Documentation for ‘Analysis Tools’ developed at Rishi Biotech, Mumbai

-Nayana Ramachandran

Introduction

At Rishi Biotech, Mumbai, we have developed five sequence analysis tools, namely, FormatSeq, ORFreader, TranslORF, InfoProt and RestrDigest. These different tools are integrated and placed on a common platform headed under ‘Analysis Tools.’

This documentation will enable a user to utilise the tools to the fullest extent thereby taking advantage of this free edition of Analysis Tools.

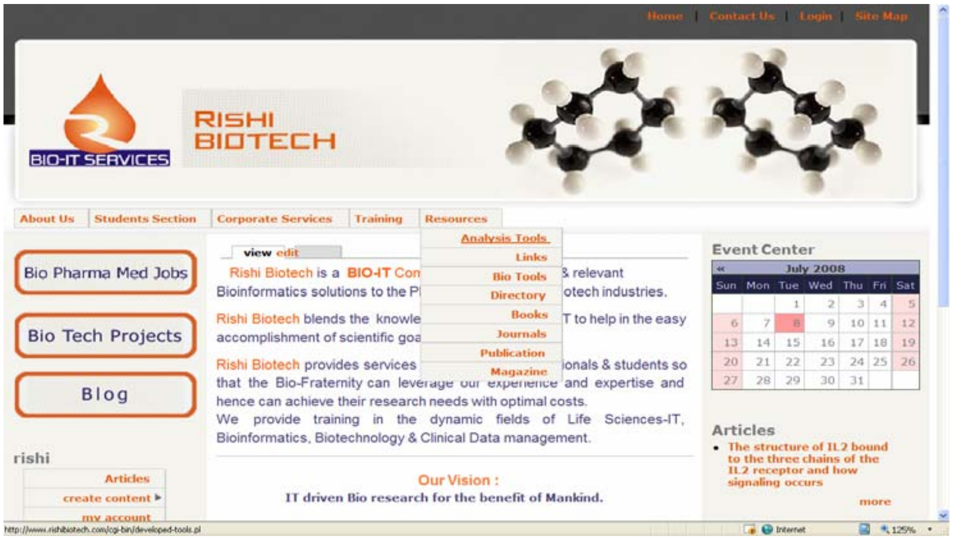
It is advised that proper formats for submitting the nucleotide or protein sequence be followed until further version of ‘Analysis Tools’ are developed which will take all the standard input formats.

The organisation of this document is based on the five tools and their working.

A] Analysis Tools

We, at Rishi Biotech, have developed and designed five tools for sequence analysis. The tools can be freely accessed through the home page for ‘Analysis Tools’ at http://www.rishibiotech.com/cgi-bin/developed-tools.pl

Alternatively, any user can visit the Rishi Biotech home page at http://www.rishibiotech.com and go to the ‘Resources’ section and access the link titled ‘Analysis Tools.’



Rishi Biotech is a BIO-IT Con & relevant Bioinformatics solutions to the P otech industries.

Rishi Biotech blends the knowle T to help in the easy accomplishment of scientific goa

Rishi Biotech provides services ionals & students so that the Bio-Fraternity can leverage our experience and expertise and hence can achieve their research needs with optimal costs.

We provide training in the dynamic fields of Life Sciences-IT, Bioinformatics, Biotechnology & Clinical Data management.

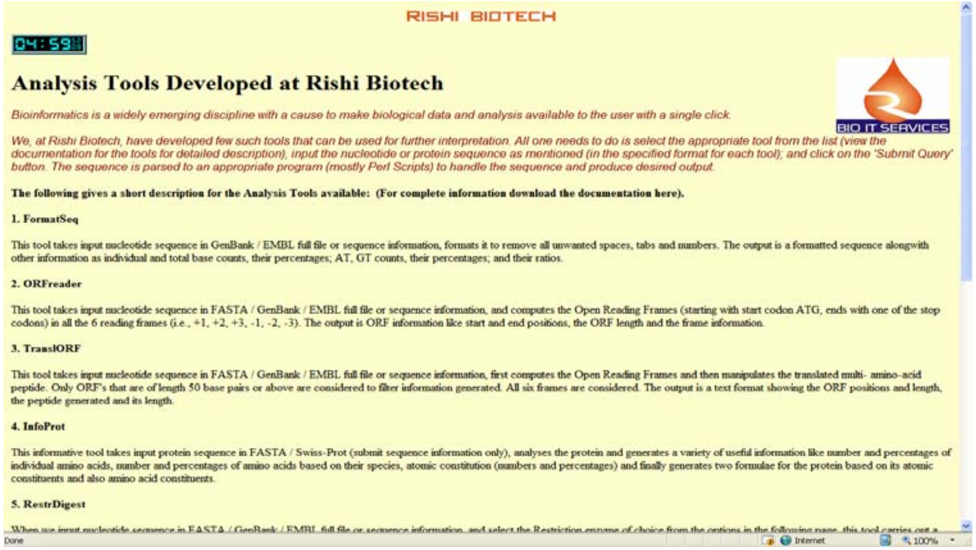
Our Vision :

IT driven Bio research for the benefit of Mankind.

