MotifMarker Documentation

Abstract
MotifMarker is a tool designed to search for motifs whose consensus are represented by regular expressions.

Correspondence/Bug reports
Cheng Wu Albert
albertwcheng@gmail.com

Program Flow
The program first reads in the fasta file containing the sequences, and loads in a consensus files containing the motifs represented in regular expression and a script templates containing the template to create scripts. The program then searches for each motif against each sequence and generate a script file for each match. The program halts after generating a report in HTML format and display the report on a browser.

Requirements
J2SE (Java 2 Standard Edition) 5.0 Runtime Environment which could be obtained from Sun Microsystems Java homepage (http://java.sun.com/j2se/1.5.0/download.jsp)

The program package
Obtain MotifMarker.zip from http://albertwcheng.inscyber.net/MotifMarker.htm
The program package comes in a zip file containing the following:

MotifMarker.jar	The java binary for this program, also containing the source
starthtml.win.bat	A BATCH script for lauching report in browser for Windows
documentation.pdf	This documentation
demo.pdf	The demo of the program
consensus/*.txt	Some sample consensus files
spdb.txt	A sample script template for use with Spdb-viewer
proteinstruct/hsa_cellcyclekinase/*.pdb	The sample pdb files

To Setup MotifMarker

1) Setup the J2SE Runtime Environment 5.0

2) Extract MotifMarker.zip

3) Done

To Run MotifMarker

1) run command in a console

java –jar MotifMarker.jar

2) Follow the instructions of the program

The Parameters by Example

Parameter	**Example**
protein sequence **(The fasta file containing the sequences)**	cellcyclekinase_hs.txt
Consensus **(The consensus file containing the regular expressions of the consensus)**	consensus/pstaire.txt
script template **(The script template file)**	spdb.txt
pdb protein structure **(The folder containing the PDB files, this is used to correct for the residue positions of the protein structure; type "_" to ignore)**	proteinstruct/hsa_cellcyclekinase

(**boldface** represents user input)

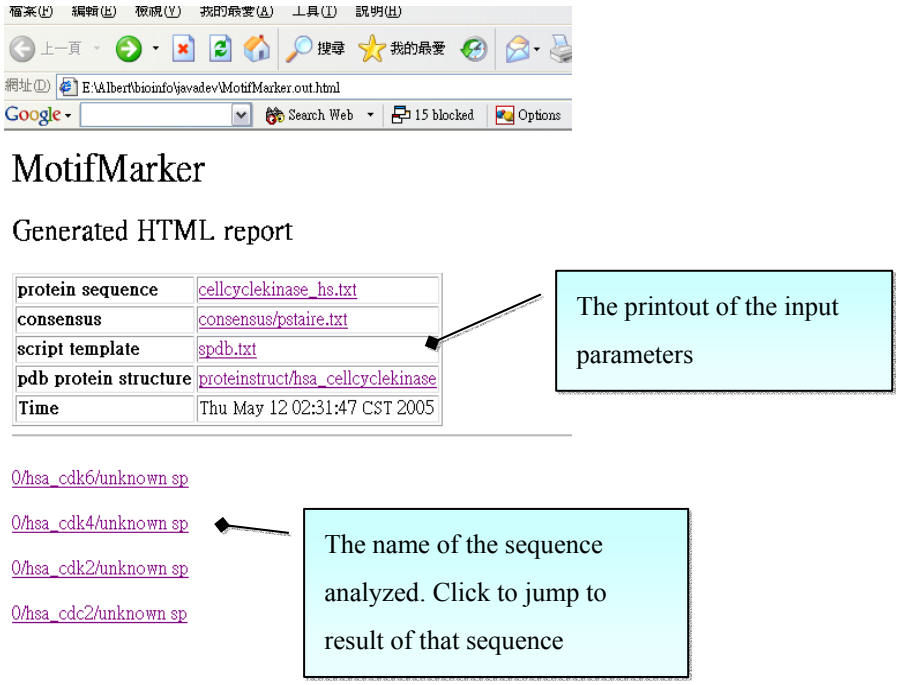
```
>java -jar MotifMarker.jar
protein sequence>cellcyclekinase_hs.txt
consensus>consensus/pstaire.txt
script temp>spdb.txt
pdb protein structure( '_' to ignore)>proteinstruct/hsa_cellcyclekinase




Processing 0/hsa_cdk6/unknown sp
-----------------------------------------------
```

