Computational Biology Research Center, AIST

Active Workflow Component Type

User Manual

CBRC

2012/11/14

Contents

1 Introduction

This manual describes Active workflow Component type developed at Computational Biology Research Center, Advanced Industrial Science and Technology (AIST).

For the installation of Active workflow Component type please refer to the installation manual available in Life Science Database Integration Web site.

Life Science Database Integration Web : http://togo.cbrc.jp/

The Active workflows run on KNIME platform.
Please refer to the KNIME site for the details of KNIME.
This manual explains how the user can work with Active workflows.

KNIME : http://www.knime.org/

2 About the Active workflow Component type

There are nine Active workflow combination types available, which are listed in the table below.

2-1 Active workflow component type list

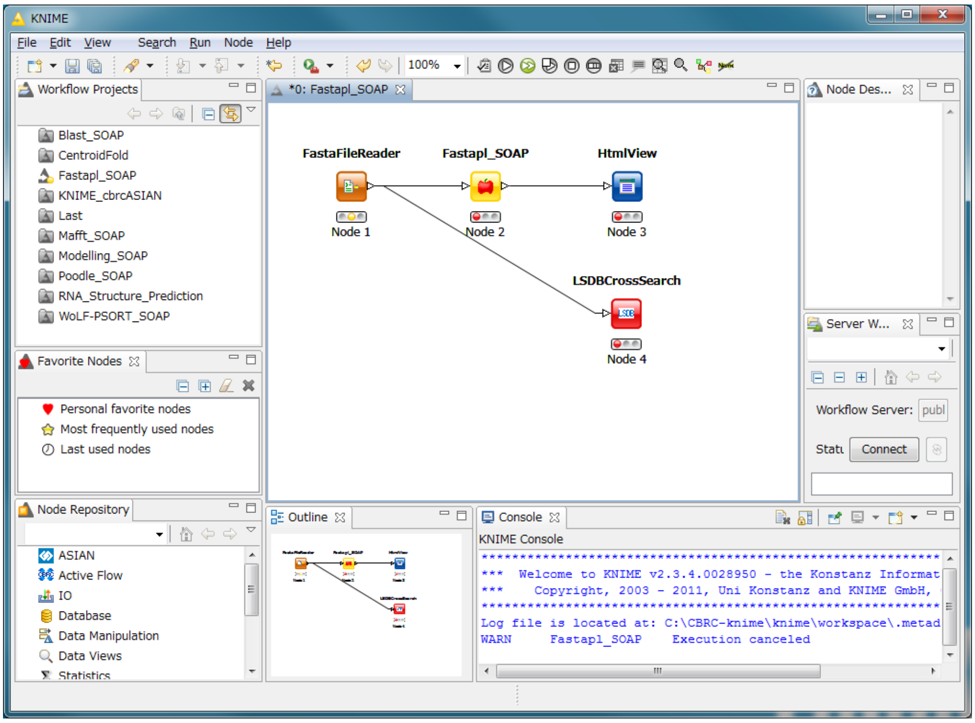
No.	Active workflow component type name	OS	Explanation
1	Fastapl Active Workflow	Windows 32bit	Workflow that performs sequence processing of FASTA form file
2	Mafft Active Workflow	Windows 32bit	Workflow that performs multiple alignments.
3	Blast Active Workflow	Windows 32bit	Workflow that performs homology search.
4	Last Active Workflow	Windows 32bit	Workflow that performs sequence comparison.
5	WolfPSORT Active Workflow	Windows 32bit	Workflow that predicts localization in cell from amino-acid sequence
6	Modelling Active Workflow	Windows 32bit	Workflow that performs homology modeling from amino-acid sequence.
7	CentroidFold Active Workflow	Windows 32bit	Workflow that predicts secondly structure from the RNA sequence.
8	POODLE Active Workflow	Windows 32bit	Workflow that predicts disorder area from amino-acid sequence
9	ASIAN Active Workflow	Windows 32bit Linux	Integrated analytical workflow using gene network inferring system.
10	AutoDock Active Workflow	Windows 32bit	Chemical compounds – protein docking workflow.

3 Common rules

Ccommon rules in all Active workflows are as follows.

1. Starting Active workflow

   Double-click on the workflow the user will use in Workflow Projects column after KNIME starts. The workflow is then shown and ready to use.

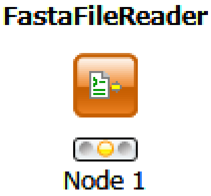


```
****************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Informat
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH,
****************************************************************
Log file is located at: C:\CBRC-knime\knime\workspace\.metad
WARN      Fastapl_SOAP    Execution canceled
```

3-1 Fastapl Active workflow (example)

2. Node

   A node is an icon that is shown in a workflow screen as follows;



3-2　Fasta File Reader Node (example)

When the node is selected, the explanation of each node is displayed in the "Node Description" column at the right of the KNIME screen.

3. Node progress

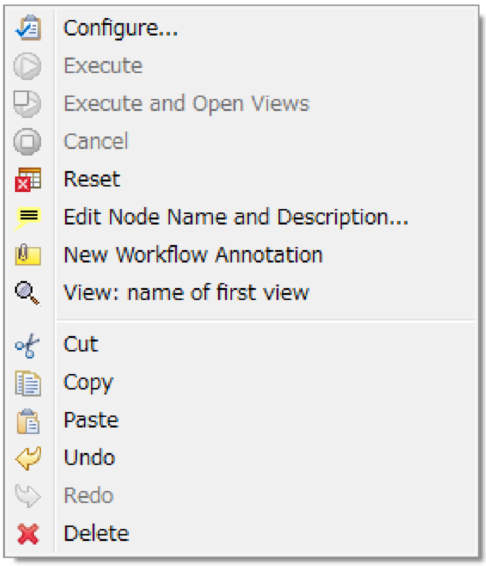
Signals below a node indicate progress as shown below.

3-3 Signal of Node progress list

signal color	color	Progress message
🔴●●	Red	Preparing execution
●🟡●	Yellow	Stand-by
●●🟢	Green	Complete
🟦	Thick blue	Executing
queued	queued	Queued

4. Node menu

A node menu is shown when right-clicking on a node as shown below.



3-4 Node menu

3-5 Node menu list

Menu command	Action	Note
Configure…	Various settings of node.	Another window is started.
Execute	Execute the node.	The node cannot be used unless the node status is yellow.
Execute and Open Views	It is an active display for the node that displays the result window. Execute a node.	The node cannot be used unless the node status is yellow.
Cancel	Cancel the execution.	The node cannot be used unless the node status is deep blue.
Reset	The setting is reset.	If the node status is green the node is active.
Edit Node Name and Description…	Use to change the node name or Description.	Another window is started.
New Workflow Annotation	Use to insert some comment.	The comment column is displayed.
View : [viewer name]	Use to display results.	Another window is started.
Cut	The node and the comment, etc. are cut.	-
Copy	The node and the comment, etc. are copied.	-
Paste	The node and the comment, etc., which are copied, are pasted.	-
Undo	Use to undo cut, copy or paste.	-
Redo	Use to cancel the action undone.	-
Delete	The node and the comment,	-

	etc. are deleted.	

5. **Execute all executable nodes**

   When all the configurations of nodes complete, all the nodes can be executed at a time.

   In that case, click on the icon in the top of the KNIME screen (shown below) after selecting the node, which is a starting point. (Execute all executable nodes (Shift+F7))

3-5 Execute all executable nodes

6. **Alert messages and Error messages**

   If an alert or an error occurred after a node is executed, a pop-up screen will appear along with messages in Console of KNIME screen. Those should be checked to resolve problems.

   Examples of the messages and measures are shown as follows:

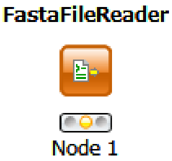
3-6 Alert messages：sample

No	Messages	Cause and method of settlement
1	Console： WARN　FastaFileReader 0:2:1 failed to apply settings: Please specify a filename.	Cause： The file is not specified. Method of settlement： Specify the file.
2	Pop up： SOAP execution error. Please resubmit again later. Console： ERROR CentroidFold_SOAP Execute failed: Error occurred.	Cause： An error occurred when SOAP is executed. Measures： Execute it again later.

7. **Operation for specifying a file or a directory in node configuration**

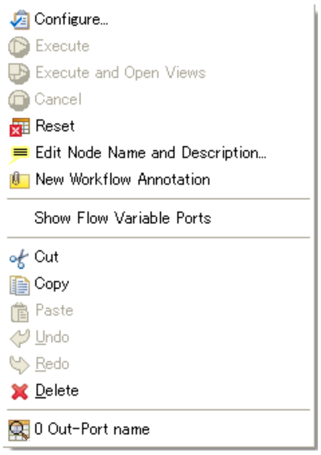
   In many nodes, a file or a directory needs to be specified as an input or an output directory. Please specify as follows:

   1) Select the icon of a node, followed by right-clicking. A menu appears.



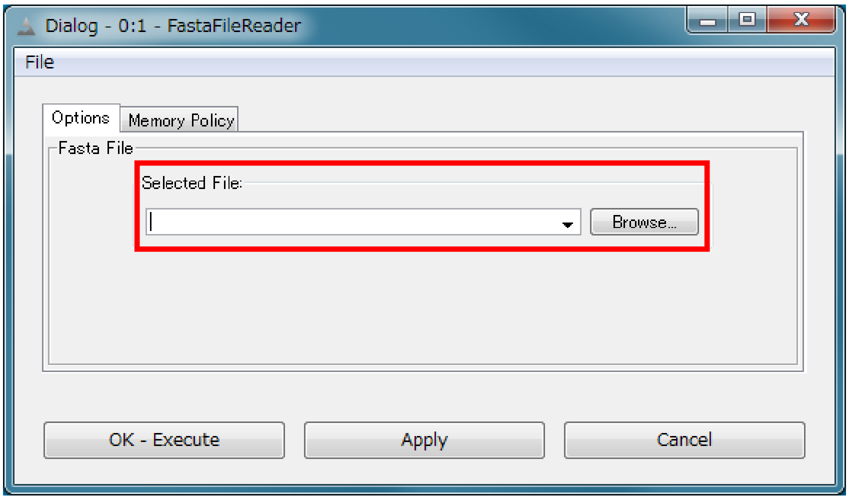
3.7 FastaFileReader Icon (example)

2) Select "Configure" from the menu.



3.8 right-click-menu

3) Select a file or a directory using “Brows” in the pop-up dialog.



3.9 FastaFileReader : Configure…

Press “OK” after selecting.

4 Use of each Active workflow

Usage of each Active workflow is explained below.

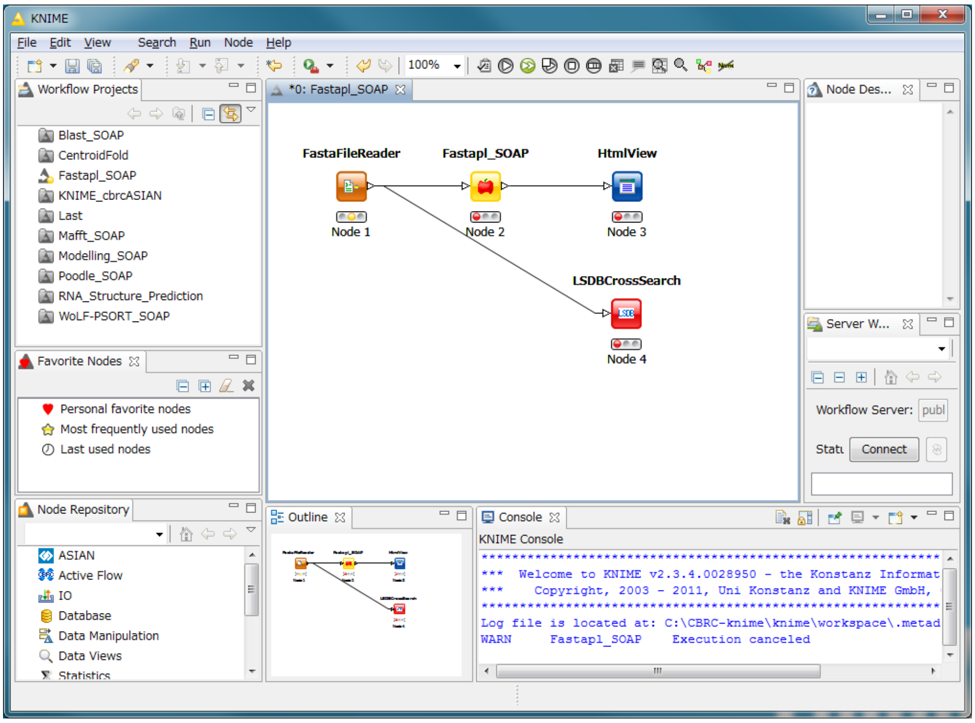
4.1 Fastapl Active Workflow

Fastapl Active Workflow performs sequence processing.

Please refer to the following sites for the explanation and the usage example of fastapl/fastqpl.

fastapl/fasqpl : http://seq.cbrc.jp/fastapl

Furthermore, this workflow can retrieve variety of related information by using node LSDBCrossSearch that performs Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



```
*********************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Informat
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH,
*********************************************************************
Log file is located at: C:\CBRC-knime\knime\workspace\.metad
WARN    Fastapl_SOAP    Execution canceled
```

4.1-1 Fastapl Active Workflow

4.1.1 Preparation

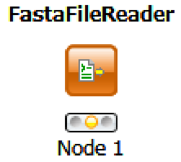
A file needed for execution is a sequence file in FASTA format. Multi-FASTA format can also be used.

File type
(Multi-)FASTA format

4.1.2 Node

There are 4 nodes.

4.1.2-1 Fastapl Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader		The FASTA format file is read.
Node 2	Fastapl_SOAP		fastapl/fastqpl executes.
Node 3	HtmlView		The prediction result is displayed.
Node4	LSDBCrossSearch		Execute LSDB cross-search.

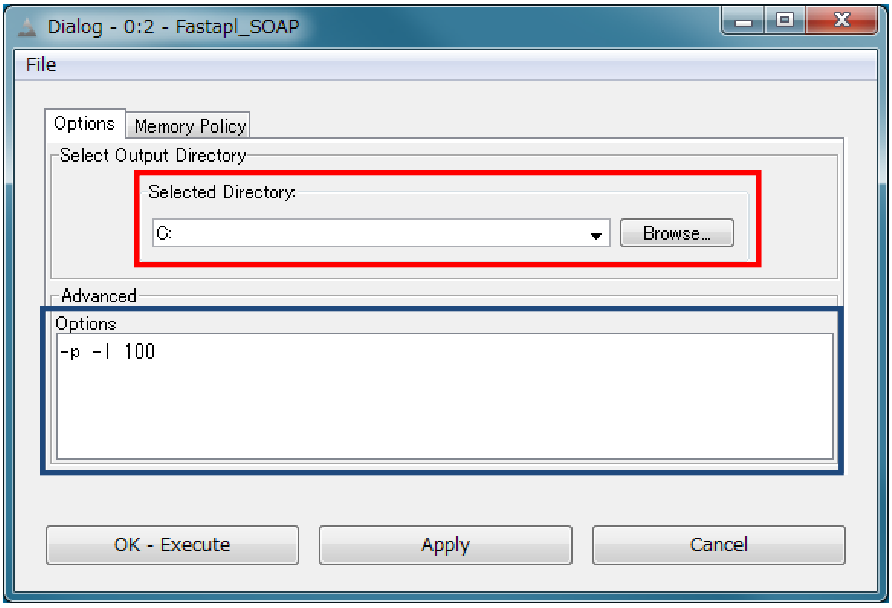
4.1.3 Step 1. Node setting

1. Node1　: FastaFileReader

   Select a FASTA file as an input using right-click-menu.

2. Node2　: Fastapl_SOAP

   Select an output directory using right-click-menu and set options if necessary.



Select Output Directory

Selected Directory: C:

Advanced

Options

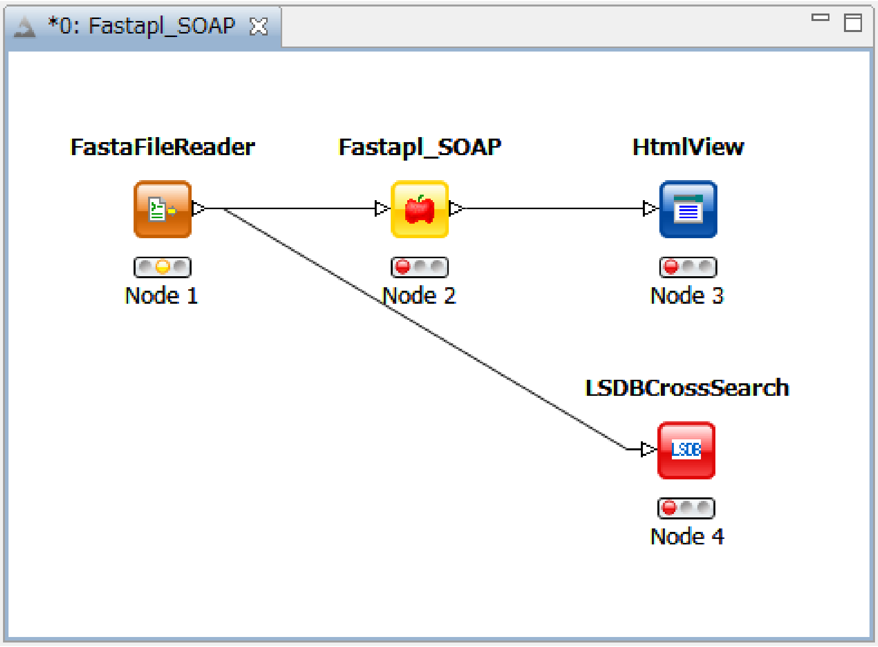
```
-p -l 100
```

4.1.3-1 Fastapl_SOAP　: Configure…

・Options tab → Advanced → Options

The default options are “-p –l 100” meaning that sequence length of the FASTA file will be adjusted to 100 characters a line.

4.1.4 Step 2. Execution



4.1.4-1 Fastapl_SOAP all Nodes

1) FastaFileReader

Select “Execute” in the right-click-menu for execution.

2) Fastapl_SOAP

Select “Execute” in the right-click-menu for execution.

3) HtmlView

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

4) LSDBCrossSearch

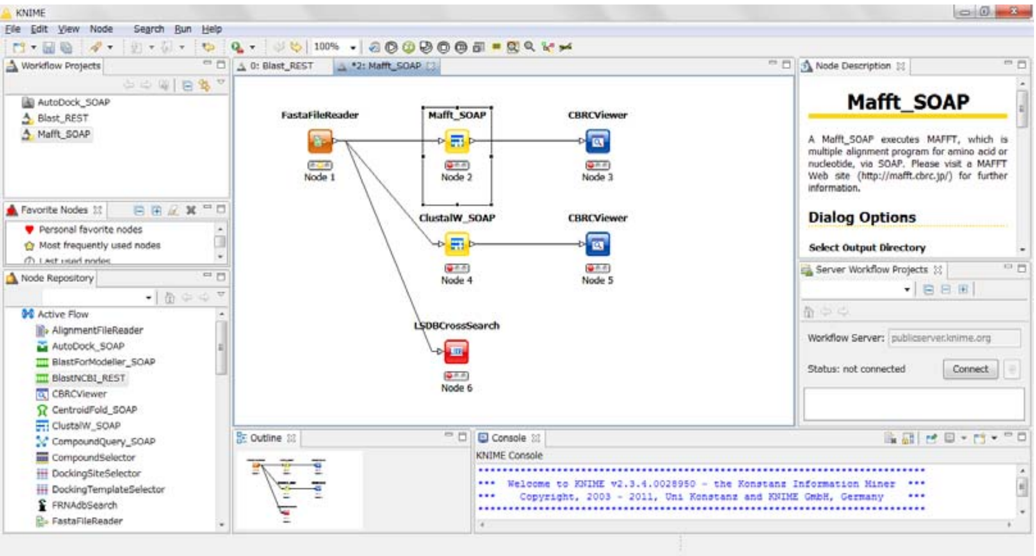
Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to the following "5.1 Appendix A : LSDBCrossSearch " for the use of the result screen.

4.2 Mafft Active Workflow

Mafft Active Workflow performs multiple alignment for nucleic acid sequences or of amino-acid sequences via SOAP. It uses ClustalW (http://www.clustal.org/) or MAFFT (http://mafft.cbrc.jp/).

This workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



Mafft_SOAP

A Mafft_SOAP executes MAFFT, which is multiple alignment program for amino acid or nucleotide, via SOAP. Please visit a MAFFT Web site (http://mafft.cbrc.jp/) for further information.

Dialog Options

Select Output Directory

Workflow Server: publicserver.knime.org

Status: not connected

KNIME Console

```
**************************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Miner  ***
***     Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany    ***
**************************************************************************
```

4.2-1 Mafft Active Workflow

4.2.1 Preparation

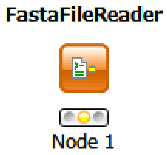
A file needed for execution is a Multi-FASTA format file containing base sequences or amino-acid sequences in FASTA format.

File type
Multi-FASTA format

4.2.2 Node

There are 6 nodes.

4.2.2-1 Mafft Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader	 FastaFileReader Node 1	The FASTA format file is read.
Node 2	Mafft_SOAP	 Mafft_SOAP Node 2	Execute Mafft.
Node 3	CBRCViewer	 CBRCViewer Node 3	The multiple alignment result is displayed.
Node4	ClustalW_SOAP	 ClustalW_SOAP Node 4	Execute ClustalW.

Node5	CBRCViewer	CBRCViewer Node 5	The multiple alignment result is displayed.
Node6	LSDBCrossSearch	LSDBCrossSearch LSDB Node 6	Execute LSDB cross-search.

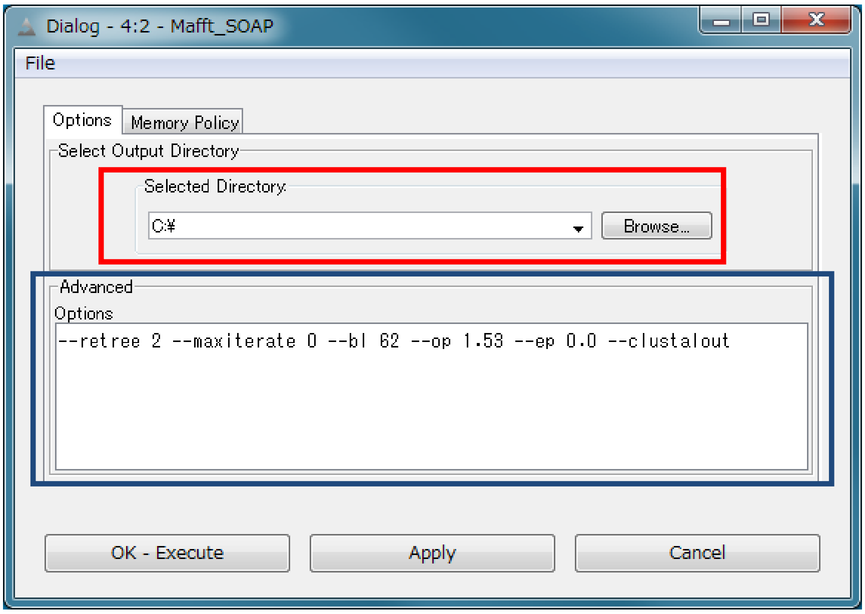
4.2.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a Multi-FASTA file as an input using right-click-menu.

2. Node2 : Mafft_SOAP

   Select an output directory using right-click-menu and set options if necessary.



Selected Directory: C:¥

Options:

```
--retree 2 --maxiterate 0 --bl 62 --op 1.53 --ep 0.0 --clustalout
```

4.2.3-1 Mafft_SOAP : Configure...

・Options tab → Advanced → Options

Options are explained below.

--op # : Gap opening penalty, default: 1.53

--ep # : Offset (works like gap extension penalty), default: 0.0

--maxiterate # : Maximum number of iterative refinement, default: 0

--clustalout : Output: clustal format, default: fasta

--reorder : Outorder: aligned, default: input order

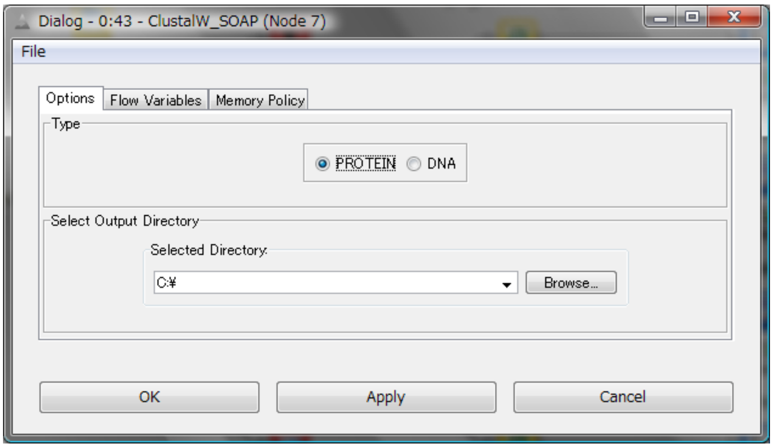
--quiet : Do not report progress

The default options are as follows.

```
--retree 2 --maxiterate 0 --bl 62 --op 1.53 --ep 0.0 --clustalout
```

3. <u>Node4 : ClustalW_SOAP</u>

Specify an absolute path of a directory to store ClustalW results, or select the output directory using "Browse…" button.



Dialog - 0:43 - ClustalW_SOAP (Node 7)

Type: ● PROTEIN ○ DNA

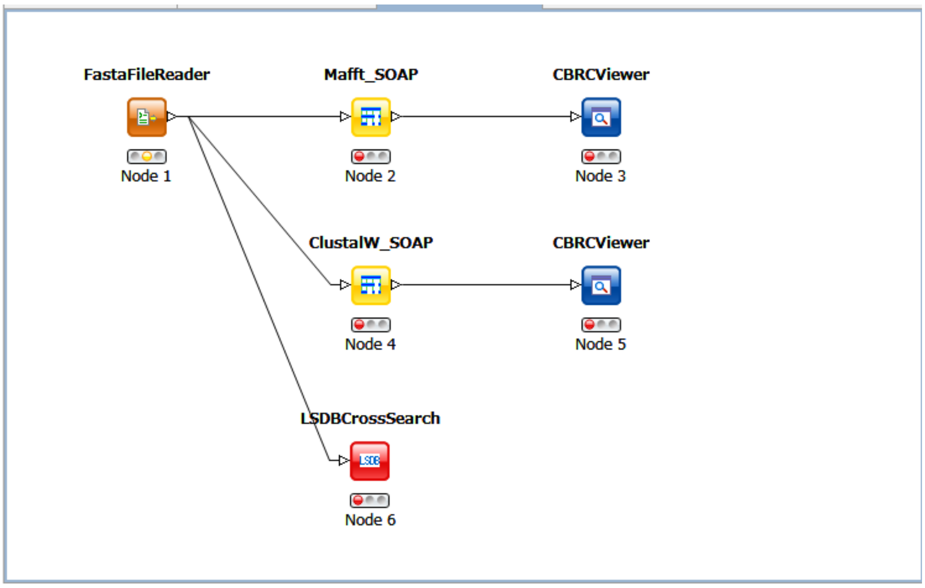
Select Output Directory

Selected Directory: C:¥

4.2.3-2 Mafft_SOAP : Configure…

Specify "PROTEIN" (for protein sequences) or "DNA" (for nucleic acid sequences) radio button.

4.2.4 Step2. Execution



4.2.4-1 Mafft_SOAP Node

Mafft or ClustalW can be selected.

1) Node1 : FastaFileReader

   Select "Execute" in the right-click-menu for execution.

2) Node2 : Mafft_SOAP

   Select "Execute" in the right-click-menu for execution.

3) Node3 : CBRCViewer

   Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

4) Node4 : ClustalW_SOAP

   Select "Execute" in the right-click-menu for execution.

5) Node5 : CBRCViewer

   Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

6) Node6 : LSDBCrossSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

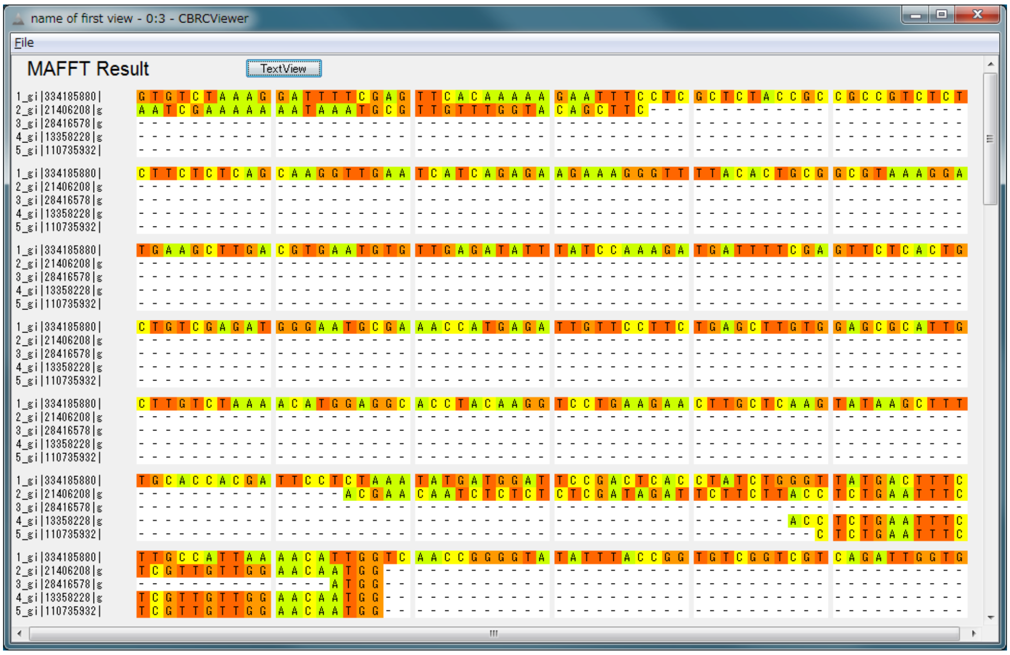
Please refer to the following "5.1 Appendix A : LSDBCrossSearch " for the use of the result screen.