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

- The structure of IL2 bound to the three chains of the IL2 receptor and how signaling occurs

On clicking the link, you will be redirected to an integrated tools page titled ‘Analysis Tools.’



Analysis Tools Developed at Rishi Biotech

Bioinformatics is a widely emerging discipline with a cause to make biological data and analysis available to the user with a single click.

We, at Rishi Biotech, have developed few such tools that can be used for further interpretation. All one needs to do is select the appropriate tool from the list (view the documentation for the tools for detailed description), input the nucleotide or protein sequence as mentioned (in the specified format for each tool); and click on the 'Submit Query' button. The sequence is parsed to an appropriate program (mostly Perl Scripts) to handle the sequence and produce desired output.

The following gives a short description for the Analysis Tools available: (For complete information download the documentation here).

1. FormatSeq

This tool takes input nucleotide sequence in GenBank / EMBL full file or sequence information, formats it to remove all unwanted spaces, tabs and numbers. The output is a formatted sequence alongwith other information as individual and total base counts, their percentages; AT, GT counts, their percentages; and their ratios.

2. ORFreader

This tool takes input nucleotide sequence in FASTA / GenBank / EMBL full file or sequence information, and computes the Open Reading Frames (starting with start codon ATG, ends with one of the stop codons) in all the 6 reading frames (i.e., +1, +2, +3, -1, -2, -3). The output is ORF information like start and end positions, the ORF length and the frame information.

3. TranslORF

This tool takes input nucleotide sequence in FASTA / GenBank / EMBL full file or sequence information, first computes the Open Reading Frames and then manipulates the translated multi- amino-acid peptide. Only ORF's that are of length 50 base pairs or above are considered to filter information generated. All six frames are considered. The output is a text format showing the ORF positions and length, the peptide generated and its length.

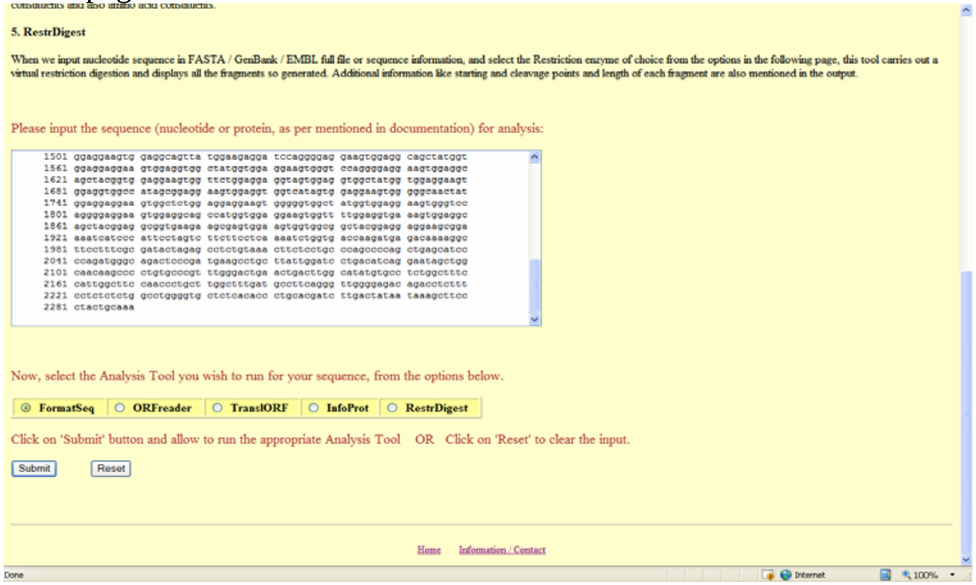
4. InfoProt

This informative tool takes input protein sequence in FASTA / Swiss-Prot (submit sequence information only), analyses the protein and generates a variety of useful information like number and percentages of individual amino acids, number and percentages of amino acids based on their species, atomic constitution (numbers and percentages) and finally generates two formulae for the protein based on its atomic constituents and also amino acid constituents.

5. RestrDigest

When we input nucleotide sequence in FASTA / GenBank / EMBL full file or sequence information, and select the Restriction enzyme of choice from the options in the following page, this tool carries out a

There is brief description about the various tools given on this page. Scroll to the bottom of the page. You will find a text-area like this:



constituents and also amino acid constituents.

5. RestrDigest

When we input nucleotide sequence in FASTA / GenBank / EMBL full file or sequence information, and select the Restriction enzyme of choice from the options in the following page, this tool carries out a virtual restriction digestion and displays all the fragments so generated. Additional information like starting and cleavage points and length of each fragment are also mentioned in the output.

