XWRAPComposer:

A Wrapper Generation System for Integrating Bioinformatic Data Sources

Ling Liu
College of Computing
Georgia Tech

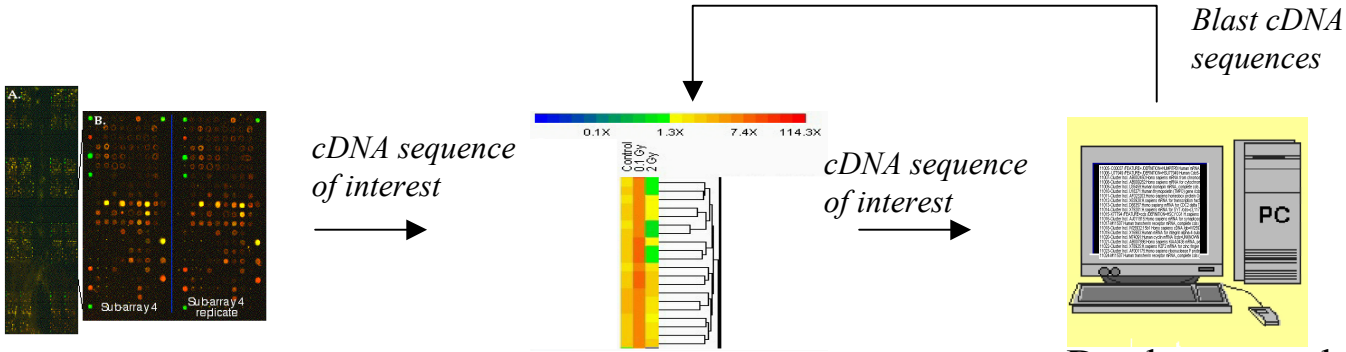
Team:
Faculty: Ling Liu, Calton Pu
Students: David Buttler, Wei Han, Henrique Paques

XWRAPComposer: A quick intro

- What is it?
  - Capable of generating complex wrappers
  - Wrappers (info. extraction programs)
    - Extract information from *multiple* Web pages connected by URLs (page links) and
    - Package the extracted information into an XML document for complex data integration
  - Extremely useful for integrating access to multiple scientific (e.g., bio-informatics) data sources

A Pilot Application Scenario

- Building and extending a promoter model
  - Provided by Matthew Coleman (LLNL)
- Study the effects of low-dose radiation on human genes
- Typical Processing Steps:
  - Microarray analysis
  - Statistical clustering analysis
    DB search for common promoter elements to link new candidate genes



Microarray analysis

cDNA Cluster

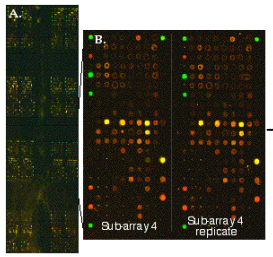
Database search for promoter identification

Technical Challenges of the Pilot Scenario

- A number of alternative solutions to implement the pilot scenario:
  - at the microarray step
    - what genes are chosen, when and what level of radiations are the genes exposed to?, and so on
    - Which types of microarry system is used?
  - at the cluster step → data integration challenge
    - what clustering techniques are used and from which bio web service providers?
  - at the search step → data integration challenge
    - what similarity criteria are used to identify more genes (blast, structure, different blast algorithms, etc.)?
    - Which bio-data service providers offer the required search functionality?

The Pilot Scenario - Use Case 1

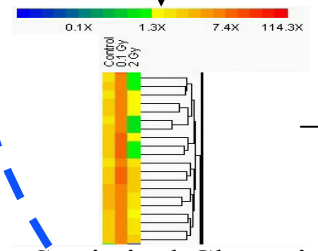
- Challenges for Heterogeneous Data Access
  - Source-specific information wrapping
  - Data integration across multiple heterogeneous sources
- A Simple Use Case:
  - Start with the initial results from a microarray analysis
    - A cDNA microarray system
  - at the clustering step
    - Start with single Web source: **clusfavor** (http://mbcr.bcm.tmc.edu/genepi/)
  - at the search step
    - Start with single bio Web source:
      - NCBI Genbank Keyword search
      - NCBI Blastn



Microarray analysis

Genes that changed significantly

XWrap Composer CLUSFAVOR



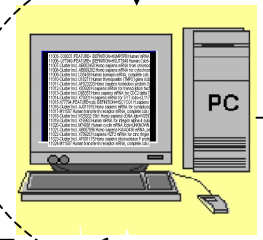
Statistical Clustering of genes

Gene ids

XWrap Composer NCBI BLAST

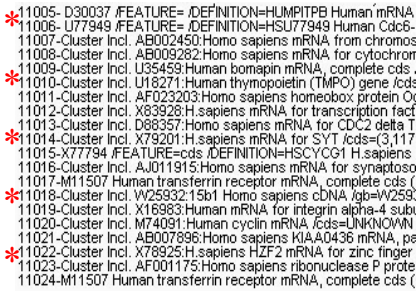
AA045112

Scenario composer



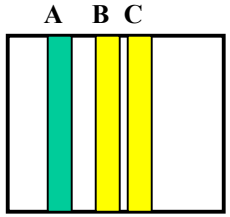
A list of related sequences

BLAST search over a variety of data sources for common promoter elements to link new candidate genes