4.2.5 Step.3 Result viewing

1) Node3 CBRCViewer – Mafft Result

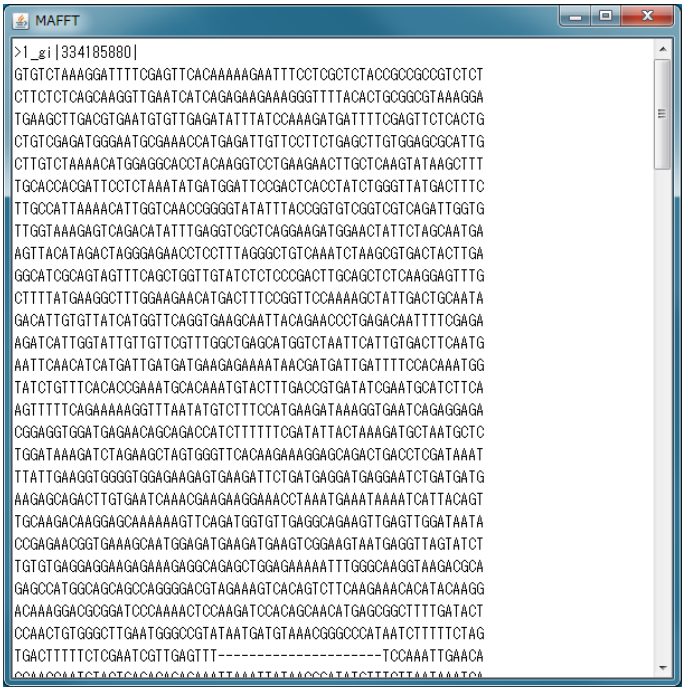
The sequence identifier used for the input is displayed on the left. The aligned sequence is shown on the right.

A text version of the results is shown by pressing "TextView" button.



```
1_gi|334185880|   GTGTCTAAAG GATTTTCGAG TTCACAAAAA GAATTTCCTC GCTCTACCGC CGCCGTCTCT
2_gi|21406208|g   AATCGAAAAA AATAAATGCG TTGTTTGGTA CAGCTTC--- ---------- ----------
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------- ----------

1_gi|334185880|   CTTCTCTCAG CAAGGTTGAA TCATCAGAGA AGAAAGGGTT TTACACTGCG GCGTAAAGGA
2_gi|21406208|g   ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------- ----------

1_gi|334185880|   TGAAGCTTGA CGTGAATGTG TTGAGATATT TATCCAAAGA TGATTTTCGA GTTCTCACTG
2_gi|21406208|g   ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------- ----------

1_gi|334185880|   CTGTCGAGAT GGGAATGCGA AACCATGAGA TTGTTCCTTC TGAGCTTGTG GAGCGCATTG
2_gi|21406208|g   ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------- ----------

1_gi|334185880|   CTTGTCTAAA ACATGGAGGC ACCTACAAGG TCCTGAAGAA CTTGCTCAAG TATAAGCTTT
2_gi|21406208|g   ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------- ----------

1_gi|334185880|   TGCACCACGA TTCCTCTAAA TATGATGGAT TCCGACTCAC CTATCTGGGT TATGACTTTC
2_gi|21406208|g   ---------- -----ACGAA CAATCTCTCT CTCGATAGAT TCTTCTTACC TCTGAATTTC
3_gi|28416578|g   ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g   ---------- ---------- ---------- ---------- -------ACC TCTGAATTTC
5_gi|110735932|   ---------- ---------- ---------- ---------- ---------C TCTGAATTTC

1_gi|334185880|   TTGCCATTAA AACATTGGTC AACCGGGGTA TATTTACCGG TGTCGGTCGT CAGATTGGTG
2_gi|21406208|g   TCGTTGTTGG AACAATGG-- ---------- ---------- ---------- ----------
3_gi|28416578|g   ---------- ----ATGG-- ---------- ---------- ---------- ----------
4_gi|13358228|g   TCGTTGTTGG AACAATGG-- ---------- ---------- ---------- ----------
5_gi|110735932|   TCGTTGTTGG AACAATGG-- ---------- ---------- ---------- ----------
```

4.2.5-1 Node3 CBRCViewer – MAFFT Result



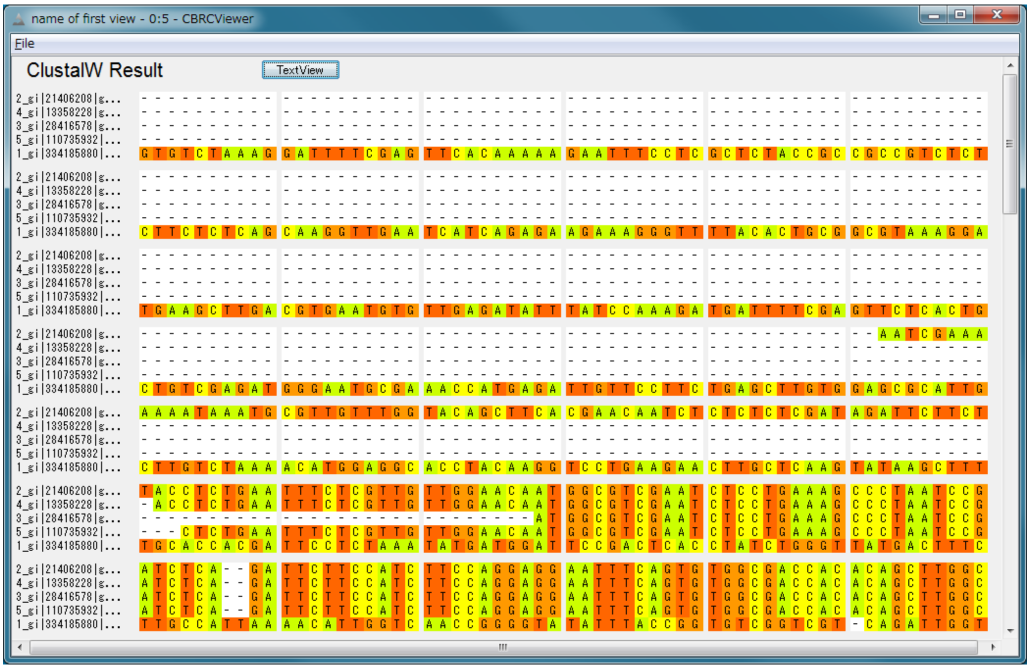
```
>1_gi|334185880|
GTGTCTAAAGGATTTTCGAGTTCACAAAAAGAATTTCCTCGCTCTACCGCCGCCGTCTCT
CTTCTCTCAGCAAGGTTGAATCATCAGAGAAGAAAGGGTTTTACACTGCGGCGTAAAGGA
TGAAGCTTGACGTGAATGTGTTGAGATATTTATCCAAAGATGATTTTCGAGTTCTCACTG
CTGTCGAGATGGGAATGCGAAACCATGAGATTGTTCCTTCTGAGCTTGTGGAGCGCATTG
CTTGTCTAAAACATGGAGGCACCTACAAGGTCCTGAAGAACTTGCTCAAGTATAAGCTTT
TGCACCACGATTCCTCTAAATATGATGGATTCCGACTCACCTATCTGGGTTATGACTTTC
TTGCCATTAAAACATTGGTCAACCGGGGTATATTTACCGGTGTCGGTCGTCAGATTGGTG
TTGGTAAAGAGTCAGACATATTTGAGGTCGCTCAGGAAGATGGAACTATTCTAGCAATGA
AGTTACATAGACTAGGGAGAACCTCCTTTAGGGCTGTCAAATCTAAGCGTGACTACTTGA
GGCATCGCAGTAGTTTCAGCTGGTTGTATCTCTCCCGACTTGCAGCTCTCAAGGAGTTTG
CTTTTATGAAGGCTTTGGAAGAACATGACTTTCCGGTTCCAAAAGCTATTGACTGCAATA
GACATTGTGTTATCATGGTTCAGGTGAAGCAATTACAGAACCCTGAGACAATTTTCGAGA
AGATCATTGGTATTGTTGTTCGTTTGGCTGAGCATGGTCTAATTCATTGTGACTTCAATG
AATTCAACATCATGATTGATGATGAAGAGAAAATAACGATGATTGATTTTCCACAAATGG
TATCTGTTTCACACCGAAATGCACAAATGTACTTTGACCGTGATATCGAATGCATCTTCA
AGTTTTTCAGAAAAAGGTTTAATATGTCTTTCCATGAAGATAAAGGTGAATCAGAGGAGA
CGGAGGTGGATGAGAACAGCAGACCATCTTTTTTCGATATTACTAAAGATGCTAATGCTC
TGGATAAAGATCTAGAAGCTAGTGGGTTCACAAGAAAGGAGCAGACTGACCTCGATAAAT
TTATTGAAGGTGGGGTGGAGAAGAGTGAAGATTCTGATGAGGATGAGGAATCTGATGATG
AAGAGCAGACTTGTGAATCAAACGAAGAAGGAAACCTAAATGAAATAAAATCATTACAGT
TGCAAGACAAGGAGCAAAAAAGTTCAGATGGTGTTGAGGCAGAAGTTGAGTTGGATAATA
CCGAGAACGGTGAAAGCAATGGAGATGAAGATGAAGTCGGAAGTAATGAGGTTAGTATCT
TGTGTGAGGAGGAAGAGAAAGAGGCAGAGCTGGAGAAAAATTTGGGCAAGGTAAGACGCA
GAGCCATGGCAGCAGCCAGGGGACGTAGAAAGTCACAGTCTTCAAGAAACACATACAAGG
ACAAAGGACGCGGATCCCAAAACTCCAAGATCCACAGCAACATGAGCGGCTTTTGATACT
CCAACTGTGGGCTTGAATGGGCCGTATAATGATGTAAACGGGCCCATAATCTTTTTCTAG
TGACTTTTTCTCGAATCGTTGAGTTT---------------------TCCAAATTGAACA
```

4.2.5-2 MAFFT Result – TextView

2) Node5 CBRCViewer –ClustalW Result

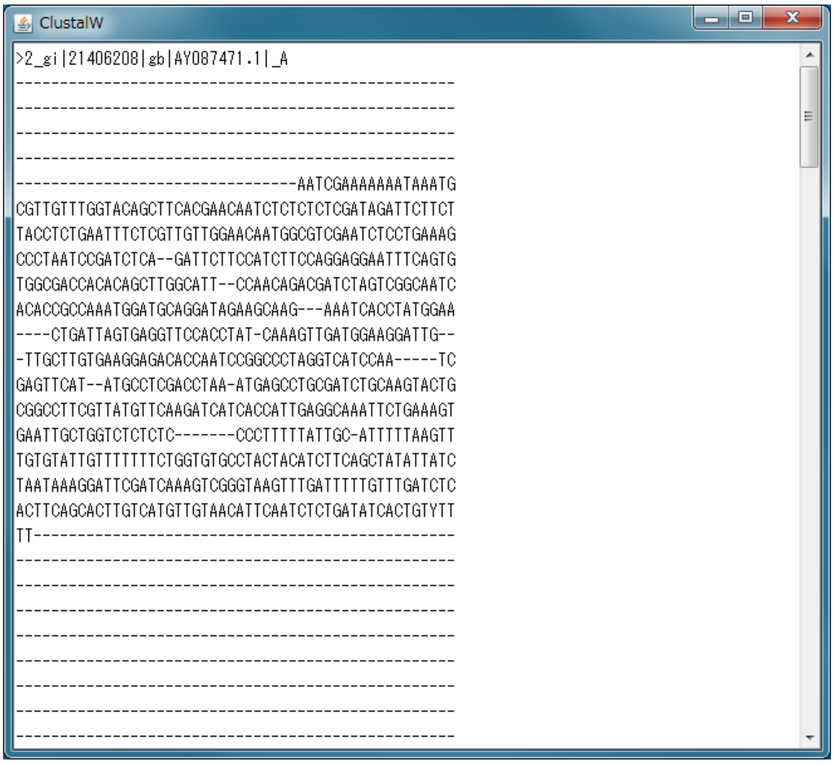
The sequence identifier used for the input is displayed on the left. The aligned sequence is shown on the right.

A text version of the results is shown by pressing "TextView" button.



```
2_gi|21406208|g...  ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g...  ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g...  ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|...  ---------- ---------- ---------- ---------- ---------- ----------
1_gi|334185880|...  GTGTCTAAAG GATTTTCGAG TTCACAAAAA GAATTTCCTC GCTCTACCGC CGCCGTCTCT

2_gi|21406208|g...  ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g...  ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g...  ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|...  ---------- ---------- ---------- ---------- ---------- ----------
1_gi|334185880|...  CTTCTCTCAG CAAGGTTGAA TCATCAGAGA AGAAAGGGTT TTACACTGCG GCGTAAAGGA

2_gi|21406208|g...  ---------- ---------- ---------- ---------- ---------- ----------
4_gi|13358228|g...  ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g...  ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|...  ---------- ---------- ---------- ---------- ---------- ----------
1_gi|334185880|...  TGAAGCTTGA CGTGAATGTG TTGAGATATT TATCCAAAGA TGATTTTCGA GTTCTCACTG

2_gi|21406208|g...  ---------- ---------- ---------- ---------- ---------- --AATCGAAA
4_gi|13358228|g...  ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g...  ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|...  ---------- ---------- ---------- ---------- ---------- ----------
1_gi|334185880|...  CTGTCGAGAT GGGAATGCGA AACCATGAGA TTGTTCCTTC TGAGCTTGTG GAGCGCATTG

2_gi|21406208|g...  AAAATAAATG CGTTGTTTGG TACAGCTTCA CGAACAATCT CTCTCTCGAT AGATTCTTCT
4_gi|13358228|g...  ---------- ---------- ---------- ---------- ---------- ----------
3_gi|28416578|g...  ---------- ---------- ---------- ---------- ---------- ----------
5_gi|110735932|...  ---------- ---------- ---------- ---------- ---------- ----------
1_gi|334185880|...  CTTGTCTAAA ACATGGAGGC ACCTACAAGG TCCTGAAGAA CTTGCTCAAG TATAAGCTTT

2_gi|21406208|g...  TACCTCTGAA TTTCTCGTTG TTGGAACAAT GGCGTCGAAT CTCCTGAAAG CCCTAATCCG
4_gi|13358228|g...  -ACCTCTGAA TTTCTCGTTG TTGGAACAAT GGCGTCGAAT CTCCTGAAAG CCCTAATCCG
3_gi|28416578|g...  ---------- ---------- --------AT GGCGTCGAAT CTCCTGAAAG CCCTAATCCG
5_gi|110735932|...  ---CTCTGAA TTTCTCGTTG TTGGAACAAT GGCGTCGAAT CTCCTGAAAG CCCTAATCCG
1_gi|334185880|...  TGCACCACGA TTCCTCTAAA TATGATGGAT TCCGACTCAC CTATCTGGGT TATGACTTTC

2_gi|21406208|g...  ATCTCA--GA TTCTTCCATC TTCCAGGAGG AATTTCAGTG TGGCGACCAC ACAGCTTGGC
4_gi|13358228|g...  ATCTCA--GA TTCTTCCATC TTCCAGGAGG AATTTCAGTG TGGCGACCAC ACAGCTTGGC
3_gi|28416578|g...  ATCTCA--GA TTCTTCCATC TTCCAGGAGG AATTTCAGTG TGGCGACCAC ACAGCTTGGC
5_gi|110735932|...  ATCTCA--GA TTCTTCCATC TTCCAGGAGG AATTTCAGTG TGGCGACCAC ACAGCTTGGC
1_gi|334185880|...  TTGCCATTAA AACATTGGTC AACCGGGGTA TATTTACCGG TGTCGGTCGT -CAGATTGGT
```

4.2.5-3 Node5 CBRCViewer – ClustalW Result



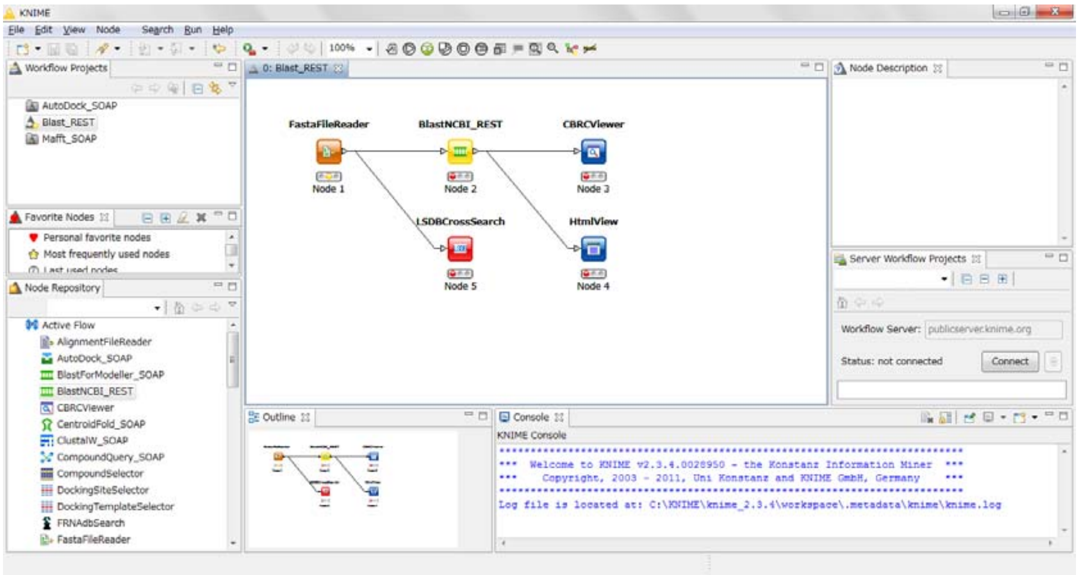
```
>2_gi|21406208|gb|AY087471.1|_A
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------AATCGAAAAAAATAAATG
CGTTGTTTGGTACAGCTTCACGAACAATCTCTCTCTCGATAGATTCTTCT
TACCTCTGAATTTCTCGTTGTTGGAACAATGGCGTCGAATCTCCTGAAAG
CCCTAATCCGATCTCA--GATTCTTCCATCTTCCAGGAGGAATTTCAGTG
TGGCGACCACACAGCTTGGCATT--CCAACAGACGATCTAGTCGGCAATC
ACACCGCCAAATGGATGCAGGATAGAAGCAAG---AAATCACCTATGGAA
----CTGATTAGTGAGGTTCCACCTAT-CAAAGTTGATGGAAGGATTG--
-TTGCTTGTGAAGGAGACACCAATCCGGCCCTAGGTCATCCAA-----TC
GAGTTCAT--ATGCCTCGACCTAA-ATGAGCCTGCGATCTGCAAGTACTG
CGGCCTTCGTTATGTTCAAGATCATCACCATTGAGGCAAATTCTGAAAGT
GAATTGCTGGTCTCTCTC-------CCCTTTTTATTGC-ATTTTTAAGTT
TGTGTATTGTTTTTTTCTGGTGTGCCTACTACATCTTCAGCTATATTATC
TAATAAAGGATTCGATCAAAGTCGGGTAAGTTTGATTTTTGTTTGATCTC
ACTTCAGCACTTGTCATGTTGTAACATTCAATCTCTGATATCACTGTYTT
TT------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
```

4.2.5-4 ClustalW Result – TextView

4.3 Blast Active Workflow

Blast Active Workflow performs homologue search via REST.
The result of BlastNCBI_REST can be viewed using CBRCViewerNode.

This workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



Workflow Server: publicserver.knime.org
Status: not connected

KNIME Console

```
*************************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Miner  ***
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany    ***
*************************************************************************
Log file is located at: C:\KNIME\knime_2.3.4\workspace\.metadata\knime\knime.log
```

4.3-1 Blast Active Workflow

4.3.1 Preparation

A file needed for execution is a file containing a nucleic acid sequence/amino acid sequence in FASTA format.

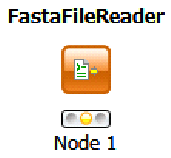
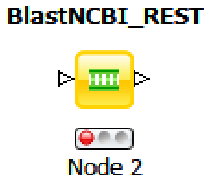
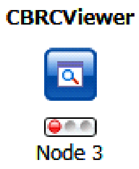
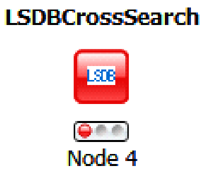
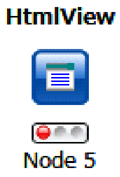
※ Multi-FASTA format cannot be used.

File type
FASTA Format file

4.3.2 Node

There are 5 nodes.

4.3.2-1 Blast Active Workflow Node list

Node ID	Node name	Icon	Explanation
Node 1	FastaFileReader	 FastaFileReader Node 1	The FASTA format file is read.
Node 2	BlastNCBI_REST	 BlastNCBI_REST Node 2	Execute Blast.
Node 3	CBRCViewer	 CBRCViewer Node 3	The Blast execution result is graphically displayed.
Node4	LSDBCrossSearch	 LSDBCrossSearch Node 4	Execute LSDB cross-search.
Node5	HtmlView	 HtmlView Node 5	The Blast execution result is displayed in text.

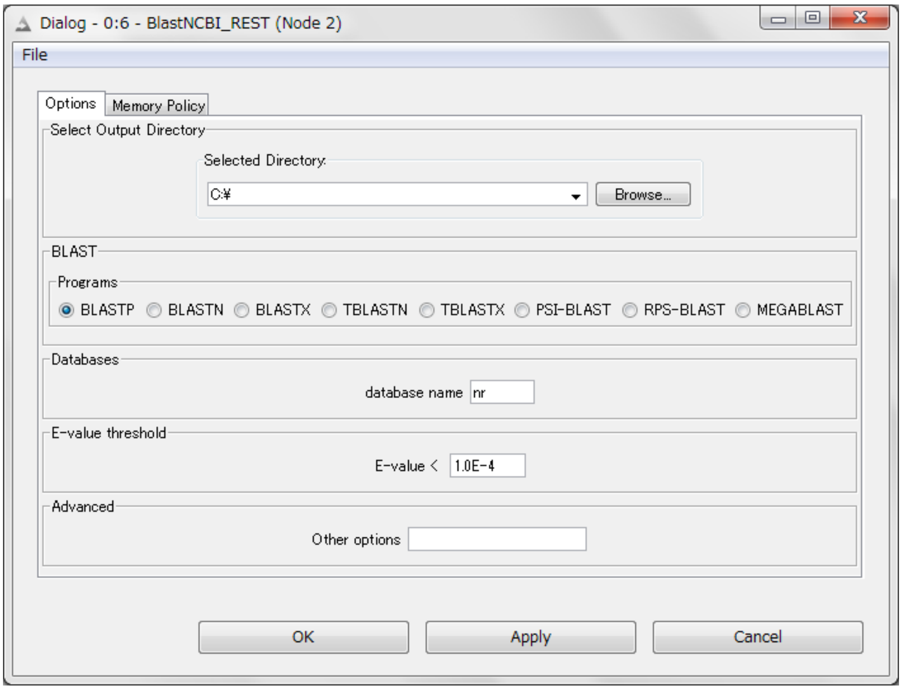
4.3.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a FASTA file as an input using right-click-menu.

2. Node2 : BlastNCBI_REST

   Specify an absolute path of a directory to store Blast Results, or select the directory using “Browse…” button.



Dialog - 0:6 - BlastNCBI_REST (Node 2)

File

Options | Memory Policy

Select Output Directory

Selected Directory: C:¥ [Browse...]

BLAST

Programs: ● BLASTP ○ BLASTN ○ BLASTX ○ TBLASTN ○ TBLASTX ○ PSI-BLAST ○ RPS-BLAST ○ MEGABLAST

Databases

database name: nr

E-value threshold

E-value < 1.0E-4

Advanced

Other options:

[OK] [Apply] [Cancel]

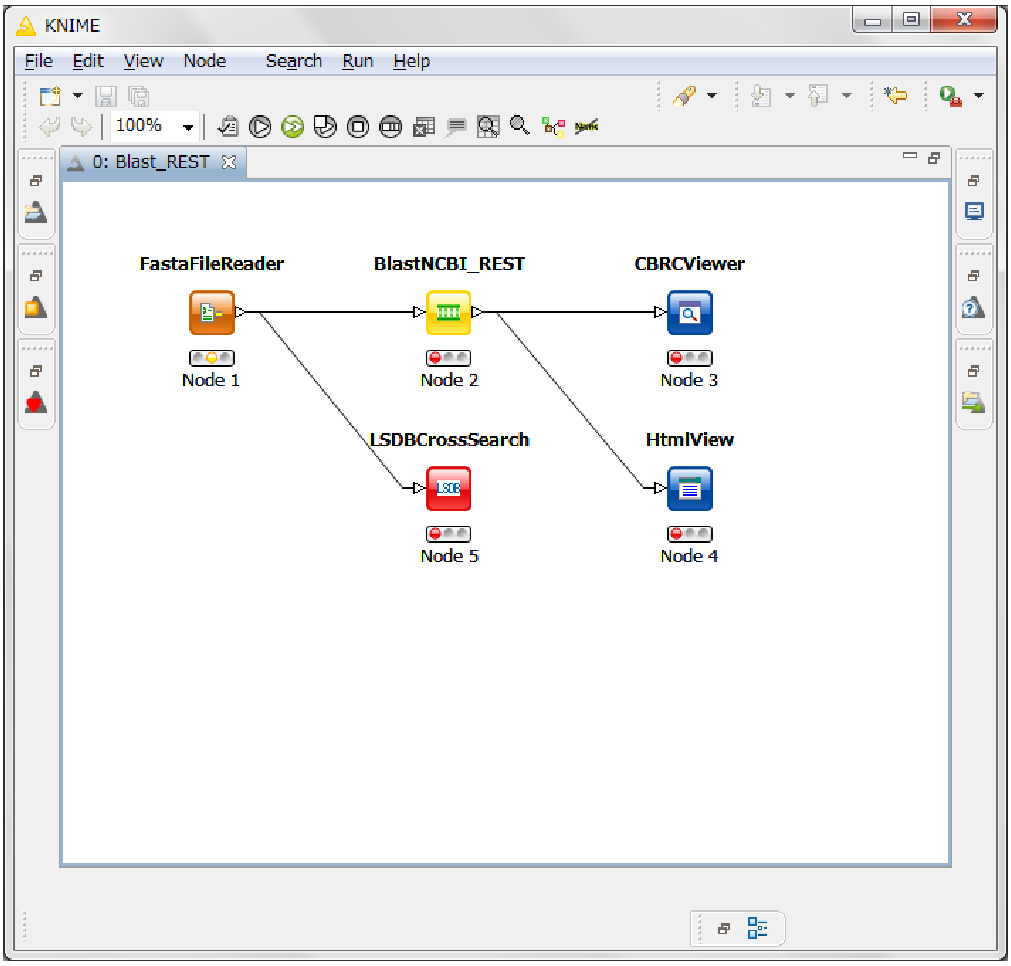
4.3.3-1 BlastNCBI_REST : Configure…

・Options tab → BLAST → Programs

Specify “Programs” (default: BLASTP), “Databases” (default: nr), “E-value Threshold” (default:1.0e-4), and “Advanced” (default: empty).

Please check a BlastNCBI_REST node description for further information.

4.3.4 Step2. Execution



4.3.4-1 Blast_REST Node

1) Node1 : FastaFileReader

   Select "Execute" in the right-click-menu for execution.

2) Node2 : BlastNCBI_REST

   Select "Execute" in the right-click-menu for execution.

3) Node3 : CBRCViewer

   Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

4) Node4 : LSDBCrossSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to the following "5.1 Appendix A：LSDBCrossSearch " for the use of the result screen.

5) Node5　: HtmlView

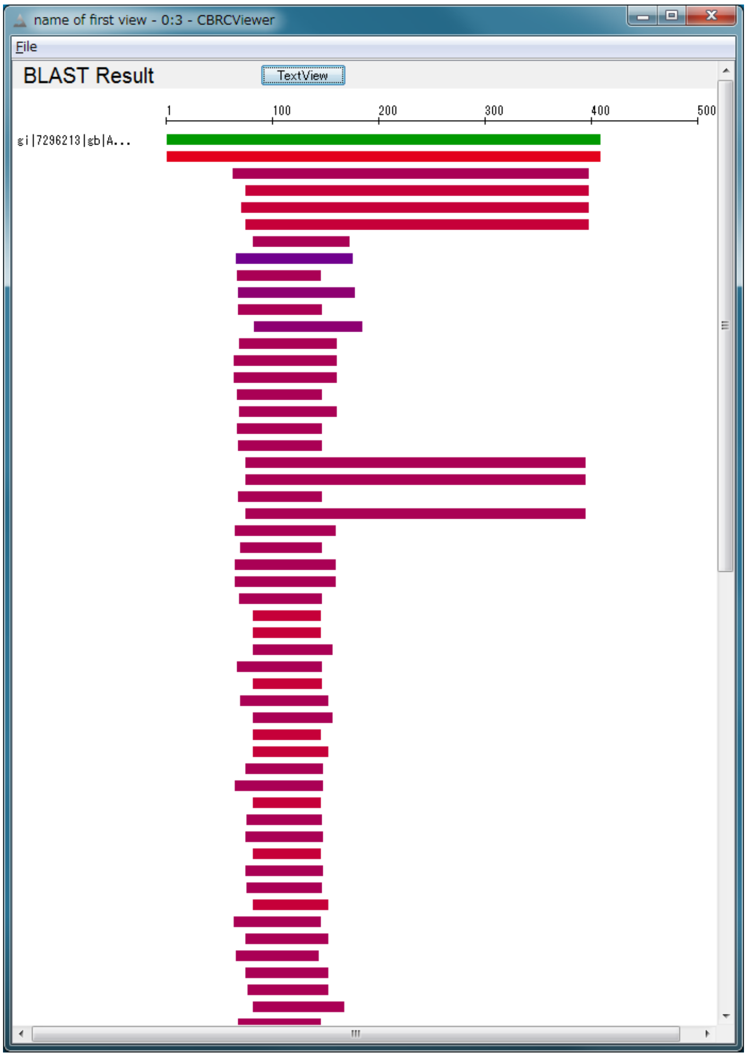
Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

4.3.5 Step.3 result viewing

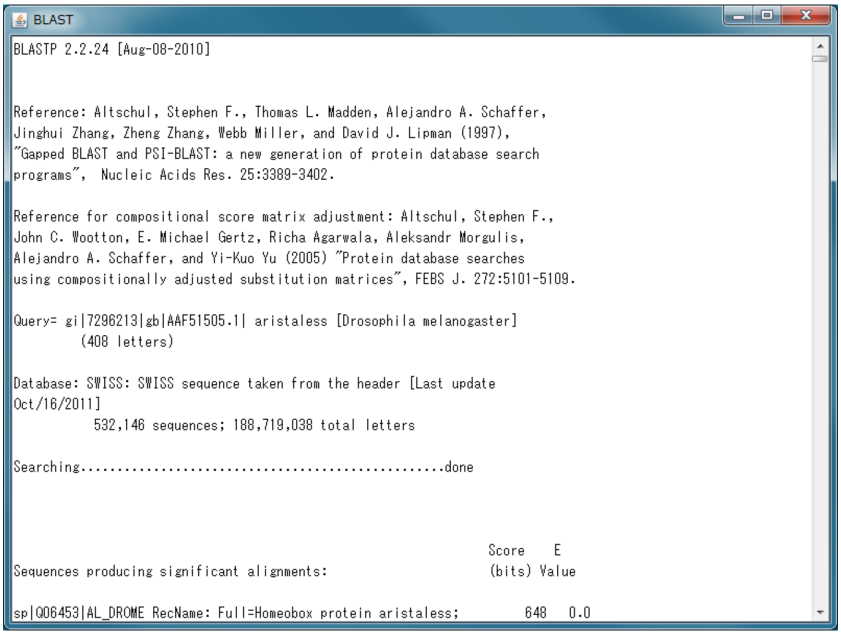
1) Node3 CBRCViewer – BLAST Result

The execution result of BlastNCBI_REST can be viewed as BLAST Result.