Please input the sequence (nucleotide or protein, as per mentioned in documentation) for analysis:

```
     1501 ggaggaagtg gaggcagtta tggaagagga tccaggggag gaagtggagg cagctatggt
     1561 ggaggaggaa gtggaggtgg ctatggtgga ggaagtgggt ccaggggagg aagtggaggc
     1621 agctacggtg gaggaagtgg ttctggagga ggtagtggag gtggctatgg tggaggaagt
     1681 ggaggtggcc atagcggagg aagtggaggt ggtcatagtg gaggaagtgg gggcaactat
     1741 ggaggaggaa gtggctctgg aggaggaagt gggggtggct atggtggagg aagtgggtcc
     1801 aggggaggaa gtggaggcag ccatggtgga ggaagtggtt ttggaggtga aagtggaggc
     1861 agctacggag gcggtgaaga agcgagtgga agtggtggcg gctacggagg aggaagcgga
     1921 aaatcatccc attcctagtc ttcttcctca aaatctggtg accaagatga gacaaaaggc
     1981 ttcctttcgc gatactagag cctctgtaaa cttctcctgc ccagccccag ctgagcatcc
     2041 ccagatgggc agactcccga tgaagcctgc ttattggatc ctgacatcag gaatagctgg
     2101 caacaagccc ctgtgcccgt ttgggactga actgacttgg catatgtgcc tctggctttc
     2161 cattggcttc caaccctgct tggctttgat gccttcaggg ttggggagac agacctcttt
     2221 cctctctctg gcctggggtg ctctcacacc ctgcacgatc ttgactataa taaagcttcc
     2281 ctactgcaaa
```

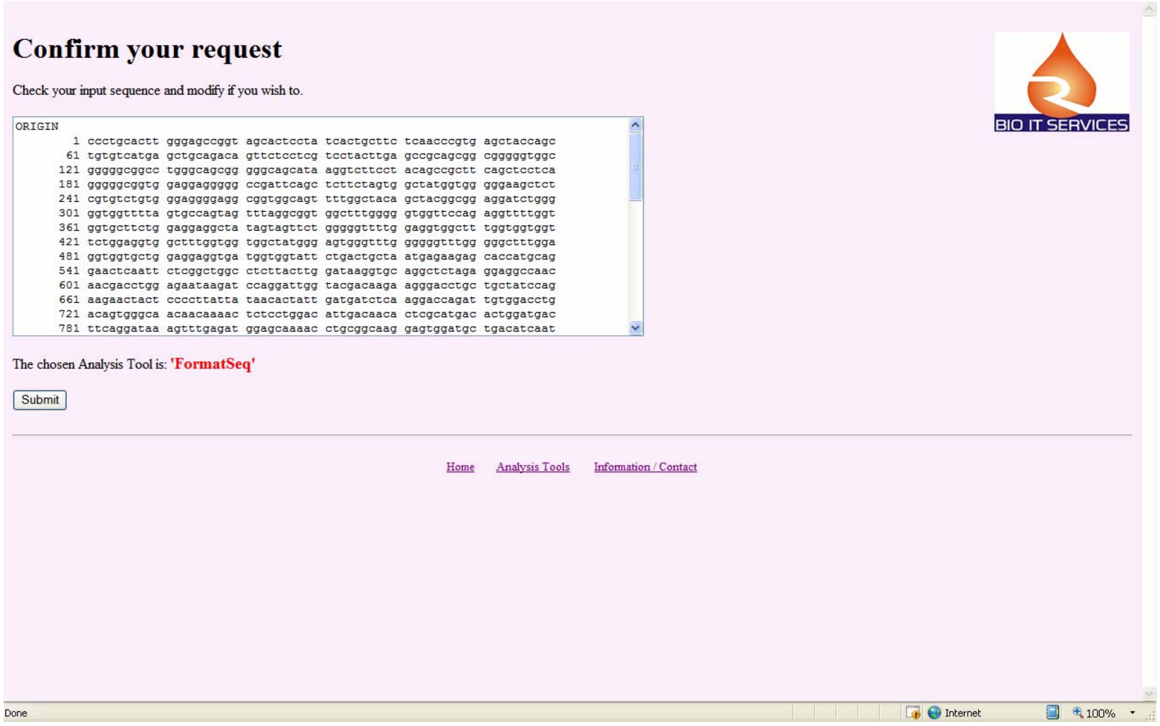
Now, select the Analysis Tool you wish to run for your sequence, from the options below.

● **FormatSeq** ○ **ORFreader** ○ **TranslORF** ○ **InfoProt** ○ **RestrDigest**

Click on 'Submit' button and allow to run the appropriate Analysis Tool OR Click on 'Reset' to clear the input.

It is here that you have to input your nucleotide or protein sequence for all the tools that you want to use. After inputting the sequence you have to choose a tool and click on ‘Submit’ button.

You will be redirected to an intermediary page where you have a chance to change your input sequence, if required. The tool that you have selected will appear in bold red font. If you want to change the tool, simply navigate backwards through the ‘back’ button on your browser and you will see that your sequence remains intact. Now, you can change your tool, if you wish.
Confirm your actions by clicking on ‘Submit.’