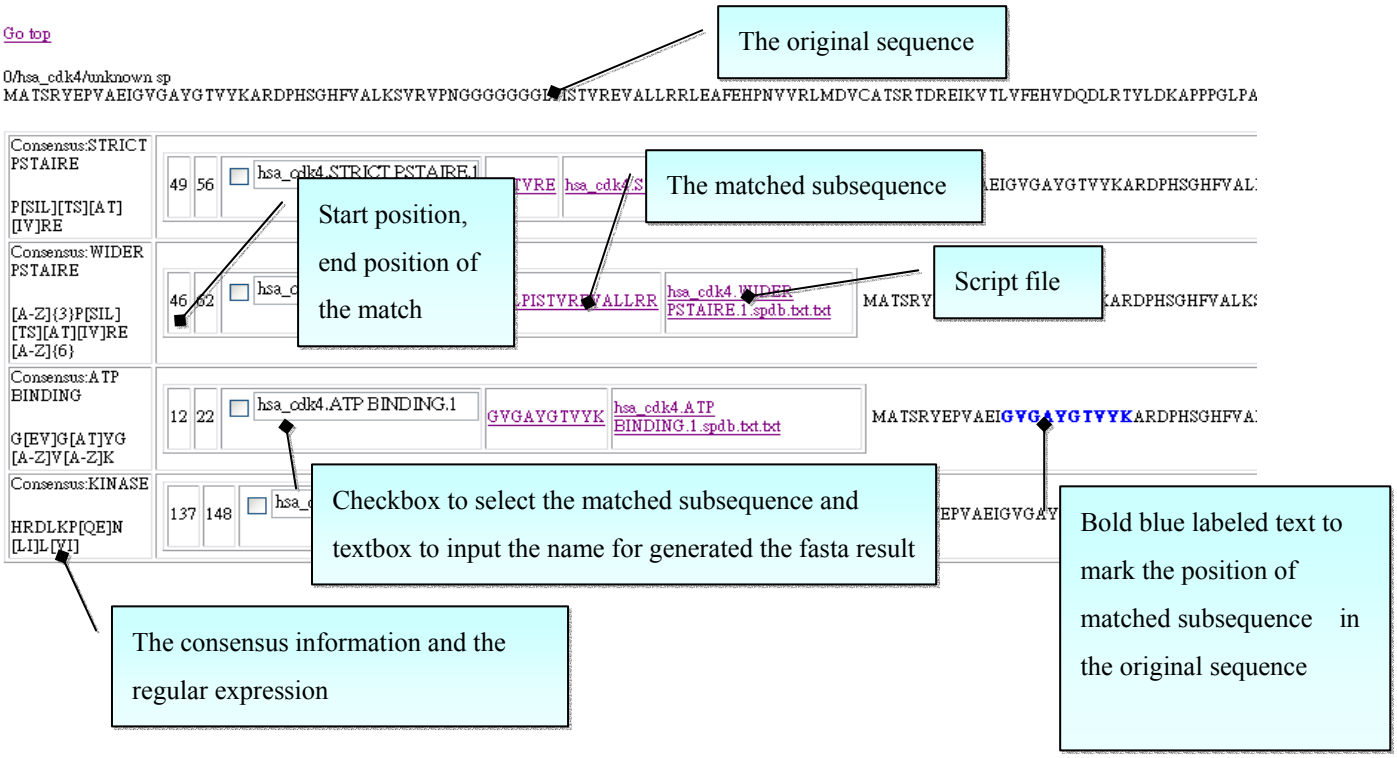
```
 1Matches found for consensusP[SIL][TS][AT][IV]RE
----------------------------------------------
 1Matches found for consensus[A-Z]{3}P[SIL][TS][AT][IV]RE[A-Z]{6}
----------------------------------------------
 1Matches found for consensusG[EV]G[AT]YG[A-Z]V[A-Z]K
----------------------------------------------
 1Matches found for consensusHRDLKP[QE]N[LI]L[VI]



Processing 0/hsa_cdk4/unknown sp
----------------------------------------------
 1Matches found for consensusP[SIL][TS][AT][IV]RE
----------------------------------------------
 1Matches found for consensus[A-Z]{3}P[SIL][TS][AT][IV]RE[A-Z]{6}
----------------------------------------------
 1Matches found for consensusG[EV]G[AT]YG[A-Z]V[A-Z]K
----------------------------------------------
 1Matches found for consensusHRDLKP[QE]N[LI]L[VI]



Processing 0/hsa_cdk2/unknown sp
----------------------------------------------
 1Matches found for consensusP[SIL][TS][AT][IV]RE
----------------------------------------------
 1Matches found for consensus[A-Z]{3}P[SIL][TS][AT][IV]RE[A-Z]{6}
----------------------------------------------
 1Matches found for consensusG[EV]G[AT]YG[A-Z]V[A-Z]K
----------------------------------------------
 1Matches found for consensusHRDLKP[QE]N[LI]L[VI]



Processing 0/hsa_cdc2/unknown sp
----------------------------------------------
 1Matches found for consensusP[SIL][TS][AT][IV]RE
----------------------------------------------
 1Matches found for consensus[A-Z]{3}P[SIL][TS][AT][IV]RE[A-Z]{6}
```