A text version of the results is shown by pressing "TextView" button.



4.3.5-1 Node3 CBRCViewer – BLAST Result

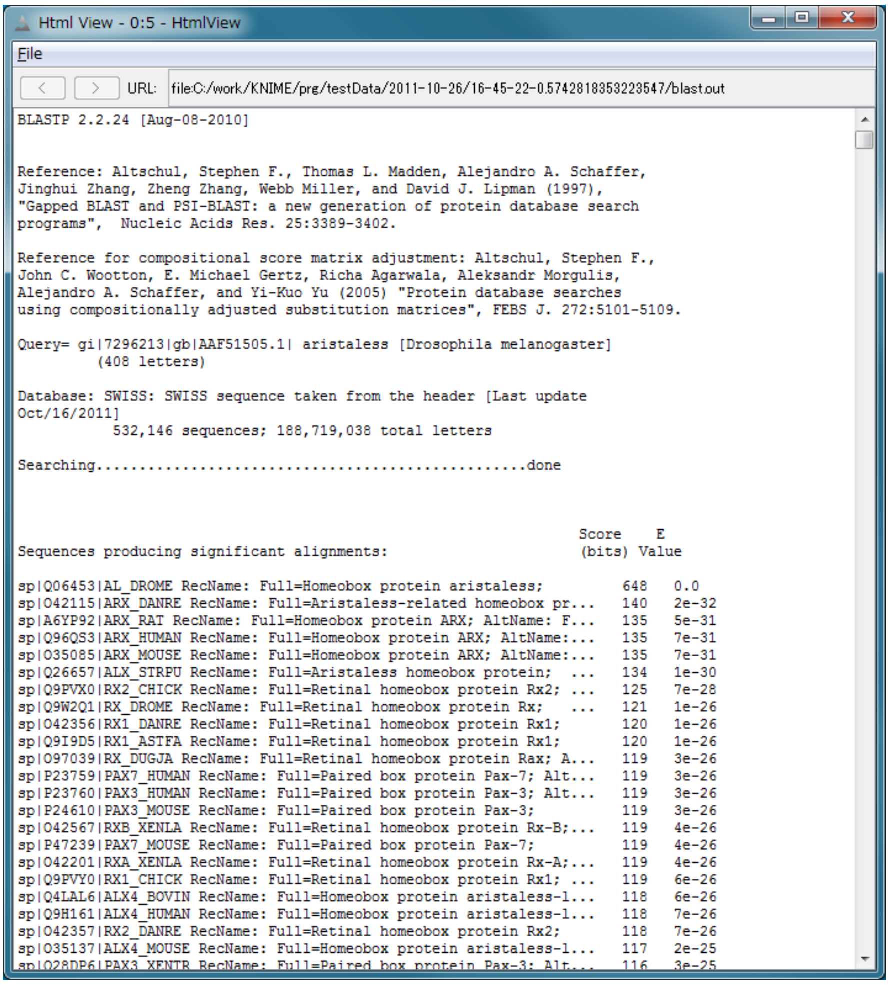


```
BLASTP 2.2.24 [Aug-08-2010]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
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.

Reference for compositional score matrix adjustment: Altschul, Stephen F.,
John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis,
Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches
using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

Query= gi|7296213|gb|AAF51505.1| aristaless [Drosophila melanogaster]
         (408 letters)

Database: SWISS: SWISS sequence taken from the header [Last update
Oct/16/2011]
           532,146 sequences; 188,719,038 total letters

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



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

sp|Q06453|AL_DROME RecName: Full=Homeobox protein aristaless;        648   0.0
```

4.3.5-2 BLAST Result – TextView

2) Node5 HtmlView – BLAST Result

The execution result of BlastNCBI_REST can be viewed as follows:.


Html View - 0:5 - HtmlView
File
URL:
file:C:/work/KNIME/prg/testData/2011-10-26/16-45-22-0.5742818353223547/blast.out

```
BLASTP 2.2.24 [Aug-08-2010]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
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.

Reference for compositional score matrix adjustment: Altschul, Stephen F.,
John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis,
Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches
using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

Query= gi|7296213|gb|AAF51505.1| aristaless [Drosophila melanogaster]
         (408 letters)

Database: SWISS: SWISS sequence taken from the header [Last update
Oct/16/2011]
           532,146 sequences; 188,719,038 total letters

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



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

sp|Q06453|AL_DROME RecName: Full=Homeobox protein aristaless;        648   0.0
sp|O42115|ARX_DANRE RecName: Full=Aristaless-related homeobox pr...   140   2e-32
sp|A6YP92|ARX_RAT RecName: Full=Homeobox protein ARX; AltName: F...   135   5e-31
sp|Q96QS3|ARX_HUMAN RecName: Full=Homeobox protein ARX; AltName:...   135   7e-31
sp|O35085|ARX_MOUSE RecName: Full=Homeobox protein ARX; AltName:...   135   7e-31
sp|Q26657|ALX_STRPU RecName: Full=Aristaless homeobox protein;  ...   134   1e-30
sp|Q9PVX0|RX2_CHICK RecName: Full=Retinal homeobox protein Rx2; ...   125   7e-28
sp|Q9W2Q1|RX_DROME RecName: Full=Retinal homeobox protein Rx;   ...   121   1e-26
sp|O42356|RX1_DANRE RecName: Full=Retinal homeobox protein Rx1;       120   1e-26
sp|Q9I9D5|RX1_ASTFA RecName: Full=Retinal homeobox protein Rx1;       120   1e-26
sp|O97039|RX_DUGJA RecName: Full=Retinal homeobox protein Rax; A...   119   3e-26
sp|P23759|PAX7_HUMAN RecName: Full=Paired box protein Pax-7; Alt...   119   3e-26
sp|P23760|PAX3_HUMAN RecName: Full=Paired box protein Pax-3; Alt...   119   3e-26
sp|P24610|PAX3_MOUSE RecName: Full=Paired box protein Pax-3;          119   3e-26
sp|O42567|RXB_XENLA RecName: Full=Retinal homeobox protein Rx-B;...   119   4e-26
sp|P47239|PAX7_MOUSE RecName: Full=Paired box protein Pax-7;          119   4e-26
sp|O42201|RXA_XENLA RecName: Full=Retinal homeobox protein Rx-A;...   119   4e-26
sp|Q9PVY0|RX1_CHICK RecName: Full=Retinal homeobox protein Rx1; ...   119   6e-26
sp|Q4LAL6|ALX4_BOVIN RecName: Full=Homeobox protein aristaless-l...   118   6e-26
sp|Q9H161|ALX4_HUMAN RecName: Full=Homeobox protein aristaless-l...   118   7e-26
sp|O42357|RX2_DANRE RecName: Full=Retinal homeobox protein Rx2;       118   7e-26
sp|O35137|ALX4_MOUSE RecName: Full=Homeobox protein aristaless-l...   117   2e-25
sp|Q28DP6|PAX3_XENTR RecName: Full=Paired box protein Pax-3; Alt...   116   3e-25
```

4.3.5-3 Node5 HtmlView – BLAST Result

4.4 Last Active Workflow

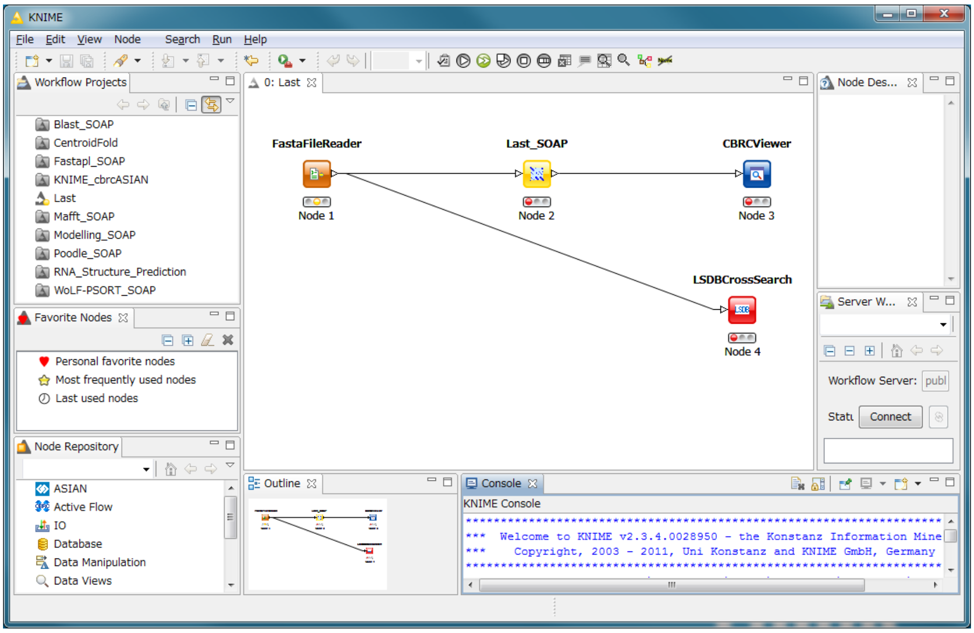
Last Active Workflow performs sequence comparison via SOAP.

The result of Last_SOAP can be viewed using CBRCViewerNode.

Please refer to the following sites for the details of Last.

LAST : http://last.cbrc.jp/

Furthermore, this workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



```
*******************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Mine
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany
*******************************************************************
```

4.4-1 Last Active Workflow

4.4.1 Preparation

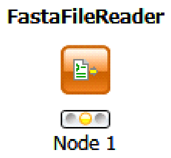
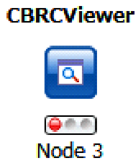
A file needed for execution is a sequence file of nuclear acid/amino acid in FASTA format. Multi-FASTA format can also be used.

File Type
(Multi-)FASTA Format File

4.4.2 Node

There are 4 nodes.

4.4.2-1 Last Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader		The FASTA format file is read.
Node 2	Last_SOAP		Execute Last.
Node 3	CBRCViewer		The Last execution result is graphically displayed.
Node4	LSDBCrossSearch		Execute LSDB cross-search.

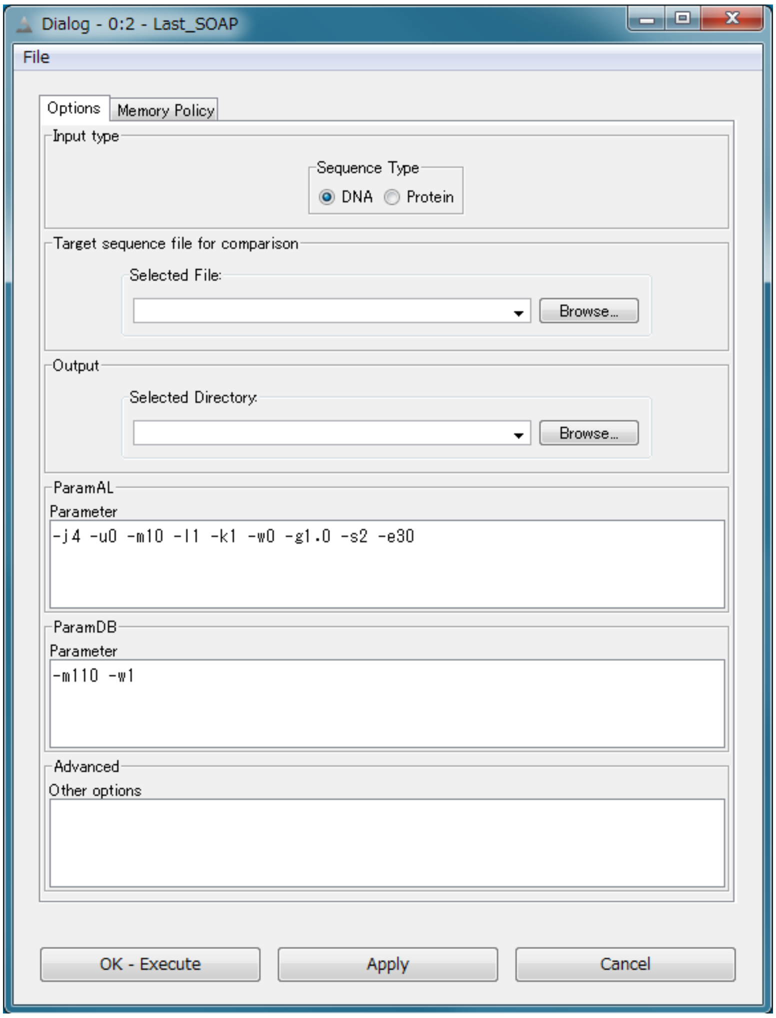
4.4.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a FASTA file as an input using right-click-menu.

2. Node2 : Last_SOAP

   Select “configure” in right-click-menu.



Sequence Type: ● DNA ○ Protein

ParamAL Parameter: `-j4 -u0 -m10 -l1 -k1 -w0 -g1.0 -s2 -e30`

ParamDB Parameter: `-m110 -w1`

4.4.3-1 Last_SOAP : Configure...

- **Options tab → Input type → Sequence Type**

  Select DNA or protein.

- **Options tab → Target sequence file for comparison → Selected File :**

  Select an input file to compare.

- **Options tab → Output → Selected Directory :**

  Select an output directory.

- **Options tab → ParamAL → Parameter**

  Enter AL parameters, if necessary.

  The default parameters are as follows:

```
-j4 -u0 -m10 -l1 -k1 -w0 -g1.0 -s2 -e30
```

- **Options tab → ParamDB → Parameter**

  Enter DB parameters, if necessary.

  The default parameters are as follows:

```
-m110 -w1
```

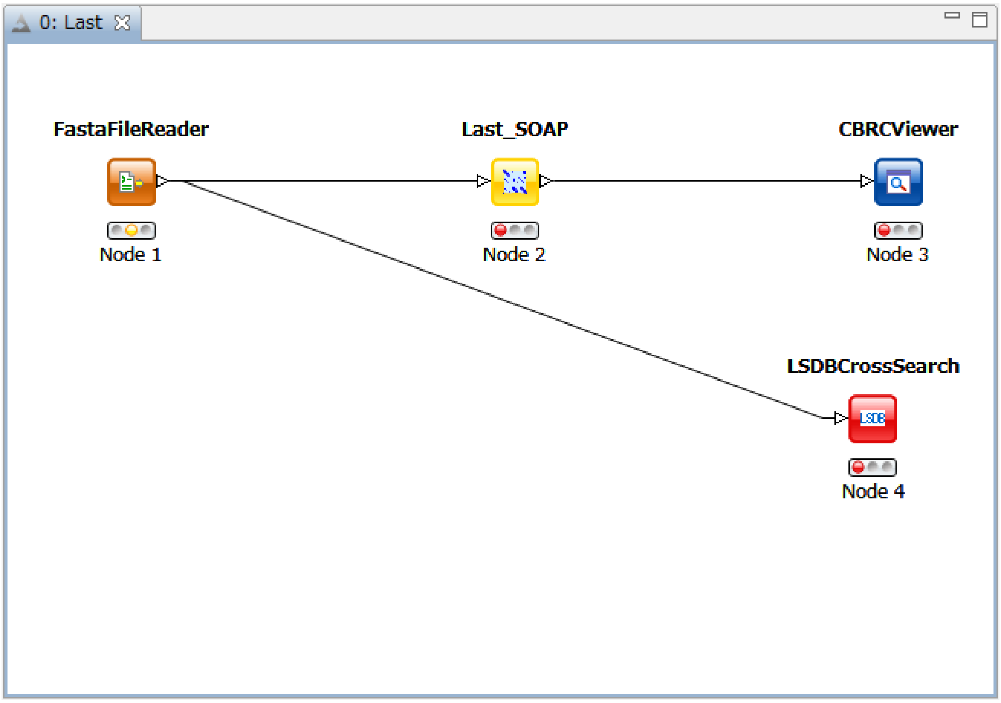
- **Options tab → Advanced → Other options**

  Enter other options, if necessary.

  Please refer to appendix B for details of the options of Last.

Press “OK” after entering.

4.4.4 Step2. Execution



4.4.4-1 Last Node

1) Node1 : FastaFileReader

Select "Execute" in the right-click-menu for execution.

2) Node2 : Last_SOAP

Select "Execute" in the right-click-menu for execution.

3) Node3 : CBRCViewer

Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

4) Node4 : LSDBCrossSearch

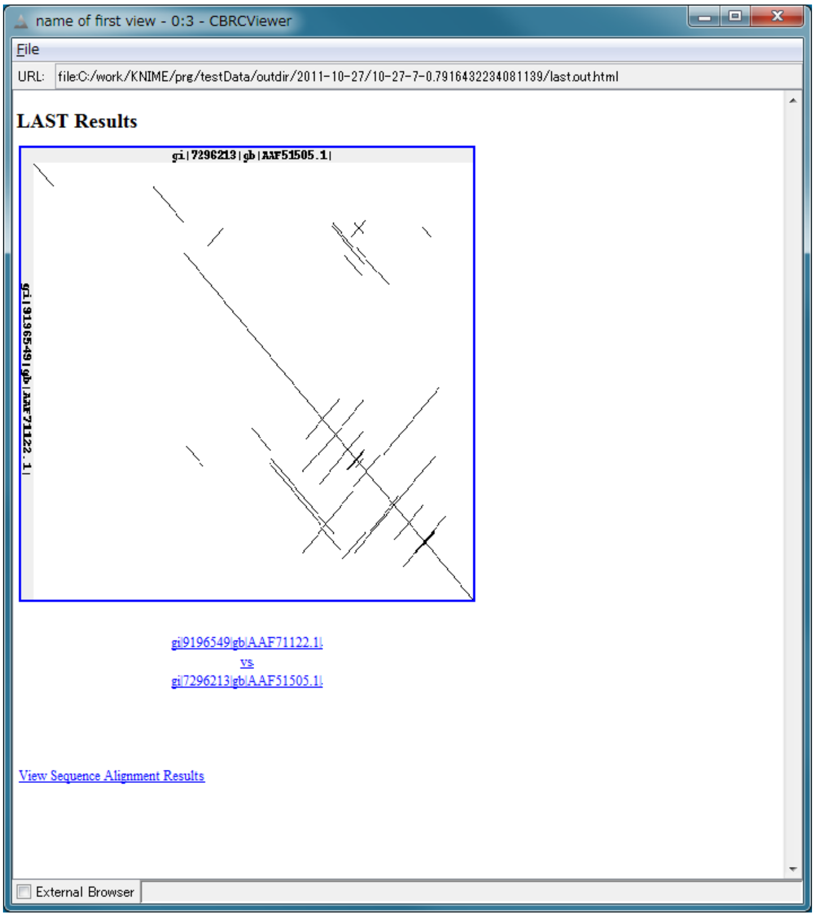
Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

Please refer to the following "5.1 Appendix A : LSDBCrossSearch " for the use of the result screen.

4.4.5 Step.3 Result viewing

1) Node3 CBRCViewer – LAST Results

The execution result of Last_SOAP can be viewed using CBRCViewerNode. A text version of the results is shown by clicking "View Sequence Alignment Results" link.


name of first view - 0:3 - CBRCViewer
File
URL:

file:C:/work/KNIME/prg/testData/outdir/2011-10-27/10-27-7-0.7916432234081139/last.out.html

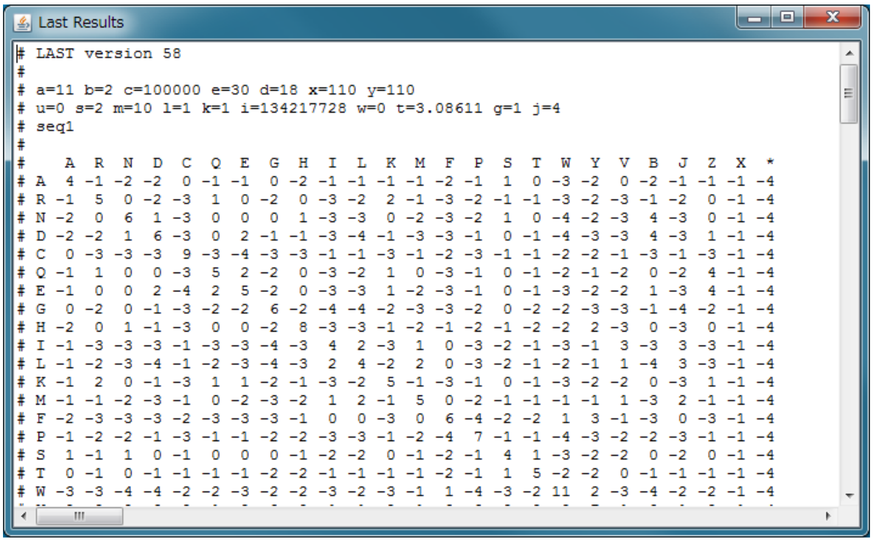
LAST Results

gi|7296213|gb|AAF51505.1|
gi|9196549|gb|AAF71122.1|

gi|9196549|gb|AAF71122.1|
vs
gi|7296213|gb|AAF51505.1|

View Sequence Alignment Results

4.4.5-1 Node3 CBRCViewer – LAST Result



```
# LAST version 58
#
# a=11 b=2 c=100000 e=30 d=18 x=110 y=110
# u=0 s=2 m=10 l=1 k=1 i=134217728 w=0 t=3.08611 g=1 j=4
# seq1
#
#    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  J  Z  X  *
# A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1 -1 -1 -4
# R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1 -2  0 -1 -4
# N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  4 -3  0 -1 -4
# D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4 -3  1 -1 -4
# C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -1 -3 -1 -4
# Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0 -2  4 -1 -4
# E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1 -3  4 -1 -4
# G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -4 -2 -1 -4
# H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0 -3  0 -1 -4
# I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3  3 -3 -1 -4
# L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4  3 -3 -1 -4
# K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0 -3  1 -1 -4
# M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3  2 -1 -1 -4
# F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3  0 -3 -1 -4
# P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -3 -1 -1 -4
# S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0 -2  0 -1 -4
# T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1 -1 -1 -4
# W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -2 -2 -1 -4
```

4.4.5-2 LAST Results – View Sequence Alignment Results

4.5 WolfPSORT Active Workflow

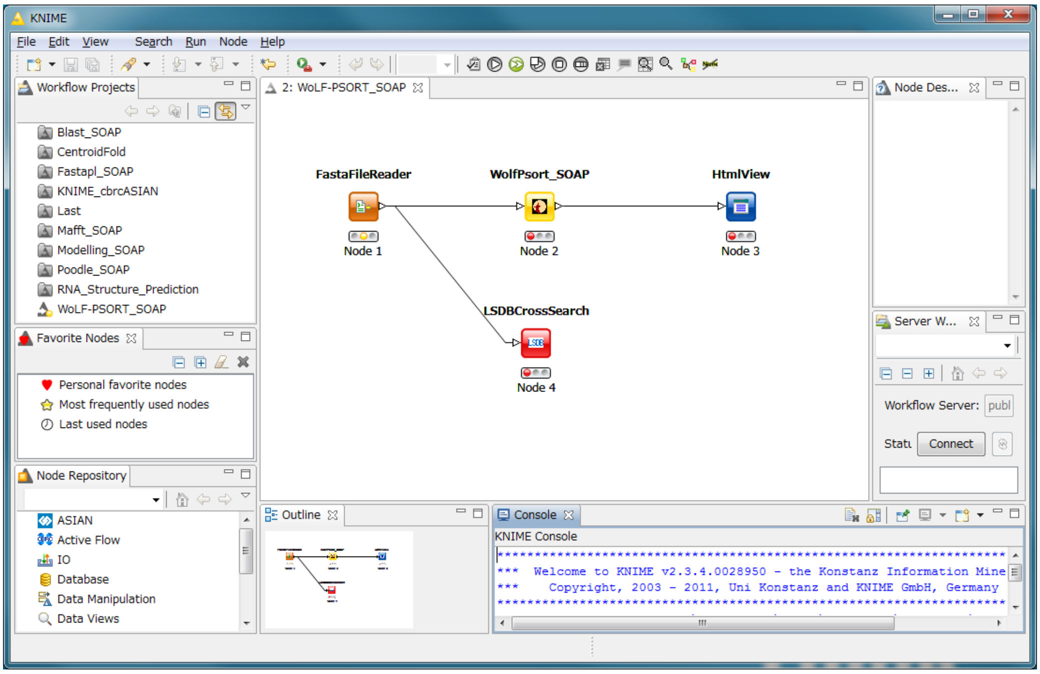
WolfPSORT Active Workflow performs cell localization prediction via SOAP.

The result of WoLF PSORT can be viewed using HtmlViewNode.

Please refer to the following sites for details of WoLF PSORT.

WoLF PSORT : http://wolfpsort.seq.cbrc.jp/

Furthermore, this workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



```
KNIME Console
*****************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Mine
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany
*****************************************************************
```

4.5-1 WolfPSORT Active Workflow

4.5.1 Preparation

A file needed for execution is an amino acid sequence file in FASTA format. Multi-FASTA format can be used.

File Type
(Multi-)FASTA Format file

4.5.2 Node

There are 4 nodes.

4.5.2-1 WolfPsort Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader	FastaFileReader Node 1	The FASTA format file is read.
Node 2	WolfPsort_SOAP	WolfPsort_SOAP Node 2	Execute WoLF PSORT.
Node 3	HtmlView	HtmlView Node 3	The WoLF PSORT execution result is displayed.
Node4	LSDBCrossSearch	LSDBCrossSearch Node 4	Execute LSDB cross-search.

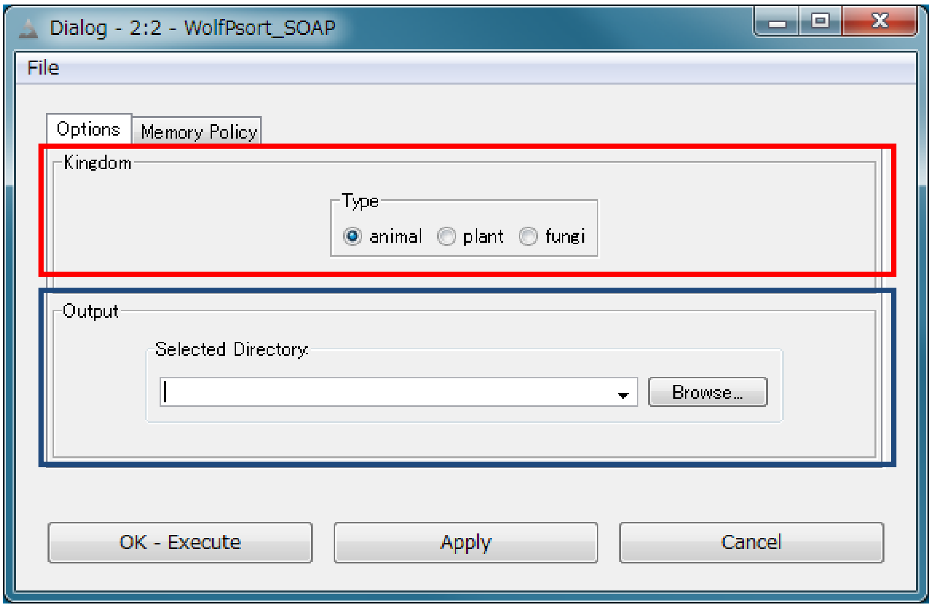
4.5.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a FASTA file as an input using right-click-menu.

2. Node2 : WolfPsort_SOAP

   Select an output directory and kingdom using right-click-menu.

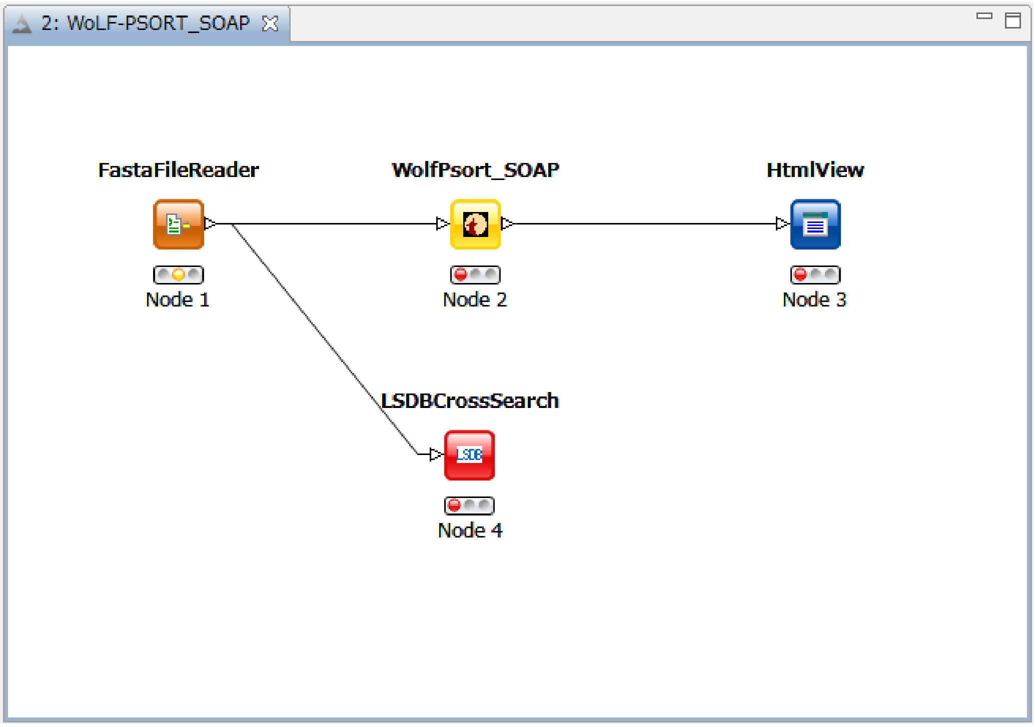


4.5.3-1 WolfPsort_SOAP : Configure…

・Options tab → Kingdom → Type

Select animal, plant or fungi.