Confirm your request

Check your input sequence and modify if you wish to.

```
ORIGIN
        1 ccctgcactt gggagccggt agcactccta tcactgcttc tcaacccgtg agctaccagc
       61 tgtgtcatga gctgcagaca gttctcctcg tcctacttga gccgcagcgg cgggggtggc
      121 gggggcggcc tgggcagcgg gggcagcata aggtcttcct acagccgctt cagctcctca
      181 gggggcggtg gaggaggggg ccgattcagc tcttctagtg gctatggtgg gggaagctct
      241 cgtgtctgtg ggaggggagg cggtggcagt tttggctaca gctacggcgg aggatctggg
      301 ggtggtttta gtgccagtag tttaggcggt ggctttgggg gtggttccag aggttttggt
      361 ggtgcttctg gaggaggcta tagtagttct gggggttttg gaggtggctt tggtggtggt
      421 tctggaggtg gctttggtgg tggctatggg agtgggtttg gggggtttgg gggctttgga
      481 ggtggtgctg gaggaggtga tggtggtatt ctgactgcta atgagaagag caccatgcag
      541 gaactcaatt ctcggctggc ctcttacttg gataaggtgc aggctctaga ggaggccaac
      601 aacgacctgg agaataagat ccaggattgg tacgacaaga agggacctgc tgctatccag
      661 aagaactact cccttatta taacactatt gatgatctca aggaccagat tgtggacctg
      721 acagtgggca acaacaaaac tctcctggac attgacaaca ctcgcatgac actggatgac
      781 ttcaggataa agtttgagat ggagcaaaac ctgcggcaag gagtggatgc tgacatcaat
```

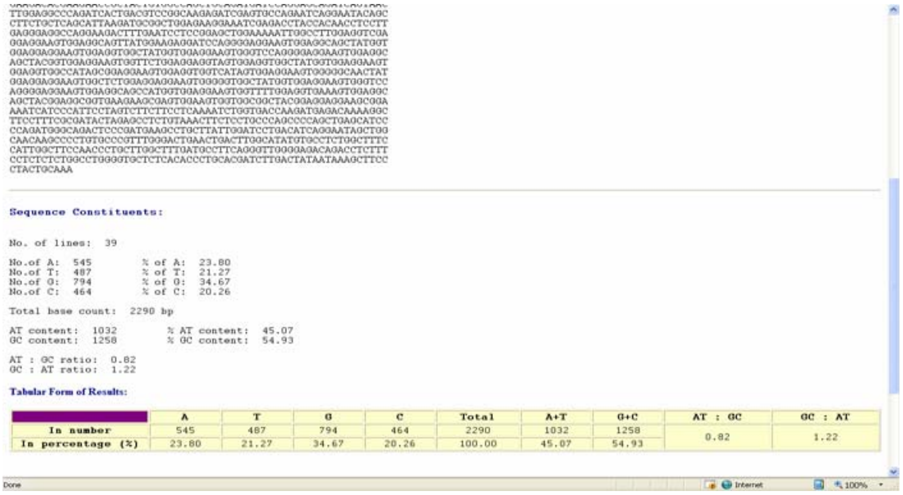
The chosen Analysis Tool is: **'FormatSeq'**

FormatSeq

This is a nucleotide sequence formatting tool.
This tool formats your output, i.e. removes all the spaces, gaps and unknown bases from your input nucleotide sequence. (Note: Any ‘U’s in the sequence will be converted to ‘T’).
This tool renders your sequence fit to apply to many tools found over the World Wide Web. However, the condition is that you have to submit an unformatted sequence, in either of these formats:

1. GenBank: full-text starting with ‘LOCUS’ or sequence information starting with ‘1’ or ‘ORIGIN.’
2. EMBL: full-text starting with ‘ID’ or sequence information starting with ‘SQ’ or the sequence itself.

