,373 total letters

   Searching.....................................done

                                                      Score    E
   Sequences producing significant alignments:        (bits) Value

   UG:Hs.279912 gnl|UG|Hs#S2139501 Homo sapien...      975   0.0
   UG:Hs.423894 gnl|UG|Hs#S4005427 00004 Human...       96   1e-18
   UG:Hs.421675 gnl|UG|Hs#S3169779 IL0-OT0123-...       92   2e-17
   UG:Hs.284275 gnl|UG|Hs#S4840949 Homo sapien...       72   2e-11

   >UG:Hs.279912 gnl|UG|Hs#S2139501 Homo sapiens KIAA0419 gene
   product(KIAA0419), mRNA /cds=(292,3267) /gb=NM_014711 /gi=7662105
   /ug=Hs.279912 /len=5399 Length = 5399

    Score =  975 bits (492), Expect = 0.0
    Identities = 512/516 (99%), Gaps = 2/516 (0%)
    Strand = Plus / Minus
*****************************************************************//
import java.io.*;
import java.util.*;
import java.sql.*;

public class Unigene extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\HS_UNIGENE";
            new Unigene().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
```

```
                DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");

            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }



      public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            String hsugAccessID = null;
            String hsugLocus = null;
            String hsugLength = null;
            String hsugDetailBlastResult = new
File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));



            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  if(hsugLength!=null) break;

                  if(line.trim().startsWith(">UG")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;
                        hsugAccessID = find2(line, ">UG");
                        if(hsugAccessID!=null)
                              hsugAccessID = "UG:"+hsugAccessID;

                        hsugLocus = getToken(line, " ", 1);
                  }

                  if(reachFirstLine && line.indexOf("Length")!=-1){
                        if(hsugLength==null)
                              hsugLength = find(line, "Length", "=");
                  }
            }
            fin.close();

            save(id, hsugAccessID, hsugLocus, hsugLength,
hsugDetailBlastResult);
```

```
	}

	void cleanDB() throws Exception{
		Statement stmt = conn.createStatement();
		stmt.execute("delete from Blast_in_HSUNIGENE");
		stmt.close();
	}

	void save(int i, String hsugAccessID, String hsugLocus,
		String hsugLength, String hsugDetailBlastResult) throws
Exception{

		String insertValue =
			"insert into Blast_in_HSUNIGENE("
			+"sampleID, "
			+"hsugAccessID, "
			+"hsugLocus, "
			+"hsugLength, "
			+"hsugDetailBlastResult) values("
			+i+", "
			+getSqlStr(hsugAccessID)+","
			+getSqlStr(hsugLocus)+","
			+hsugLength+","
			+getSqlStr(hsugDetailBlastResult)+")";

		System.out.println(insertValue+"\n\n");

		Statement stmt = conn.createStatement();
		stmt.execute(insertValue);
		stmt.close();

	}

}
```

```
//*****************************************************************
```

This is an example of BLSTA result in TIGRHGI database

```
    BLASTN 2.2.3 [May-13-2002]

    Query= s1 Temp_71_188538__068, 590 bases, A006594B checksum.
             (590 letters)

    Database: tigrhgi 673,992 sequences; 341,075,042 total letters

    Searching.....................................done

                                                      Score    E
    Sequences producing significant alignments:      (bits) Value

    NP209996|NM_014711.1|NP_055526.1 KIAA0419...      975   0.0
    THC737506 KIAA0419^^KIAA0420^^KIAA0419 ge...      975   0.0
    AA312040                                          484   e-135
    AF074665                                          363   8e-99
    THC771578                                          98   8e-19

    >NP209996|NM_014711.1|NP_055526.1 KIAA0419 gene product
              Length = 2976

     Score =  975 bits (492), Expect = 0.0
     Identities = 512/516 (99%), Gaps = 2/516 (0%)
     Strand = Plus / Minus
 ******************************************************************//
```

```
import java.io.*;
import java.util.*;
import java.sql.*;

public class Tigrhgi extends Finder{
      public static void main(String[] args) throws Exception{
            BufferedReader reader = new BufferedReader(new
InputStreamReader(System.in));

            String folderName = "C:\\documents and settings\\jeibo
lu\\my documents\\auf1 target\\hgmp result\\TIGRHGI";
            new Tigrhgi().processFolder(folderName);
      }

      Connection conn;


      public void processFolder(String folderName) throws Exception{
            File folder = new File(folderName);
            if(!folder.isDirectory()){
                  System.out.println("Input folder name invalid!");
                  System.exit(0);
            }

            File[] files = folder.listFiles();