4.5.4 Step2. Execution and result viewing



4.5.4-1 WoLF-PSORT_SOAP Node

1) Node1 : FastaFileReader

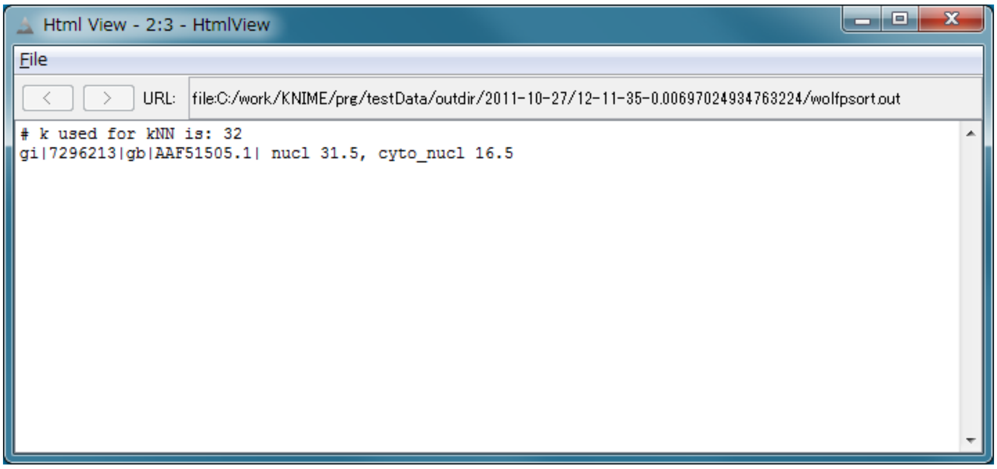
   Select “Execute” in the right-click-menu for execution.

2) Node2 : WolfPsort_SOAP

   Select “Execute” in the right-click-menu for execution.

3) Node3 : HtmlView

   Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.



file:C:/work/KNIME/prg/testData/outdir/2011-10-27/12-11-35-0.00697024934763224/wolfpsort.out

```
# k used for kNN is: 32
gi|7296213|gb|AAF51505.1| nucl 31.5, cyto_nucl 16.5
```

4.5.4-2 Node3 HtmlView– WoLF PSORT Result

4) Node4 : LSDBCrossSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

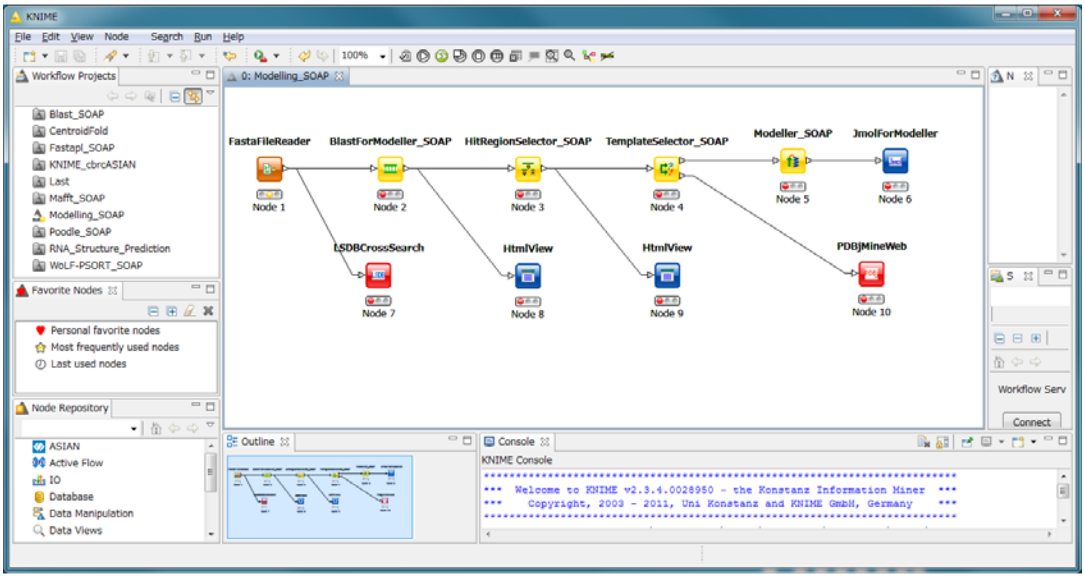
Please refer to the following "5.1 Appendix A : LSDBCrossSearch " for the use of the result screen.

4.6 Modelling Active Workflow

Modelling_SOAP performs 3D structure modeling of a protein via SOAP.

First, BLAST/PSI-BLAST is carried out to search similar regions against PDB database (http://www.rcsb.org/). If similar regions are found, a program called MODELLER (http://salilab.org/modeller/) models the query protein based on the similar regions as a template. A key license is required to run MODELLER.

Furthermore, this workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



```
******************************************************************************
***   Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Miner   ***
***     Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany     ***
******************************************************************************
```

4.6-1 Modelling Active Workflow

4.6.1 Preparation

A file needed for execution is an amino acid sequence file in FASTA format.

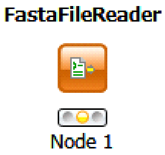
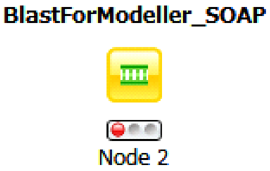
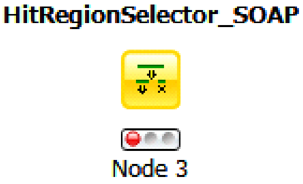
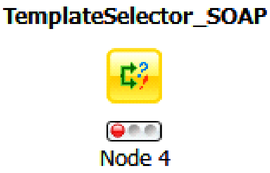
※　Multi-Fasta format cannot be used.

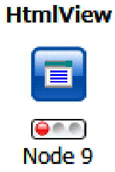
File type
FASTA format amino acid sequence file

4.6.2 Node

There are 10 nodes.

4.6.2-1 Modelling Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader		The FASTA format file is read.
Node 2	BlastForModeller_SOAP		Execute BLAST or PSI-BLAST.
Node 3	HitRegionSelector_SOAP		3D structural hit area is extracted from the execution result of BLAST or PSI-BLAST.
Node4	TemplateSelector_SOAP		A template of 3D structure modeling is selected.

Node5	Modeller_SOAP		Execute MODELLER.
Node6	JmolForModeller		Protein 3D structures are displayed using Jmol.
Node7	LSDBCrossSearch		Execute LSDB cross-search.
Node8	HtmlView		The execution result of BlastForModeller_SOAP is displayed.
Node9	HtmlView		The execution result of HitRegionSelector_SOAP is displayed.
Node10	PDBjMineWeb		Known 3D structure information is displayed by PDBj Mine.

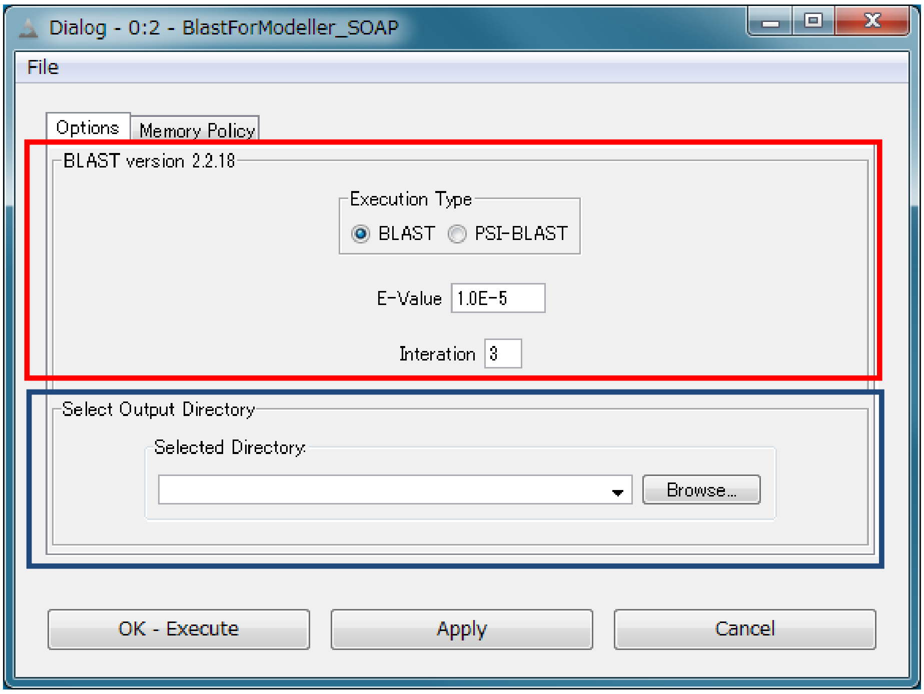
4.6.3 Step 1. Node setting

1. Node1　: FastaFileReader

   Select a FASTA file as an input in “Configure” using the right-click-menu.

2. Node2　: BlastForModeller_SOAP

   Select an output directory and set options in “Configure” using the right-click-menu.



Dialog - 0:2 - BlastForModeller_SOAP

4.6.3-1 BlastForModeller_SOAP　: Configure…

- **Options　tab → BLAST version 2.2.18 → Execution Type**

  Select BLAST or PSI-BLAST.

- **Options　tab → BLAST version 2.2.18 → E-Value**

  Enter a E-Value, which is used as a threshold when BLAST or PSI-BLAST is performed.

  The default value is 1.0E-5.

- **Options　tab → BLAST version 2.2.18 → Interation**

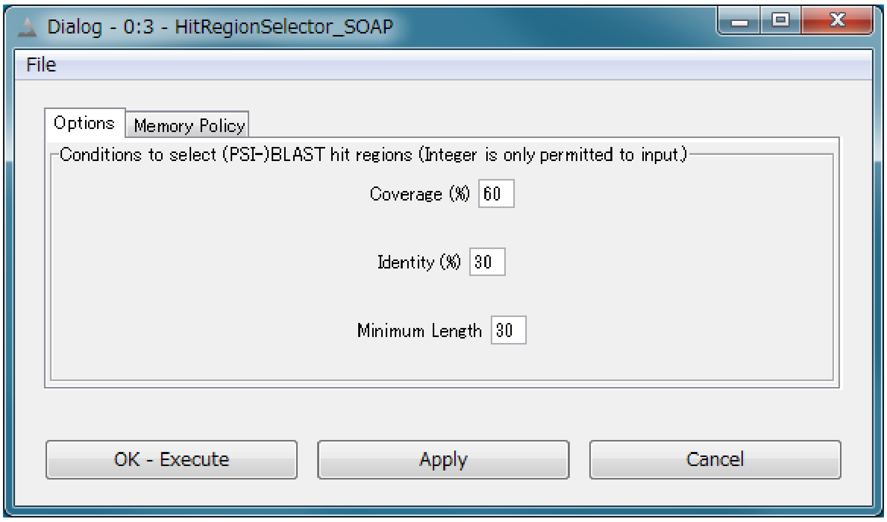
  Enter a value for iteration for PSI-BLAST.

The default value is 3.

3. <u>Node3 : HitRegionSelector_SOAP</u>

Set conditions for BLAST or PSI-BLAST.

1) Select "Configure" in the right-click-menu.



Conditions to select (PSI-)BLAST hit regions (Integer is only permitted to input)

Coverage (%) 60

Identity (%) 30

Minimum Length 30

4.6.3-2 HitRegionSelector_SOAP : Configure…

・Options tab → Condition to select (PSI-)BLAST hit regions (Integer is only permitted to input) → Coverage(%)

Set coverage.

Coverage is a ratio in a hit area against the total length of the protein structure hit.

The default value is 60.

The range of the value is below.

50 < Coverage(%) < 100 : Integer

・Options tab → Condition to select (PSI-)BLAST hit regions (Integer is only permitted to input) → Identity(%)

Set identity.

Identity is an amino acid matching rate in the hit area between the query and the target.

The default value is 30.

The range of the value is below.

10 < Identity(%) < 100　: Integer

・Options　tab → Condition to select (PSI-)BLAST hit regions (Integer is only permitted to input) → Minimum Length

Set Minimum Length.

Minimum Length is a value of minimum length of amino acid of the hit area.

The default value is 30.

The range of the value is below.

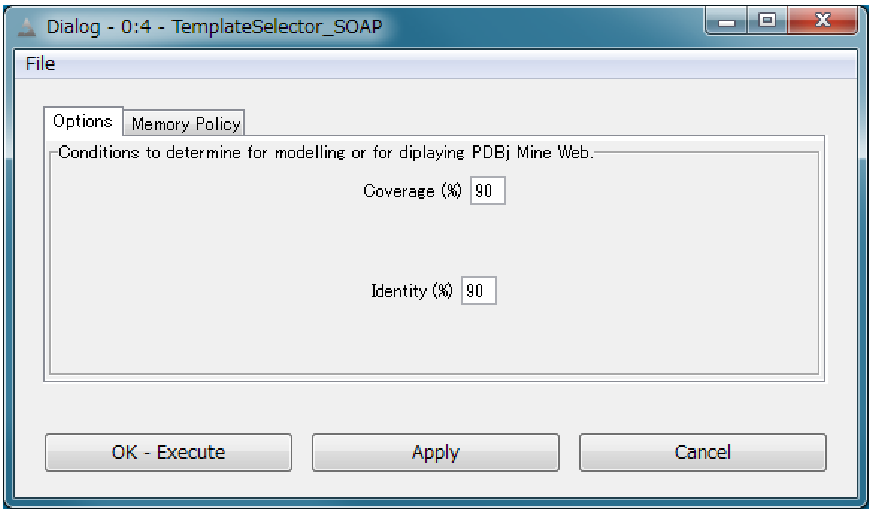
26 < Minimum Length < Input amino acid sequence length　: Integer

Press "OK" after entering.

4. Node4　: TemplateSelector_SOAP

Set conditions for a template for 3D structure modeling.

1) Select "Configure" in the right-click-menu.



Conditions to determine for modelling or for diplaying PDBj Mine Web.

Coverage (%) 90

Identity (%) 90

4.6.3-3 TemplateSelector_SOAP：Configure…

・Options　tab → Condition to determine for modelling or for displaying PDBj Mine Web. → Coverage(%), Identity(%)

Set Coverage and Identity.

Coverage is a ratio in a hit area against the total length of the protein structure hit.

Identity is an amino acid matching rate in the hit area between the query and the target.

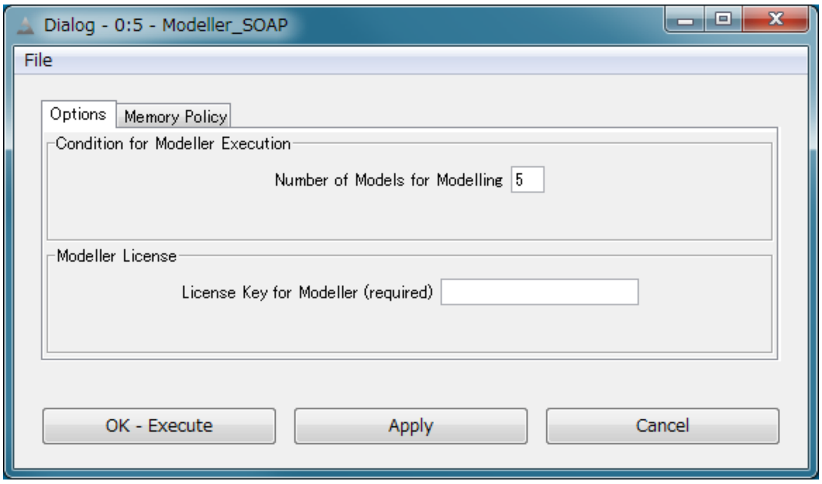
The default value of Coverage is 90 %, and of Identity 90 %.

Only integer can be used.

5. Node5 : Modeller_SOAP

Set a license key and a number of models to generate for MODELLER.

1) Select "Configure" in the right-click-menu.



Dialog - 0:5 - Modeller_SOAP

File

Options | Memory Policy

Condition for Modeller Execution

Number of Models for Modelling: 5

Modeller License

License Key for Modeller (required):

4.6.3-4 Modeller_SOAP：Configure…

・Options tab → Condition for Modeller Execution → Number of Models for Modelling

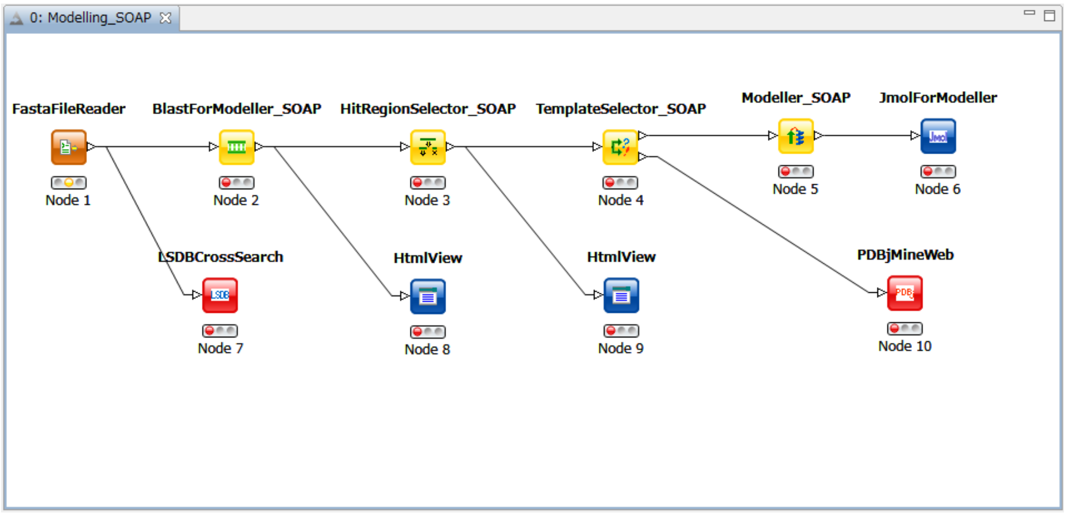
Enter a number of Models to generate.

The value range is 1-10.

・Options tab → Modeller License → License Key for Modeller (required)

Enter a License Key for Modeller (required).

4.6.4 Step2. Execution



4.6.4-1 Modelling _SOAP Node

1) Node1 : FastaFileReader

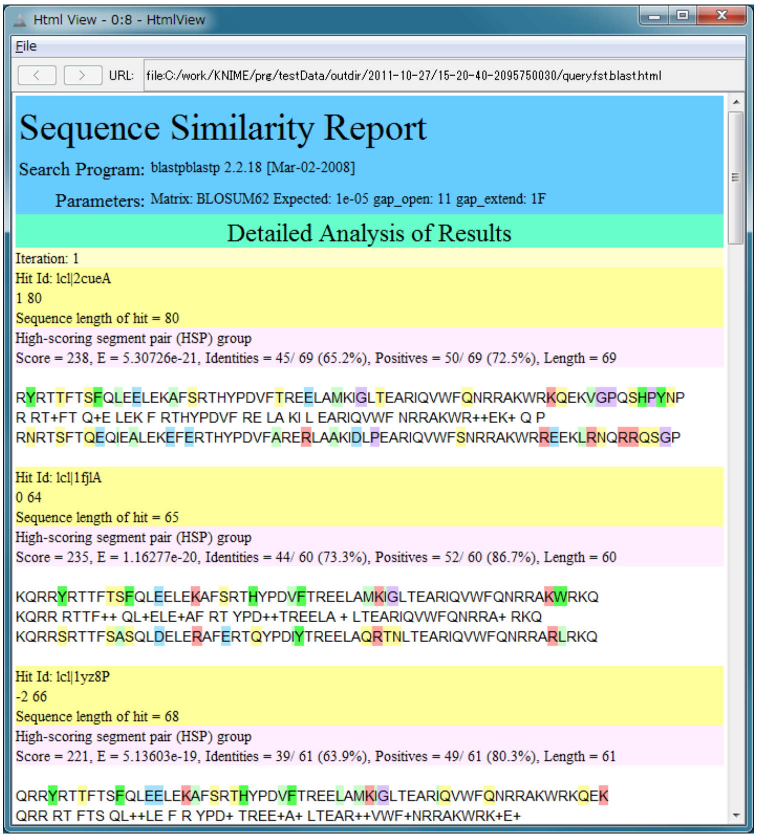
Select “Execute” in the right-click-menu for execution.

2) Node2 : BlastForModeller_SOAP

Select “Execute” in the right-click-menu for execution.

3) Node8 : HtmlView

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.



file:C:/work/KNIME/prg/testData/outdir/2011-10-27/15-20-40-2095750030/query.fst.blast.html

Sequence Similarity Report

Search Program: blastpblastp 2.2.18 [Mar-02-2008]

Parameters: Matrix: BLOSUM62 Expected: 1e-05 gap_open: 11 gap_extend: 1F

Detailed Analysis of Results

```
Iteration: 1
Hit Id: lcl|2cueA
1 80
Sequence length of hit = 80
High-scoring segment pair (HSP) group
Score = 238, E = 5.30726e-21, Identities = 45/ 69 (65.2%), Positives = 50/ 69 (72.5%), Length = 69

RYRTTFTSFQLEELEKAFSRTHYPDVFTREELAMKIGLTEARIQVWFQNRRAKWRKQEKVGPQSHPYNP
R RT+FT Q+E LEK F RTHYPDVF RE LA KI L EARIQVWF NRRAKWR++EK+ Q P
RNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQSGP

Hit Id: lcl|1fjlA
0 64
Sequence length of hit = 65
High-scoring segment pair (HSP) group
Score = 235, E = 1.16277e-20, Identities = 44/ 60 (73.3%), Positives = 52/ 60 (86.7%), Length = 60

KQRRYRTTFTSFQLEELEKAFSRTHYPDVFTREELAMKIGLTEARIQVWFQNRRAKWRKQ
KQRR RTTF++ QL+ELE+AF RT YPD++TREELA + LTEARIQVWFQNRRA+ RKQ
KQRRSRTTFSASQLDELERAFERTQYPDIYTREELAQRTNLTEARIQVWFQNRRARLRKQ

Hit Id: lcl|1yz8P
-2 66
Sequence length of hit = 68
High-scoring segment pair (HSP) group
Score = 221, E = 5.13603e-19, Identities = 39/ 61 (63.9%), Positives = 49/ 61 (80.3%), Length = 61

QRRYRTTFTSFQLEELEKAFSRTHYPDVFTREELAMKIGLTEARIQVWFQNRRAKWRKQEK
QRR RT FTS QL++LE F R YPD+ TREE+A+ LTEAR++VWF+NRRAKWRK+E+
```

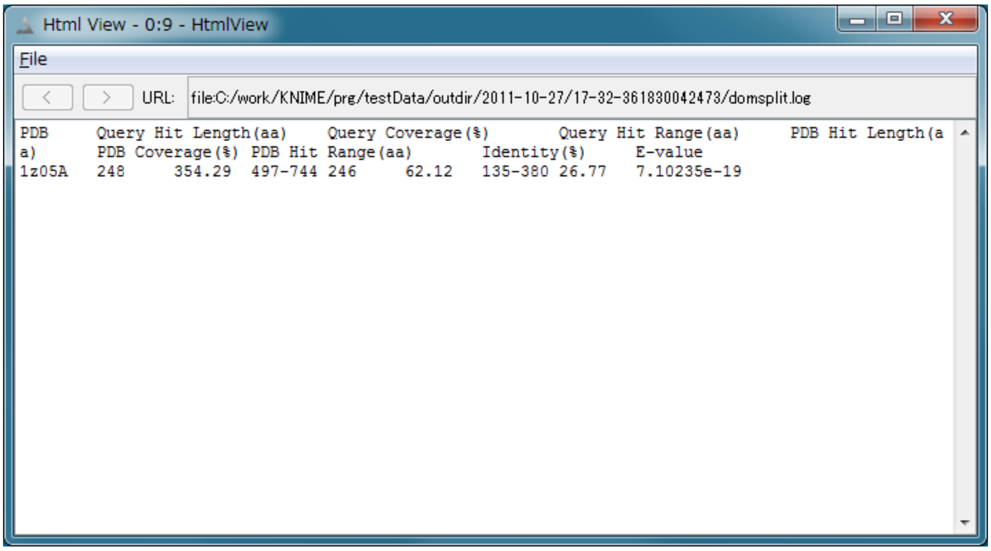
4.6.4-2 BlastForModeller_SOAP Result view(HtmlView)

4) Node3 : HitRegionSelector_SOAP

Select “Execute” in the right-click-menu for execution.

5) Node9 : HtmlView

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.



```
file:C:/work/KNIME/prg/testData/outdir/2011-10-27/17-32-361830042473/domsplit.log
```

```
PDB     Query Hit Length(aa)    Query Coverage(%)       Query Hit Range(aa)     PDB Hit Length(a
a)      PDB Coverage(%) PDB Hit Range(aa)       Identity(%)     E-value
1z05A   248     354.29  497-744 246     62.12   135-380 26.77   7.10235e-19
```

4.6.4-3 HitRegionSelector_SOAP Result View(HtmlView)

6) Node4 : TemplateSelector_SOAP

Select "Execute" in the right-click-menu for execution.

7) Node10 : PDBjMineWeb

Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

Please refer to following description "4.6.5 Step.3 ResultStep.3 " for the use of PDBj Mine.

8) Node5 : Modeller_SOAP

Select "Execute" in the right-click-menu for execution.

9) Node6 : JmolForModeller

Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

10) Node7 : LSDBCrossSearch

Select "Execute and Open Views" in the right-click-menu for execution and viewing the results.

Please refer to "5.1 Appendix A : LSDBCrossSearch" for the use of the result screen.

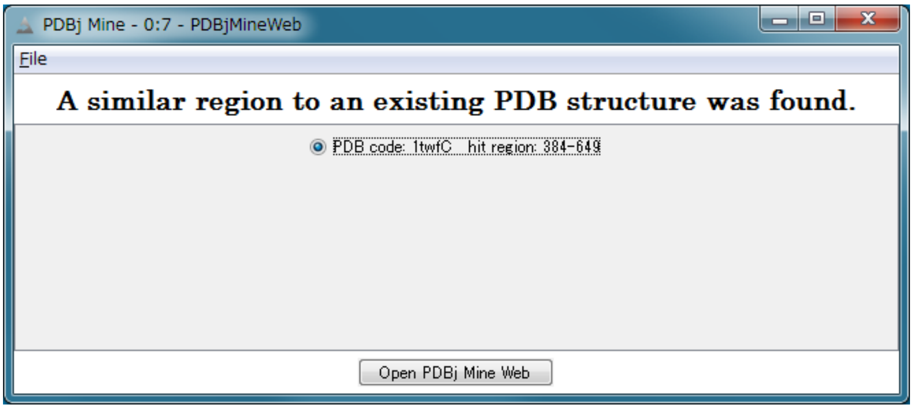
4.6.5 Step.3 Result viewing

1) Node10 : PDBjMineWeb – PDBj Mine

The execution result of TemplateSelector_SOAP of Node4 can be viewed by PDBjMineWeb node.

This window shows a list of known 3D structure information (PDB code + chain identifier) for each hit region.

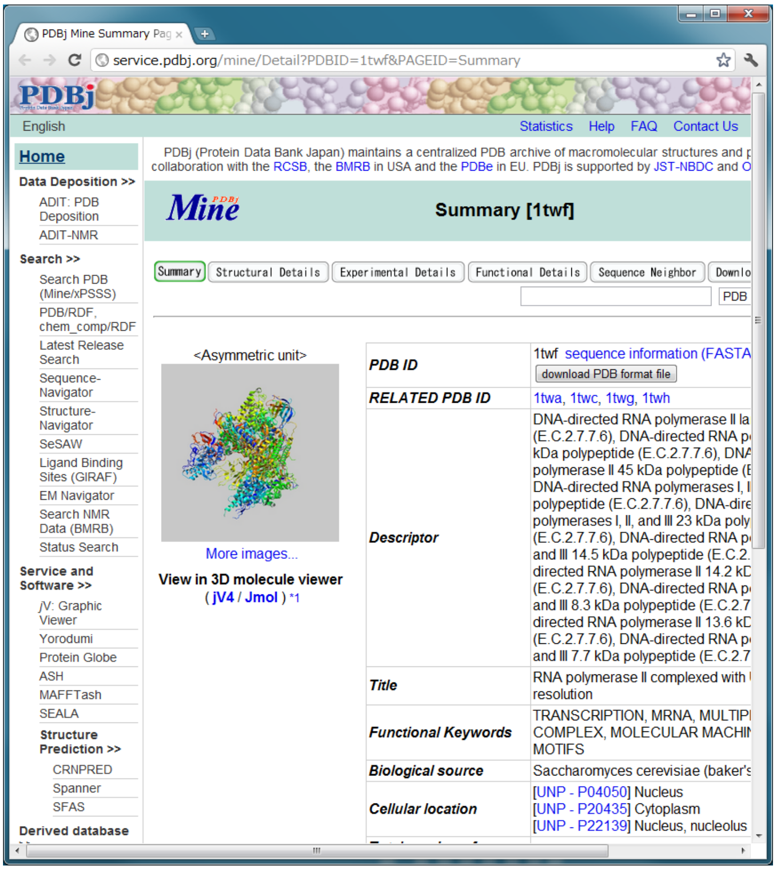
3D structure information stored in PDBj Mine of PDBJ is shown by selecting from the list.



A similar region to an existing PDB structure was found.