The output displays various information like: Formatted sequence, base constituents, their percentages, AT: GC and GC:AT ratio, which can be analysed for comparing between two sequences.



```
TTGGAGGCCCAGATCACTGACGTCCGGCAAGAGATCGAGTGCCAGAATCAGGAATACAGC
CTTCTGCTCAGCATTAAGATGCGGCTGGAGAAGGAAATCGAGACCTACCACAACCTCCTT
GAGGGAGGCCAGGAAGACTTTGAATCCTCCGGAGCTGGAAAAATTGGCCTTGGAGGTCGA
GGAGGAAGTGGAGGCAGTTATGGAAGAGGATCCAGGGGAGGAAGTGGAGGCAGCTATGGT
GGAGGAGGAAGTGGAGGTGGCTATGGTGGAGGAAGTGGGTCCAGGGGAGGAAGTGGAGGC
AGCTACGGTGGAGGAAGTGGTTCTGGAGGAGGTAGTGGAGGTGGCTATGGTGGAGGAAGT
GGAGGTGGCCATAGCGGAGGAAGTGGAGGTGGTCATAGTGGAGGAAGTGGGGGCAACTAT
GGAGGAGGAAGTGGCTCTGGAGGAGGAAGTGGGGGTGGCTATGGTGGAGGAAGTGGGTCC
AGGGGAGGAAGTGGAGGCAGCCATGGTGGAGGAAGTGGTTTTGGAGGTGAAAGTGGAGGC
AGCTACGGAGGCGGTGAAGAAGCGAGTGGAAGTGGTGGCGGCTACGGAGGAGGAAGCGGA
AAATCATCCCATTCCTAGTCTTCTTCCTCAAAATCTGGTGACCAAGATGAGACAAAAGGC
TTCCTTTCGCGATACTAGAGCCTCTGTAAACTTCTCCTGCCCAGCCCCAGCTGAGCATCC
CCAGATGGGCAGACTCCCGATGAAGCCTGCTTATTGGATCCTGACATCAGGAATAGCTGG
CAACAAGCCCCTGTGCCCGTTTGGGACTGAACTGACTTGGCATATGTGCCTCTGGCTTTC
CATTGGCTTCCAACCCTGCTTGGCTTTGATGCCTTCAGGGTTGGGGAGACAGACCTCTTT
CCTCTCTCTGGCCTGGGGTGCTCTCACACCCTGCACGATCTTGACTATAATAAAGCTTCC
CTACTGCAAA
```

Sequence Constituents:

```
No. of lines:  39

No.of A:  545          % of A:  23.80
No.of T:  487          % of T:  21.27
No.of G:  794          % of G:  34.67
No.of C:  464          % of C:  20.26

Total base count:  2290 bp

AT content:  1032          % AT content:  45.07
GC content:  1258          % GC content:  54.93

AT : GC ratio:  0.82
GC : AT ratio:  1.22
```

Tabular Form of Results:

	A	T	G	C	Total	A+T	G+C	AT : GC	GC : AT
In number	545	487	794	464	2290	1032	1258	0.82	1.22
In percentage (%)	23.80	21.27	34.67	20.26	100.00	45.07	54.93		

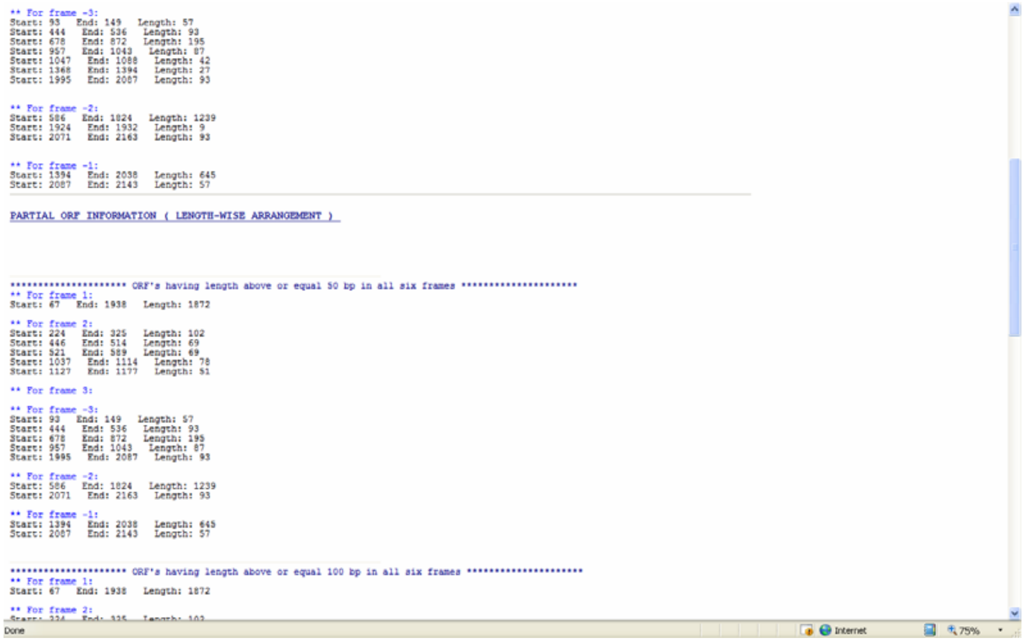
ORFreader

ORFreader scans the input nucleotide sequence for the Open Reading Frames in all the six frames. The ORF's start with a start codon ATG (or AUG in RNA) and end with one of the three stop codons: TAA, TGA or TAG.

The tool takes nucleotide input as mentioned for FormatSeq as well as an additional FASTA input. The information is displayed in two levels:

1. Complete ORF information which segregates the ORF's frame-wise for all the six frames.
2. Partial ORF information which segregates the ORF with respect to its length, in six levels, of or above 50, 100, 150, 200, 250 and 300 base pairs.