            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            conn =
```

```
                DriverManager.getConnection("jdbc:odbc:MS Access
Database;DBQ=C:\\documents and settings\\jeibo lu\\my documents\\blast-
auf1.mdb", "", "");

            cleanDB();

            for(int i=0; i<files.length; i++){
                  String fileName = files[i].getAbsolutePath();
                  processFile(i, fileName);
            }
            conn.close();
      }
      public void processFile(int i, String fileName) throws Exception{

            boolean reachFirstLine = false;
            String tigrhgiAccessID = null;
            String tigrhgiLength = null;
            String tigrhgiScore = null;
            String tigrhgiEvalue = null;
            String tigrhgiStrand = null;
            String tigrhgiDetailResult = new
File(fileName).getName()+"#"+
                  new File(fileName).getAbsolutePath()+"##";

            String fName = new File(fileName).getName();
            int dotIdx = fName.indexOf(".");
            int id = Integer.parseInt(fName.substring(1, dotIdx));


            BufferedReader fin = new BufferedReader(new
FileReader(fileName));
            String line;
            while ((line=fin.readLine()) != null){
                  if(tigrhgiStrand!=null) break;

                  if(line.trim().startsWith(">")){
                        if(reachFirstLine) break;
                        reachFirstLine = true;

                        tigrhgiAccessID = line.substring(1).trim();
                        while ((line=fin.readLine()) != null){
                              if(line.indexOf("Length")!=-1)
                                    break;
                              tigrhgiAccessID = tigrhgiAccessID+"
"+line.trim();
                        }
                        if(line == null) break;
                  }

                  if(reachFirstLine && line.indexOf("Length")!=-1){
                        if(tigrhgiLength==null)
                              tigrhgiLength = find(line, "Length",
"=");
                  }

                  if(reachFirstLine && line.indexOf("Score")!=-1){
                        if(tigrhgiScore==null)
```

```
                              tigrhgiScore = find(line, "Score", "=");
                        if(tigrhgiEvalue==null)
                              tigrhgiEvalue = find(line, "Expect",
"=");
                  }

                  if(reachFirstLine && line.indexOf("Strand")!=-1){
                        if(tigrhgiStrand==null)
                              tigrhgiStrand = find(line, "Strand",
"=");
                  }

            }
            fin.close();

            save(id, tigrhgiAccessID, tigrhgiLength, tigrhgiScore,
                        tigrhgiEvalue, tigrhgiStrand,
tigrhgiDetailResult);

      }

      void cleanDB() throws Exception{
            Statement stmt = conn.createStatement();
            stmt.execute("delete from Blast_in_TIGRHGI");
            stmt.close();
      }

      void save(int i, String tigrhgiAccessID, String tigrhgiLength,
String tigrhgiScore,
            String tigrhgiEvalue, String tigrhgiStrand, String
tigrhgiDetailResult) throws Exception{

            String insertValue =
                  "insert into Blast_in_TIGRHGI("
                  +"SampleID, "
                  +"tigrhgiAccessID, "
                  +"tigrhgiLength, "
                  +"tigrhgiScore, "
                  +"tigrhgiEvalue, "
                  +"tigrhgiStrand, "
                  +"tigrhgiDetailResult) values("
                  +i+", "
                  +getSqlStr(tigrhgiAccessID)+","
                  +tigrhgiLength+","
                  +getSqlStr(tigrhgiScore)+","
                  +getSqlStr(tigrhgiEvalue)+","
                  +getSqlStr(tigrhgiStrand)+","
                  +getSqlStr(tigrhgiDetailResult)+")";

            System.out.println(insertValue+"\n\n");

            Statement stmt = conn.createStatement();
            stmt.execute(insertValue);
            stmt.close();

      }
}
```

REFERENCES

1. J. Guhaniyogi and G. Brewer, "Regulation of mRNA stability in mammalian cells," *Gene*, vol. 265, pp. 11-23, 2001.

2. C. T. DeMaria and G. Brewer, "AUF1 binding affinity to A+U-rich elements correlates with rapid mRNA degradation," *The Journal of Biological Chemistry*, vol. 271, pp. 12179-12184, 1996.

3. G. M. Wilson and G. Brewer, "Identification and characterization of proteins binding A + U-rich elements," *Methods*, vol. 17, pp. 74-83, 1999.

4. A. Pende, K. D. Tremmel, C. T. DeMaria, B. C. Blaxall, W. A. Minobe, J. A. Shermant, J. D. Bisognano, M. R. Bristow, G. Brewer, and J. D. Port, "Regulation the mRNA-binding protein AUF1 by activation of β–adrenergic receptor signal transduction pathway," *The Journal of Biological Chemistry*, vol. 271, pp. 8493-8501, 1996.

5. B. C. Blaxall, A. C. Pellett, S. C. Wu, A. Pende, and J. D. Port, "Purification and characterization of β–adrenergic receptor mRNA-binding proteins," *The Journal of Biological Chemistry*, vol. 274, pp. 4290-4297, 2000.

6. M. J. Lohse, S. Engelhardt, and T. Eschenhagen, "What is the role of ß-adrenergic signaling in heart failure?" *Circulation Research*, vol. 93, pp. 896-906, 2003.