● PDB code: 1twfC hit region: 384-649

4.6.5-1 Node10 PDBjMineWeb – PDBj Mine



PDBj (Protein Data Bank Japan) maintains a centralized PDB archive of macromolecular structures and p
collaboration with the RCSB, the BMRB in USA and the PDBe in EU. PDBj is supported by JST-NBDC and O

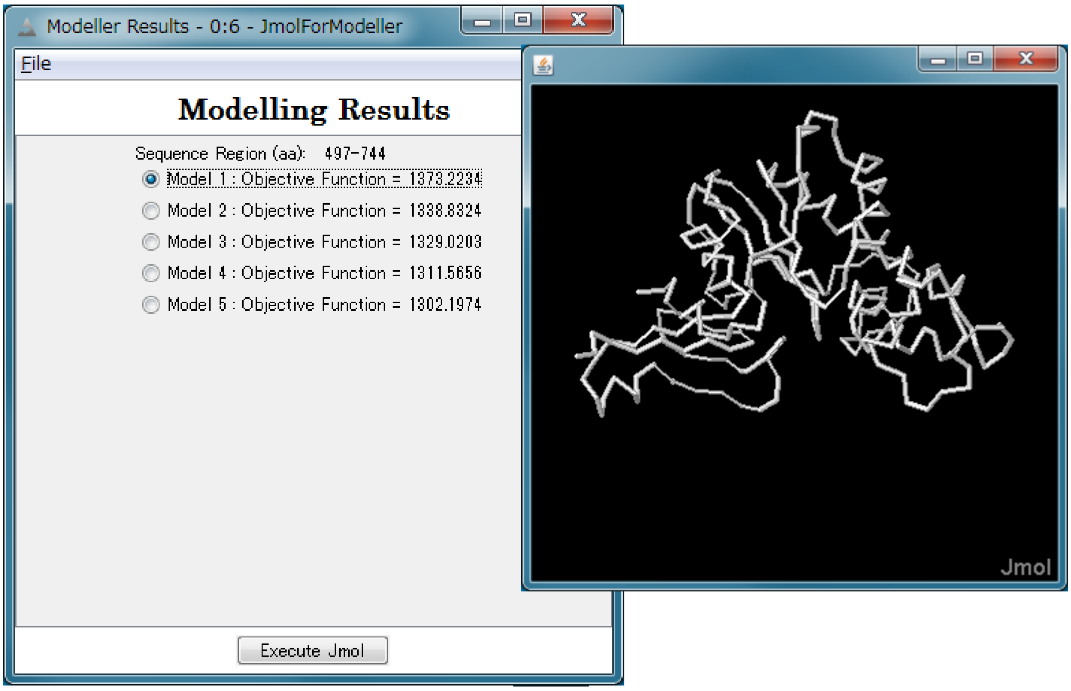
Summary [1twf]

PDB ID	1twf sequence information (FASTA download PDB format file
RELATED PDB ID	1twa, 1twc, 1twg, 1twh
Descriptor	DNA-directed RNA polymerase Ⅱ la (E.C.2.7.7.6), DNA-directed RNA p kDa polypeptide (E.C.2.7.7.6), DNA polymerase Ⅱ 45 kDa polypeptide (E DNA-directed RNA polymerases I, I polypeptide (E.C.2.7.7.6), DNA-dire polymerases I, II, and Ⅲ 23 kDa poly (E.C.2.7.7.6), DNA-directed RNA p and Ⅲ 14.5 kDa polypeptide (E.C.2. directed RNA polymerase Ⅱ 14.2 kD (E.C.2.7.7.6), DNA-directed RNA p and Ⅲ 8.3 kDa polypeptide (E.C.2.7 directed RNA polymerase Ⅱ 13.6 kD (E.C.2.7.7.6), DNA-directed RNA p and Ⅲ 7.7 kDa polypeptide (E.C.2.7
Title	RNA polymerase Ⅱ complexed with resolution
Functional Keywords	TRANSCRIPTION, MRNA, MULTIP COMPLEX, MOLECULAR MACHI MOTIFS
Biological source	Saccharomyces cerevisiae (baker's
Cellular location	[UNP - P04050] Nucleus [UNP - P20435] Cytoplasm [UNP - P22139] Nucleus, nucleolus

4.6.5-2 Node10 PDBjMineWeb – PDBj Mine

2) Node6 : JmolForModeller – Modeller Results

The execution result of Modeller_SOAP of Node5 can be viewed as Modeller Results by JmolForModellerNode.



Modelling Results

Sequence Region (aa): 497-744

● Model 1 : Objective Function = 1373.2234
○ Model 2 : Objective Function = 1338.8324
○ Model 3 : Objective Function = 1329.0203
○ Model 4 : Objective Function = 1311.5656
○ Model 5 : Objective Function = 1302.1974

4.6.5-3 Node6 JmolForModeller – Modeller Results

The Modeller Results displays the resulting protein structures by Jmol.

Once a model in the list is selected, Jmol screen with a structure appears by pressing “Execute Jmol” button.

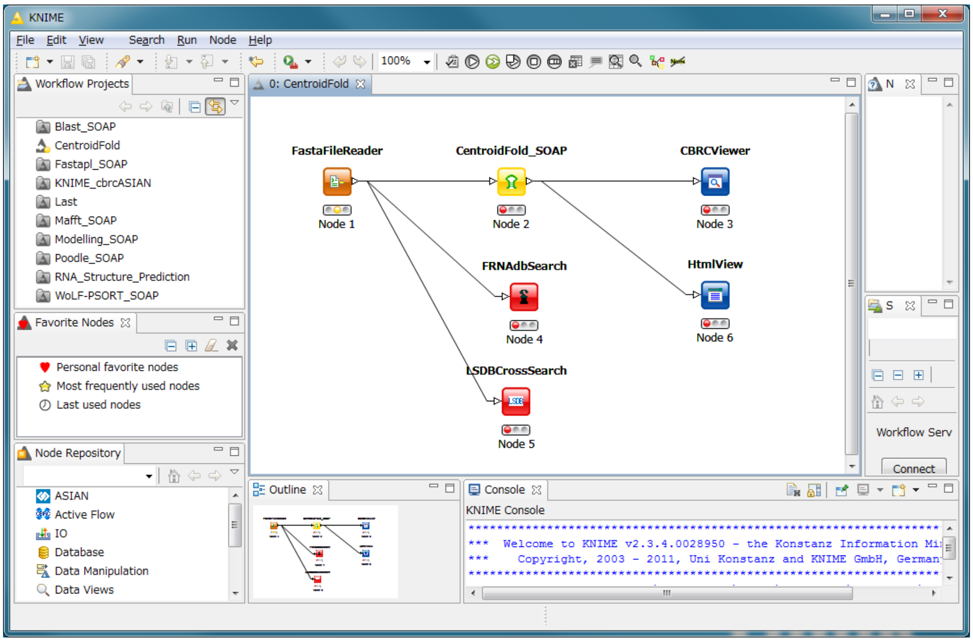
Please refer to the following for the details of Jmol.

Jmol : http://jmol.sourceforge.net/

4.7 CentroidFold Active Workflow

CentroidFold Active Workflow performs prediction of RNA secondary structure from a RNA sequence via SOAP.

Furthermore, this workflow can retrieve a variety of related information by using node LSDBCrossSearch that executes Life Science DataBase cross-search (http://lifesciencedb.jp/dbsearch/) with regard to the input sequence.



```
*****************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Mi
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, German
*****************************************************************
```

4.7-1 CentroidFold Active Workflow

4.7.1 Preparation

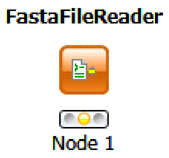
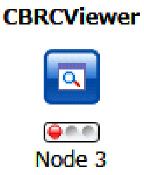
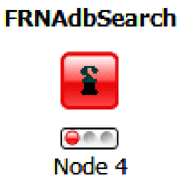
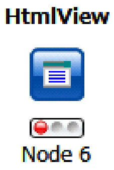
A file needed for execution is an RNA sequence file in FASTA format or an RNA sequence of alignment result file (.aln) of ClustalW. Multi-FASTA can also be used.

File type
(Multi-)FASTA Format File
ClustalW ALN File

4.7.2 Node

There are 6 nodes.

4.7.2-1 CentroidFold Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader	 FastaFileReader Node 1	The FASTA format file is read.
Node 2	CentroidFold_SOAP	 CentroidFold_SOAP Node 2	Execute CentroidFold.
Node 3	CBRCViewer	 CBRCViewer Node 3	The CentroidFold execution result is displayed.
Node 4	FRNAdbSearch	 FRNAdbSearch Node 4	Execute fRNAdb search .
Node 5	LSDBCrossSearch	 LSDBCrossSearch Node 4	Execute LSDB cross-search.
Node 6	HtmlView	 HtmlView Node 6	The CentroidFold execution result is displayed。

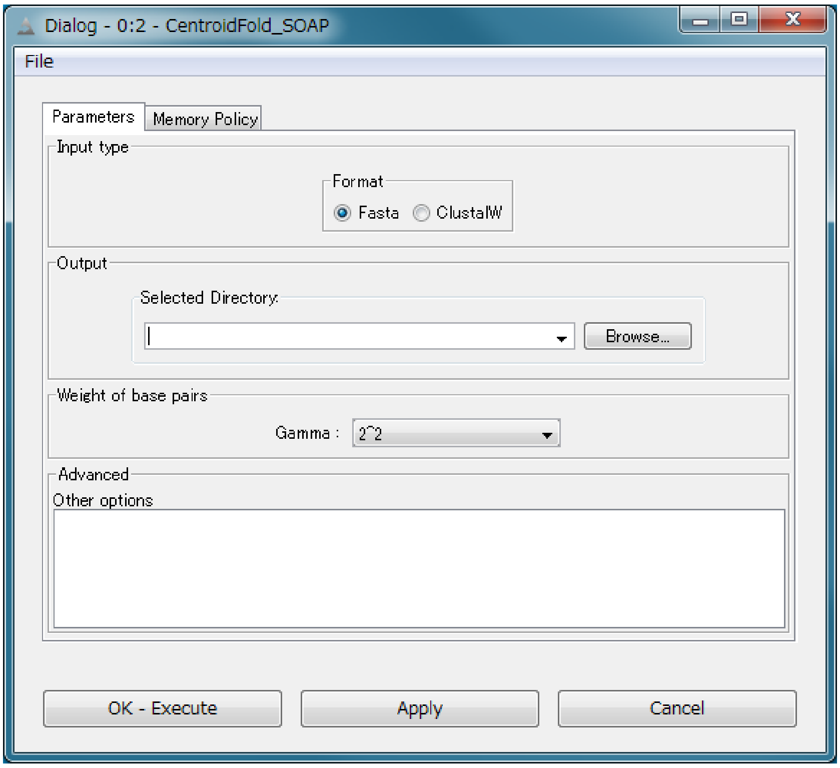
4.7.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a FASTA file as an input in "Configure" using the right-click-menu.

2. Node2 : CentroidFold_SOAP

   Select an output directory and format in the right-click-menu.



Format: ● Fasta ○ ClustalW

Gamma : 2^2

4.7.3-1 CentroidFold_SOAP : Configure…

- **Options tab → Input type → Format**

  Select FASTA or ClustalW as a format.

- **Options tab → Weight of base pairs → Gamma:**

  Select a value from the pull-down menu.

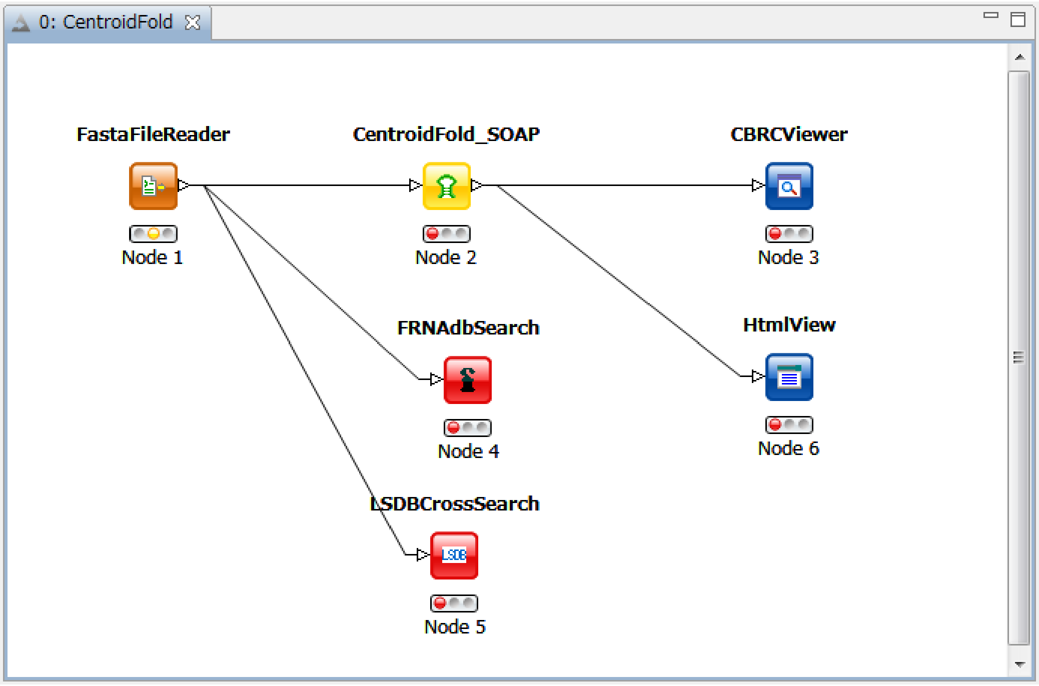
- **Options tab → Advanced → Other options**

  Enter other options, if necessary.

Please refer to the following sites for details of CentroidFold.

CentroidFold : http://www.ncrna.org/centroidfold/software/centroidfold

4.7.4 Step2. Execution



4.7.4-1 CentroidFold Node

It executes it from left FastaFileReaderNode.

1) Node1 : FastaFileReader

Select “Execute” in the right-click-menu for execution.

2) Node2 : CentroidFold_SOAP

Select “Execute” in the right-click-menu for execution.

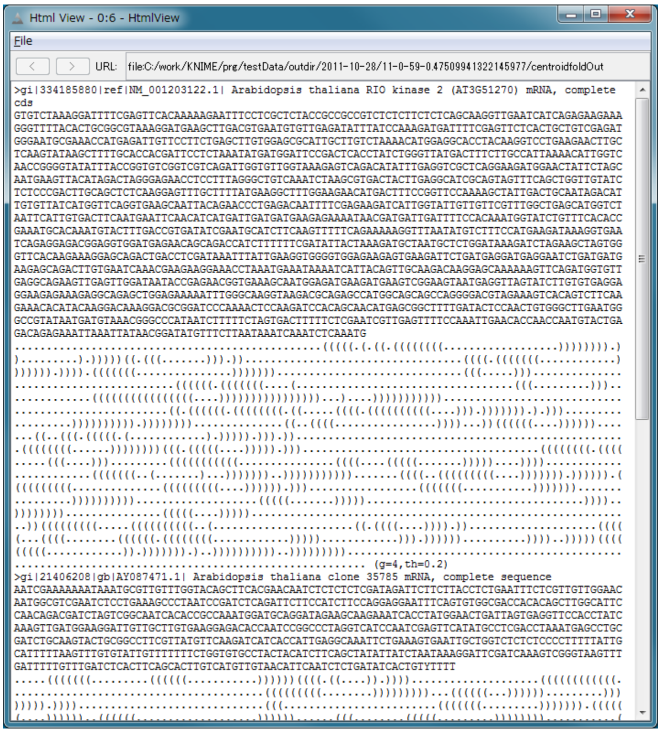
3) Node3 : CBRCViewer

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

4) Node6 : HtmlView

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to "4.7.5 Step.3 Result " for the details.



file:C:/work/KNIME/prg/testData/outdir/2011-10-28/11-0-59-0.47509941322145977/centroidfoldOut

```
>gi|334185880|ref|NM_001203122.1| Arabidopsis thaliana RIO kinase 2 (AT3G51270) mRNA, complete
cds
GTGTCTAAAGGATTTTCGAGTTCACAAAAAGAATTTCCTCGCTCTACCGCCGCCGTCTCTCTTCTCTCAGCAAGGTTGAATCATCAGAGAAGAAA
GGGTTTTACACTGCGGCGTAAAGGATGAAGCTTGACGTGAATGTGTTGAGATATTTATCCAAAGATGATTTTCGAGTTCTCACTGCTGTCGAGAT
GGGAATGCGAAACCATGAGATTGTTCCTTCTGAGCTTGTGGAGCGCATTGCTTGTCTAAAACATGGAGGCACCTACAAGGTCCTGAAGAACTTGC
TCAAGTATAAGCTTTTGCACCACGATTCCTCTAAATATGATGGATTCCGACTCACCTATCTGGGTTATGACTTTCTTGCCATTAAAACATTGGTC
AACCGGGGTATATTTACCGGTGTCGGTCGTCAGATTGGTGTTGGTAAAGAGTCAGACATATTTGAGGTCGCTCAGGAAGATGGAACTATTCTAGC
AATGAAGTTACATAGACTAGGGAGAACCTCCTTTAGGGCTGTCAAATCTAAGCGTGACTACTTGAGGCATCGCAGTAGTTTCAGCTGGTTGTATC
TCTCCCGACTTGCAGCTCTCAAGGAGTTTGCTTTTATGAAGGCTTTGGAAGAACATGACTTTCCGGTTCCAAAAGCTATTGACTGCAATAGACAT
TGTGTTATCATGGTTCAGGTGAAGCAATTACAGAACCCTGAGACAATTTTCGAGAAGATCATTGGTATTGTTGTTCGTTTGGCTGAGCATGGTCT
AATTCATTGTGACTTCAATGAATTCAACATCATGATTGATGATGAAGAGAAAATAACGATGATTGATTTTCCACAAATGGTATCTGTTTCACACC
GAAATGCACAAATGTACTTTGACCGTGATATCGAATGCATCTTCAAGTTTTTCAGAAAAAGGTTTAATATGTCTTTCCATGAAGATAAAGGTGAA
TCAGAGGAGACGGAGGTGGATGAGAACAGCAGACCATCTTTTTTCGATATTACTAAAGATGCTAATGCTCTGGATAAAGATCTAGAAGCTAGTGG
GTTCACAAGAAAGGAGCAGACTGACCTCGATAAATTTATTGAAGGTGGGGTGGAGAAGAGTGAAGATTCTGATGAGGATGAGGAATCTGATGATG
AAGAGCAGACTTGTGAATCAAACGAAGAAGGAAACCTAAATGAAATAAAATCATTACAGTTGCAAGACAAGGAGCAAAAAAGTTCAGATGGTGTT
GAGGCAGAAGTTGAGTTGGATAATACCGAGAACGGTGAAAGCAATGGAGATGAAGATGAAGTCGGAAGTAATGAGGTTAGTATCTTGTGTGAGGA
GGAAGAGAAAGAGGCAGAGCTGGAGAAAAATTTGGGCAAGGTAAGACGCAGAGCCATGGCAGCAGCCAGGGGACGTAGAAAGTCACAGTCTTCAA
GAAACACATACAAGGACAAAGGACGCGGATCCCAAAACTCCAAGATCCACAGCAACATGAGCGGCTTTTGATACTCCAACTGTGGGCTTGAATGG
GCCGTATAATGATGTAAACGGGCCCATAATCTTTTTCTAGTGACTTTTTCTCGAATCGTTGAGTTTTCCAAATTGAACACCAACCAATGTACTGA
GACAGAGAAATTAAATTATAACGGATATGTTTCTTAATAAATCAAATCTCAAATG
.........................................(((((.(.((.((((((((..................)))))))).)
)........).)))))((.(((.......))).)).................................((((.(((((((............)
)))))).)))).(((((((................)))))))...........................(((.....)))............
.........................((((((.(((((((....(....................................(((.........)))..
............(((((((((((((((((....))))))))))))))))...)....))))))))))........................
.......................((.((((((.((((((((.((.....((((.((((((((((....))).))))))).).)))........
.........)))))))))).)))))))).............((..((((.................))))...))((((((....))))))....
...((..(((.(((((.(.............).))))).))).))...........................................
.((((((((......))))))))(((.(((((....))))).)))................................((((((((.((((
.....(((....)))........(((((((((((...............((((....(((((.......)))))....))))............
............((((((..(........)...)))))))..))))))))))).......((((..((((((((((....))))))).)))))).(
(((((((((..............((((((((((....)))))).)))..................((((((((..........)))))))).......
.........)))))))))..................(((((.......))))).................................))))..
))))))))..............(((((....)))))..................................................
..))((((((((((.....((((((((((..(.....................((.((((....)))).))...................((((
(...((((........((((((.((((((((............))))))............)))).))))))..........))))..)))))((((
((((.............)).))))))))).)..))))))))))..))))))))).............................................
.................................................. (g=4,th=0.2)
>gi|21406208|gb|AY087471.1| Arabidopsis thaliana clone 35785 mRNA, complete sequence
AATCGAAAAAAATAAATGCGTTGTTTGGTACAGCTTCACGAACAATCTCTCTCTCGATAGATTCTTCTTACCTCTGAATTTCTCGTTGTTGGAAC
AATGGCGTCGAATCTCCTGAAAGCCCTAATCCGATCTCAGATTCTTCCATCTTCCAGGAGGAATTTCAGTGTGGCGACCACACAGCTTGGCATTC
CAACAGACGATCTAGTCGGCAATCACACCGCCAAATGGATGCAGGATAGAAGCAAGAAATCACCTATGGAACTGATTAGTGAGGTTCCACCTATC
AAAGTTGATGGAAGGATTGTTGCTTGTGAAGGAGACACCAATCCGGCCCTAGGTCATCCAATCGAGTTCATATGCCTCGACCTAAATGAGCCTGC
GATCTGCAAGTACTGCGGCCTTCGTTATGTTCAAGATCATCACCATTGAGGCAAATTCTGAAAGTGAATTGCTGGTCTCTCTCCCCTTTTTATTG
CATTTTTAAGTTTGTGTATTGTTTTTTTCTGGTGTGCCTACTACATCTTCAGCTATATTATCTAATAAAGGATTCGATCAAAGTCGGGTAAGTTT
GATTTTTGTTTGATCTCACTTCAGCACTTGTCATGTTGTAACATTCAATCTCTGATATCACTGTYTTTT
.....(((((((.........((((((...........))))))((((.((....)).))))....................((((((((((((.
...................................(((((((((........)))))))))...((((((...))))))..........)))
))))).))))...........................(((.........................(((((((.........))))))).(((((
(....))))))..((((((.....................))))))......((.........(((((..........)))))))..........(
```

4.7.4-2 Node6 HtmlView – CentroidFold Results

5) Node4 : FRNAdbSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to "4.7.5 Step.3 Result " for the details.

6) Node5 : LSDBCrossSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to the following "5.1 Appendix A : LSDBCrossSearch" for the use of the result screen.

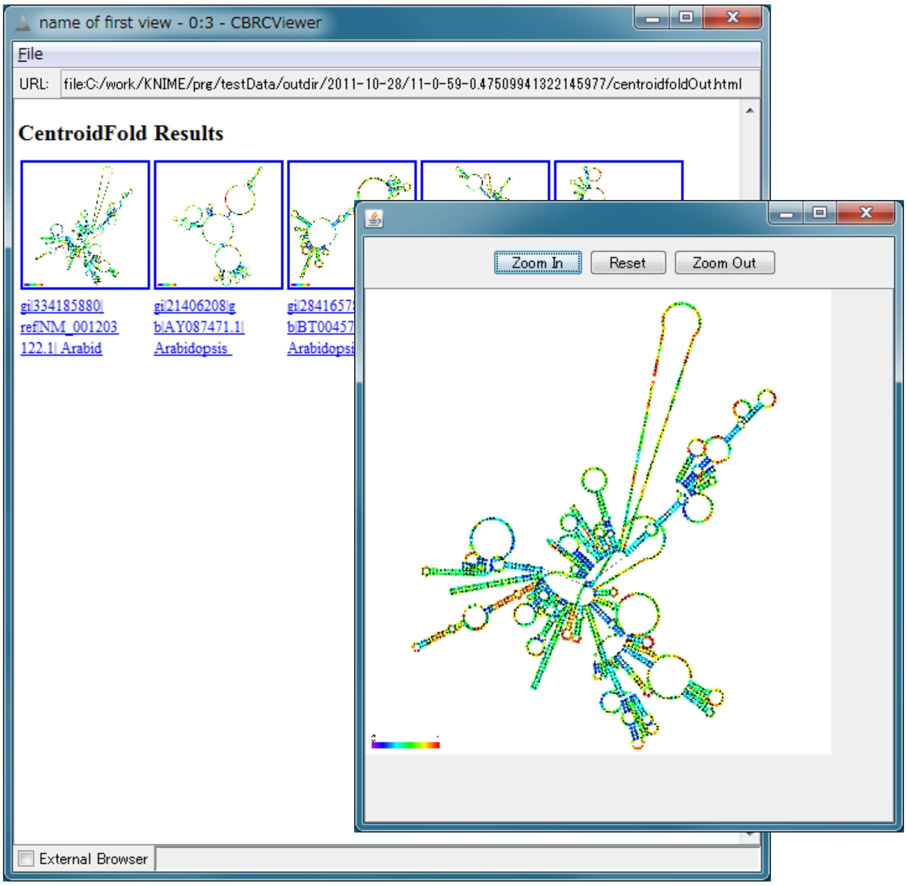
4.7.5 Step.3 Result viewing

1) Node3 : CBRCViewer – CentroidFold Resuls

The execution result of CentroidFold_SOAP of Node2 can be viewed as CentroidFold Results by CBRCViewer.

Please refer to the following site for details of CentroidFold.

CentroidFold : http://www.ncrna.org/centroidfold/software/centroidfold



file:C:/work/KNIME/prg/testData/outdir/2011-10-28/11-0-59-0.47509941322145977/centroidfoldOut.html

CentroidFold Results

gi|334185880|refNM_001203122.1| Arabid

gi|21406208|gb|AY087471.1| Arabidopsis

gi|2841657 b|BT00457 Arabidopsi

4.7.5-1 Node3 CBRCViewer – CentroidFold Results

2) Node4 : FRNAdbSearch

FRNAdbSearch displays a retrieval screen to fRNAdb.

If the input RNA sequence file is in FASTA format, the header line of the FASTA format is displayed in the FASTA Header Lists column.

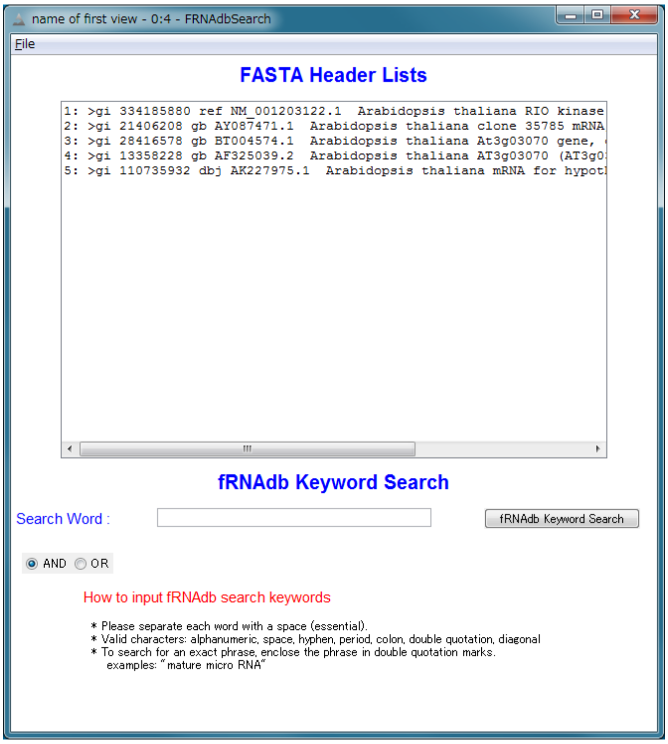
If the input RNA sequence file is in ALN format, this column is blank.

A search keyword(s) to fRNAdb should be entered in the text box at the center of the window. A search can be carried out by pressing "fRNAdb Keyword Search" button.

The result of the retrieval is displayed in another window as shown in figures 16.1-2.

Please refer to the following site for details of fRNAdb.

fRNAdb : http://www.ncrna.org/frnadb/index.html

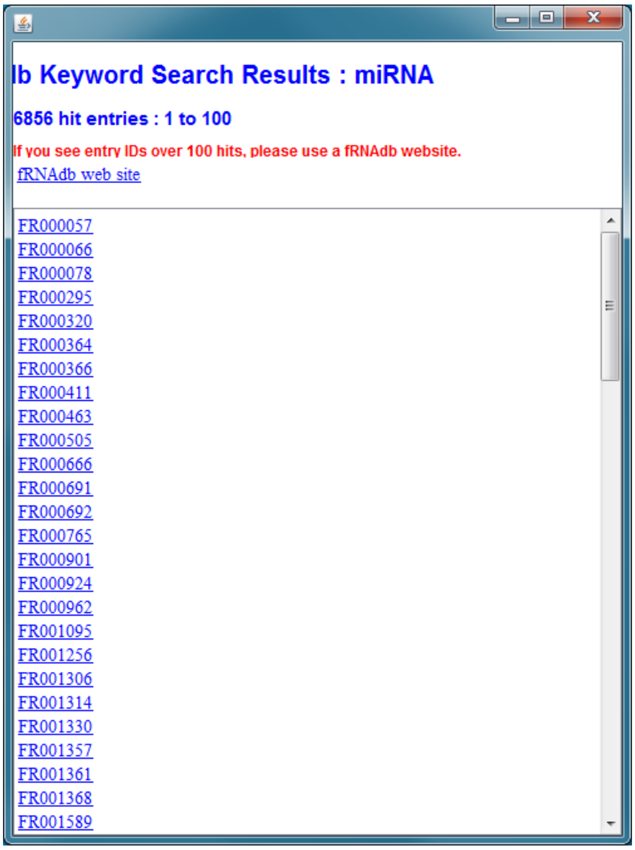


```
1: >gi 334185880 ref NM_001203122.1  Arabidopsis thaliana RIO kinase
2: >gi 21406208 gb AY087471.1  Arabidopsis thaliana clone 35785 mRNA
3: >gi 28416578 gb BT004574.1  Arabidopsis thaliana At3g03070 gene, 
4: >gi 13358228 gb AF325039.2  Arabidopsis thaliana AT3g03070 (AT3g0
5: >gi 110735932 dbj AK227975.1  Arabidopsis thaliana mRNA for hypot
```

How to input fRNAdb search keywords

＊ Please separate each word with a space (essential).
＊ Valid characters: alphanumeric, space, hyphen, period, colon, double quotation, diagonal
＊ To search for an exact phrase, enclose the phrase in double quotation marks.
examples: "mature micro RNA"

4.7.5-2 Node4 fRNAdbSearch – fRNAdb Keyword Search



lb Keyword Search Results : miRNA

6856 hit entries : 1 to 100

If you see entry IDs over 100 hits, please use a fRNAdb website.

fRNAdb web site

FR000057
FR000066
FR000078
FR000295
FR000320
FR000364
FR000366
FR000411
FR000463
FR000505
FR000666
FR000691
FR000692
FR000765
FR000901
FR000924
FR000962
FR001095
FR001256
FR001306
FR001314
FR001330
FR001357
FR001361
FR001368
FR001589

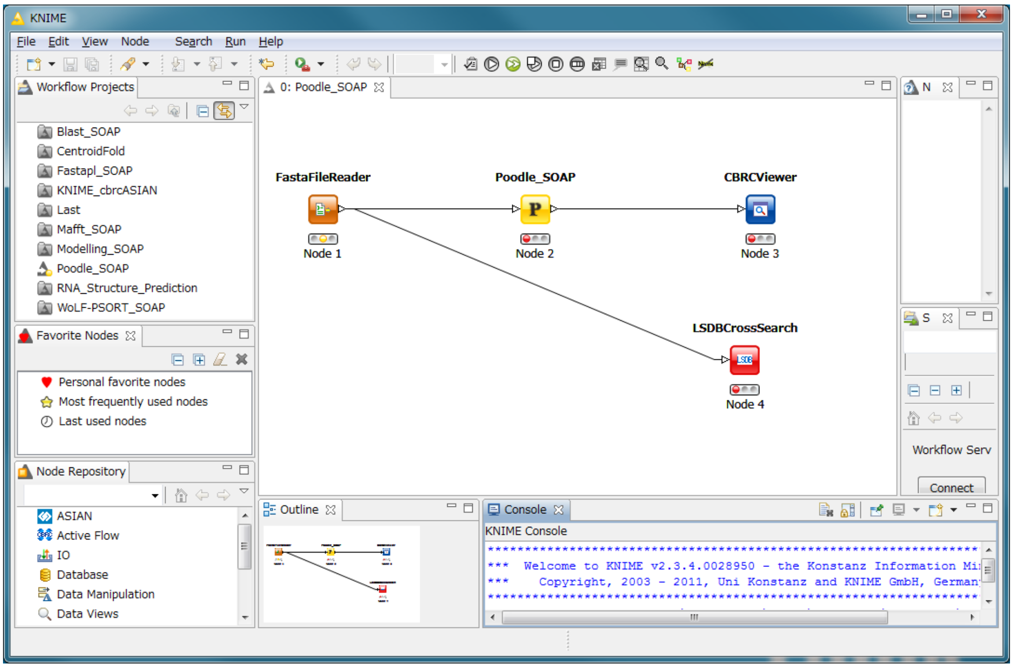
4.7.5-3 Node4 fRNAdbSearch – Search results

4.8 POODLE Active Workflow

POODLE (Prediction Of Order and Disorder by machine LEarning) developed at CBRC predicts disorder regions from an amino-acid sequence. POODLE has 2 types, POODLE-L, which is optimized for longer disorder regions (> 40 a.a.), and POODLE-S, which is optimized for shorter disorder regions.