This enables the user to view selectively the ORF's and use them for further analysis.



```
** For frame -3:
Start: 93   End: 149   Length: 57
Start: 444   End: 536   Length: 93
Start: 678   End: 872   Length: 195
Start: 957   End: 1043   Length: 87
Start: 1047   End: 1088   Length: 42
Start: 1368   End: 1394   Length: 27
Start: 1995   End: 2087   Length: 93

** For frame -2:
Start: 586   End: 1824   Length: 1239
Start: 1924   End: 1932   Length: 9
Start: 2071   End: 2163   Length: 93

** For frame -1:
Start: 1394   End: 2038   Length: 645
Start: 2087   End: 2143   Length: 57

PARTIAL ORF INFORMATION ( LENGTH-WISE ARRANGEMENT )



******************** ORF's having length above or equal 50 bp in all six frames ********************
** For frame 1:
Start: 67   End: 1938   Length: 1872

** For frame 2:
Start: 224   End: 325   Length: 102
Start: 446   End: 514   Length: 69
Start: 521   End: 589   Length: 69
Start: 1037   End: 1114   Length: 78
Start: 1127   End: 1177   Length: 51

** For frame 3:

** For frame -3:
Start: 93   End: 149   Length: 57
Start: 444   End: 536   Length: 93
Start: 678   End: 872   Length: 195
Start: 957   End: 1043   Length: 87
Start: 1995   End: 2087   Length: 93

** For frame -2:
Start: 586   End: 1824   Length: 1239
Start: 2071   End: 2163   Length: 93

** For frame -1:
Start: 1394   End: 2038   Length: 645
Start: 2087   End: 2143   Length: 57


******************** ORF's having length above or equal 100 bp in all six frames ********************
** For frame 1:
Start: 67   End: 1938   Length: 1872

** For frame 2:
Start: 224   End: 325   Length: 102
```

The positions and the length of ORF are sufficient for the user to analyze the sequence, especially genomic sequences.

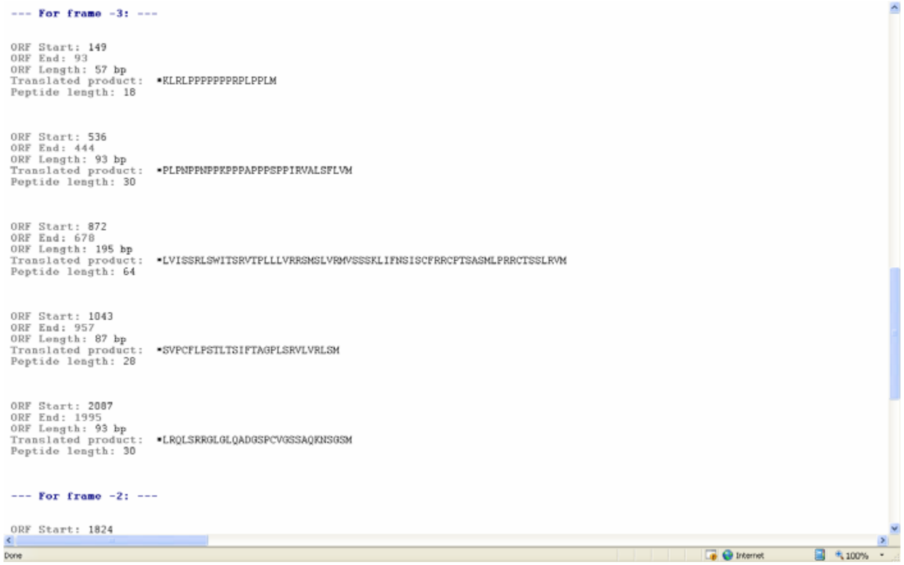
In further versions of ORFreader, we propose to include a graphical interface to view the ORF's.

TranslORF

This tool is an extension of ORFreader and translates the given sequences wherever there is an ORF.

The input format is same as above, namely, GenBank, EMBL or FASTA nucleotide sequence.

TranslORF considers only those ORF's which are equal to or more than 50 base pairs in length as only these can form stable peptides or proteins. The output is organized frame-wise with all the lengths of ORF equal to or above 50 bp. The details of output are: frame information, ORF start, end and length information, the translated product and the peptide length.



```
--- For frame -3: ---


ORF Start: 149
ORF End: 93
ORF Length: 57 bp
Translated product:  *KLRLPPPPPPPRPLPPLM
Peptide length: 18


ORF Start: 536
ORF End: 444
ORF Length: 93 bp
Translated product:  *PLPNPPNPPKPPPAPPPSPPIRVALSFLVM
Peptide length: 30


ORF Start: 872
ORF End: 678
ORF Length: 195 bp
Translated product:  *LVISSRLSWITSRVTPLLLVRRSMSLVRMVSSSKLIFNSISCFRRCPTSASMLPRRCTSSLRVM
Peptide length: 64


ORF Start: 1043
ORF End: 957
ORF Length: 87 bp
Translated product:  *SVPCFLPSTLTSIFTAGPLSRVLVRLSM
Peptide length: 28


ORF Start: 2087
ORF End: 1995
ORF Length: 93 bp
Translated product:  *LRQLSRRGLGLQADGSPCVGSSAQKNSGSM
Peptide length: 30


--- For frame -2: ---


ORF Start: 1824
```

In future versions of TranslORF we plan to include gene organization information into introns and exons.