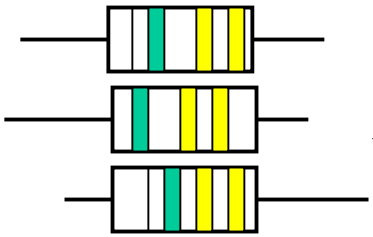
*- New candidate target genes

Consensus sequences

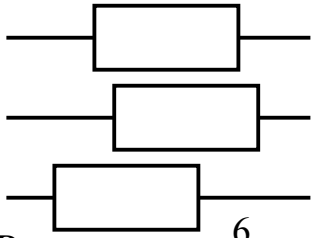


Promoter model

http://genomatix.ge

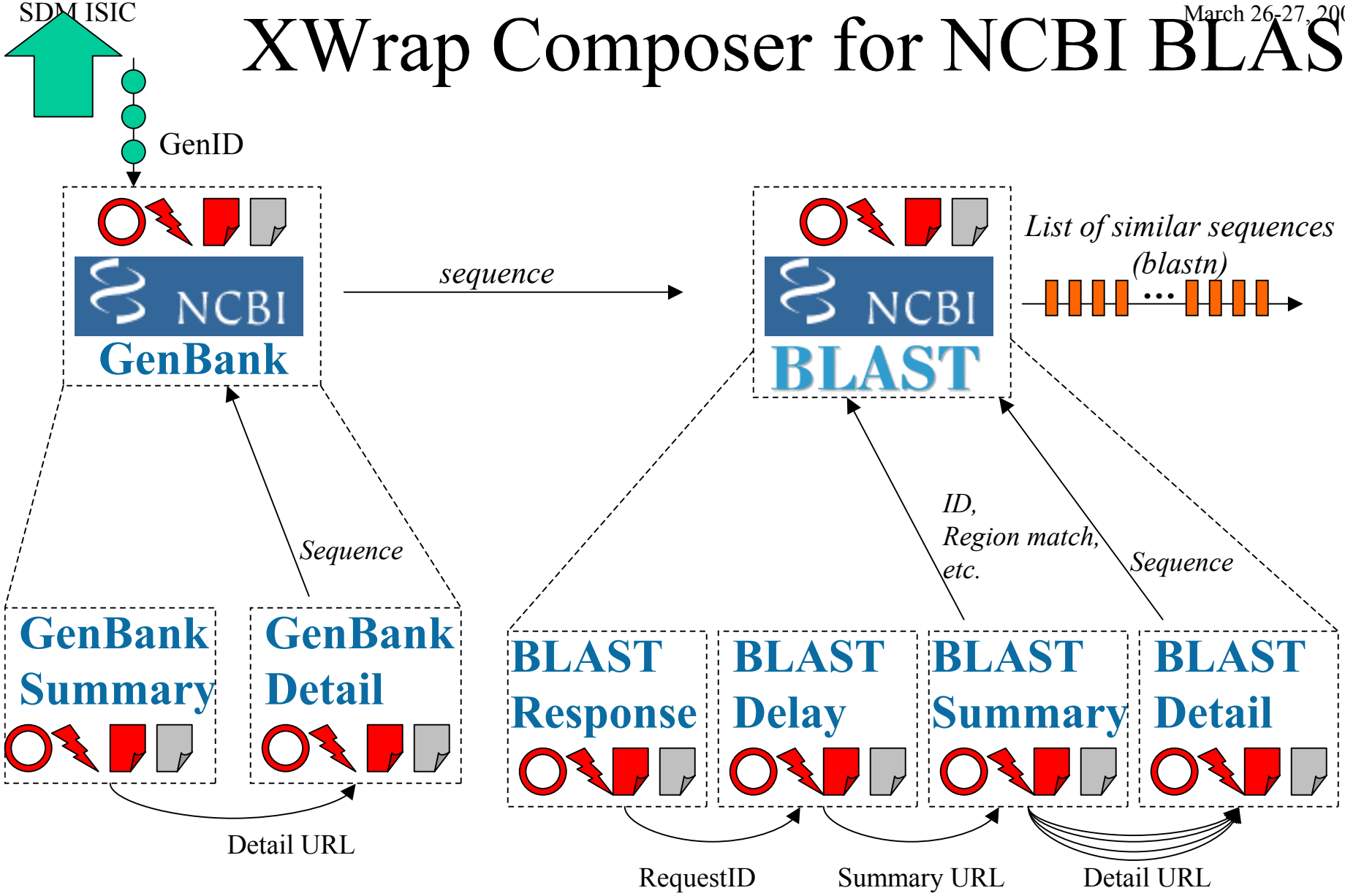


Common promoter alignment



Promoter sequences

XWrap Composer for NCBI BLAST



Research & Development Plan

- Components of XWRAPComposer
  - Interface language
    - Naming Space
    - Specialization and Reuse of Wrappers
    - GUI for XWRApComposer Interface Design
  - Composer scripting language
    - Merging several single-webpage data extractors into one complex XML page composer wrapper.
  - XWRAPComposer Code Library
    - Java Library
    - C Library
  - Self-configuring and self-tuning
    - performance optimization, handling various unexpected delays, failover solutions

Development Plan

- XWRAPComposer - Initial tookit release
  - for identified bio Web sources
    - such as NCBI blast, NCBI Genbank search, PDB keyword, blast search, etc.
  - Testing locally
    - Using AQR
    - Using WebCQ
  - Testing within SDM
    - LLNL, SDSC, NCSU, NWU, and other teams
- First official release (w/o optimization)
  - By end of 2002
  - Next spring/summer on the Web
- Next Release
  - Self-recoverable from crash
  - Secure
  - Performance optimization …

Scenario 1 Interface

Input: Filename or URL of Clusfavor Output

Output:

```
<BlastN source="NCBi" url="" queryString="">
   <objects desc="">
      <object>
         <genid>gen id</genid>
         <desc>description</desc>
         …
         <topmatchedsequence> matched sequence </topmatchedsequence>
      </object>
   </objects>
</BlastN>
```

Template: matexample.ext_template

Scenario 1 Composer Script

```
<XWRAPCOMPOSER:PageComposer wrappername="matexample">
  <XWRAPCOMPOSER:RunWrapper name=”clusfavor” type=”spacedelimited” inputurl=”&url;” queryString=”” 
    extractionTemplate=”clusfavor.ext_template”>
  <!-- if RunWrapper does not specify a style file, it should have an element of style that describes the style.-->

    <XWRAPCOMPOSER:style>
      <xsl:for-each select=”ResultSet/Object”>
        <XWRAPCOMPOSER:RunWrapper name=” NCBiGenBank” type=”PageComposer” 
  inputurl=”blahblahblah&listid={sequenceid}” extractionTemplate="NCBiGenBank.ext_template">
        <!-- if RunWrapper does not specify a style file, it should have an element of style that describes the style. -->

          <XWRAPCOMPOSER:style>
            <XWRAPCOMPOSER:RunWrapper name=”BlastN” type=”PageComposer” 
  inputurl=”blastnurl+{sequence}” style=”default” extractionTemplate="BlastN.ext_template">
            </XWRAPCOMPOSER:RunWrapper>
          </XWRAPCOMPOSER:style>

        </XWRAPCOMPOSER:RunWrapper>
      </xsl:for-each>
    </XWRAPCOMPOSER:style>

  </XWRAPCOMPOSER:RunWrapper>
</XWRAPCOMPOSER:PageComposer>
```

CLUSFAVOR: Table Delimited Wrapper

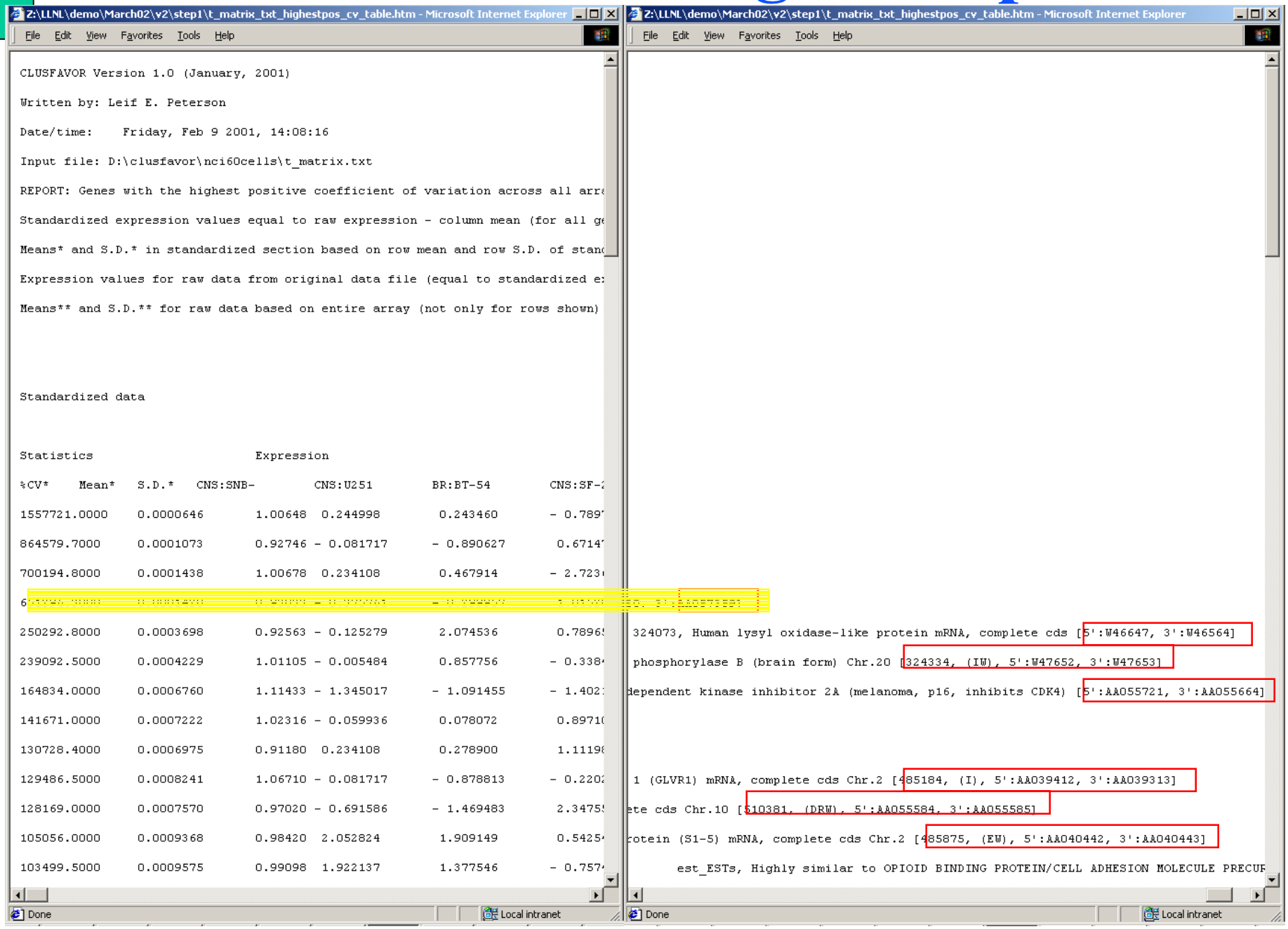
Input:

Filename or URL of Clusfavor Output

Output:

```
<ResultSet>
   <object>
      <sequenceid>sequence1</sequenceid>
      <sequenceid>sequence2</sequenceid>
   </object>
</ResultSet>
```

Clusfavor Browsing Example



Z:\LLNL\demo\March02\v2\step1\t_matrix_txt_highestpos_cv_table.htm - Microsoft Internet Explorer

File Edit View Favorites Tools Help

CLUSFAVOR Version 1.0 (January, 2001)

Written by: Leif E. Peterson

Date/time: Friday, Feb 9 2001, 14:08:16

Input file: D:\clusfavor\nci60cells\t_matrix.txt

REPORT: Genes with the highest positive coefficient of variation across all arr

Standardized expression values equal to raw expression - column mean (for all g

Means* and S.D.* in standardized section based on row mean and row S.D. of stan

Expression values for raw data from original data file (equal to standardized e

Means** and S.D.** for raw data based on entire array (not only for rows shown)

Standardized data

Statistics			Expression			
%CV*	Mean*	S.D.*	CNS:SNB-	CNS:U251	BR:BT-54	CNS:SF-
1557721.0000	0.0000646	1.00648	0.244998	0.243460	- 0.789	
864579.7000	0.0001073	0.92746	- 0.081717	- 0.890627	0.6714	
700194.8000	0.0001438	1.00678	0.234108	0.467914	- 2.723	
6						
250292.8000	0.0003698	0.92563	- 0.125279	2.074536	0.7896	
239092.5000	0.0004229	1.01105	- 0.005484	0.857756	- 0.338	
164834.0000	0.0006760	1.11433	- 1.345017	- 1.091455	- 1.402	
141671.0000	0.0007222	1.02316	- 0.059936	0.078072	0.8971	
130728.4000	0.0006975	0.91180	0.234108	0.278900	1.1119	
129486.5000	0.0008241	1.06710	- 0.081717	- 0.878813	- 0.220	
128169.0000	0.0007570	0.97020	- 0.691586	- 1.469483	2.3475	
105056.0000	0.0009368	0.98420	2.052824	1.909149	0.5425	
103499.5000	0.0009575	0.99098	1.922137	1.377546	- 0.757	

Done Local intranet

Z:\LLNL\demo\March02\v2\step1\t_matrix_txt_highestpos_cv_table.htm - Microsoft Internet Explorer

File Edit View Favorites Tools Help

324073, Human lysyl oxidase-like protein mRNA, complete cds [5':W46647, 3':W46564]

phosphorylase B (brain form) Chr.20 [324334, (IW), 5':W47652, 3':W47653]

dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4) [5':AA055721, 3':AA055664]

1 (GLVR1) mRNA, complete cds Chr.2 [485184, (I), 5':AA039412, 3':AA039313]

te cds Chr.10 [510381, (DRW), 5':AA055584, 3':AA055585]

rotein (S1-5) mRNA, complete cds Chr.2 [485875, (EW), 5':AA040442, 3':AA040443]

est_ESTs, Highly similar to OPIOID BINDING PROTEIN/CELL ADHESION MOLECULE PRECUR

Done Local intranet

Clusfavor Wrapped Data

```
<ResultSet>
    <object>
        <sequenceid>T98316</sequenceid>
        <sequenceid>T98261</sequenceid>
    </object>
    <object>
        <sequenceid>AA045112</sequenceid>
    </object>
    <object>
        <sequenceid>W44378</sequenceid>
        <sequenceid>W45731</sequenceid>
    </object>
    ...
</ResultSet>
```

Clusfavor Composer Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”Clusfavor” releaseDate="" owner="">
  <XWRAPCOMPOSER:RunTableDelimition>
    <XWRAPCOMPOSER:Delimitors>
      <XWRAPCOMPOSER:Delimitor>tab</XWRAPCOMPOSER:Delimitor>
      <XWRAPCOMPOSER:Delimitor>comma</XWRAPCOMPOSER:Delimitor>
      <XWRAPCOMPOSER:Delimitor>colon</XWRAPCOMPOSER:Delimitor>
    </XWRAPCOMPOSER:Delimitors>
    <XWRAPCOMPOSER:IgnoredCharacters>
      <XWRAPCOMPOSER:Character>singlequote</XWRAPCOMPOSER:Character>
      <XWRAPCOMPOSER:Character>[</XWRAPCOMPOSER:Character>
      <XWRAPCOMPOSER:Character>]</XWRAPCOMPOSER:Character>
    </XWRAPCOMPOSER:IgnoredCharacters>
    <XWRAPCOMPOSER:columns>
      <XWRAPCOMPOSER:column name=”5inchsequenceid” columnnumber="126"/>
      <XWRAPCOMPOSER:column name=”3inchsequenceid” columnnumber="128"/>
    </XWRAPCOMPOSER:columns>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <ResultSet>
          <xsl:for-each select="table/row”>
            <object>
              <sequence><xsl:value-of select="5inchsequenceid"/></sequence>
              <sequence><xsl:value-of select="3inchsequenceid"/></sequence>
            </object>
          </xsl:for-each>
        </ResultSet>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunTableDelimition >
</XWRAPCOMPOSER:PageExtractor>
```

GenBank Keyword Search:

Input:

A URL that contains a sequenceid in the queryString

Output:

<sequence> seq </sequence>

GenBank Composer Script

```
<XWRAPCOMPOSER:PageComposer exe_wrapper_name=”NCBiGenBank” source_code=java
    releaseDate="" owner="">
  <XWRAPCOMPOSER:RunPageExtractor exe_name="NCBiGenBankSummary“ source_code=java
    code_generator="XWRAPElite" input_url=”&url;” query_string=””
    extraction_template="NCBiGenBankSummary.ext_template">
  <!-- NCBiGenBankSummary.ext_template should contain enough information to produce the
    pageextractor as well as other description, such as the release date and the owner. -->

    <XWRAPCOMPOSER:parameters></XWRAPCOMPOSER:parameters>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">

        <xsl:for-each select=”ResultSet/object”>
          <XWRAPCOMPOSER:RunPageExtractor name="NCBiGenBankDetail"
    type=”KeywordExtraction” inputurl=”{detailpagelink}” querystring=””
    extractionTemplate=”NCBiGenBankDetail.ext_template” style="default">
            <!-- the default style refers to the predefined output format of NCBiGenBankDetail. -->
          </XWRAPCOMPOSER:RunPageExtractor>
        </xsl:for-each>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunPageExtractor>
</XWRAPCOMPOSER:PageComposer>
```

GenBank Wrapped Data

```
<sequence>
CACCTGGAGAAACTTCTGCACTGGCACTGTGTTCCNAGAGCTCCTTCTATGCGTCCCTCC
CAAGTGATTTAATTTCAGCTGATTGGACTACGAATTCACAAGGCAGAAAAGTCAAGGTCA
TTTGGNATCTGGAGACAGGAGAACTCAAGGAACCNAAAGGACT
</sequence>
```

GenBank Summary Interface

Input:

A URL of the GenBank summary page

Output:

```
<ResultSet>
   <structure>
      <detailpagelink>link</detailpagelink>
   </structure>
</ResultSet>
```

Template:

NCBiGenBankSummary.ext_template (generated by XWRAPElite)

GenBank Summary Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”NCBiGenBankSummary” releaseDate="" owner="">
  <XWRAPCOMPOSER:RunXWRAPEliteExtraction>
    <XWRAPCOMPOSER:parameters>
      <XWRAPCOMPOSER:para name=”tagElementSeparator” value=”td”/>
      …
    </XWRAPCOMPOSER:parameters>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <ResultSet>
          <xsl:for-each select=”ResultSet/object”>
          <structure>
            <detailpagelink><xsl:value-of select="element2/link"/>
            </detailpagelink>
          </structure>
          </xsl:for-each>
        </ResultSet>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunXWRAPEliteExtraction>
</XWRAPCOMPOSER:PageExtractor>
```

GenBank Summary Data

<ResultSet source="NCBiGenBankSummary">
<structure>
<detailPageLink>
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=1523314&dopt=GenBank
</detailPageLink>
</structure>
</ResultSet>

GenBank Detail Page Interface

Input:

A URL of a GenBank detail page

Output:

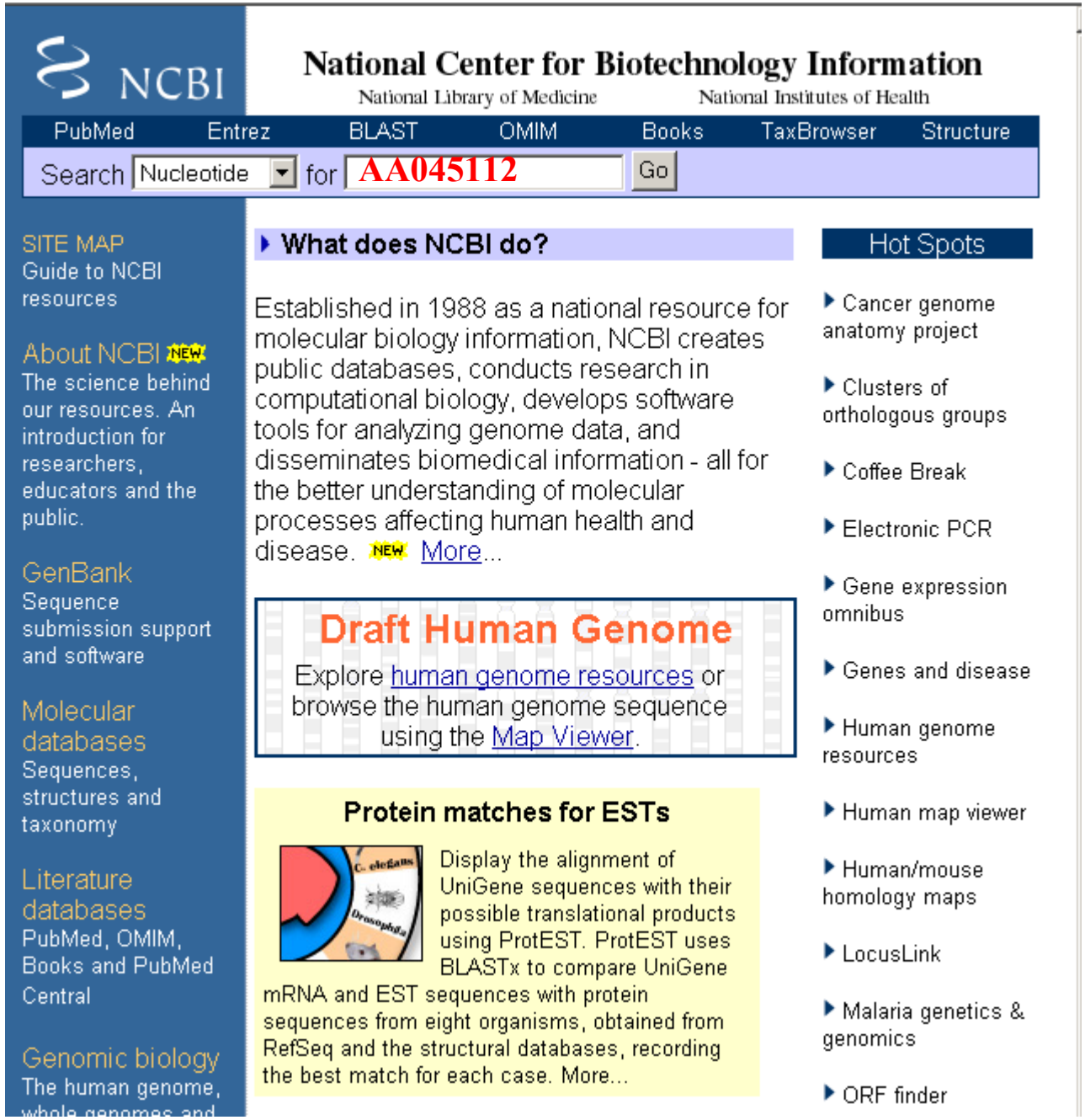
<sequence> seq </sequence>

Template:

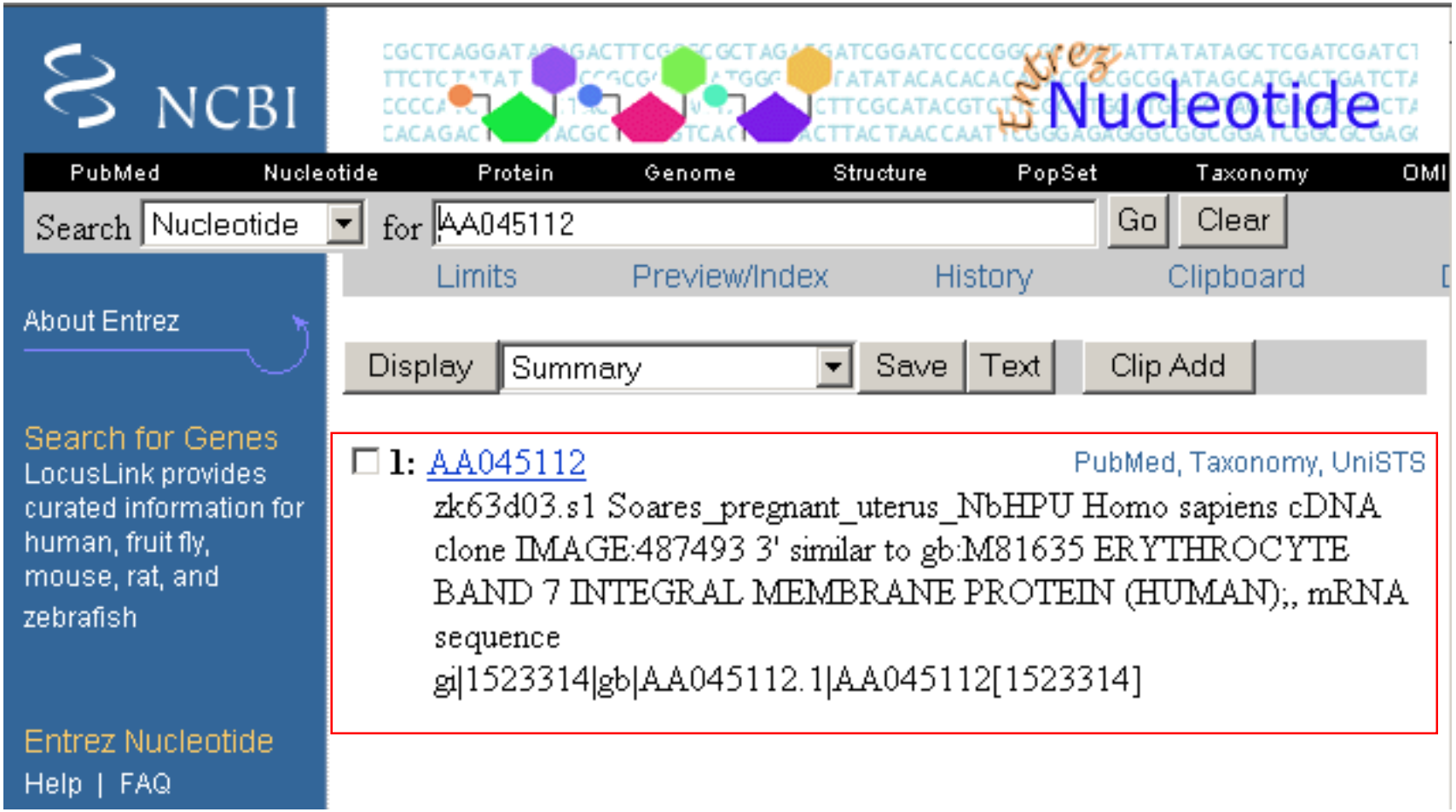
NCBiGenBankDetail.ext_template

GenBank Detail Page Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”NCBiGenBankDetail” releaseDate=""
    owner="">
  <XWRAPCOMPOSER:RunKeywordExtraction>
    <XWRAPCOMPOSER:variables>
      <XWRAPCOMPOSER:variable name=”sequence” BeginMatch=”&lt;B&gt;
    SEQUENCE&lt;/B&gt;” EndMatch=”Quality:” />
    </XWRAPCOMPOSER:variables>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <sequence><xsl:value-of select="sequence"/>
        </sequence>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunKeywordExtraction>
</XWRAPCOMPOSER:PageExtractor>
```


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
Entrez
BLAST
OMIM
Books
TaxBrowser
Structure
Search
Nucleotide
for
AA045112
Go
SITE MAP
Guide to NCBI resources
About NCBI
NEW
The science behind our resources. An introduction for researchers, educators and the public.
GenBank
Sequence submission support and software
Molecular databases
Sequences, structures and taxonomy
Literature databases
PubMed, OMIM, Books and PubMed Central
Genomic biology
The human genome,
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.
NEW
More...
Draft Human Genome
Explore human genome resources or browse the human genome sequence using the Map Viewer.
Protein matches for ESTs
C. elegans
Drosophila
Display the alignment of UniGene sequences with their possible translational products using ProtEST. ProtEST uses BLASTx to compare UniGene mRNA and EST sequences with protein sequences from eight organisms, obtained from RefSeq and the structural databases, recording the best match for each case. More...
Hot Spots
Cancer genome anatomy project
Clusters of orthologous groups
Coffee Break
Electronic PCR
Gene expression omnibus
Genes and disease
Human genome resources
Human map viewer
Human/mouse homology maps
LocusLink
Malaria genetics & genomics
ORF finder

GenBank Summary Example



GenBank Detail Browsing Example

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMI

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard

Cubby: not logged in

Display default Save Text Add to Clipboard

1: AA045112. zk63d03.s1 Soares...[gi:1523314] PubMed, Taxonomy, UniSTS, LinkOut

IDENTIFIERS

dbEST Id: **660769**

EST name: zk63d03.s1

GenBank Acc: AA045112

GenBank gi: 1523314

GDB Id: 3761207

CLONE INFO

Clone Id: IMAGE:487493 (3')

Source: IMAGE Consortium, LLNL

Insert length: 416

DNA type: cDNA

PRIMERS

Sequencing: -40M13 fwd. from Amersham

PolyA Tail: Unknown

SEQUENCE

CACCTGGAGAAACTTCTGCACTGGCACTGTGTTCCNAGAGCTCCTTCTATGCGTCCCTCC

CAAGTGATTTAATTTCAGCTGATTGGACTACGAATTCACAAGGCAGAAAAGTCAAGGTCA

TTTGGNATCTGGAGACAGGAGAACTCAAGGAACCNAAAGGACT

Quality: High quality sequence stops at base: 105

Entry Created: Sep 19 1996

Last Updated: May 11 1997

COMMENTS

This clone is available royalty-free through LLNL ; contact the IMAGE Consortium (info@image.llnl.gov) for further information.

PUTATIVE ID Assigned by submitter

gb:M81635 ERYTHROCYTE BAND 7 INTEGRAL MEMBRANE PROTEIN (HUMAN);

LIBRARY

Lib Name: Soares_pregnant_uterus_NbHPU

Organism: Homo sapiens

Sex: female

Organ: uterus

Develop. stage: adult

Lab host: DH10B

Vector: pT7T3-Pac

R. Site 1: Not I

R. Site 2: Eco RI

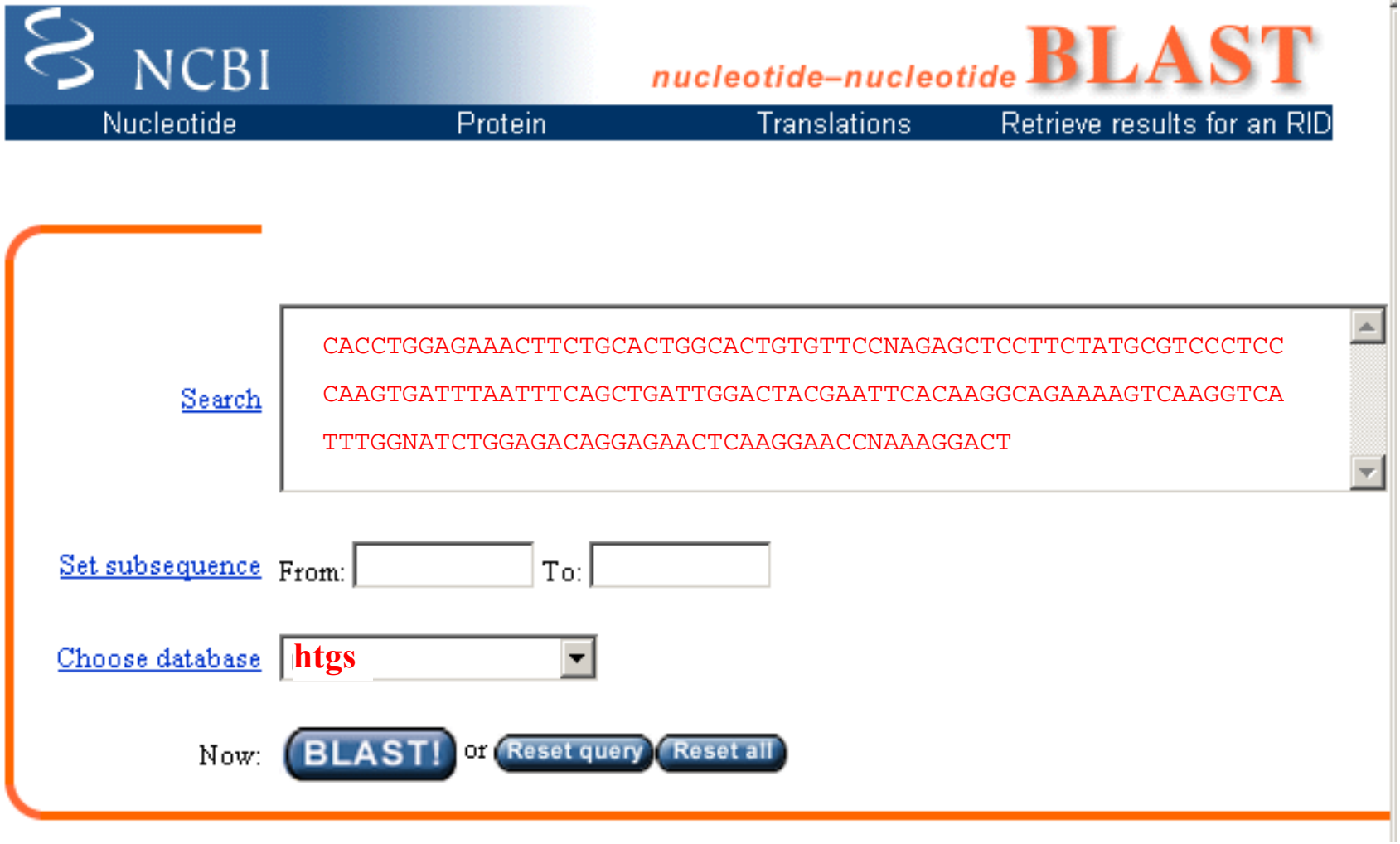
GenBank Detail Data

```
<sequence>

CACCTGGAGAAACTTCTGCACTGGCACTGTGTTCCNAGAGCTCCTTCTATGCGTCCCTCC
CAAGTGATTTAATTTCAGCTGATTGGACTACGAATTCACAAGGCAGAAAAGTCAAGGTCA
TTTGGNATCTGGAGACAGGAGAACTCAAGGAACCNAAAGGACT

</sequence>
```

BLAST Example



BlastResponse Browsing Example

Your request has been successfully submitted and put into the Blast Queue.

Query = (163 letters)

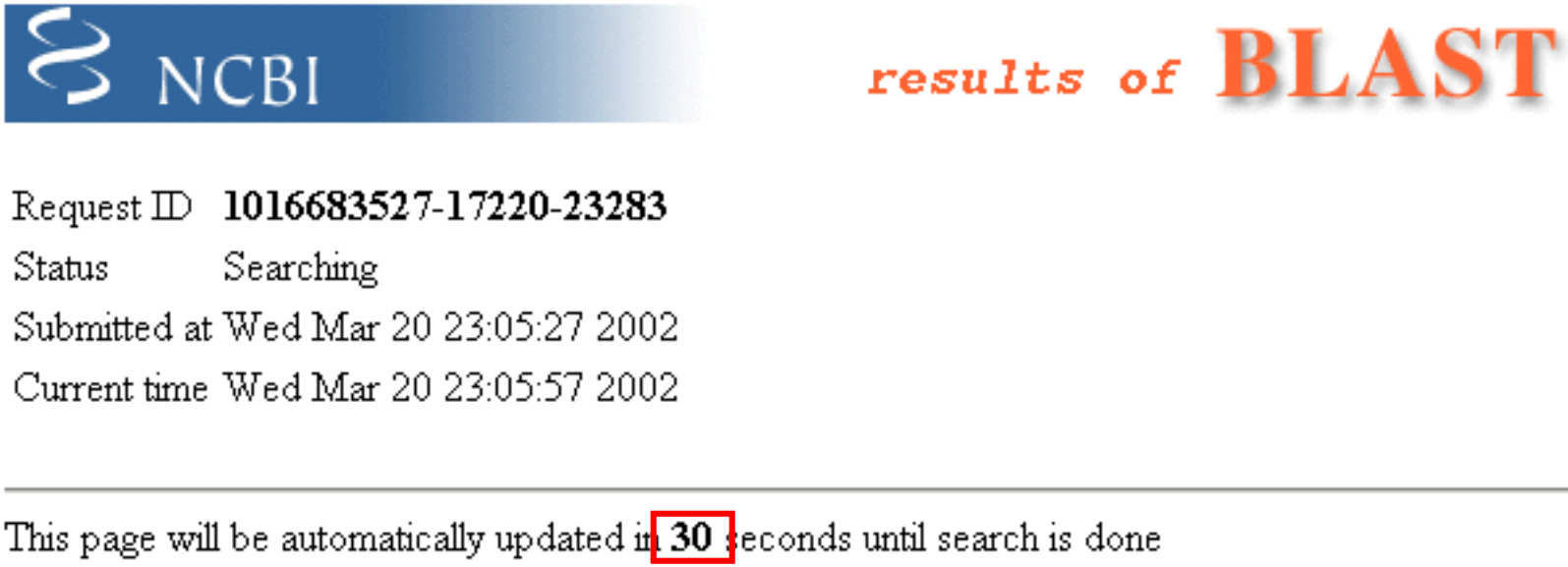
The request ID is 1016683527-17220-23283

Format! or Reset all

The results are estimated to be ready in 24 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

BlastDelay Browsing Example



BLAST Summary Example

NCBI *results of* **BLAST**

BLASTN 2.2.2 [Dec-14-2001]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1016683527-17220-23283

Query=
(163 letters)

Database: Unfinished High Throughput Genomic Sequences;
Sequences: phases 0,1 and 2
47,855 sequences; 5,955,708,580 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

Taxonomy reports

Distribution of 11 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

Sequences producing significant alignments:		Score (bits)	E Value
gi|7630668|gb|AC011969.3|AC011969	Homo sapiens chromosome 4...	297	2e-78
gi|18701932|gb|AC111982.1|	Rattus norvegicus clone CH230-12...	38	2.8
gi|18860211|gb|AC109888.2|	Rattus norvegicus clone CH230-31...	38	2.8
gi|18643512|gb|AC110160.1|	Mus musculus clone RP23-66A20, L...	38	2.8
gi|18369970|gb|AC108124.1|	Homo sapiens chromosome 5 clone ...	38	2.8

Alignments

```
>gi|7630668|gb|AC011969.3|AC011969  Homo sapiens chromosome 4 clone RP11-520J8 :
            SEQUENCE, 13 unordered pieces
          Length = 193168

 Score =  297 bits (150), Expect = 2e-78
 Identities = 156/159 (98%)
 Strand = Plus / Minus


Query: 1      cacctggagaaacttctgcactggcactgtgttccnagagctccttctatgcgtccctcc 60
               |||||||||||||||||||||||||||||||||| |||||||||||||||||||||||
Sbjct: 10906 cacctggagaaacttctgcactggcactgttcctagagctccttctatgcgtccctcc 10847


Query: 61     caagtgatttaatttcagctgattggactacgaattcacaaggcagaaaagtcaaggtca 120
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10846 caagtgatttaatttcagctgattggactacgaattcacaaggcagaaaagtcaaggtca 10787


Query: 121    tttggnatctggagacaggagaactcaaggaaccnaaag 159
               ||||| |||||||||||||||||||||||||| ||||
Sbjct: 10786 tttggtatctggagacaggagaactcaaggaaccaaaag 10748
```

```
>gi|18701932|gb|AC111982.1|  Rattus norvegicus clone CH230-122A17, *** SEQUENCI
            53 unordered pieces
          Length = 120045

 Score = 38.2 bits (19), Expect = 2.8
 Identities = 19/19 (100%)
 Strand = Plus / Minus


Query: 98     acaaggcagaaaagtcaag 116
               ||||||||||||||||||
Sbjct: 56116 acaaggcagaaaagtcaag 56098
```

```
>gi|18860211|gb|AC109888.2|  Rattus norvegicus clone CH230-31P8, *** SEQUENCIN
            unordered pieces
          Length = 176267

 Score = 38.2 bits (19), Expect = 2.8
 Identities = 19/19 (100%)
 Strand = Plus / Minus


Query: 5      tggagaaacttctgcactg 23
               ||||||||||||||||||
Sbjct: 42352 tggagaaacttctgcactg 42334
```

BLAST Detail Example

Entrez Nucleotide

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard D

Cubby: not logged in

Display default Save Text Add to Clipboard

1: AC011969. Homo sapiens chro...[gi:7630668]

Taxonomy, UniSTS, LinkOut

```
LOCUS       AC011969             193168 bp    DNA     linear   HTG 21-APR-2000
DEFINITION  Homo sapiens chromosome 4 clone RP11-520J8 map 4, WORKING DRAFT
            SEQUENCE, 13 unordered pieces.
ACCESSION   AC011969
VERSION     AC011969.3  GI:7630668
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT.
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 193168)
  AUTHORS   Birren,B., Linton,L., Nusbaum,C. and Lander,E.
  TITLE     Homo sapiens chromosome 4, clone RP11-520J8
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 193168)
  AUTHORS   Birren,B., Linton,L., Nusbaum,C., Lander,E., Allen,N., Anderson,M.,
            Baldwin,J., Barna,N., Beckerly,R., Boguslavkiy,L., Boukhgalter,B.,
            Brown,A., Castle,A., Colangelo,M., Collins,S., Collymore,A.,
            Cooke,P., DeArellano,K., Dewar,K., Domino,M., Donelan,L., Doyle,M.,
            Ferreira,P., FitzHugh,W., Forrest,C., Funke,R., Gage,D.,
            Galagan,J., Gardyna,S., Grant,G., Hagos,B., Heaford,A., Horton,L.,
```

Need 1000 base upstream

```
                     /note="assembly_fragment"
BASE COUNT     55031 a   40516 c   42876 g   53468 t    1277 others
ORIGIN
        1 tgtaggtgcc tggatgctta agtctttcta actctggggg ttacatattt aggaggccca
       61 gggtttgcag atataagcca cattctccaa tatcagcttt aactataatg aaagtaatat
      121 ttaccaccct ctgccttcct taccccttg tccaatttac caattggttc agattttgga
      181 cagtcaaagg gggtactagt tcttgatctc cattcaaaag ccctaacata aaccccaatg
      241 ttgctactgc cagaccagat acaagggtga tgtgaccctt ggatccactg gggatgcatt
      301 aagcccaaac agcagcactc tcttgtcctc tgtgatacag agaatgagtt atgagatcta
      361 gatctgcctt catgttacat agtcatccta tgagcctccc actagtcact taaaacacca
      421 atccattgtc cacacaccat cattcaaaat actagtccat ccgaatacta gaatgagcaa
      481 cttctgattt aaccttattt taactacttt atcttaaaaa aaggactatg ttagcagtaa
      541 gccacattcg gtggctgagt aaatgtagat gaaaagagaa gataaaaaaa ttaaaagaaa
      601 [illegible] [illegible]taaa gactattttt ttagagaagt tttatgttca cagaaaaatt
      661 gcgccaaaag tacagagttc tcatatacct cctgctccca catttgggac agccccaccc
      721 ctccactatc aacatccctg caccagagtg atacatttgt tacaatcgat gaagctacac
      781 tgacatatca ctatcaccca acatccatag tttacattag ggttcactct tggtgtggta
      841 cattctaagg gtttaacaaa agtataatga catgtatcca ccattatagt atcatacaaa
      901 atagcatcac tgccctaaaa attctctgag ctttgctcat taatccctgc ctccctgaac
      961 aaccctgaac aatcactgac ctttttactg tctccatagt tttacctttt ccagaatgtc
     1021 atataagcac acagcatgta gccttttcag tgggcttctt ttacttagta acatatcttt
     1081 tcatgacttg atannnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
     1141 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnngcggtta
     1201 tgtcacagat tgcagaacag tcatacaatc ggctnggggg gtttatattt tttttatcct
     1261 atggacaggg gaaaggaggg ggtgaactgg agaaaacctc atggggnttg gcaattatga
     1321 aataaagctt ctgtaaatat ccatgtgcag gttttttttgt agacataagt ttttaactca
     1381 tttgggtgaa taccaaggag tacgattgct ggatcgtgtg gtaaaaatat atttagtttt
     1441 gtaagaaact gccaaattgt cttccaaagt ggctgtacca ttttgcattc ctaccagcaa
```

BLAST Interface

Input:

- The URL of BlastN answer page
- QueryString

Output:

```
<BlastN source="NCBi" url="" queryString="">
   <objects desc="">
      <object>
         <genid>gen id</genid>
         <desc>description</desc>
         …
         <topmatchedsequence> matched sequence </topmatchedsequence>
      </object>
   </objects>
</BlastN>
```

Template: BlastN.ext_template

BLAST Composer Script

BlastN.ext_template

```
<XWRAPCOMPOSER:PageComposer wrappername=”BlastN” releaseDate=""
      owner="">
  <XWRAPCOMPOSER:RunPageExtractor name=”NCBiBlastResponse”
      inputurl=”&queryurl;” queryString=”&queryString;”
      extractionTemplate=”NCBiBlastResponse.ext_template”>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">

        <XWRAPCOMPOSER:RunPageExtractor name=”NCBiBlastDelay”
      inputurl=”&BlastCGI;” queryString=”&BlastQueryString;{requestid}”
      extractionTemplate=”NCBiBlastDelay.ext_template”>
          <XWRAPCOMPOSER:style>
            <xsl:template match="/SummaryPageLink/">
            <xsl:when test=”{waitinginterval}">
              <!-- if waitinginverval has a value. -->
                <XWRAPCOMPOSER:Refresh interval=”waitinginterval}” />
                <!-- Refresh will run the page extractor after specified interval
      again.-->
              </xsl:when>
              <xsl:otherwise>
                <XWRAPCOMPOSER:style src=”summaryanddetail.style” />
                <!-- apply the style in the file that src references to. →
              </xsl:otherwise>
            </xsl:template>
          </XWRAPCOMPOSER:style>
        </XWRAPCOMPOSER:RunPageExtractor>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunPageExtractor>
</XWRAPCOMPOSER:PageComposer>
```

summaryAndDetail.style

```
<XWRAPCOMPOSER:style>
  <xsl:template match="/ ">
    <XWRAPCOMPOSER:RunPageExtractor name=”NCBiBlastSummary” inputurl=”{url}”
      queryString=”{queryString}” extractionTemplate=”NCBiBlastSummary.ext_template”>
      <XWRAPCOMPOSER:style>
        <xsl:template match="/">
          <BlastN source=”NCBI” url=”&realurl;” queryString=”&realQueryString;”>
            <objects desc=”&description;”>
              <xsl:for-each select="ResultSet/object/">
                <object>
                  <genid><xsl:value-of select="genid"/></genid>
                  <desc><xsl:value-of select="desc"/></desc>
                  <length><xsl:value-of select="length"/></length>
                  <score><xsl:value-of select="score"/></score>
                  …
<XWRAPCOMPOSER:RunPageExtractor name=”NCBiBlastDetail” inputurl=”{sequencelink}”
      queryString=”” style=”topmatched.style” extractionTemplate=”NCBiBlastDetail.ext_template”
      >
  <XWRAPCOMPOSER:parameters>
    <XWRAPCOMPOSER:para name=”startline” value=”{startline}”/>
    <XWRAPCOMPOSER:para name=”endline” value=”{endline}”/>
  </XWRAPCOMPOSER:parameters>
</XWRAPCOMPOSER:RunPageExtractor>
                </object>
              </xsl:for-each>
            </objects>
          </BlastN>
        </xsl:template>
      </XWRAPCOMPOSER:style>
    </XWRAPCOMPOSER:RunPageExtractor>
  </xsl:template>
</XWRAPCOMPOSER:style>
```

topMatched.style

```
<XWRAPCOMPOSER:style>
  <xsl:template match="/"><topMatchedSequence><xsl:value-of select="ResultSet/object/topMatchedSequence"/> </topMatchedSequence></xsl:template>
</XWRAPCOMPOSER:style>
```

BLAST Wrapped Data

```
<BlastN source="NCBi" url="..." queryString="...">
    <objects desc="...">
      <object>
        <genid>gi|7630668|gb|AC011969.3|AC011969</genid>
        <desc>Homo sapiens chromosome 4 clone RP11-520J8 map 4, WORKING DRAFT SEQUENCE, 13 unordered pieces</desc>
        <length>193168 </length>
        <score>297 bits (150) </score>
        <expect>2e-78</expect>
        <identities>156/159 (98%)</identities>
        <strand>Plus / Minus </strand>
        <topmatchedsequence>
10381 catttgtaac atttcctctt tgagactctg agttcaccta gagaagtcta agcataacag
10441 ctttctttcc cagcacgagc ctttatagct ctctttagct caaccactct gtccatccag
...
11161 ttccctgggg agtttcaaga tccacacaca ccctccacca ccacaaagct ttaactgact
        </topmatchedsequence>
      </object>
      <object>
        <genid>gi|18701932|gb|AC111982.1|</genid>
        <desc>Rattus norvegicus clone CH230-122A17, *** SEQUENCING IN PROGRESS ***, 53 unordered pieces </desc>
        <length>120045 </length>
        <score>38.2 bits (19)</score>
        <expect>2.8</expect>
        <identities>19/19 (100%)</identities>
        <strand>Plus / Minus</strand>
        <topmatchedsequence>
55681 cccatgtcga aggttcccag catcctgcca catccctctt tcttccttct gcatgctttc
55741 tccatctcct tagtctgctt ggatgtgatt acagagcttt tgcatcagct ctgtcggaat
...
56521 agaaagtaac tggagaaagt tctgtttgtc tccttcgtta gagcatgagt gcatttgcta
        </topmatchedsequence>
      </object>
...
    </objects>
</BlastN>
```

BlastResponse Interface

Input:

A URL

Output:

<requestid>…</requestid>

Template: BlastNResponse.ext_template

BlastResponse Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”BlastNResponse” releaseDate=""
    owner="">
  <XWRAPCOMPOSER:RunKeywordExtraction>
    <XWRAPCOMPOSER:variables>
      <XWRAPCOMPOSER:variable name=”requestid” BeginMatch=”The request ID is
    &lt;input name=&quot;RID&quot; size=&quot;50&quot; type=&quot;text&quot;
    value=&quot;” EndMatch=”&quot;&gt;” />
      <!-- &quot; refers to a character of quote. →
    </XWRAPCOMPOSER:variables>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <requestid><xsl:value-of select="requestid"/>
        </requestid>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunKeywordExtraction>
</XWRAPCOMPOSER:PageExtractor>
```

BlastResponse Wrapped Data

<requestid>

1016683527-17220-23283

</requestid>

BlastDelay Interface

Input: A URL and a QueryString

Output:

<SummaryPageLink>
  <url>summaryurl</url>
  <queryString>queryString</queryString>
  <wait>waitinginterval</wait>
</SummaryPageLink>

Template: BlastNDelay.ext_template

BlastDelay Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”BlastNDelay” releaseDate="" owner="">
  <XWRAPCOMPOSER:RunKeywordExtraction>
    <XWRAPCOMPOSER:variables>
      <XWRAPCOMPOSER:variable name=”waitinginterval” BeginMatch=”This page will be
      automatically updated in &lt;b&gt;” EndMatch=”&lt;/b&gt; seconds until search is done” />
      <!-- &quot; refers to a character of quote. →
    </XWRAPCOMPOSER:variables>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <SummaryPageLink>
          <url>&realurl;</url>
          <queryString>&realQueryString;</queryString>
          <wait><xsl:value-of select="waitinginterval"/></wait>
        </SummaryPageLink>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunKeywordExtraction>
</XWRAPCOMPOSER:PageExtractor>
```

BlastDelay Wrapped Data

<SummaryPageLink>
<url>
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
</url>
<queryString>
ALIGNMENT=50…&RID= 1016683527-17220-23283&…CMD=get
</queryString>
<wait>30</wait>
</SummaryPageLink>

BLAST Summary Interface

Input:

<url>The URL of BLAST Summary Page</url>

Output:

```
<ResultSet>
  <object>
    <genid>gen id</genid>
    <sequencelink>sequencelink</sequencelink>
    <desc>description</desc>
    <length>length</length>
    <score>score </score>
    <expect>expect</expect>
    <identities> identities </identities>
    <strand> strand </strand>
    <startline>startline</startline>
    <endline>endline</endline>
  </object>
</ResultSet>
```

Template: NCBiBlastSummary.ext_template

BLAST Summary Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”NCBiBlastSummary” releaseDate="" owner="">
  <XWRAPCOMPOSER:RunXWRAPEliteExtraction>
    <XWRAPCOMPOSER:parameters>
      <XWRAPCOMPOSER:para name=”tagElementSeparator” value=”td”/>
      …
    </XWRAPCOMPOSER:parameters>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <ResultSet>
          <xsl:for-each select=”ResultSet/object”>
          <object>
            <genid><xsl:value-of select="element1"/></genid>
            ...
          </object>
          </xsl:for-each>
        </ResultSet>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunXWRAPEliteExtraction>
</XWRAPCOMPOSER:PageExtractor>
```

BLAST Summary Data

```
<ResultSet>
  <object>
    <genid>gi|7630668|gb|AC011969.3|AC011969</genid>
     <sequencelink>http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=07630668&
     dopt=GenBank</sequencelink>
    <desc>Homo sapiens chromosome 4 clone RP11-520J8 map 4, WORKING DRAFT SEQUENCE, 13 unordered
     pieces</desc>
    <length>193168 </length>
    <score>297 bits (150) </score>
    <expect>2e-78</expect>
    <identities>156/159 (98%)</identities>
    <strand>Plus / Minus </strand>
    <startline>10906</startline>
    <endline>10786</endline>
  </object>
  <object>
     ...
  </object>
  ...
</ResultSet>
```

BLAST Detail Interface

Input:

The URL of The BLAST Detail Page

Output:

```
<ResultSet>
   <object>
      <topMatchedSequence>Topmatched Sequence
      </topMatchedSequence>
   </object>
</ResultSet>
```

Template:NCBiBlastNDetail.ext_template

BLAST Detail Script

```
<XWRAPCOMPOSER:PageExtractor wrappername=”NCBiBlastDetail” releaseDate="" owner="">
  <XWRAPCOMPOSER:RunLineExtraction>
    <XWRAPCOMPOSER:parameters>
      <XWRAPCOMPOSER:para name=”startline” value=”&startline;”/>
      <XWRAPCOMPOSER:para name=”endline” value=”&endline”/>
      <!—the two parameter values should be obtained from input/environments. →
    </XWRAPCOMPOSER:parameters>
    <XWRAPCOMPOSER:style>
      <xsl:template match="/">
        <ResultSet>
          <object>
            <topmatchedsequence>
              <xsl:for-each select=”ResultSet/object/”>
              <!—the output of lineextraction would be ResultSet/object/line/content. →
                <xsl:value-of select="line"/>
              </xsl:for-each>
            </topmatchedsequence>
          </object>
        </ResultSet>
      </xsl:template>
    </XWRAPCOMPOSER:style>
  </XWRAPCOMPOSER:RunLineExtraction>
</XWRAPCOMPOSER:PageExtractor>
```

BLAST Detail Data

<ResultSet source="NCBI_Blastn" search_seq="AA045112", …>
  <object>
    <topmatchedsequence>
      <sequence_id>AC011969 </sequence_id>
      <seq_fragment_matched>
        10381 catttgtaac atttcctctt tgagactctg agttcaccta gagaagtcta agcataacag
        10441 ctttctttcc cagcacgagc ctttatagct ctctttagct caaccactct gtccatccag
        10501 ccaatggatg tcccttcccc tgtaccccaa tttcaagctt attttaggaa gccttgaact
        10561 accatgtatc ctggctccta gctgagttta ttagaggtat ggagcagtgc aacttaaact
        10621 caagttgcac ttacattttg aattttaaaa tgatggtttt atctgttgtg tgaagtggtt
        10681 caccccttgag gaccaggagc ctccatatcc tgactgaaaa ccttttctga gacttagagt
        10741 aacagtactt ttggttcctt gagttctcct gtctccagat accaaatgac cttgactttt
        10801 ctgccttgtg aattcgtagt ccaatcagct gaaattaaat cacttgggag ggacgcatag
        10861 aaggagctct aggaacacag tgccagtgca gaagtttctc caggtggcct ccctttccaa
        10921 caatgtacat aataaagtgt atgcactttc actaatattt ttggggtgag agtctgtttc
        10981 ggcctgtatt gaatgtctgt ggatttccgt tttcagaagt agtacattag atcctccggt
        11041 tctgagctgg ctggttggct tcttcgtgtg ctttgtgggc caggggaagg ggacaggctg
        11101 ctgtgggcca tctgctgtct cccaggtcca ggcaccctct ggtgcactgg cccacacatt
        11161 ttccctgggg agtttcaaga tccacacaca ccctccacca ccacaaagct ttaactgact
      <seq_fragment_matched>
    </topmatchedsequence>
  </object>
  <object>
    …..
  </object>
  ……
</ResultSet>

Questions?