POODLE results can be viewed in line-plot format.

POODLE : http://mbs.cbrc.jp/poodle/



KNIME Console

```
**************************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Mi
***     Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, German
**************************************************************************
```

4.8-1 POODLE Active Workflow

4.8.1 Preparation

A file needed for execution is an amino-acid sequence file in FASTA format.

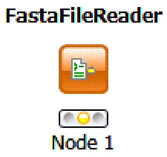
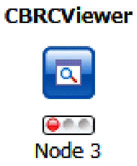
※ Multi-FASTA format file cannot be used.

File type
FASTA Format File

4.8.2 Node

There are 4 nodes.

4.8.2-1 Poodle Active Workflow Node list

Node ID	Node name	Icon	explanation
Node 1	FastaFileReader	 FastaFileReader Node 1	The FASTA format file is read.
Node 2	Poodle_SOAP	Poodle_SOAP Node 2	Execute POODLE.
Node 3	CBRCViewer	 CBRCViewer Node 3	The POODLE execution result is displayed.
Node4	LSDBCrossSearch	LSDBCrossSearch Node 4	Execute LSDB cross-search.

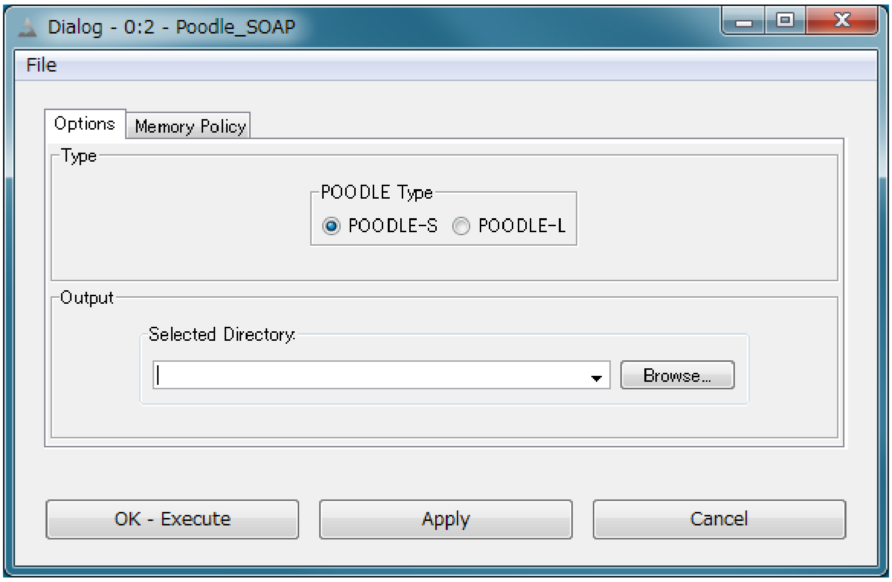
4.8.3 Step 1. Node setting

1. Node1 : FastaFileReader

   Select a FASTA file as an input in “Configure” using the right-click-menu.

2. Node2 : Poodle_SOAP

   Select an output directory and program type in “Configure” using the right-click-menu.



Dialog - 0:2 - Poodle_SOAP

POODLE Type: ● POODLE-S ○ POODLE-L

Selected Directory: |

4.8.3-1 Poodle_SOAP : Configure…

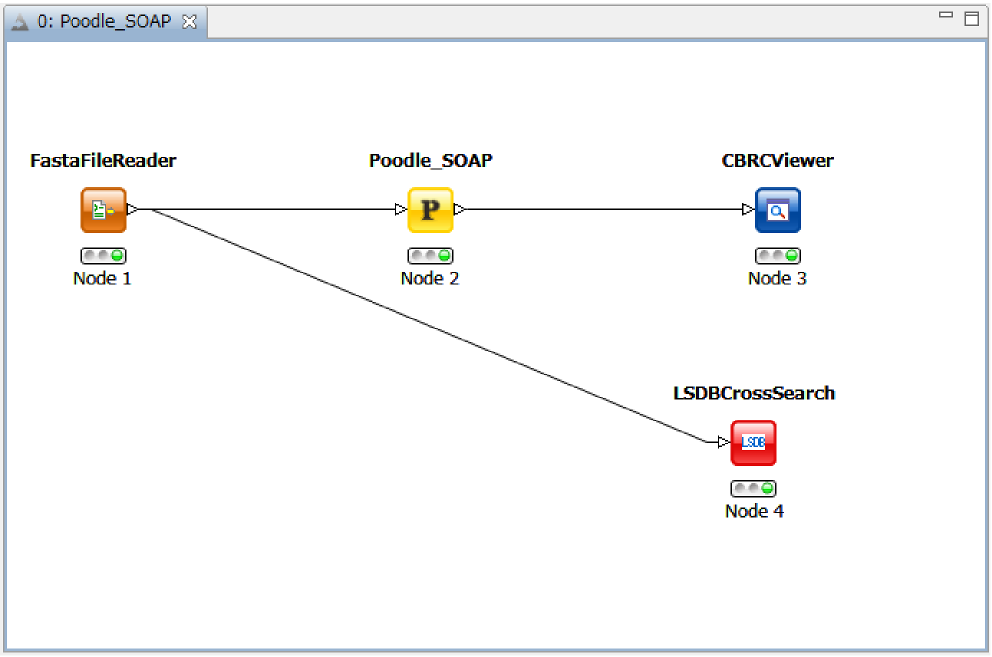
- **Options tab → Type → POODLE Type**

  Select type POODLE-S or POODLE-L.

  POODLE-S predicts shorter disorder regions.

  POODLE-L predicts longer disorder regions (> 40 a.a.).

4.8.4 Step2. Execution



4.8.4-1 Poodle_SOAPNode

1) Node1 : FastaFileReader

Select “Execute” in the right-click-menu for execution.

2) Node2 : Poodle_SOAP

Select “Execute” in the right-click-menu for execution.

3) Node3 : CBRCViewer

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

4) Node4 : LSDBCrossSearch

Select “Execute and Open Views” in the right-click-menu for execution and viewing the results.

Please refer to the following "5.1 Appendix A : LSDBCrossSearch" for the use of the result screen.

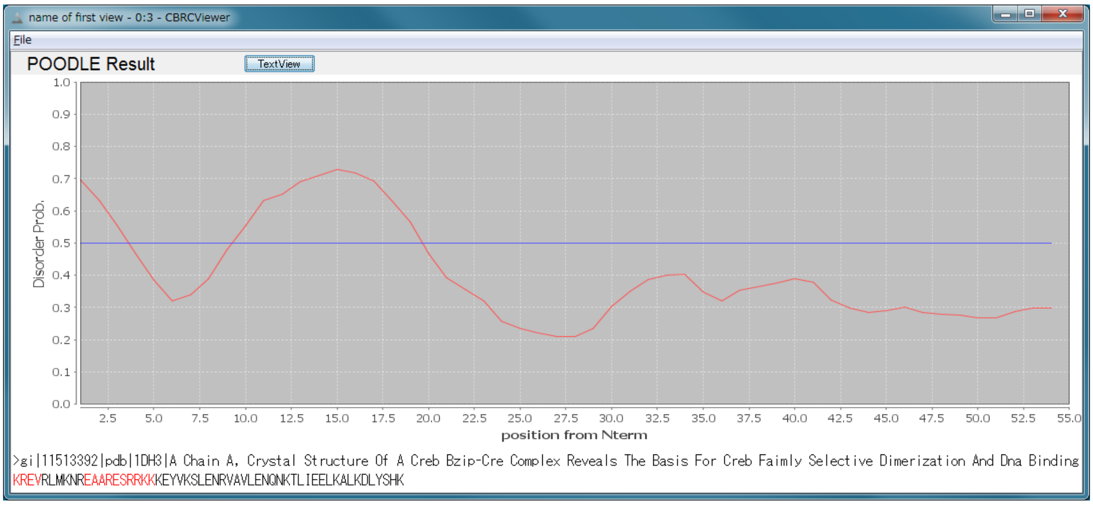
4.8.5 Step.3 Result viewing

1) Node3 CBRCViewer – POODLE Result

The execution result of Poodle_SOAP can be viewedto as POODLE Result by CBRCViewer node.

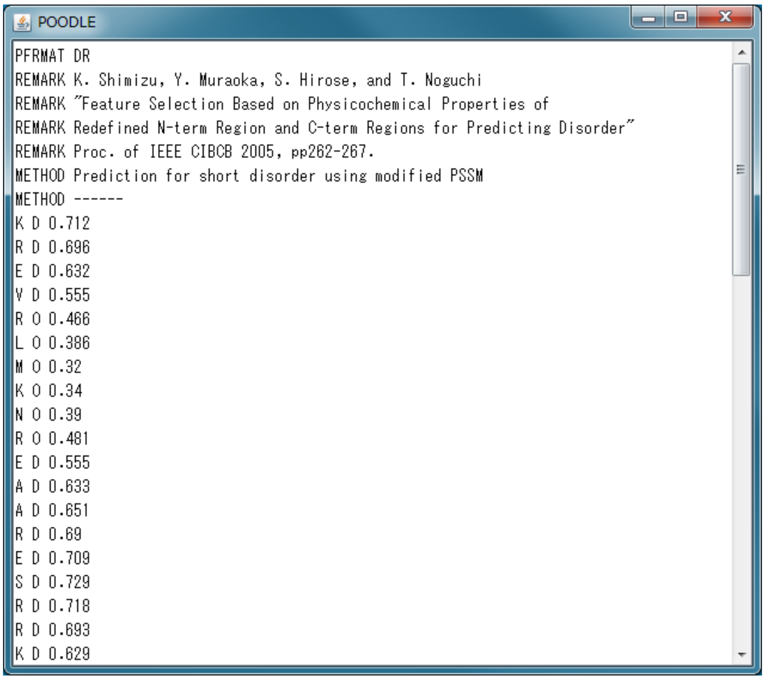
This screen displays the disorder prediction results of POODLE-S or POODLE-L as a plot. The vertical axis indicates disorder probability and the horizontal axis indicates residue numbers. Amino acids in red indicate disorder-predicted.

The text version of the results can be shown by pressing TextView button.



```
>gi|11513392|pdb|1DH3|A Chain A, Crystal Structure Of A Creb Bzip-Cre Complex Reveals The Basis For Creb Faimly Selective Dimerization And Dna Binding
KREVRLMKNREAARESRRKKKEYVKSLENRVAVLENQNKTLIEELKALKDLYSHK
```

4.8.5-1 Node3 CBRCViewer – POODLE Result



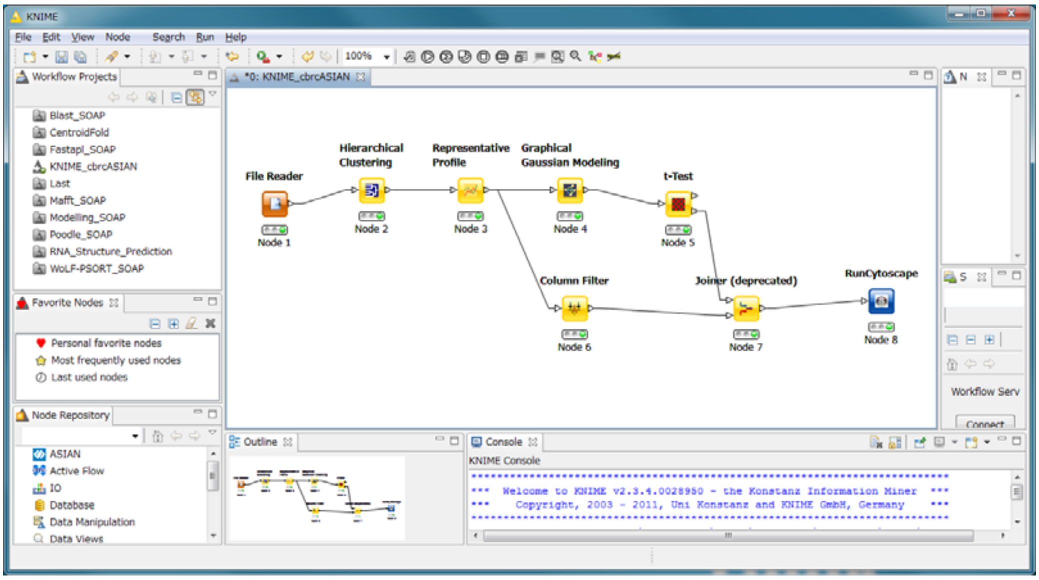
```
PFRMAT DR
REMARK K. Shimizu, Y. Muraoka, S. Hirose, and T. Noguchi
REMARK ″Feature Selection Based on Physicochemical Properties of
REMARK Redefined N-term Region and C-term Regions for Predicting Disorder″
REMARK Proc. of IEEE CIBCB 2005, pp262-267.
METHOD Prediction for short disorder using modified PSSM
METHOD ------
K D 0.712
R D 0.696
E D 0.632
V D 0.555
R O 0.466
L O 0.386
M O 0.32
K O 0.34
N O 0.39
R O 0.481
E D 0.555
A D 0.633
A D 0.651
R D 0.69
E D 0.709
S D 0.729
R D 0.718
R D 0.693
K D 0.629
```

4.8.5-2 Node3 POODLE Result - TextView

4.9 ASIAN Active Workflow

ASIAN (Automatic System for Inferring A Network) developed at CBRC is a network inferring tool that combines a hierarchical clustering with graphical Gaussian modeling (GGM).

Please refer to the ASIAN web site for the details. http://eureka.cbrc.jp/asian/



```
************************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Miner  ***
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany    ***
************************************************************************
```

4.9-1 ASIAN Active Workflow

4.9.1 Preparation

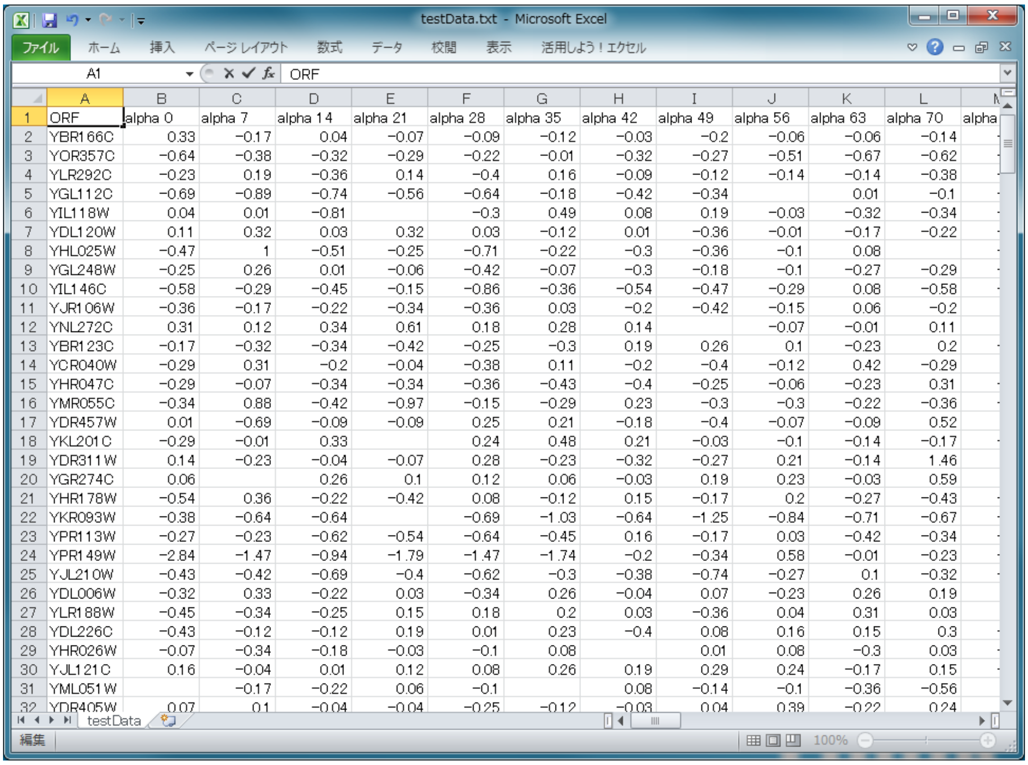
A file needed for execution is a file of matrix form of the gene appearance data.

In ASIAN Active Workflow, a variable to be analyzed is treated by each line. Therefore, the vector of one variable is described in the line.

File Type
Gene appearance data file of matrix format

Gene appearance data of Yeast is shown as an example.

In this example, the experiment name of microarray is described as ORF name and a column name of Yeast ID of the line.



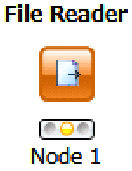
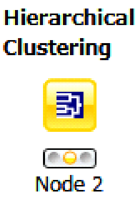
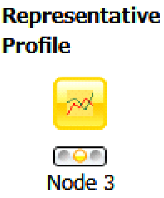
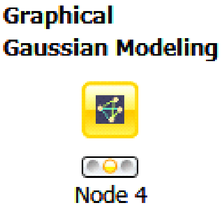
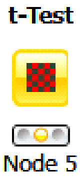
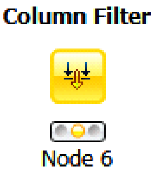
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ORF	alpha 0	alpha 7	alpha 14	alpha 21	alpha 28	alpha 35	alpha 42	alpha 49	alpha 56	alpha 63	alpha 70	alpha
2	YBR166C	0.33	−0.17	0.04	−0.07	−0.09	−0.12	−0.03	−0.2	−0.06	−0.06	−0.14	
3	YOR357C	−0.64	−0.38	−0.32	−0.29	−0.22	−0.01	−0.32	−0.27	−0.51	−0.67	−0.62	
4	YLR292C	−0.23	0.19	−0.36	0.14	−0.4	0.16	−0.09	−0.12	−0.14	−0.14	−0.38	
5	YGL112C	−0.69	−0.89	−0.74	−0.56	−0.64	−0.18	−0.42	−0.34		0.01	−0.1	
6	YIL118W	0.04	0.01	−0.81		−0.3	0.49	0.08	0.19	−0.03	−0.32	−0.34	
7	YDL120W	0.11	0.32	0.03	0.32	0.03	−0.12	0.01	−0.36	−0.01	−0.17	−0.22	
8	YHL025W	−0.47	1	−0.51	−0.25	−0.71	−0.22	−0.3	−0.36	−0.1	0.08		
9	YGL248W	−0.25	0.26	0.01	−0.06	−0.42	−0.07	−0.3	−0.18	−0.1	−0.27	−0.29	
10	YIL146C	−0.58	−0.29	−0.45	−0.15	−0.86	−0.36	−0.54	−0.47	−0.29	0.08	−0.58	
11	YJR106W	−0.36	−0.17	−0.22	−0.34	−0.36	0.03	−0.2	−0.42	−0.15	0.06	−0.2	
12	YNL272C	0.31	0.12	0.34	0.61	0.18	0.28	0.14		−0.07	−0.01	0.11	
13	YBR123C	−0.17	−0.32	−0.34	−0.42	−0.25	−0.3	0.19	0.26	0.1	−0.23	0.2	
14	YCR040W	−0.29	0.31	−0.2	−0.04	−0.38	0.11	−0.2	−0.4	−0.12	0.42	−0.29	
15	YHR047C	−0.29	−0.07	−0.34	−0.34	−0.36	−0.43	−0.4	−0.25	−0.06	−0.23	0.31	
16	YMR055C	−0.34	0.88	−0.42	−0.97	−0.15	−0.29	0.23	−0.3	−0.3	−0.22	−0.36	
17	YDR457W	0.01	−0.69	−0.09	−0.09	0.25	0.21	−0.18	−0.4	−0.07	−0.09	0.52	
18	YKL201C	−0.29	−0.01	0.33		0.24	0.48	0.21	−0.03	−0.1	−0.14	−0.17	
19	YDR311W	0.14	−0.23	−0.04	−0.07	0.28	−0.23	−0.32	−0.27	0.21	−0.14	1.46	
20	YGR274C	0.06		0.26	0.1	0.12	0.06	−0.03	0.19	0.23	−0.03	0.59	
21	YHR178W	−0.54	0.36	−0.22	−0.42	0.08	−0.12	0.15	−0.17	0.2	−0.27	−0.43	
22	YKR093W	−0.38	−0.64	−0.64		−0.69	−1.03	−0.64	−1.25	−0.84	−0.71	−0.67	
23	YPR113W	−0.27	−0.23	−0.62	−0.54	−0.64	−0.45	0.16	−0.17	0.03	−0.42	−0.34	
24	YPR149W	−2.84	−1.47	−0.94	−1.79	−1.47	−1.74	−0.2	−0.34	0.58	−0.01	−0.23	
25	YJL210W	−0.43	−0.42	−0.69	−0.4	−0.62	−0.3	−0.38	−0.74	−0.27	0.1	−0.32	
26	YDL006W	−0.32	0.33	−0.22	0.03	−0.34	0.26	−0.04	0.07	−0.23	0.26	0.19	
27	YLR188W	−0.45	−0.34	−0.25	0.15	0.18	0.2	0.03	−0.36	0.04	0.31	0.03	
28	YDL226C	−0.43	−0.12	−0.12	0.19	0.01	0.23	−0.4	0.08	0.16	0.15	0.3	
29	YHR026W	−0.07	−0.34	−0.18	−0.03	−0.1	0.08		0.01	0.08	−0.3	0.03	
30	YJL121C	0.16	−0.04	0.01	0.12	0.08	0.26	0.19	0.29	0.24	−0.17	0.15	
31	YML051W		−0.17	−0.22	0.06	−0.1		0.08	−0.14	−0.1	−0.36	−0.56	
32	YDR405W	0.07	0.1	−0.04	−0.04	−0.25	−0.12	−0.03	0.04	0.39	−0.22	0.24	

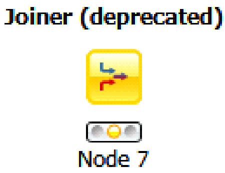
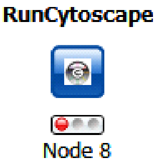
4.9.1-1 ASIAN Active Workflow：sample matrix file

4.9.2 Node

There are 8 nodes.

4.9.2-1 ASIAN Active Workflow Node list

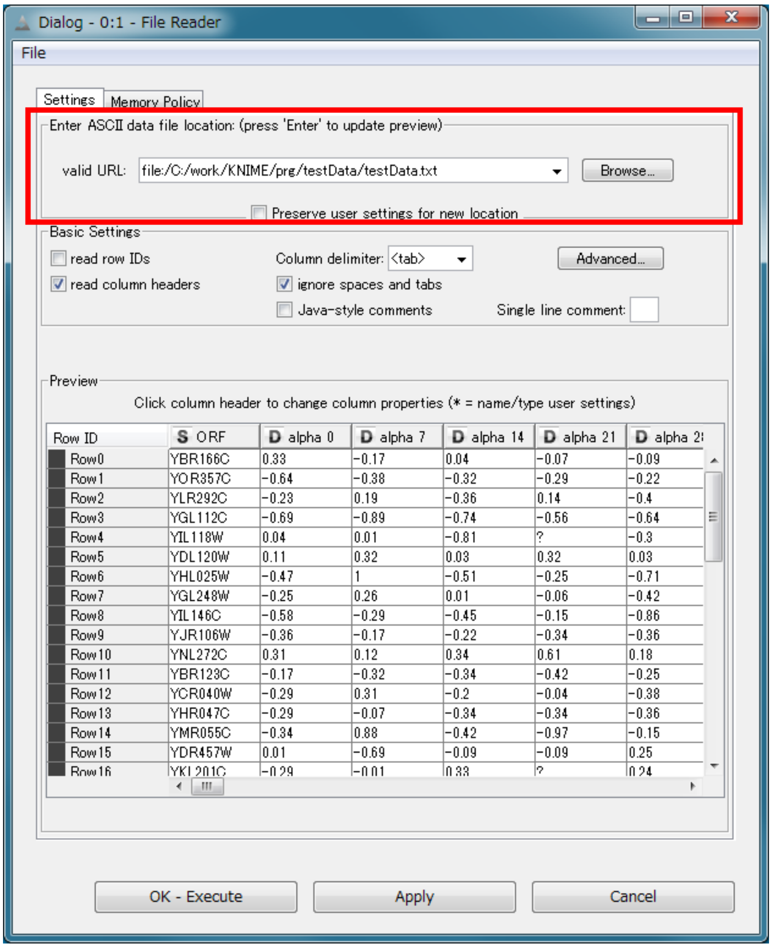
Node ID	Node name	Icon	explanation
Node 1	File Reader		The matrix file is read.
Node 2	Hierarchical Clustering		Execute Hierarchical Clustering.
Node 3	Representative Profile		change the profile data to the representative.
Node 4	Graphical Gaussian Modeling		Execute GGM.
Node 5	t-Test		Execute t-test.
Node 6	Column Filter		Execute column filter.

Node 7	Joiner (deprecated)	 Joiner (deprecated) Node 7	Execute uniting columns.
Node 8	RunCytoscape	 RunCytoscape Node 8	Execute Cytoscape.

4.9.3 Configuring running environment

1. Node1　:　File Reader

Select a matrix file of gene appearance data as an input in “Configure” in the right-click-menu.



Enter ASCII data file location: (press 'Enter' to update preview)

valid URL: file:/C:/work/KNIME/prg/testData/testData.txt

□ Preserve user settings for new location

Basic Settings

□ read row IDs

☑ read column headers

Column delimiter: <tab>

☑ ignore spaces and tabs

□ Java-style comments

Single line comment:

Preview

Click column header to change column properties (* = name/type user settings)

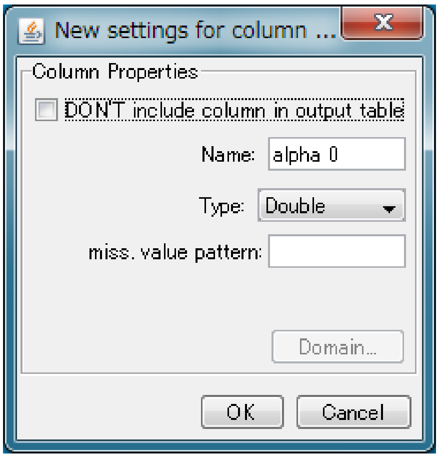
Row ID	S ORF	D alpha 0	D alpha 7	D alpha 14	D alpha 21	D alpha 2
Row0	YBR166C	0.33	-0.17	0.04	-0.07	-0.09
Row1	YOR357C	-0.64	-0.38	-0.32	-0.29	-0.22
Row2	YLR292C	-0.23	0.19	-0.36	0.14	-0.4
Row3	YGL112C	-0.69	-0.89	-0.74	-0.56	-0.64
Row4	YIL118W	0.04	0.01	-0.81	?	-0.3
Row5	YDL120W	0.11	0.32	0.03	0.32	0.03
Row6	YHL025W	-0.47	1	-0.51	-0.25	-0.71
Row7	YGL248W	-0.25	0.26	0.01	-0.06	-0.42
Row8	YIL146C	-0.58	-0.29	-0.45	-0.15	-0.86
Row9	YJR106W	-0.36	-0.17	-0.22	-0.34	-0.36
Row10	YNL272C	0.31	0.12	0.34	0.61	0.18
Row11	YBR123C	-0.17	-0.32	-0.34	-0.42	-0.25
Row12	YCR040W	-0.29	0.31	-0.2	-0.04	-0.38
Row13	YHR047C	-0.29	-0.07	-0.34	-0.34	-0.36
Row14	YMR055C	-0.34	0.88	-0.42	-0.97	-0.15
Row15	YDR457W	0.01	-0.69	-0.09	-0.09	0.25
Row16	YKL201C	-0.29	-0.01	0.33	?	0.24

4.9.3-1 FastaFileReader　: Configure…

・Settings tab → Enter ASCII data file location: (press ‘Enter’ to update preview) → valid URL:

Enter the location of an input file. "Browse…" can be used for browing a file. After a file is specified, the read file is displayed in a lower Preview column.

When the column header in the Preview column is pressed, the following screens are displayed.



New settings for column ...

Column Properties

□ DON'T include column in output table

Name: alpha 0

Type: Double

miss. value pattern:

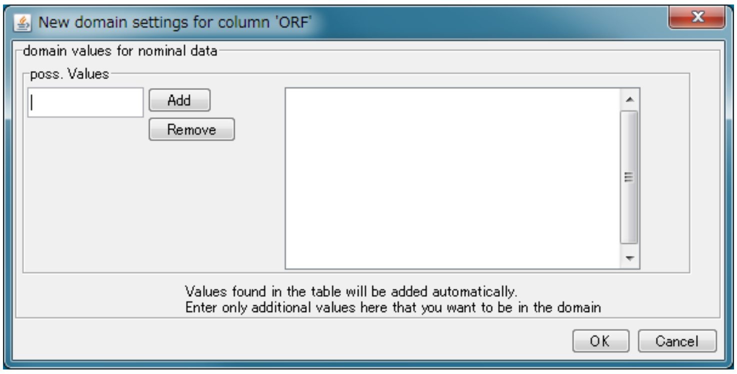
4.9.3-2 Configure… → Column Properties

In this window, whether the output file contains the column name, etc. is configured.