InfoProt

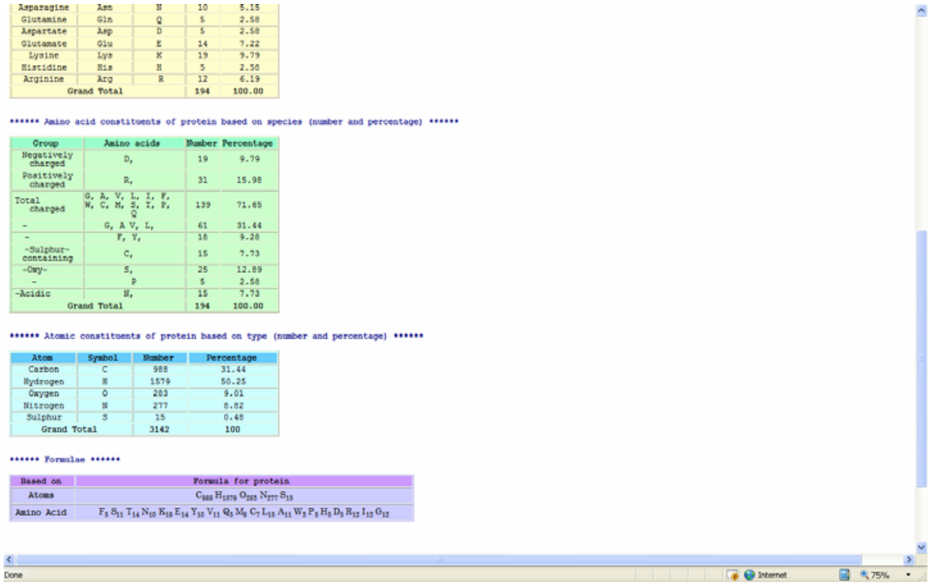
This is currently the sole protein sequence information tool at Rishi Biotech. This tool requires protein sequence as input in the following formats:

1. Swiss-Prot sequence format (consisting of sequence, gaps and numbers)
2. FASTA protein sequence format

The output is a text-cum-tabulated format displaying sequence information, number of amino acids, the peptide constituents in terms number and percentages of amino acids, atoms and species.

At the end, there are two formulae assigned to the protein:

a. Based on atomic information
b. Based on amino acid information



Asparagine	Asn	N	10	5.15
Glutamine	Gln	Q	5	2.58
Aspartate	Asp	D	5	2.58
Glutamate	Glu	E	14	7.22
Lysine	Lys	K	19	9.79
Histidine	His	H	5	2.58
Arginine	Arg	R	12	6.19
Grand Total			194	100.00

****** Amino acid constituents of protein based on species (number and percentage) ******

Group	Amino acids	Number	Percentage
Negatively charged	D,	19	9.79
Positively charged	R,	31	15.98
Total charged	G, A, V, L, I, F, W, C, M, S, T, P, Q	139	71.65
-	G, A V, L,	61	31.44
-	F, Y,	18	9.28
-Sulphur-containing	C,	15	7.73
-Oxy-	S,	25	12.89
-	P	5	2.58
-Acidic	N,	15	7.73
Grand Total		194	100.00

****** Atomic constituents of protein based on type (number and percentage) ******

Atom	Symbol	Number	Percentage
Carbon	C	988	31.44
Hydrogen	H	1579	50.25
Oxygen	O	283	9.01
Nitrogen	N	277	8.82
Sulphur	S	15	0.48
Grand Total		3142	100

****** Formulae ******

Based on	Formula for protein
Atoms	$C_{988} H_{1579} O_{283} N_{277} S_{15}$
Amino Acid	$F_5 S_{11} T_{14} N_{10} K_{19} E_{14} Y_{10} V_{11} Q_5 M_5 C_7 L_{15} A_{11} W_5 P_5 H_5 D_5 R_{12} I_{12} G_{12}$

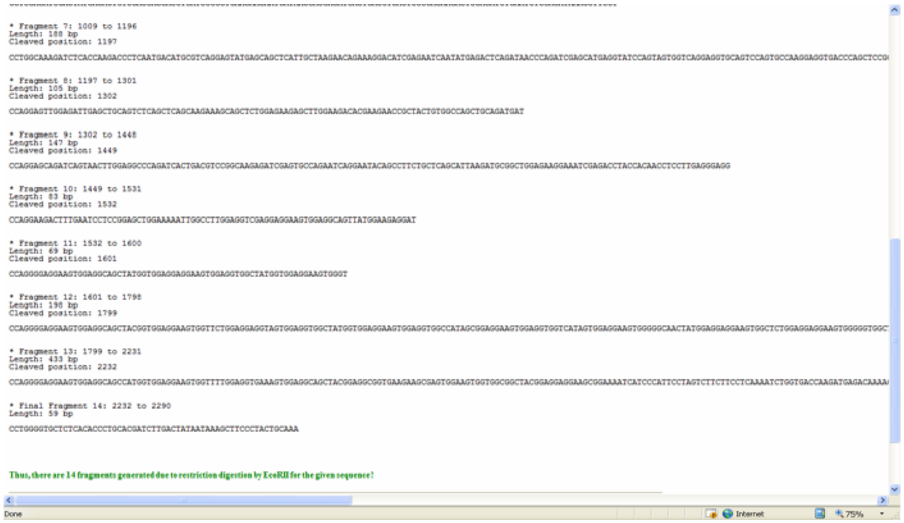
RestrDigest

This is a very useful tool which can be used to predict restriction digestion by restriction endonucleases. If you choose this tool on ‘Analysis Tools’ page, on the confirmatory page you will have to select the restriction enzyme with which you want to predict the digestion.

The user can select from a range of twenty enzymes available (in future versions, all enzymes will be made available) and proceed with ‘Submit.’

The result page will display the fragments created due to restriction digestion and fragment information like the positions, length, point of cleavage, etc.

If there is no restriction digestion, only a single fragment will be displayed to the user.



```
• Fragment 7: 1009 to 1196
Length: 188 bp
Cleaved position: 1197

CCTGGCAAAGATCTCACCAAGACCCTCAATGACATGCGTCAGGAGTATGAGCAGCTCATTGCTAAGAACAGAAAGGACATCGAGAATCAATATGAGACTCAGATAACCCAGATCGAGCATGAGGTATCCAGTAGTGGTCAGGAGGTGCAGTCCAGTGCCAAGGAGGTGACCCAGCTCCG

• Fragment 8: 1197 to 1301
Length: 105 bp
Cleaved position: 1302

CCAGGAGTTGGAGATTGAGCTGCAGTCTCAGCTCAGCAAGAAAGCAGCTCTGGAGAAGAGCTTGGAAGACACGAAGAACCGCTACTGTGGCCAGCTGCAGATGAT

• Fragment 9: 1302 to 1448
Length: 147 bp
Cleaved position: 1449

CCAGGAGCAGATCAGTAACTTGGAGGCCCAGATCACTGACGTCCGGCAAGAGATCGAGTGCCAGAATCAGGAATACAGCCTTCTGCTCAGCATTAAGATGCGGCTGGAGAAGGAAATCGAGACCTACCACAACCTCCTTGAGGGAGG

• Fragment 10: 1449 to 1531
Length: 83 bp
Cleaved position: 1532

CCAGGAAGACTTTGAATCCTCCGGAGCTGGAAAAATTGGCCTTGGAGGTCGAGGAGGAAGTGGAGGCAGTTATGGAAGAGGAT

• Fragment 11: 1532 to 1600
Length: 69 bp
Cleaved position: 1601

CCAGGGGAGGAAGTGGAGGCAGCTATGGTGGAGGAGGAAGTGGAGGTGGCTATGGTGGAGGAAGTGGGT

• Fragment 12: 1601 to 1798
Length: 198 bp
Cleaved position: 1799

CCAGGGGAGGAAGTGGAGGCAGCTACGGTGGAGGAAGTGGTTCTGGAGGAGGTAGTGGAGGTGGCTATGGTGGAGGAAGTGGAGGTGGCCATAGCGGAGGAAGTGGAGGTGGTCATAGTGGAGGAAGTGGGGGCAACTATGGAGGAGGAAGTGGCTCTGGAGGAGGAAGTGGGGGTGGC

• Fragment 13: 1799 to 2231
Length: 433 bp
Cleaved position: 2232

CCAGGGGAGGAAGTGGAGGCAGCCATGGTGGAGGAAGTGGTTTTGGAGGTGAAAGTGGAGGCAGCTACGGAGGCGGTGAAGAAGCGAGTGGAAGTGGTGGCGGCTACGGAGGAGGAAGCGGAAAATCATCCCATTCCTAGTCTTCTTCCTCAAAATCTGGTGACCAAGATGAGACAAAA

• Final Fragment 14: 2232 to 2290
Length: 59 bp

CCTGGGGTGCTCTCACACCCTGCACGATCTTGACTATAATAAAGCTTCCCTACTGCAAA
```

Thus, there are 14 fragments generated due to restriction digestion by EcoRII for the given sequence!

Thus, the user can carry out in-silico restriction digestion with any given sequence. The user can also navigate back and forth and try different restriction enzymes on the same sequence.

CONCLUSION

The analysis tools have been designed using Perl 5.8.8 as the programming language. They are under constant upgradation and development.
We plan to release the future versions of ‘Analysis Tools’ by improving on the existing tools and adding new tools to the website.

If any queries or suggestions regarding the tools, please e-mail to:
info@rishibiotech.com