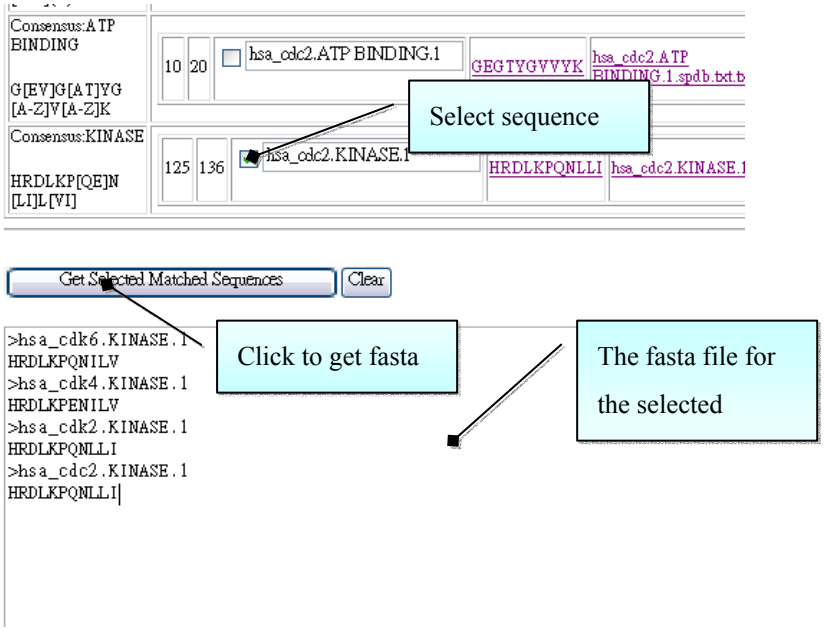
```
-----------------------------------------------

 1Matches found for consensusG[EV]G[AT]YG[A-Z]V[A-Z]K

-----------------------------------------------

 1Matches found for consensusHRDLKP[QE]N[LI]L[VI]

Done! Thanks for using

Done! Display report(only for windows)? [Y/N]Y
```

An HTML report MotifMarker.out.html is generated and displayed on the default browser

The HTML report




Go top
0/hsa_cdk4/unknown sp
MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPA
Consensus:STRICT PSTAIRE
P[SIL][TS][AT][IV]RE
49 56
hsa_cdk4.STRICT PSTAIRE.1
Consensus:WIDER PSTAIRE
[A-Z]{3}P[SIL][TS][AT][IV]RE[A-Z]{6}
46 62
hsa_cdk4.WIDER PSTAIRE.1.spdb.txt.txt
MATSRY
Consensus:ATP BINDING
G[EV]G[AT]YG[A-Z]V[A-Z]K
12 22
hsa_cdk4.ATP BINDING.1
GVGAYGTVYK
hsa_cdk4.ATP BINDING.1.spdb.txt.txt
MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVA
Consensus:KINASE
HRDLKP[QE]N[LI]L[VI]
137 148
EPVAEIGVGAY
The original sequence
The matched subsequence
Start position, end position of the match
Script file
Checkbox to select the matched subsequence and textbox to input the name for generated the fasta result
Bold blue labeled text to mark the position of matched subsequence in the original sequence
The consensus information and the regular expression


Consensus:ATP BINDING
G[EV]G[AT]YG[A-Z]V[A-Z]K
10 20
hsa_cdc2.ATP BINDING.1
GEGTYGVVYK
hsa_cdc2.ATP BINDING.1.spdb.txt.t
Consensus:KINASE
HRDLKP[QE]N[LI]L[VI]
125 136
hsa_cdc2.KINASE.1
HRDLKPQNLLI
hsa_cdc2.KINASE.1
Select sequence
Get Selected Matched Sequences
Clear
>hsa_cdk6.KINASE.1
HRDLKPQNILV
>hsa_cdk4.KINASE.1
HRDLKPENILV
>hsa_cdk2.KINASE.1
HRDLKPQNLLI
>hsa_cdc2.KINASE.1
HRDLKPQNLLI
Click to get fasta
The fasta file for the selected

The consensus file format

Consensus file contains the regular expression of the consensus. Each consensus is written in a line and the consensus file can contain infinite number of lines of the following format (where _ means tab "\t" <> means required fields, [] means optional, … means none or many times of the previous argument)

<consensus name>_<regular expression>_[<script replace key>_<script replace value>]_....

Parameter	Example
Consensus Name **(A fancy description of the consensus)**	ATP BINDING
Regular Expression **(The regular expression to represent the consensus) please refer to java.util.Regex.Pattern in J2SE documentation for more information**[see ref #]	G[EV]G[AT]YG[A-Z]V[A-Z]K (for consensus GE/VGA/TYGXVXK)
script replace key script replace value (replace every occurrence of script replace key by script replace value in the script file)	color0 RED (to replace all the occurance of *color0* in script files by *RED*

The script template file format

The script template file is an ordinary script file (in Rasmol format or Deepview format) with labels left out for MotifMarker to fill in and replace.

Some of the labels are replaced by MotifMarker, some of the labels are replaced by those defined in the consensus file, e.g., color0 is replaced by red in the script file in the example above

Some reserved label of MotifMarker

Parameter	Meaning
start0	The first residue of the motif, e.g., 49
end0	The last residue of the motif, e.g., 56

For example for a script in Deepview

```
$start=start0;
$end=end0;
$color=color0;
```

will be replaced to

```
$start=49;
$end=56;
$color=RED;
```

That's it. Enjoy!
Thanks,
Albert

References

Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis* 18**:** 2714-2723.

Guex N, and Schwede T. (n.d.) DeepView Scripting Language. (Online) http://tw.expasy.org/spdbv/text/script.htm [Available: 12 May 2005]

Peitsch, M. C. (1995) Protein modeling by E-mail *Bio/Technology* 13**:** 658-660.

Schwede T, Kopp J, Guex N, and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* 31**:** 3381-3385.

[#] Pattern (Java 2 Platform SE 5.0) (online) http://java.sun.com/j2se/1.5.0/docs/api/java/util/regex/Pattern.html [Available: 12 May 2005]