- DON'T include column in output table

  --- Tick the check-box if the output file does not include column names.
- Name --- The column name is to change.
- Type --- The type of data in the column is to change.
- miss. value pattern --- Enter a value, which is not included in analysis.

・Domain… --- Enter a domain name in the dialog below, which is added to the column.



New domain settings for column 'ORF'

domain values for nominal data

poss. Values

Values found in the table will be added automatically.
Enter only additional values here that you want to be in the domain

4.9.3-3 Column Properties → Domain…

・Settings tab → Enter ASCII data file location: (press 'Enter' to update preview) → Preserve user settings for new location

Tick the check-box if the user settings are to preserve in figure 4.9.3-2.

・Settings tab → Basic Settings

In Basic Settings at the center of Settings tab, basic settings need to be done.

・Read row IDs ：Row IDs are to be read.

・Column delimiter ：Select a delimiter in the input file from the pull-down menu.

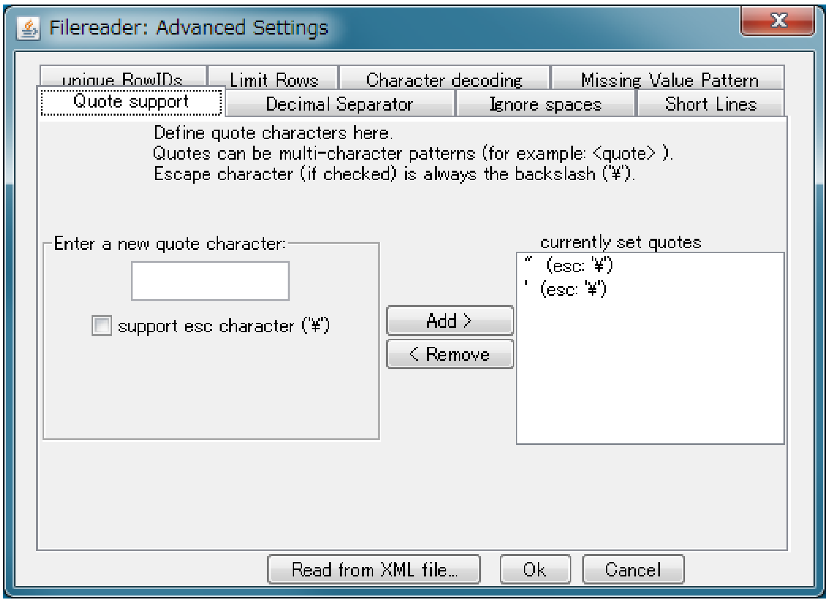
・Read column headers ：Column header of the input file is to be read.

・Ignore spaces and tabs ：Space and tab are to be disregarded.

・The comment on the Java-style comments ：Java style is to be read.

・Single line comment ：A key to the line comment is to be set.

・Advanced… ：In addition, to do detailed settings, the following screen appears.



Define quote characters here.
Quotes can be multi-character patterns (for example: <quote>).
Escape character (if checked) is always the backslash ('¥').

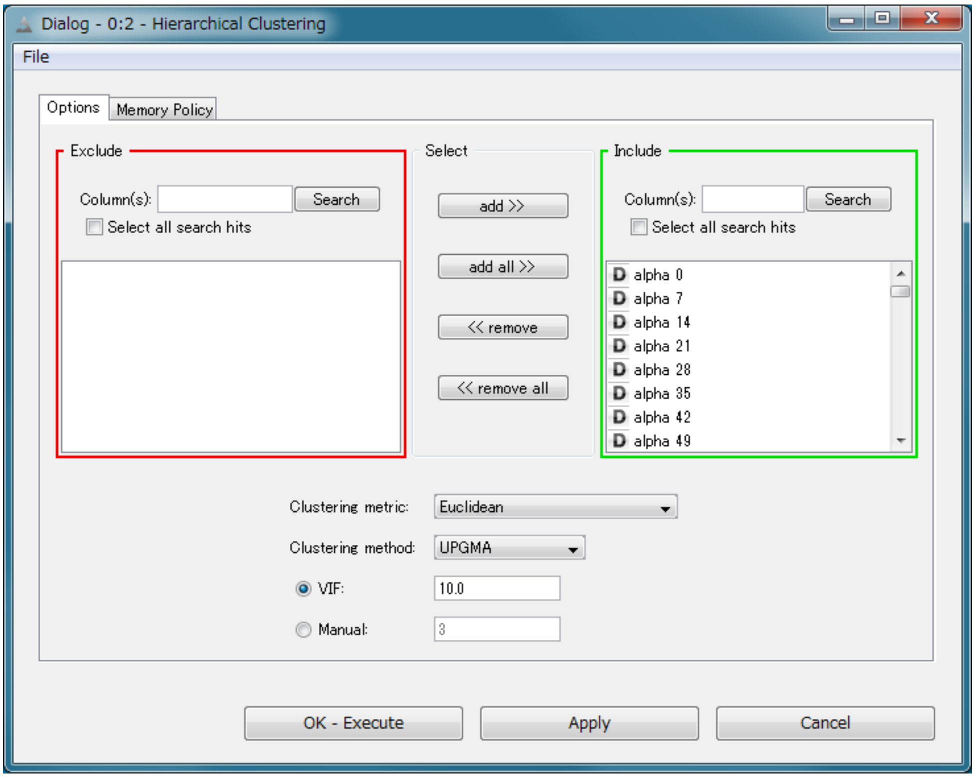
4.9.3-4 Basic Settings → Advanced…

Press “OK” after specifying.

Select "Execute" in the right-click-menu for execution.

2. Node2 : Hierarchical Clustering

Set a hierarchical clustering parameter between variables in “Configure” using the right-click-menu.



Include:
- D alpha 0
- D alpha 7
- D alpha 14
- D alpha 21
- D alpha 28
- D alpha 35
- D alpha 42
- D alpha 49

Clustering metric: Euclidean

Clustering method: UPGMA

● VIF: 10.0

○ Manual: 3

4.9.3-5 Hierarchical Clustering : Configure...

・Options tab

Select columns for hierarchical clustering by adding to “Include” section. In default, all columns will be processed.

Set parameters for execution.

・Clustering metric : Select from the following.

- Euclidean (Euclidean distance)
- Pearson Correlation Coefficient (Pearson correlation coefficient.)
- Eisen Correlation Coefficient (Correlation coefficient.)
- Euclidean between Correlations (Euclidean distance between correlation coefficient vectors.)

・Clustering method : Select from the following.

- Single Linkage
- Complete Linkage

- UPGMA
- WPGMA
- Wards
  - Big N (requires less memory using Reciprocal nearest neighbor method, however, requires more time. Its results are the same as Wards method.)

・VIF：Enter a numerical value.
A number of clusters based on Variance Inflation Factor is inferred. The default value is 10.

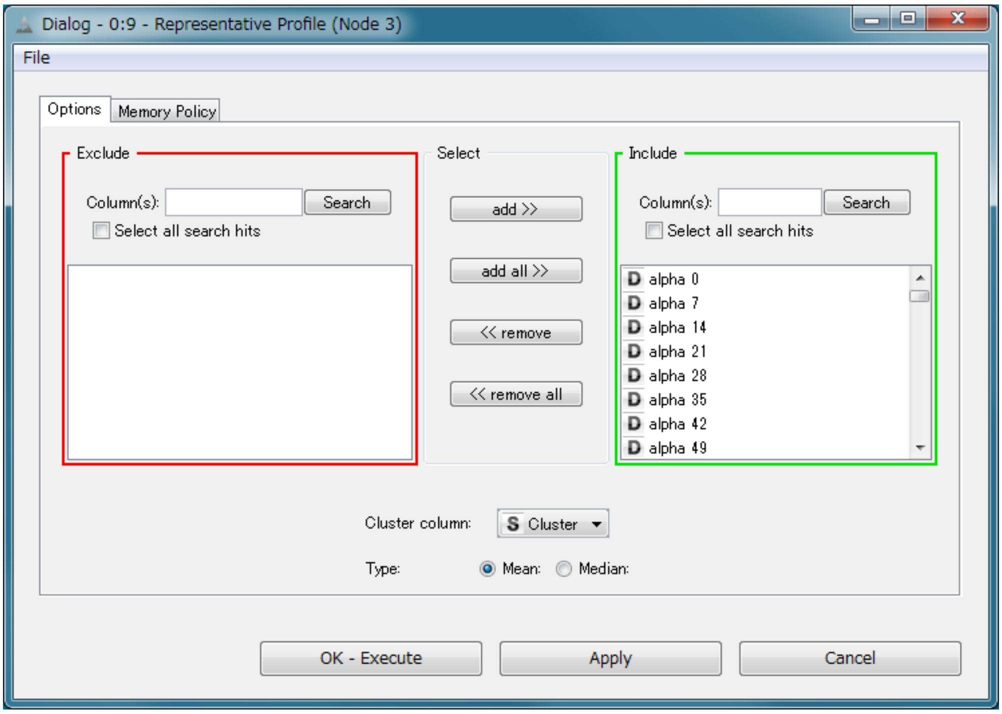
・Manual：Enter a number of clusters.
Wards method and Big N method can be used only in Euclidean distance. The default value is 3.

Press “OK” after specifying.

Select “Execute” in the right-click-menu for execution.

3. Node3 : Representative Profile

Set options for representative profile in “Configure” using the right-click-menu.



4.9.3-6 Representative Profile : Configure…

・**Options　tab**

Select columns for representative by adding to “Include” section. In default, all columns will be processed.

Set an option for representative.

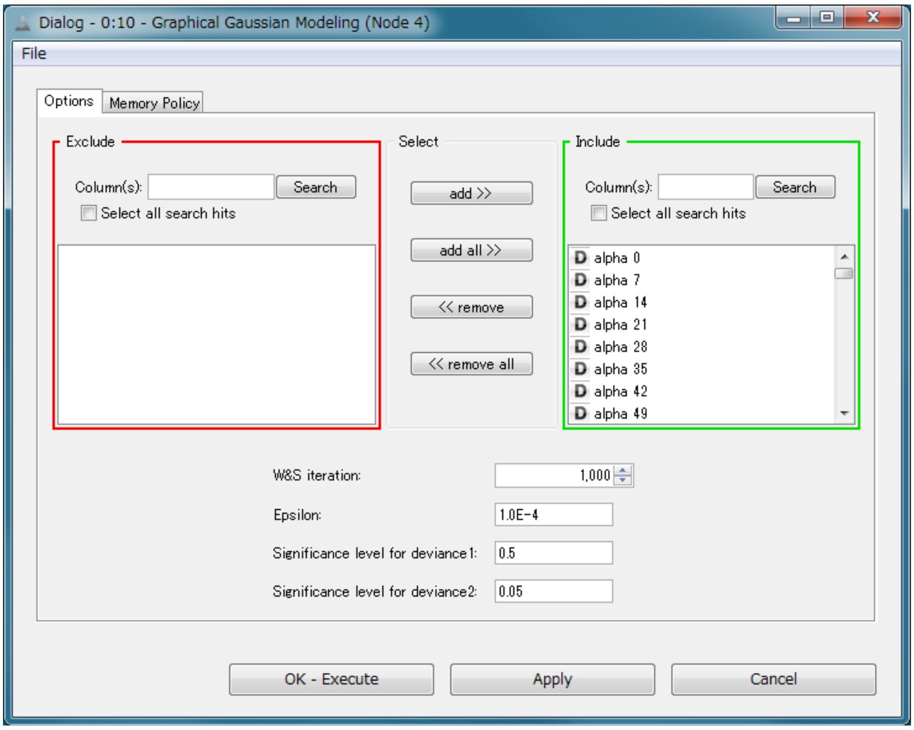
・Type：Select mean or median.

Press “OK” after selecting.

Select “Execute” in the right-click-menu for execution.

4. <u>Node4　: Graphical Gaussian Modeling</u>

Set options for Graphical Gaussian Modeling(GGM) in “Configure” using the right-click-menu.



Dialog - 0:10 - Graphical Gaussian Modeling (Node 4)

File

Options | Memory Policy

Exclude

Column(s): Search

□ Select all search hits

Select

add >>

add all >>

<< remove

<< remove all

Include

Column(s): Search

□ Select all search hits

- alpha 0
- alpha 7
- alpha 14
- alpha 21
- alpha 28
- alpha 35
- alpha 42
- alpha 49

W&S iteration: 1,000

Epsilon: 1.0E-4

Significance level for deviance1: 0.5

Significance level for deviance2: 0.05

OK - Execute | Apply | Cancel

4.9.3-7 Graphical Gaussian Modeling：Configure…

・**Options　tab**

Select columns for GGM by adding to “Include” section. In default, all columns will be processed.

Set options.

- W&S iteration :

  Enter a value of iteration for Wermuth/Scheidt algorithm. The default value is 1000.

- Epsilon :

  Enter a value of Epsilon. The default value is 1e-4.

- Significance level for deviance1 :

  Enter a value of Significance level for deviance1. The default value is 0.5.

- Significance level for deviance2 :

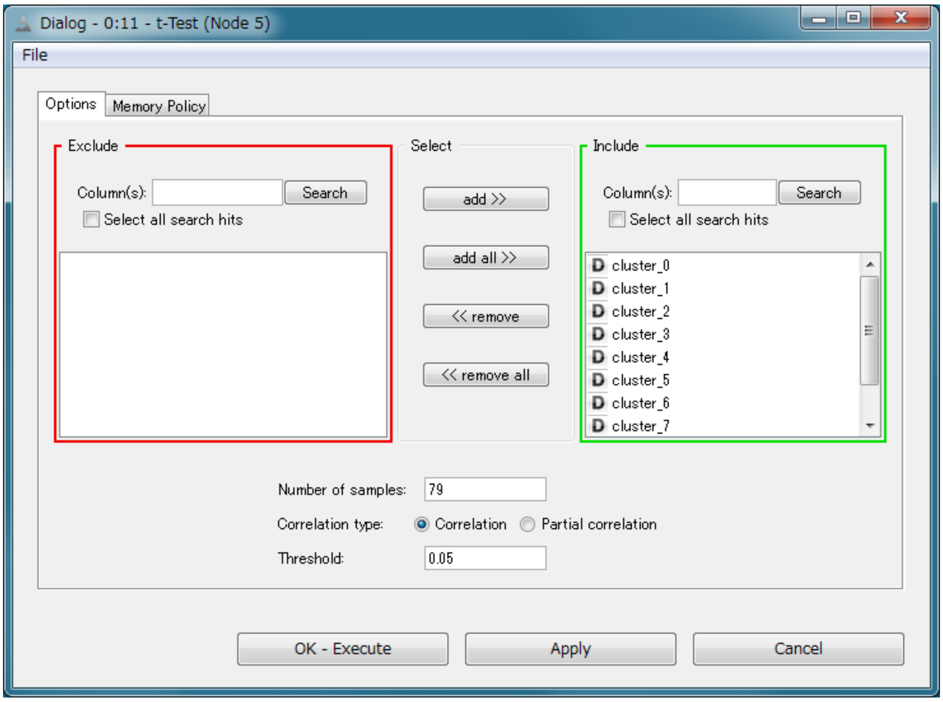
  Enter a value of Significance level for deviance2. The default value is 0.01.

Press “OK” after entering values.

Select “Execute” in the right-click-menu for execution.

5. Node5　: t-Test

Set options for t-Test in “Configure” using the right-click-menu.



Include: cluster_0, cluster_1, cluster_2, cluster_3, cluster_4, cluster_5, cluster_6, cluster_7

Number of samples: 79

Correlation type: ● Correlation ○ Partial correlation

Threshold: 0.05

4.9.3-8 t-Test：Configure...

・Options tab

Select columns for t-Test by adding to “Include” section. In default, all columns will be processed.

Set parameters for t-Test.

・Number of samples：

Enter a value. The default value is 79.

・Correlation type：

Select either correlation coefficient (Correlation) or partial correlation coefficient (Partial correlation).

・Threshold：

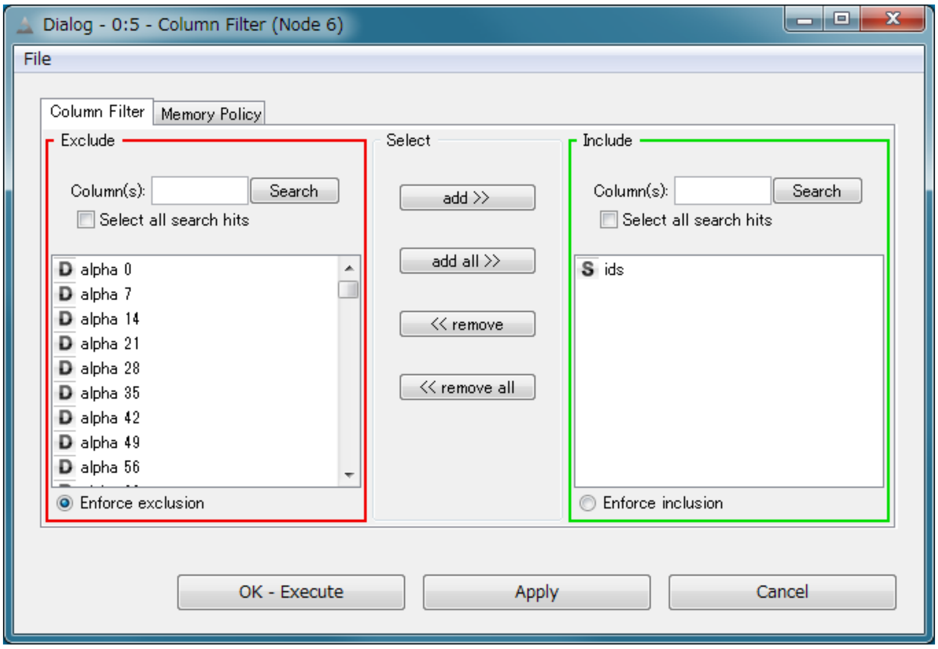
Enter a value for significant level. The default value is 0.05.

Press “OK” after completing.

Select “Execute” in the right-click-menu for execution.

6. <u>Node6 ：Column Filter</u>

Set column filter in “Configure” using the right-click-menu.



Dialog - 0:5 - Column Filter (Node 6)

Exclude

Column(s): Search

□ Select all search hits

D alpha 0
D alpha 7
D alpha 14
D alpha 21
D alpha 28
D alpha 35
D alpha 42
D alpha 49
D alpha 56

● Enforce exclusion

Include

Column(s): Search

□ Select all search hits

S ids

○ Enforce inclusion

4.9.3-9 Column Filter：Configure…

・**Column Filter　tab**

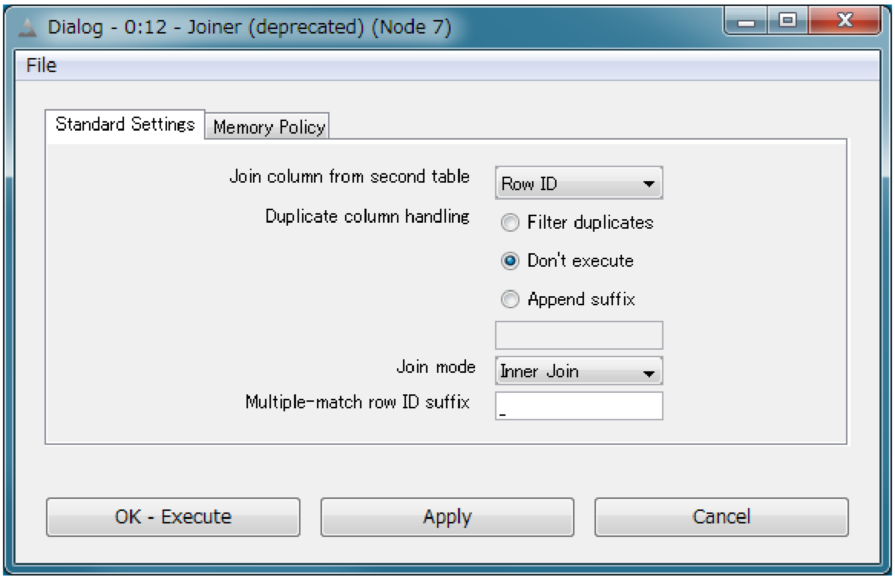
Select columns for t-Test by adding to “Include” section.

Press “OK” after selecting.

Select “Execute” in the right-click-menu for execution.

7. Node7　: Joiner(deprecated)

Set column join in “Configure” using the right-click-menu.



Join column from second table: Row ID

Duplicate column handling:
○ Filter duplicates
● Don't execute
○ Append suffix

Join mode: Inner Join

Multiple-match row ID suffix: _

4.9.3-10 Joiner(deprecated)：Configure…

・**Standart Settings　tab**

・Join column from second table --- Select Row ID or ids.

・Duplicate column handling --- Select Fileter duplicates, Don’t execute or Append suffix. Enter suffix in case of Append suffix.

・Join mode --- Select either Inner Join, Left Outer Join, Right Outer Join or Full Outer Join

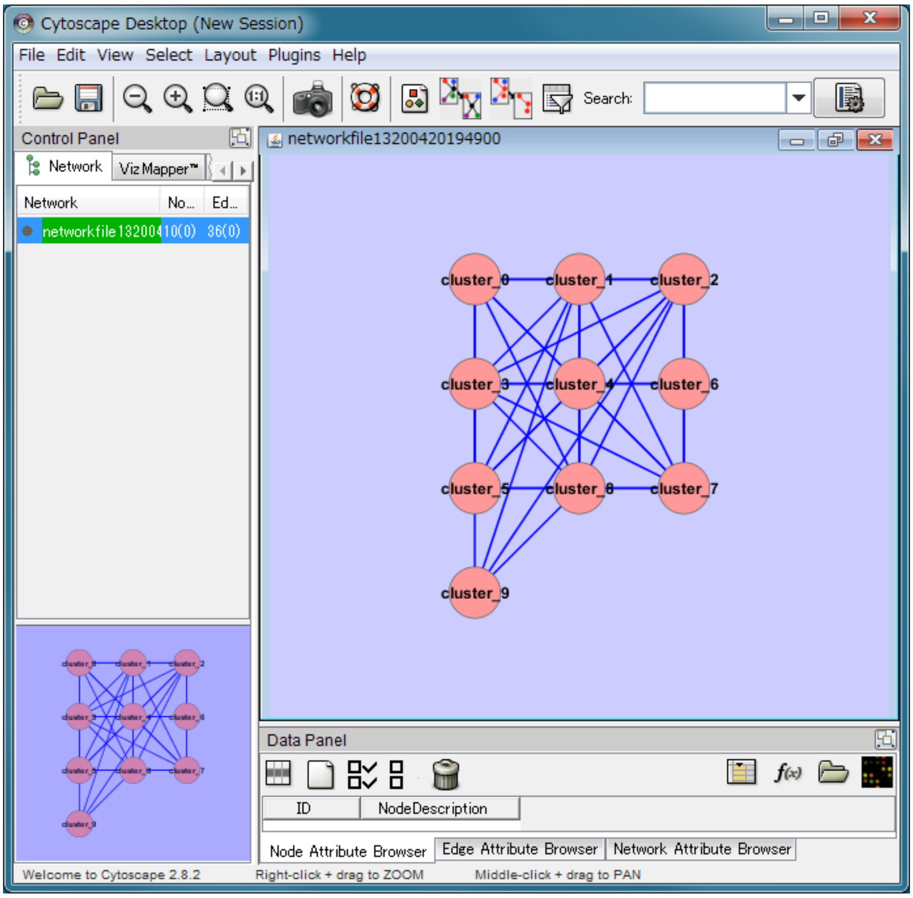
・Multiple-match row ID suffix --- Enter Suffix for multiple-joined Row ID.

Press “OK” after completing.

Select “Execute” in the right-click-menu for execution.

8. Node8 : RunCytoscape

Select “Execute and Open Views” in the right-click-menu to execute Cytoscape.


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search:
Control Panel
Network
VizMapper™
Network No... Ed...
networkfile13200410(0) 36(0)
networkfile13200420194900
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4
cluster_6
cluster_5
cluster_8
cluster_7
cluster_9
Data Panel
ID
NodeDescription
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

4.9.3-11 Cytoscape

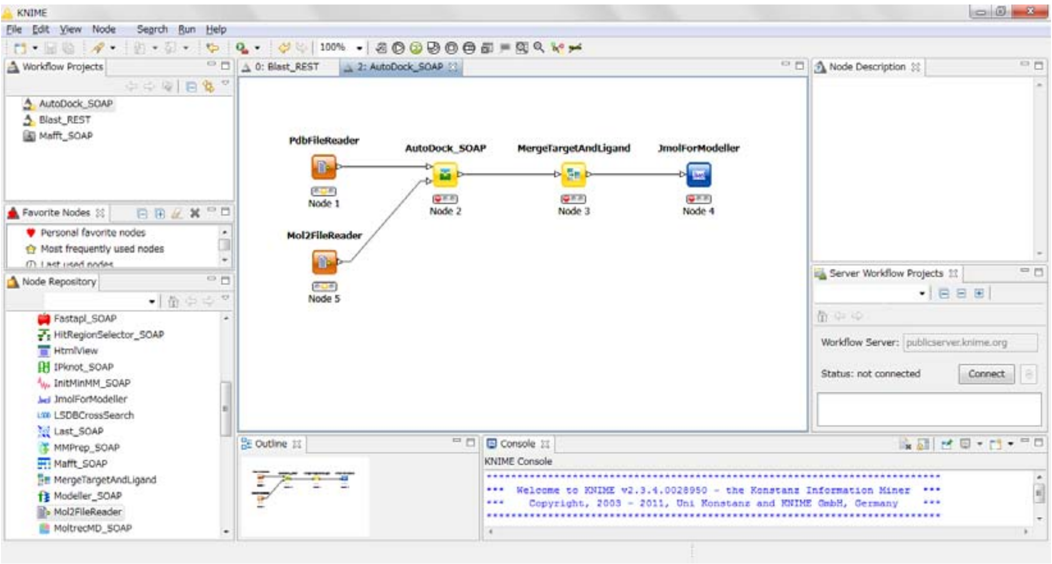
Please refer to the following sites for the details of Cytoscape.

Cytoscape : http://www.cytoscape.org/

4.10 AutoDock Active Workflow

AutoDock_SOAP executes AUTODOCK, which is widely used protein-ligand docking software developed at Scripps Institute (http://autodock.scripps.edu), via SOAP. The user needs to provide two things. A target protein PDB file (a single chain protein NOT a protein complex) without bound ligands and a MOL2-formatted molecule file. The program will automatically identify potential binding sites and calculate binding energy.

AutoDock : http://autodock.scripps.edu



```
*************************************************************************
***  Welcome to KNIME v2.3.4.0028950 - the Konstanz Information Miner  ***
***    Copyright, 2003 - 2011, Uni Konstanz and KNIME GmbH, Germany     ***
*************************************************************************
```

4.10-1 AutoDock Active Workflow

4.10.1 Preparation

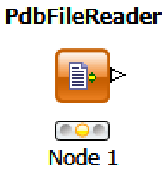
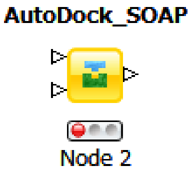
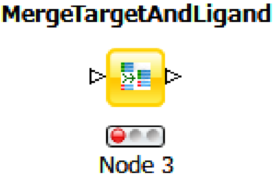
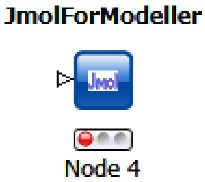
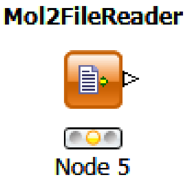
This node requires two files, PDB format file and MOL2 format file.

File Type
PDB format file
MOL2 format file

4.10.2 Node

There are 5 nodes.

4.10.2-1 AutoDock Active Workflow

Node	Name	Icon	Description
Node 1	PdbFileReader	 PdbFileReader Node 1	Read PDB format file.
Node 2	AutoDock_SOAP	 AutoDock_SOAP Node 2	Execute AutoDock via SOAP.
Node 3	MergeTargetAndLigand	 MergeTargetAndLigand Node 3	Merge PDB format file and AutoDock results file.
Node4	JmolForModeller	 JmolForModeller Node 4	Launch Jmol.
Node 5	Mol2FileReader	 Mol2FileReader Node 5	Read MOL2 format file.

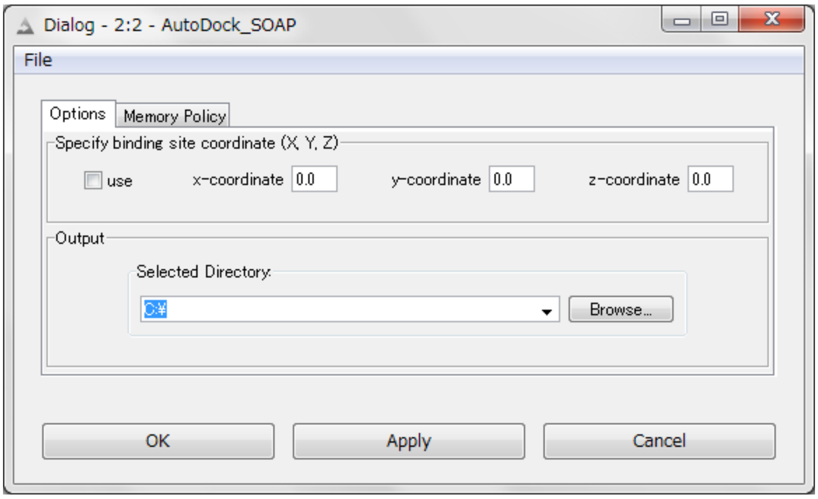
4.10.3 Step1. Node setting

1. Node1 : PdbFileReader

   Select a PDB file as an input using right-click-menu.

2. Node2 : AutoDock_SOAP

   Specify an absolute path of a directory to store AutoDock results, or select the directory using “Browse…” button.



Specify binding site coordinate (X, Y, Z)

□ use x-coordinate 0.0 y-coordinate 0.0 z-coordinate 0.0

Output

Selected Directory: C:¥

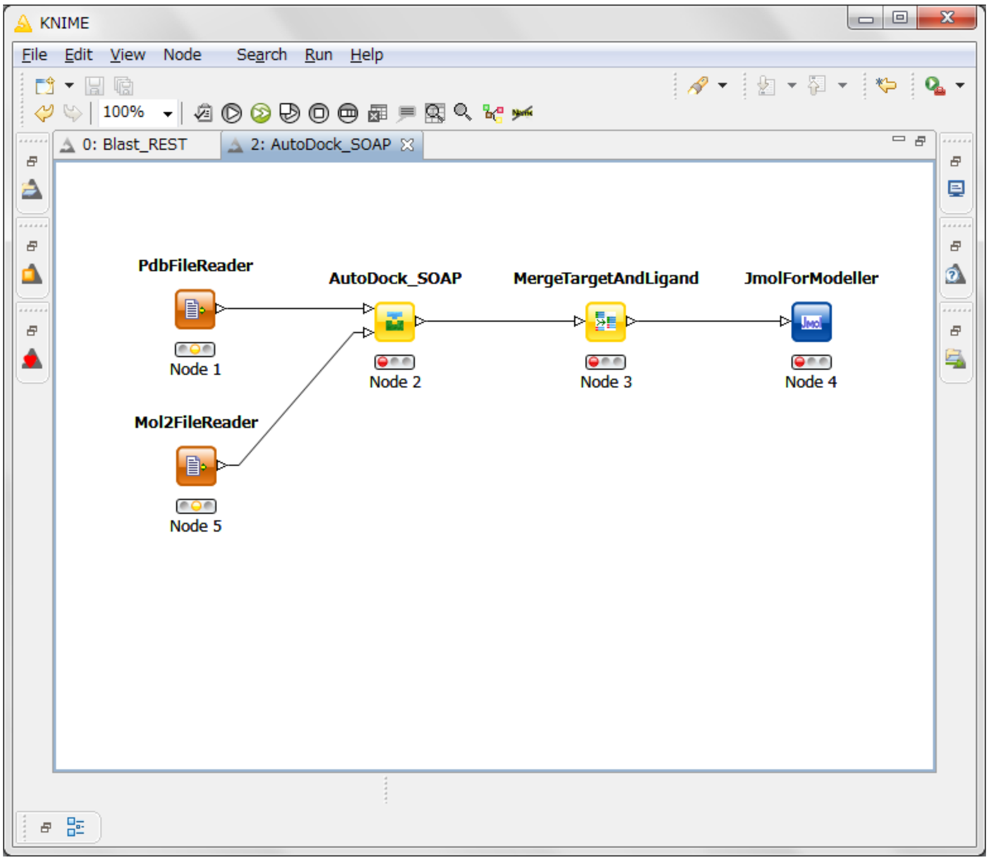
4.10.3-1 AutoDock_SOAP : Configure…

If you specify binding site coordinate, check a “use” and input coordinates in XYZ coordinates text boxes.

3. Node5 : Mol2FileReader

   Select a MOL2 file as an input using right-click-menu.

4.10.4 Step2. Execution



4.10.4-1 AutoDock_SOAP workflow

AutoDock_SOAP workflow is executed according to the following steps.

1) Node1 : PdbFileReader

   If the node is yellow, the node is ready to be executed. Right-click on the node, and select "Execute" from the menu.

2) Node2 : AutoDock_SOAP

   If the node is yellow, the node is ready to be executed. Right-click on the node, and select "Execute" from the menu.

3) Node3 : MergeTargetAndLigand

   If the node is yellow, the node is ready to be executed. Right-click on the node, and select "Execute" from the menu.

4) Node4 : JmolForModeller

If the node is yellow, the node is ready to be executed. Right-click on the node, and select “Execute” from the menu.

If the status light changes to green, the node is successfully finished. Right-click on the node, and select ”View：name of first view” from the menu.

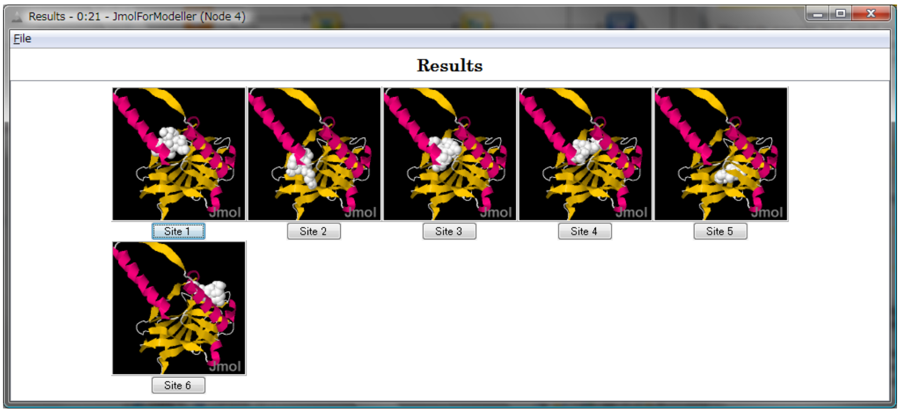
5) Node5 : Mol2FileReader

If the node is yellow, the node is ready to be executed. Right-click on the node, and select “Execute” from the menu.

4.10.5 Step.3 Execution results

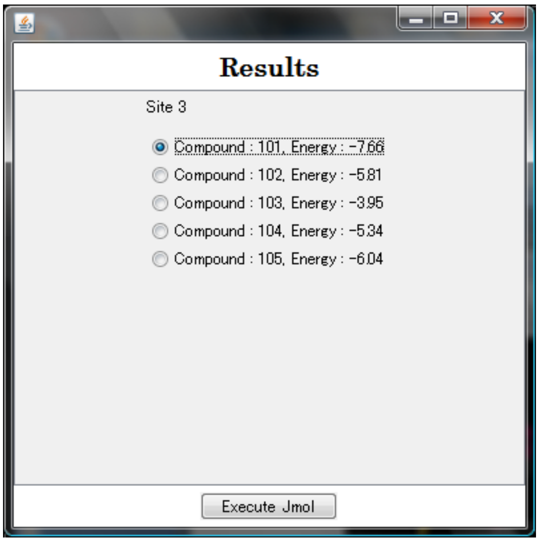
1) Node4 JmolForModeller – Result

Execution results of AutoDock_SOAP are displayed using JmolForModeller node.



4.10.5-1 Node4 JmolForModeller – Results

JmolForModeller executes Jmol, which is an application of molecule viewer. In the case of AutoDock_SOAP, there are some docking results in each docking site of a template protein structure (Figure 4.10.5-1). To display these results, click a "Site" button located under each image (Figure 4.10.5-2), and a docking result menu is opened.



Site 3

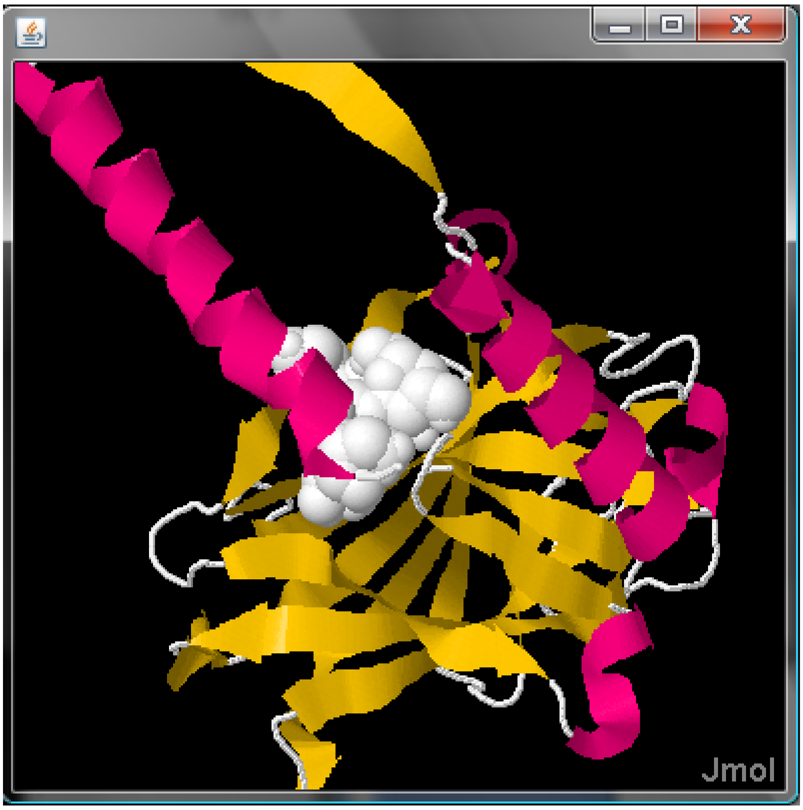
● Compound : 101, Energy : -7.66
○ Compound : 102, Energy : -5.81
○ Compound : 103, Energy : -3.95
○ Compound : 104, Energy : -5.34
○ Compound : 105, Energy : -6.04

4.10.5-2 Docking Result menu

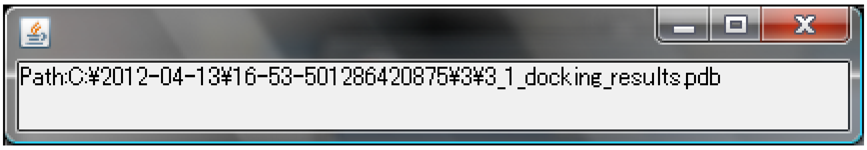
Select a radio button corresponding to each docking result and click "Execute Jmol" button. Jmol is launched and selected docking result is displayed (Figure 4.10.5-3). At a time, a pop up window is opened. This window displays an absolute path of the docking result file (Figure 4.10.5-4).

Please visit a Jmol web site for further information.

Jmol：http://jmol.sourceforge.net/



4.10.5-3 Jmol



```
Path:C:¥2012-04-13¥16-53-501286420875¥3¥3_1_docking_results.pdb
```

4.10.5-4 Pop up window to display an absolute path of a docking file

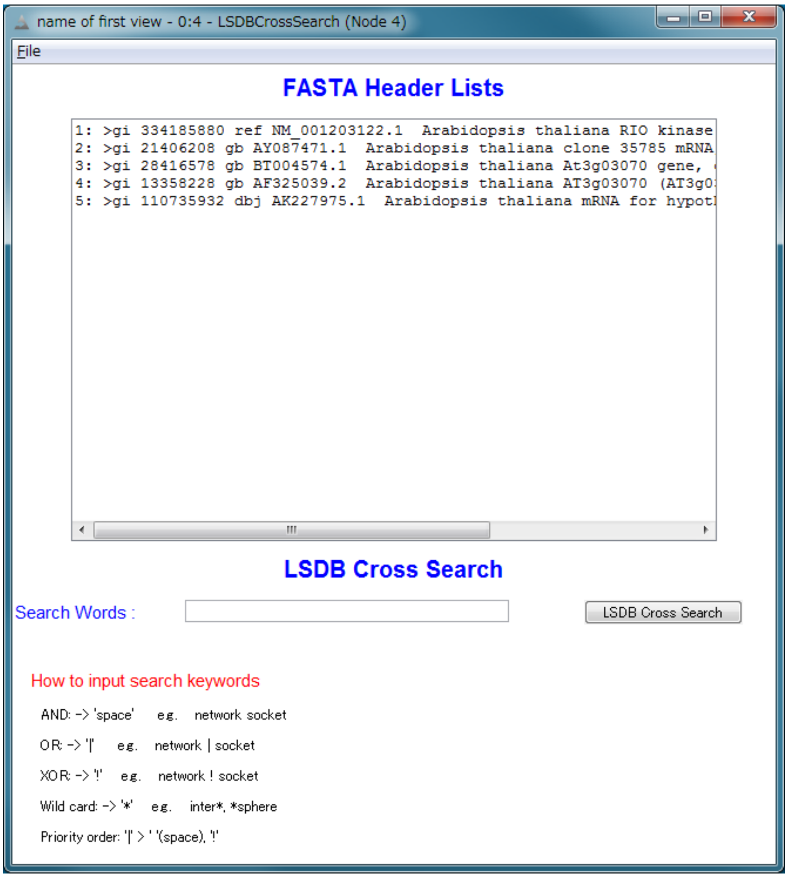
5 Appendix

5.1 Appendix A : LSDBCrossSearch

Life Science DataBase cross-search can be executed in green node status after executing LSDBCrossSearch node.

Life Science DataBase cross-search site was developed in the Database Integration project.promoted by Ministry of Education, Culture, Sports, Science and Technology.

If “View” is selected in right-click-menu on LSDBCrossSearch node, View window of LSDBCrossSearch node will appear.



```
1: >gi 334185880 ref NM_001203122.1  Arabidopsis thaliana RIO kinase
2: >gi 21406208 gb AY087471.1  Arabidopsis thaliana clone 35785 mRNA
3: >gi 28416578 gb BT004574.1  Arabidopsis thaliana At3g03070 gene, 
4: >gi 13358228 gb AF325039.2  Arabidopsis thaliana AT3g03070 (AT3g0
5: >gi 110735932 dbj AK227975.1  Arabidopsis thaliana mRNA for hypot
```

How to input search keywords

AND: -> 'space' e.g. network socket

OR: -> '|' e.g. network | socket

XOR: -> '!' e.g. network ! socket

Wild card: -> '*' e.g. inter*, *sphere

Priority order: '|' > ' '(space), '!'

5.1-1 LSDBCrossSearch View window

Headers of the FASTA file used for LSDBCrossSearch node are shown in FASTA Header Lists.

A keyword(s) for cross-search should be entered in the text box.

For a combined search, the following symbols should be used:

・AND retrieval: Space " "

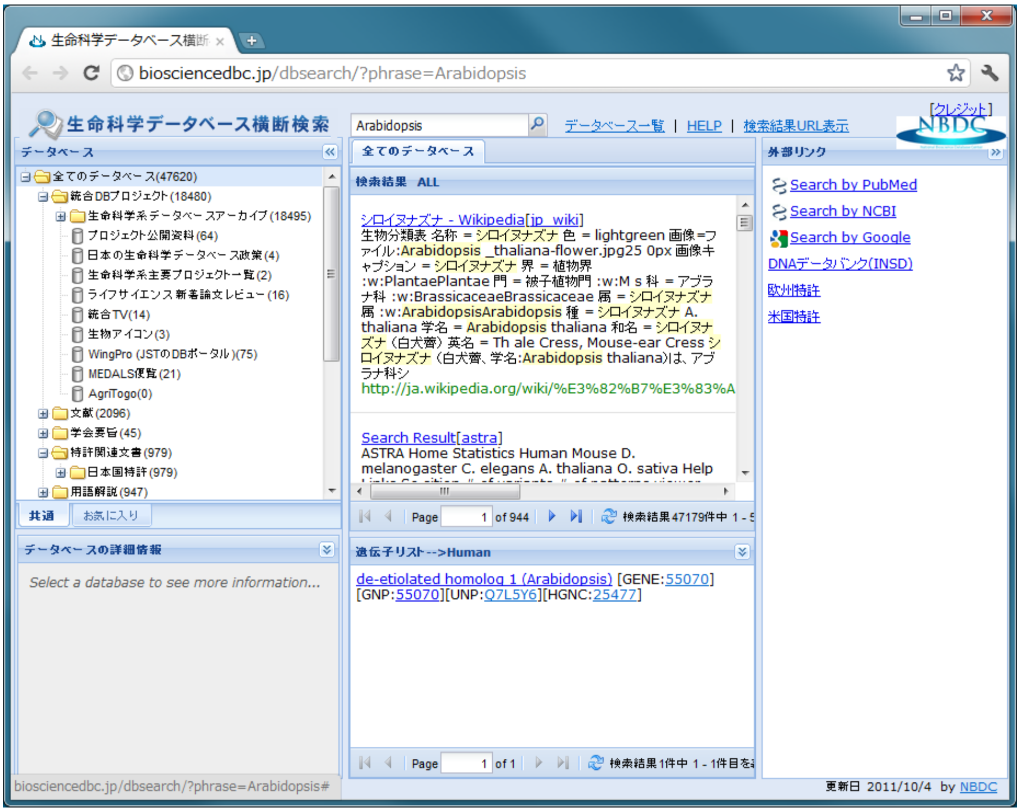
・OR retrieval: Pipe "|　“

・Exclusive-OR retrieval: Exclamation mark ""

・Wildcard search: Asterisk "* "

OR has the highest priority.

Cross-search will be carried out by clicking LSDB Cross Search button, and a Web browser of life science database cross-search will appear as shown below.



シロイヌナズナ - Wikipedia[jp_wiki]
生物分類表 名称 = シロイヌナズナ 色 = lightgreen 画像=ファイル:Arabidopsis _thaliana-flower.jpg25 0px 画像キャプション = シロイヌナズナ 界 = 植物界 :w:PlantaePlantae 門 = 被子植物門 :w:M s 科 = アブラナ科 :w:BrassicaceaeBrassicaceae 属 = シロイヌナズナ属 :w:ArabidopsisArabidopsis 種 = シロイヌナズナ A. thaliana 学名 = Arabidopsis thaliana 和名 = シロイヌナズナ（白犬薺）英名 = Th ale Cress, Mouse-ear Cress シロイヌナズナ（白犬薺、学名:Arabidopsis thaliana)は、アブラナ科シ
http://ja.wikipedia.org/wiki/%E3%82%B7%E3%83%A

Search Result[astra]
ASTRA Home Statistics Human Mouse D. melanogaster C. elegans A. thaliana O. sativa Help

de-etiolated homolog 1 (Arabidopsis) [GENE:55070]
[GNP:55070][UNP:Q7L5Y6][HGNC:25477]

5.1-2 LSDB window

Please refer to the life science database cross-search site for the details.

Life Science DataBase Site：　http://biosciencedbc.jp/dbsearch/

5.2 Appendix B：Last parameter

5.2.1 lastal parameter

Option description for LAST has been taken from LAST web site.

```
Options
-------
Cosmetic Options
~~~~~~~~~~~~~~~~

  -h  Show all options and their default settings.
  -v  Be verbose: write messages about what lastal is doing.
  -o FILE
      Write output to the specified file, instead of the screen.
  -f NUMBER
      Choose the output format: 0 means tabular and 1 means MAF.  MAF
      format looks like this:
        a score=15
        s chr3L        19433515 23 + 24543557 TTTGGGAGTTGAAGTTTTCGCCC
        s H04BA01F1907        2 21 +       25 TTTGGGAGTTGAAGGTT--GCCC
      Lines starting with "s" contain: the sequence name, the start
      coordinate of the alignment, the number of sequence letters
      spanned by the alignment, the strand, the sequence length, and
      the aligned letters.  The start coordinates are zero-based.  If
      the strand is "-", the start coordinate is in the reverse
      strand.
      The same alignment in tabular format looks like this:
        15 chr3L 19433515 23 + 24543557 H04BA01F1907 2 21 + 25 17,2:0,4
      The final column shows the sizes and offsets of gapless blocks
      in the alignment.  In this case, we have a block of size 17,
      then an offset of size 2 in the upper sequence and 0 in the
      lower sequence, then a block of size 4.


Score Options
~~~~~~~~~~~~~

  -r SCORE
```

```
    Match score.
-q COST
    Mismatch cost.
-p FILE
    Obtain match and mismatch scores from the specified file.
    Options -r and -q will be ignored.  For an example of the
    format, see hoxd70.mat in the examples directory.  Any letters
    that aren't in the file will get the lowest score in the file
    when aligned to anything.  Asymmetric scores are allowed: query
    letters correspond to columns and reference letters correspond
    to rows.  Other options can be specified on lines starting with
    "#last", but command line options override them.
-a COST
    Gap existence cost.
-b COST
    Gap extension cost.  A gap of size k costs: a + b*k.
-c COST
    This option allows use of "generalized affine gap costs" (SF
    Altschul 1998, Proteins 32(1):88-96).  Here, a "gap" may consist
    of unaligned regions of both sequences.  If these unaligned
    regions have sizes j and k, where j <= k, the cost is: a +
    b*(k-j) + c*j.  If c >= a + 2b (the default), it reduces to
    standard affine gaps.
-F COST
    Align DNA queries to protein reference sequences, using the
    specified frameshift cost.  A value of 15 seems to be
    reasonable.  The output looks like this:
      a score=108
      s myprot 422  40 +  649 FLLQAVKLQDP-STPHQIVPSP-VSDLIATHTLCPRMKYQDD
      s mydna  878 117 + 1000 FFLQ-IKLWDP¥STPH*IVSSP/PSDLISAHTLCPRMKSQDN
    The "¥" indicates a forward shift by one nucleotide, and the "/"
    indicates a reverse shift by one nucleotide.  The "*" indicates
    a stop codon.  The same alignment in tabular format looks like
    this:
      108 myprot 422 40 + 649 mydna 878 117 + 1000 4,1:0,6,0:1,10,0:-1,19
    The "-1" in the final column indicates the reverse frameshift.
```

```
  -x DROP
      Maximum score drop for gapped alignments.  Gapped alignments are
      forbidden from having any internal region with score < -DROP.
      This serves two purposes: accuracy (avoid spurious internal
      regions in alignments) and speed (the smaller the faster).
  -y DROP
      Maximum score drop for gapless alignments.
  -z DROP
      Maximum score drop for final gapped alignments.
  -d SCORE
      Minimum score for gapless alignments.
  -e SCORE
      Minimum score for gapped alignments.

Miscellaneous Options
~~~~~~~~~~~~~~~~~~~~~


  -s STRAND
      Specify which query strand should be used: 0 means reverse only,
      1 means forward only, and 2 means both.
  -m MULTIPLICITY
      Maximum multiplicity for initial matches.  Each initial match is
      lengthened until it occurs at most this many times in the
      reference.
      If the reference was split into volumes by lastdb, then lastal
      uses one volume at a time.  The maximum multiplicity then
      applies to each volume, not the whole reference.  This is why
      voluming changes the results.
  -l LENGTH
      Minimum length for initial matches.  Length means the number of
      letters spanned by the match.
  -n COUNT
      Maximum number of gapless alignments per query position.  When
      lastal extends gapless alignments from initial matches that
      start at one query position, if it gets COUNT successful
      extensions, it skips any remaining initial matches starting at
```

```
    that position. This option has no effect unless COUNT is less
    than MULTIPLICITY.
-k STEP
    Look for initial matches starting only at every STEP-th position
    in the query.  This makes lastal faster but less sensitive.
-i BYTES
    Search queries in batches of at most this many bytes.  If a
    single sequence exceeds this amount, however, it is not split.
    You can use suffixes K, M, and G to specify KibiBytes,
    MebiBytes, and GibiBytes.  This option has no effect on the
    results (apart from their order), unless k>1.
    If the reference was split into volumes by lastdb, then each
    volume will be read into memory once per query batch.
-u NUMBER
    Specify treatment of lowercase letters when extending
    alignments.  0 means do not mask them; 1 means mask them for
    gapless extensions; 2 means mask them for gapless and gapped
    extensions but not final extensions; 3 means mask them at all
    stages.  "Mask" means change their match/mismatch scores to
    min(unmasked score, 0).  This option performs not affect treatment
    of lowercase for initial matches.
-w DISTANCE
    This option is a kludge to avoid catastrophic time and memory
    usage when self-comparing a large sequence.  If the sequence
    contains a tandem repeat, we may get a gapless alignment that is
    slightly offset from the main self-alignment.  In that case, the
    gapped extension might "discover" the main self-alignment and
    extend over the entire length of the sequence.
    To avoid this problem, gapped alignments are not triggered from
    any gapless alignment that:
    * is contained, in both sequences, in the "core" of another
      alignment
    * has start coordinates offset by DISTANCE or less relative to
      this core
    Use -w0 to turn this off.
```

```
-G FILE
    Use an alternative genetic code in the specified file.  For an
    example of the format, see vertebrateMito.gc in the examples
    directory.  By default, the standard genetic code is used.  This
    option has no effect unless DNA-versus-protein alignment is
    selected with option -F.
-t TEMPERATURE
    Parameter for converting between scores and likelihood ratios.
    This affects the column ambiguity estimates.  A score is
    converted to a likelihood ratio by this formula: exp(score /
    TEMPERATURE).  The default value is 1/lambda, where lambda is
    the scale factor of the scoring matrix, which is calculated by
    the method of Yu and Altschul (YK Yu et al. 2003, PNAS
    100(26):15688-93).
-g GAMMA
    This option affects gamma-centroid and LAMA alignment only.
    Gamma-centroid alignments minimize the ambiguity of paired
    letters.  In fact, this method aligns letters whose column error
    probability is less than GAMMA/(GAMMA+1).  When GAMMA is low, it
    aligns confidently-paired letters only, so there tend to be many
    unaligned letters.  When GAMMA is high, it aligns letters more
    liberally.
    LAMA (Local Alignment Metric Accuracy) alignments minimize the
    ambiguity of columns (both paired letters and gap columns).
    When GAMMA is low, this method produces shorter alignments with
    more-confident columns, and when GAMMA is high it produces
    longer alignments including less-confident columns.
    In summary: to get the most accurately paired letters, use
    gamma-centroid.  To get accurately placed gaps, use LAMA.
    Note that the reported alignment score is that of the ordinary
    gapped alignment before realigning with gamma-centroid or LAMA.
-j NUMBER
    Output type: 0 means counts of initial matches (of all lengths);
    1 means gapless alignments; 2 means gapped alignments before
    non-redundantization; 3 means gapped alignments after
    non-redundantization; 4 means alignments with ambiguity
    estimates; 5 means gamma-centroid alignments; 6 means LAMA
    alignments.  Match counts (-j0) respect the minimum length
    option but not the maximum multiplicity option.  It's a bad idea
    to try -j0 when comparing a large sequence to itself.
-Q NUMBER
    This option allows lastal to use sequence quality scores, or
    PSSMs, for the queries.  0 means read queries in fasta format
    (without quality scores); 1 means fastq-sanger format; 2 means
    fastq-solexa format; 3 means fastq-illumina format; 4 means prb
    format; 5 means read PSSMs.
    The fastq formats look like this:
      @mySequenceName
      TTTTTTTTGCCTCGGGCCTGAGTTCTTAGCCGCG
      +
      55555555*&5-/55*5//5(55,5#&$)$)*+$
    The "+" may optionally be followed by a name (ignored), and the
    sequence and quality codes are allowed to wrap onto more than
    one line.  For fastq-sanger, the quality scores are obtained by
    subtracting 33 from the ASCII values of the characters below the
    "+".  For fastq-solexa and fastq-illumina, they are obtained by
    subtracting 64.
    prb format stores four quality scores (A, C, G, T) per position,
    with one sequence per line, like this:
      -40  40  -40  -40      -12   1  -12  -3      -10  10  -40  -40
    Since prb performs not store sequence names, lastal uses the line
    number (starting from 1) as the name.
    In fastq-sanger and fastq-illumina format, the quality scores
    are related to error probabilities like this: qScore =
    -10log10[p].  In fastq-solexa and prb, however, qScore =
    -10log10[p/(1-p)].  In lastal's MAF output, the quality scores
    are written on lines starting with "q".  For fastq, they are
    written with the same encoding as the input.  For prb, they are
    written in the fastq-solexa (ASCII-64) encoding.
    Finally, PSSM means "position-specific scoring matrix".  The
    format is:
      myLovelyPSSM
```

```
          A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
    1 M  -2 -2 -3 -4 -2 -1 -3 -3 -2  1  2 -2  8 -1 -3 -2 -1 -2 -2  0
    2 S   0 -2  0  1  3 -1 -1 -1 -2 -3 -3 -1 -2 -3 -2  5  0 -4 -3 -2
    3 D  -1 -2  0  7 -4 -1  1 -2 -2 -4 -4 -2 -4 -4 -2 -1 -2 -5 -4 -4
  The sequence appears in the second column, and columns 3 onwards
  contain the position-specific scores.  Any letters not specified
  by any column will get the lowest score in each row.  This
  format is a simplified version of PSI-BLAST's ASCII format: the
  non-simplified version is allowed too.  If you use PSSMs,
  options -r -q and -p are mostly ignored, except that they
  determine the default value of -y.
```

5.2.2 lastdb parameter

Option description for LAST has been taken from LAST web site.

```
Main Options
~~~~~~~~~~~~

  -h  Show all options and their default settings.
  -p  Interpret the sequences as proteins.  The default is to interpret
      them as DNA.
  -c  Soft-mask lowercase letters.  This means that, when we compare
      these sequences to some other sequences using lastal, lowercase
      letters will be excluded from initial matches.  This will apply
      to lowercase letters in both sets of sequences.


Advanced Options
~~~~~~~~~~~~~~~~

  -s BYTES
      Limit memory usage, by splitting the output files into smaller
      ″volumes″ if necessary.  This will limit the memory usage of
      both lastdb and lastal, but it will make lastal slower.  It is
      also likely to change the exact results found by lastal.
      BYTES should be slightly less than the amount of real memory on
      your computer.  You can use suffixes K, M, and G to specify
      KibiBytes, MebiBytes, and GibiBytes.  For example, ″-s 5G″ has
```

worked well with 6G, and "-s 1280M" has worked well with 2G. However, the output for one sequence is never split. Since the output files are several-fold bigger than the input, this means that mammalian chromosomes cannot be processed using much less than 2G.

There is a hard upper limit of about 4 billion sequence letters per volume. Together with the previous point, this means that lastdb will refuse to process any single sequence longer than about 4 billion.

-m PATTERN

Specify a spaced seed pattern, for example "-m 110101". In this example, mismatches will be allowed at every third and fifth position out of six in initial matches.

This option performs not constrain the length of initial matches. The pattern will get cyclically repeated as often as necessary to cover any length.

Although the 0 positions allow mismatches, they exclude non-standard letters (e.g. non-ACGT for DNA). If option -c is used, they also exclude lowercase letters.

-u FILE

Specify a subset seed file. The -m option will then be ignored. For an example of the format, see yass.seed in the examples directory.

-w STEP

Allow initial matches to start only at every STEP-th position in each of the sequences given to lastdb. This reduces the memory usage of lastdb and lastal, and it makes lastdb faster. Its effect on the speed and sensitivity of lastal is not entirely clear. To emulate BLAT, use "-w 11".

-a SYMBOLS

Specify your own alphabet, e.g. "-a 0123". The default (DNA) alphabet is equivalent to "-a ACGT". The protein alphabet (-p) is equivalent to "-a ACDEFGHIKLMNPQRSTVWY". Non-alphabet letters are allowed in sequences, but by default they are excluded from initial matches and get the mismatch score when aligned to anything. If -a is specified, -p is ignored.

```
-b DEPTH
    Specify the depth of "buckets" used to accelerate initial match
    finding.  Larger values increase the memory usage of lastdb and
    lastal, make lastal faster, and have no effect on lastal's
    results.  The default is to use the maximum depth that consumes
    at most one byte per possible match start position.
-x  Just count sequences and letters.  This is much faster, and the
    results are useful with lastex.  Letter counting is never
    case-sensitive.
-v  Be verbose: write messages about what lastdb is doing.
```

6 Contact

Please send your queries or comments, if you have, to the address below.

workflow@cbrc.jp

Computational Biology Research Center of AIST plans to listen to user's requests positively, and to make the system better.

Computational Biology Research Center (CBRC)
Advanced Industrial Science and Technology (AIST)
http://togo.cbrc.jp

AIST Tokyo Waterfront Bio-IT Research Building
2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan