

ResCon

User Guide

Table of Contents

| | |
|--|----|
| License | 4 |
| Requirements | 4 |
| For running ResCon software: | 4 |
| For running ResCon script: | 4 |
| Installation Instructions..... | 5 |
| For Windows: | 5 |
| For Mac: | 5 |
| For Ubuntu: | 5 |
| What the hell is ResCon? | 6 |
| Mismatch analyzer | 8 |
| Demo 1: Mismatch analysis (<i>Clustal Omega required</i>):..... | 12 |
| Demo 2: Mismatch analysis (<i>Clustal Omega not required</i>): | 13 |
| Overview of output files:..... | 14 |
| 1. Formatted_Alignment.html | 14 |
| 2. Mismatches_Tabulated.csv | 16 |
| 3. Mismatches_Detailed.txt: | 17 |
| 4. log.txt: | 17 |
| ResCon – More tools | 18 |
| 1. Subtree Sequences Extractor: | 18 |
| Demo:..... | 19 |
| Tips and Tricks / Troubleshooting: | 21 |
| 2. GenBank/GenPept to fasta converter: | 23 |
| Demo:..... | 25 |

| | |
|---|----|
| 3. Filter fasta by Sequences' ID | 26 |
| Demo - 1 (Using <i>Complete</i> sequence IDs to filter fasta sequences):..... | 27 |
| Demo - 2 (Using <i>Partial</i> sequence IDs to filter fasta sequences):..... | 28 |
| 4. Filter fasta by Blast's E-value: | 29 |
| Demo | 30 |
| Tips and Tricks / Troubleshooting: | 31 |
| 5. Filter fasta by Sequences' Description:..... | 32 |
| Demo 1 – Extracting complete Identifiers: | 38 |
| Demo 2 – Extracting Partial Descriptions: | 39 |
| 6. Fasta Description/ID Extractor: | 36 |
| Demo 1 - Using Complete Descriptions from a text file: | 34 |
| Demo 2 - Using Partial Descriptions as a list: | 35 |
| Appendix A: Understanding fasta format..... | 40 |
| Appendix B: Log file | 41 |
| Appendix C: Clustal Omega Parameters..... | 43 |
| Appendix D: Installing Clustal Omega..... | 47 |
| For Windows: | 47 |
| For UNIX based OS (Mac & Ubuntu):..... | 48 |

License

ResCon is licensed under GNU Lesser General Public License.

Authors: Manavalan Gajapathy, Joseph D. Ng

Copyright (C) <2015>

Requirements

For running ResCon software:

Now Clustal omega can be used via webserver in addition to using it locally.

- Install ResCon specific to your platform.
- Clustal Omega installation is optional ~~but it is required for complete use of ResCon.~~ See [Appendix D](#) for installation instructions.
- Many Windows OS computers may already have ‘*Microsoft Visual C++ 2008 Redistributable Package*’ installed. If it is not installed already, you may have to install it from <http://www.microsoft.com/en-us/download/details.aspx?id=29> for 32 bit Windows or ~~<http://www.microsoft.com/en-us/download/details.aspx?id=15336>~~ for 64 Windows.

For running ResCon script:

- Python 2.7

Verify modules reqd. Now also requires natsort (can be turned off though)

Download source: <https://www.python.org/download/releases/2.7/>

- Python module ‘Bio’ is required.

Biopython version 1.63

Download source: <http://biopython.org/wiki/Download>

- Though installation of Clustal omega is required for complete experience of ResCon, *it is not necessary if clustal alignment will not be run through ResCon.* See [Appendix D](#) for installation instructions.

Installation Instructions

For Windows:

Run the setup file for Windows and install it. You may need administrator password for installation.

Note: Many Windows OS computers may already have '*Microsoft Visual C++ 2008 Redistributable Package*' installed. If it is not installed already, you may have to install it from <http://www.microsoft.com/en-us/download/details.aspx?id=29> for 32 bit Windows or <http://www.microsoft.com/en-us/download/details.aspx?id=15336> for 64 Windows.

For Mac:

Unzip the file provided for Mac and move the ResCon.app file in to 'Applications' folder or save it a location of your choice.

Note:

1. When you start ResCon, it will start minimized. Go to the dock, double click ResCon icon and it should work just fine now.
2. If ResCon says that 'Clustal omega is not installed on your Mac' when it is actually properly installed, then you have to open ResCon app from terminal. Open 'Terminal'. Type open and then drag and drop ResCon.app file into the terminal and then press Enter. It should work just fine now.

For Ubuntu:

Unzip installation file provided for Ubuntu OS and then move the folder to location of your choice. Double click the executable file to run ResCon.

What the hell is ResCon?

ResCon houses tools as many as the number of continents. It was built to analyze residue conservation in protein sequences at positions of your choice. Following provides a brief introduction for each of those tools.

1. [Mismatch analyzer:](#)

This tool compares your protein sequence with other bunch of sequences and assists in analyzing conservation of residues at the positions you are interested in, based on multiple sequence alignment. Further, this tool helps you introducing taxonomy in to this analysis, provided you have taxonomy information in the sequence identifier. Works for both protein and DNA/RNA sequences.

2. [Subtree Sequences Extractor:](#)

This tool can read phylogenetic tree and then extract only the sequences corresponding to a subtree of your interest, based on that subtree's branch length. This tool extracts sequence IDs corresponding to subtree of your interest and then filters sequences of that subtree from a fasta file containing sequence data.

3. [GenBank/GenPept to fasta converter:](#)

This tool converts GenBank or GenPept files to fasta format while conserving taxonomy information in fasta sequences' header.

4. [Filter fasta by Sequences' ID:](#)

This tool extracts sequences from a fasta file based on the sequence IDs provided by you. You may provide either complete or partial sequence ID.

5. **Filter fasta by Blast's E-value:**

This tool reads a file of sequences in fasta format and a XML file containing BLAST results and then extract the sequences below or above E-value threshold of your choice.

6. **Filter fasta by Sequences' Description:**

This tool extracts sequences from a fasta file based on the sequence descriptions provided by you. You may provide either complete or partial sequence description.

7. **Fasta Description/ID Extractor:**

This tool enables you to extract descriptions or identifiers, complete or partial, from sequence headers of fasta sequences.

Mismatch analyzer

ResCon's 'Mismatch analyzer' is a tool that compares your protein sequence with other bunch of sequences and assists in analyzing conservation of residues at the positions you are interested in, based on multiple sequence alignment. Further, this tool helps you introducing taxonomy in to this analysis, provided you have taxonomy information in the sequence identifier. Now DNA/RNA mode is added.

Sequences file:

This file should have all your target sequences in fasta format. Each sequence record must have a title. If title has pipe or vertical bar symbol '|', ResCon will treat it as delimiter and divide it into individual elements when written in to output files.

Sequences file may or may not have reference sequence included in it. If reference sequence is already present in 'Sequences file', its sequence and title must be identical in both files. Number of sequences allowed is not restricted but larger number of sequences may result in longer processing time for clustal alignment.

Now, you can define which character acts as a separator.
Goto File -> Edit settings

Reference file:

This file should have your reference sequence in fasta format. Only one sequence is allowed in this file.

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'Sequences file' is located. You may choose to change this to different folder, though.

Residue positions:

This refers to the residue positions, based on reference sequence, of your interest for which you would like to perform mismatch/match analysis. If you are interested in more than one residue position, enter them as comma separated numbers.

Now, tab can be used as separator in addition to comma character

Check 'All' button if all residues of Reference sequence needs to be analyzed

Clustal alignment required - Checkbox:

- *If checked:*

This indicates that you need ResCon to perform multiple sequence alignment using Clustal Omega. Note that, in this case, you ~~need~~ to have Clustal omega installed in your computer. See [Appendix D](#) for Clustal Omega installation instructions.

Now clustal omega can be used via webserver; installation on your computer is now optional

Clustal-O command:

This is where you enter the parameters required to execute clustal omega. Clustal Omega is a command-based program and it allows changing parameters to build multiple sequence alignment. By default, ResCon uses following command:

```
{'outfmt': 'clu', 'auto': 'True', 'force': 'True', 'outfile': 'default.aln'}
```

Now following formats are supported: clustal, fasta, philip, vie

You may change this to fit your requirements. If you have to edit this default command, follow these instructions:

1. Command must open and close with curly braces.
2. Each set of parameter and value must be separated by comma.
3. Parameters and their values, each, must be inside single or double quotes and separated by colon (:) symbol.

Parameters available for Clustal omega are shown in [appendix C](#) and the instructions below explain its structure.

```
  
_Option(["--outfmt", "outfmt"],  
        "MSA output file format:"  
        " a2m=fa[sta],clu[stal],msf,phy[lip],selex,st[ockholm],vie[nna]"  
        " (default: fasta).",  
  
_Switch(["-v", "--verbose", "verbose"],  
        "Verbose output"),
```

Yellow highlighted text signifies the parameters that you may use for command line in ResCon. If a parameter belongs to 'Option' category (as marked

with red arrow in above figure), then you should provide value for that parameter. (for example, 'outfmt': 'clu'). If they belong to 'Switch' category (blue arrow marked) instead, their value is provided as either 'True' or 'False' (for example, 'auto': 'True').

Tips and tricks / Troubleshooting:

1. ~~Parameter 'outfile' must always be included as part of command line. By default, value for 'outfile' is set to 'default.aln', meaning that output alignment file will be named by ResCon. If you like to provide your own filename instead, replace 'default.aln' with filename of your choice along with its file path. For example: 'outfile': 'C:/phys/chem/new_name.aln.~~
2. If you need Clustal omega to build phylogenetic tree as output, include
`'guidetree_out': 'default.newick'`
as part of the command and this will instruct ResCon to name and save the phylogenetic tree in the output folder chosen. If you like to provide your own filename instead, replace 'default. newick' with filename of your choice along with its file path. For example: 'guidetree_out': 'C:/phys/chem/new_name.newick'.
3. 'Log file' stored in output folder shows the command directed to your computer terminal. If you run into error with clustal omega command line, see log file for clues on what is going wrong. See [Appendix B](#) to read about log file.

- ***If unchecked:***

This indicates that you need ResCon to use a pre-aligned file for mismatch analysis. Note that, in this case, your multiple sequence alignment file must have reference sequence already as a part of it and ID of that sequence should be same as that of sequence in 'Reference file'.

Clustal Aligned file:

This file should have multiple sequence alignment data in clustal '.aln' format (with or without numbers). If your alignment is in any other format, alignment format converters are available online:

1. ALTER:

<http://sing.ei.uvigo.es/ALTER/>

2. Format Converter v2.3.5:

http://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERTER/form.html

Now supports following formats:
Fasta, Clustal, Nexus, Phylip4, Ig, Stockholm

HTML formatting - Checkbox:

If checked, ResCon will color-code multiple sequence alignment at the residue positions requested for mismatch analysis and this color-coded html file will be saved in the folder selected at 'Output folder' box. Note that, your multiple sequence alignment ~~must~~ in clustal '.aln' format (with or without numbers) for this to work.

Now supports following formats: Fasta, Clustal, Nexus, Phylip4, Ig, Stockholm

Now can use clustal omega via webserver (optionl)

Demo 1: Mismatch analysis (*Clustal Omega required*):

Use the demo files from folder 'Demo files\Mismatch analyzer\Alignment required' to get a feel of how ResCon works. ~~You will need to have Clustal Omega installed in your computer.~~ Here you will analyze whether residues at positions 30, 60, and 106 in reference sequence are conserved or not in target sequences.

File Help

Sequences file Browse

Reference file Browse

Output folder Browse

Residue positions

☒ Clustal Alignment required?

Clustal-O Command

☒ HTML Formatting?

Submit job

More Tools

Now, tab can be used as separator in addition to comma character

Now 'All' check button is added. Check it to analyze all residues of Reference sequence

Click 'Submit job' after filling in the required fields as shown in above figure. When job is done, files will be written in to the output folder selected. See '[Overview of output files](#)' to understand how to use output files.

Demo 2: Mismatch analysis (Clustal Omega not required):

Use the demo files from folder 'Demo files\Mismatch analyzer\Pre-aligned\'.

You do not need to have Clustal omega installed in your computer for this to work. Note that you are using pre-aligned multiple sequence alignment in clustal format as input and this alignment already has reference sequence as part of it. Here you will analyze whether residues at positions 30, 60, and 106 in reference sequence are conserved or not in target sequences used in multiple sequence alignment.

The screenshot shows the ResCon web application interface. The window has a title bar with '76' and 'ResCon'. The menu bar includes 'File' and 'Help'. The main form contains the following fields and controls:

- Sequences file:** A text input field containing '/Mismatch analysis/Alignment required/Sequences.fasta' and a 'Browse' button.
- Reference file:** A text input field containing 's/Mismatch analysis/Alignment required/Reference.fasta' and a 'Browse' button.
- Output folder:** A text input field containing 'D:/Demo files/Mismatch analysis/Pre-aligned/Output/' and a 'Browse' button.
- Residue positions:** A text input field containing '30,60,106'.
- Clustal Alignment required?:** An unchecked checkbox.
- Clustal Aligned file:** A text input field containing 's/Mismatch analysis/Pre-aligned/Clustal_pre_aligned.aln' and a 'Browse' button.
- HTML Formatting?:** A checked checkbox.
- Submit job:** A large blue button.
- More Tools:** A blue button.

A red arrow points from a text box to the 'Clustal Aligned file' field. The text box contains the following text:

Now supports following formats: Fasta, Clustal, Nexus, Phylip4, Ig, Stockholm

Click 'Submit job' after filling in the required fields as shown in above figure.

When job is done, files will be written in to the output folder selected. See '[Overview of output files](#)' to understand how to use output files.

Overview of output files:

ResCon now shows a bar chart representing conservation of residues (not shown here).

1. Formatted_Alignment.html

- This html file contains multiple sequence alignment, color-coded at the sites requested for mismatch analysis. Use any internet browser to view this file.

| Position | 30 K | 60 N | 106 L |
|--------------|---------|---------|----------|
| % Identity | 100.0 | 25.0 | 25.0 |
| % Similarity | 100.0 | 25.0 | 100.0 |

1

Now added feature that labels the residue(s) under testing. (Not shown here)

CLUSTAL O(1.2.0) multiple sequence alignment

WP_0121692i|Bacteria|Azorhizobium|caulinodans|
WP_0098740i|Bacteria|Buchnera|aphidicola|
WP_0196459i|Bacteria|Novispirillum|itersonii|
WP_0027768i|Bacteria|Campylobacter|coli|
WP_0202604i|Bacteria|Escherichia|coli|

2

WP_0121692i|Bacteria|Azorhizobium|caulinodans|
WP_0098740i|Bacteria|Buchnera|aphidicola|
WP_0196459i|Bacteria|Novispirillum|itersonii|
WP_0027768i|Bacteria|Campylobacter|coli|
WP_0202604i|Bacteria|Escherichia|coli|

WP_0121692i|Bacteria|Azorhizobium|caulinodans|
WP_0098740i|Bacteria|Buchnera|aphidicola|
WP_0196459i|Bacteria|Novispirillum|itersonii|
WP_0027768i|Bacteria|Campylobacter|coli|
WP_0202604i|Bacteria|Escherichia|coli|

MRIDAIPVGKPNPHEVNVVIEVPIGGEPIKYEMDKDAGALFVDRFLYTAMRYPGNYGFIP
MNLNKIVAGKDLPEDIYVVIEIPANADPIKYEIDKESGALFVDRFMSTAMFYPCNYGYIN
MDIKKIPIGKDAPRDVNVIEIPLRSDPVKYEVDKESGAMYVDRFLHTAMHYPCNYGFIP
MDLSKIKIG-DIPNKINAVIEIPYG-SSIKYEIDKDSGAIMVDRVMASAMFYPCNYGFIA
MSLLNVPAGKDLPEDIYVVIEIPANADPIKYEIDKESGALFVDRFMSTAMFYPCNYGYIN
* : : * : * . . . : . : * * * : . : * * * * * : * * * : * * * * * *

HTLSGDGDPDVLVANTRAIVPGAVMSVRPVGVLMEDESGVDEKIIA--VPGPKLTCKRY
HTLSLDGDPDVLVPSHYPIKSSCVIHCKPIGILNMQDESGDDAKIIA--VPKSKICQEY
HTLSDDGDPDVMVGNMPVAVGSMRVRPVGVMYMEDESGGDEKIIA--VPHSKLHPYH
NTLADDGDPVDILVLNEYPIQAGAVIPCRILGVLIMEDESGMDEKIIA--VPNSKIDARY
HTLSLDGDPDVLVPTPYPLQPGSVIRCRPVGVLKMTDEAGEDAKIIAVAVPHSKLSKEY
. * : * * * * * : * * : . . : : : * * : * * * * * : * * * : * * * : *

DKVLSYKDLADITLKQIEHFFEHYKDLE--
KNINDISDISELLKKQITHFFQNYKTLENK
DNVKNYTDLRDVLRQIEHFFVHYKDLE--
DNIKTYTDLPAATLNKIKNFFETYLE--
DHIKDVNDLPELLKAQIAHFFEHYKDLE--
.. : . * : : : * . * * * * *

Input files used:

D:/Demo files/Mismatch analysis/Output/Aligned_Clustalo-Sequences_Reference_SeqAdded.aln
D:/Demo files/Mismatch analysis/Reference.fasta

3

- This html file has three parts (shown in yellow circles):
 - Table shows requested sites and their % identity and % similarity.
 - Cyan highlighted ID refers to reference sequence
 - Shows input files used to get this output
- Residue Sites under mismatch analysis are highlighted with three different colors as follows:

Now, you can choose colors as you like.
Go to File --> Edit settings

Green : Matching residue
Yellow : *Mismatch* but a *similar* residue
Pink : *Mismatch* and a *non-similar* residue

Essentially 'green' highlighted are matching residues whereas both 'yellow' and 'pink' highlighted are mismatching residues. Similarity of residues is defined on the basis of their physiochemical properties (same as Clustal Omega's coloring scheme) as shown below.

| Residue | Property |
|----------|---|
| AVFPMILW | Small (small + hydrophobic (including aromatic -Y)) |
| DE | Acidic |
| RK | Basic - H |
| STYHCNGQ | Hydroxyl + sulfhydryl + amine + G |

Now, you can define your own set of similarity among residues.
Go to File -> Edit Settings

2. Mismatches_Tabulated.csv

- This is a csv (comma separated values) file that can be opened using any spreadsheet software. If your spreadsheet software asks for delimiter information, use comma as delimiter.
- This has 3 sections:
 - Shows details of sequence records that have at least one mismatch at the residue positions requested.
 - Shows unique residues present at each site along with their count and fraction at the requested sites in the alignment. Also, % identity and ~~% similarity~~ are shown. *Note that Reference sequence is not included in this calculation.*
 - Shows details of sequence records that does not have mismatch at any of the residue positions requested.

Now replaced with 'Conservation score' calculated by Liu08 method

Now this gets sorted by no. of mismatches

| | | | | | | | | | |
|---|------------------|----------------|---------------|------------------|-----------------|--|--------|--------|---------|
| Data below is obtained from file: "D:/Demo files/Mismatch analysis/Output/Sequences_Reference_SeqAdded.fasta" | | | | | | | | | |
| Reference file used: "D:/Demo files/Mismatch analysis/Reference.fasta" | | | | | | | | | |
| Number of sequences in alignment: 5 | | | | | | | | | |
| *** Records that have mismatches in at least one of the query sites *** | | | | | | | | | |
| S.No | Title_1 | Title_2 | Title_3 | Title_4 | Sequence Length | No. of Mismatches | 30 "K" | 60 "N" | 106 "L" |
| 1 | WP_0121692i | Bacteria | Azorhizobium | caulinodans | 146 | 2 | = | P | I |
| 2 | WP_0098740i | Bacteria | Buchnera | aphidicola | 148 | 1 | = | = | I |
| 3 | WP_0196459i | Bacteria | Novispirillum | itersonii | 146 | 2 | = | P | I |
| 4 | WP_0027768i | Bacteria | Campylobacter | coli | 144 | 1 | = | A | = |
| *** Unique residues seen at the query sites and their count. *** | | | | | | | | | |
| ** (Note: Calculation doesn't include Reference sequences's residue at that position) ** | | | | | | | | | |
| | Expected Residue | Identity_count | % Identity | Similarity_count | % Similarity | Unique residues' count and fraction | | | |
| | 30 "K" | 4 | 100 | 4 | 100 | K: 4 (100.0%) | | | |
| | 60 "N" | 1 | 25 | 1 | 25 | P: 2 (50.0%) N: 1 (25.0%) A: 1 (25.0%) | | | |
| | 106 "L" | 1 | 25 | 4 | 100 | I: 3 (75.0%) L: 1 (25.0%) | | | |
| *** Records that Do Not have mismatches at any of the query sites *** | | | | | | | | | |
| S.No | Title_1 | Title_2 | Title_3 | Title_4 | Sequence Length | No. of Mismatches | 30 "K" | 60 "N" | 106 "L" |
| 1 | WP_0202604i | Bacteria | Escherichia | coli | 148 | 0 | = | = | = |

This part is now replaced with 'Conservation score' calculated by Liu08 method

Details shown for sequence record:

- In section 1 and 3, result for each sequence record is shown. Pipe symbol “|” in sequence ID is a delimiting character and is used to divide the sequence id into parts and are then labeled as Title_1, Title_2, etc.. For example:

For ID: “WP_0121692i|Bacteria|Azorhizobium|caulinodans|”, it is shown as:

| S.No | Title_1 | Title_2 | Title_3 | Title_4 | Sequence Length | No. of Mismatches | 30 "K" | 60 "N" | 106 "L" |
|------|-------------|----------|--------------|-------------|-----------------|-------------------|--------|--------|---------|
| 1 | WP_0121692i | Bacteria | Azorhizobium | caulinodans | 146 | 2 | = | P | I |

- In cases where all sequences do not have same number of pipe symbols, missing sections will be mentioned as “-na-” in output csv and txt files.
- *Symbols* used here:
 - Equal sign = Matching residue
 - Star sign * Gap in alignment (star symbol used instead of hyphen symbol to improve visibility in spreadsheet software)

~~3. Mismatches_Detailed.txt:-~~

~~-~~

~~Data presented in this file is same as that shown in csv file except that, here, it is presented in text format and details for each requested sites are provided separately for each site.~~

4. log.txt:

See [Appendix B](#).

ResCon – More tools

1. Subtree Sequences Extractor:

This tool can read phylogenetic tree (in newick format) and then extract only the sequences corresponding to a subtree of your interest, based on that subtree's branch length. Newick phylogenetic tree contains only the sequence IDs but not sequence data itself. This tool 'Subtree Sequences Extractor' extracts sequence IDs corresponding to subtree of your interest and then filters sequences of that subtree from a fasta file containing sequence data.

Newick file:

This must be a phylogenetic tree in newick format. Software and online converters are available elsewhere to aid in conversion of other formats to newick format.

FASTA file:

This must be in fasta file format (preferably the same fasta file used towards making the phylogenetic tree used in 'Newick file').

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'Newick file' is located. You may choose to change this to different folder, though.

Clade's Branch Length:

This may contain one or more branch length values. *If more than one, they should be comma separated.* If your branch length is 0.123456, you may enter this in one of following formats:

1) 0.123456

2) 123456E-6

3) 123456e-6

Obtaining correct branch length is little tricky; so read the following before you proceed further with this tool:

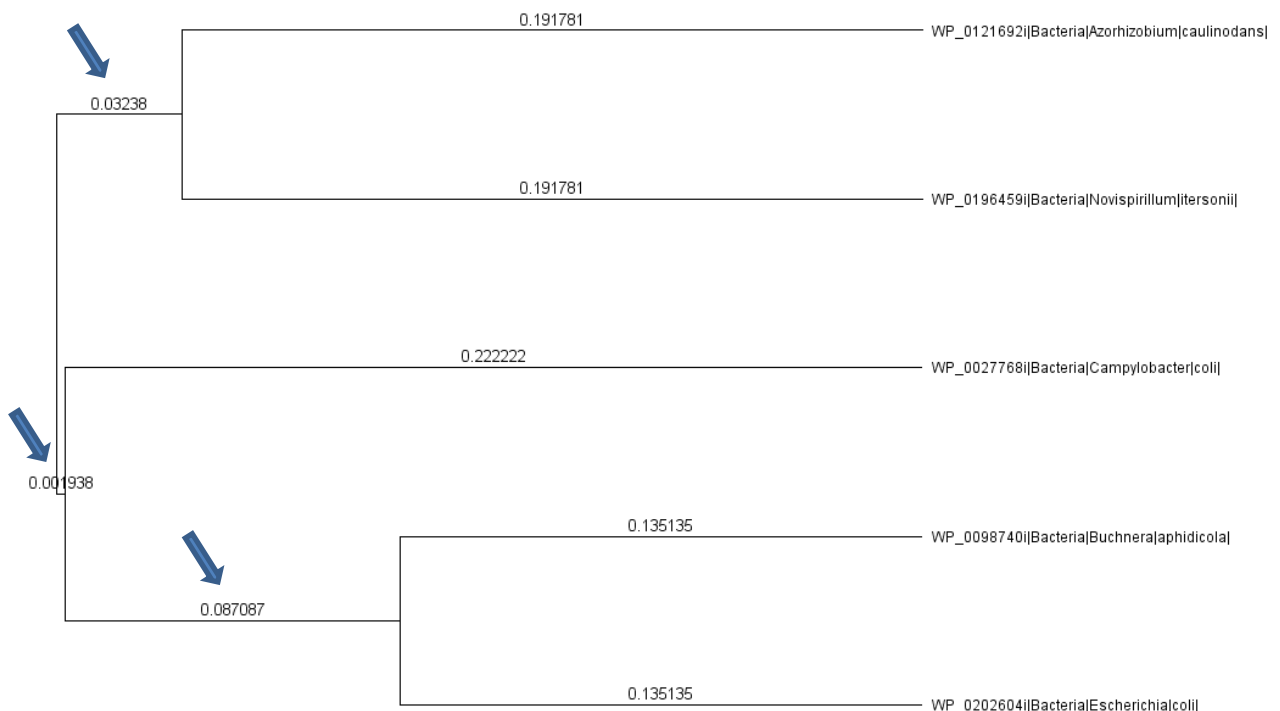
This tool requires using complete branch length number, not a rounded number. Unfortunately, many phylogenetic tree viewers do not show complete branch length but instead show them rounded to three/four decimal points. For example, in the demo files used, one of the branch lengths is '0.0323797' but a tree viewer may show this as '0.03238' or '0.0324'. Subtree Sequences Extractor will process properly only if '0.0323797' is used. In such instances, following work-around helps.

~~Work-around: Open newick tree file in a internet browser or in a text editor and search for '0.0323' using search feature (Ctrl + F). Go through all the hits and now you will be able to decipher which number you have to use.~~

Not necessary anymore. ResCon now suggests you the possible branch length numbers you are actually interested in.

Demo:

Use the demo files from folder 'Demo files\1. Subtree Sequences Extractor\' to get a feel of how tool 'Subtree Sequences Extractor' works.



Above figure shows the phylogenetic tree from demo file '*Tree_demo.newick*'. Arrow marks show branch lengths of subtrees. In the demo here, we will extract sequences corresponding to subtree with branch length '0.001938'.

The screenshot shows a software window titled "ResCon: More Tools" with a red close button in the top right corner. Inside the window, there are several tool buttons arranged in a grid: "Subtree Sequences Extractor" (highlighted in blue), "GenBank/GenPept To fasta Converter", "Filter fasta by Sequences' ID", "Filter fasta by Blast's E-value", "Name Extractor", and "Filter fasta by Sequences' Name". Below these buttons is a form with four input fields, each with a "Browse" button to its right: "Newick file" with the path "mo files/Subtree Sequences Extractor/Tree_demo.newick", "fasta file" with the path "o files/Subtree Sequences Extractor/All_Sequences.fasta", "Output folder" with the path "D:/Demo files/Subtree Sequences Extractor/Output/", and "Clade's Branch Length" with the value "0.001938" (which is highlighted in yellow). At the bottom right of the form is a blue button labeled "Extract now!".

Click 'Extract now' after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder.

Not necessary anymore. ResCon now suggests you the possible branch length numbers you are actually interested in.

Tips and Tricks / Troubleshooting:

1. *Error: Verify branch length entered. No such branch length is found in newick tree file:*

See the solution provided as '~~Work-around~~' under 'Clade's Branch Length'.

2. *Error: More than one instances of such branch length found:*

In some extreme cases, you may have same branch length for more than one subtree. You may choose to use following work-around to edit branch lengths, assuming that proper care is taken not to use that edited file elsewhere where branch length data is necessary for further studies. We recommend to work on a duplicate file to avoid any potential issues.

Solution: Open the newick file in a text editor, search for that branch length using inbuilt search feature and edit one of the branch lengths to a different number. Save it as same file or a new file and then use 'ResCon – Subtree Sequences Extractor' to extract the sequences and it should work now.

Now, choice is yours to let ResCon do this automatically

3. *Number of sequences extracted is less than the expected number of sequences:*

One possible culprit here in such case could be that your newick file has some branch lengths with negative values. Negative branch lengths (even negative zero ie. -0) spell trouble when using 'ResCon – Subtree Sequences Extractor'. The work-around for this problem involves editing the newick file and hence care must be taken not to use edited file where branch lengths are necessary. We recommend working on a duplicate file to avoid any potential issues.

~~Solution: Open the newick file in a text editor, search for ':' (i.e. colon followed by dash) using inbuilt find and replace feature, and then use 'replace all' feature to replace all ':' with just a colon ':'. Now save this file and try extracting sequences using this edited file in 'ResCon – Subtree Sequences Extractor'.~~

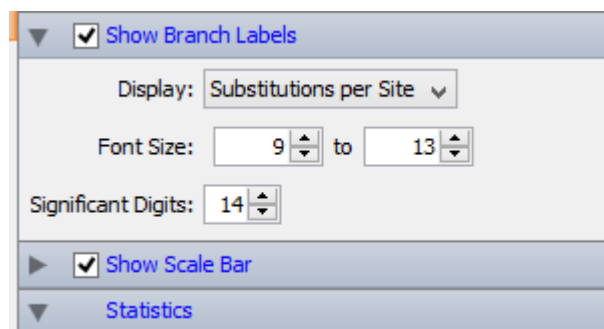
4. 'ResCon – Subtree Sequences Extractor' cannot extract if the clade has only one sequence ID in it. Use search feature in a text editor and copy it manually.

5. Sequence IDs used to build phylogenetic tree in newick format cannot have following sequences:

- Round braces (parentheses)
- Square braces
- Comma
- Single quote
- Colon
- Semi-colon

Any phylogenetic tree viewer or processor will most likely have trouble reading newick file with above symbols. You may use any text editor's find and replace feature to replace above symbols with some other character (for example, underscore).

6. Unfortunately not all phylogenetic tree viewers can show branch lengths. We recommend software 'Geneious' to visualize phylogenetic tree files as it offers an option to show complete branch length with all digits. This doesn't work all the time but it is better than most other software we have tried. If you use Geneious, change 'significant digits' under 'show branch labels' to a number greater than 10 as shown in the figure below and this will work for most of the time. This feature works even after trial period of Geneious is expired.



Following online phylogenetic tree viewers show branch lengths but number of digits they show are limited:

1. <http://evolgenius.info/evolview/#mytrees/qq/q>
2. <http://phylogeny.lirmm.fr/phylo.cgi/index.cgi>

2. GenBank/GenPept to fasta converter:

This tool converts GenBank or GenPept files to fasta format while conserving taxonomy information in fasta sequences' header. Fasta files available from ncbi database do not offer any taxonomy data except genus and species information. GenPept or GenBank files have such taxonomy information and this tool aids in creating fasta data while conserving taxonomy information.

GenBank/GenPept file:

This must be in either GenBank or GenPept format. Number of sequence records data they may have are not limited.

Output folder:


This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'GenBank/GenPept file' is located. You may choose to change this to different folder, though.

Fasta ID length:

This indicates the maximum length of identifier in fasta header (see [Appendix A](#)). Certain applications may restrict length of fasta identifier to certain length and this feature is helpful in those cases. Default value is 127 as Clustal omega can handle up to 127 characters in fasta header. You may change this to suit your needs.

Header options:

You may check the fields you may like to be included in fasta identifier (see [Appendix A](#)). We recommend selecting at least one option that is unique to a sequence (for example, Locus name and GI are unique values). Selected fields appear in the order shown (from left to right).



Now, you may choose which separator symbol to use to connect these fields. Default separator is Pipe symbol.

Now, you can choose which symbol(s) to use. Default is Underscore symbol. Go to File -> Edit Settings

Retrieval options:

1. *Replace Newick-sensitive symbols in header id with underscore symbol?*

Check this box if the converted output fasta sequences will be used to build a phylogenetic tree in newick format. Newick phylogenetic tree format will result in error if sequence ID contains symbols such as colon, semicolon, comma, single quote, parenthesis and square brackets as part of sequence's ID.

Now you get to choose which symbols need to be replaced.

2. *Ignore incomplete/partial sequence records?*

Check this option if you prefer not to extract sequences annotated as partial/incomplete in to output fasta file. Note that not all incomplete or partial sequences are annotated as such.

Period may cause issue if there is more than one period in a sequence identifier

Demo:

The screenshot shows a software window titled "ResCon: More Tools" with a red close button in the top right corner. The window contains several tool buttons arranged in a grid: "Subtree Sequences Extractor", "GenBank/GenPept To fasta Converter" (highlighted in blue), "Filter fasta by Sequences' ID", "Filter fasta by Blast's E-value", "Name Extractor", and "Filter fasta by Sequences' Name". Below these buttons is a configuration section for the "GenBank/GenPept To fasta Converter". It includes a text field for "GenBank/GenPept file" with the path "D:/Demo files/GenPept to fasta/GenPept_demo.gp" and a "Browse" button. The "Output folder" field shows "D:/Demo files/GenPept to fasta/Output/" with another "Browse" button. The "Fasta ID length" is set to "127" in a yellow-highlighted field. Under "Header options", there are three columns of checkboxes: "Locus Name", "GI", and "Domain" (all checked); "Phylum", "Class", and "Genus" (all checked); and "Species", "Seq Length", and "Seq Name" (all unchecked). Under "Retrieval options", there are two checked checkboxes: "Replace Newick-sensitive symbols in header id with underscore?" and "Ignore incomplete/partial sequence records?". A blue "Convert now!" button is located at the bottom right of the configuration section.

76 ResCon: More Tools

Subtree Sequences Extractor *GenBank/GenPept To fasta Converter*

Filter fasta by Sequences' ID Filter fasta by Blast's E-value

Name Extractor Filter fasta by Sequences' Name

GenBank/GenPept file D:/Demo files/GenPept to fasta/GenPept_demo.gp Browse

Output folder D:/Demo files/GenPept to fasta/Output/ Browse

Fasta ID length 127

Header options

| | | |
|--|---|--|
| <input checked="" type="checkbox"/> Locus Name | <input type="checkbox"/> GI | <input checked="" type="checkbox"/> Domain |
| <input checked="" type="checkbox"/> Phylum | <input checked="" type="checkbox"/> Class | <input checked="" type="checkbox"/> Genus |
| <input checked="" type="checkbox"/> Species | <input type="checkbox"/> Seq Length | <input type="checkbox"/> Seq Name |

Retrieval options

- ☒ Replace Newick-sensitive symbols in header id with underscore?
- ☒ Ignore incomplete/partial sequence records?

Convert now!

Use the demo files from folder 'Demo files\2. GenPept to fasta\' to get a feel of how tool 'ResCon - GenBank/GenPept to fasta converter' works. Click 'Convert now' after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder.

3. Filter fasta by Sequences' ID

This tool reads a file of fasta sequences and then extracts the sequences based on sequence IDs provided by you. You may provide either complete sequence ID or part of the sequence ID. See [Appendix A](#) to learn more about fasta sequence ID.

FASTA file:

This must be a fasta file. Number of sequences in it are not limited. Of course, sequences must have identifiers (IDs) in their header.

File with IDs:

This should be a text file containing only one sequence ID per line. If 'IDs are partial' checkbox is selected, each line in text file should have a partial ID. Number of IDs allowed in this file is not restricted.

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'FASTA file' is located. You may choose to change this to different folder, though.

Checkboxes – 'Include IDs in list' / 'Exclude IDs in list':

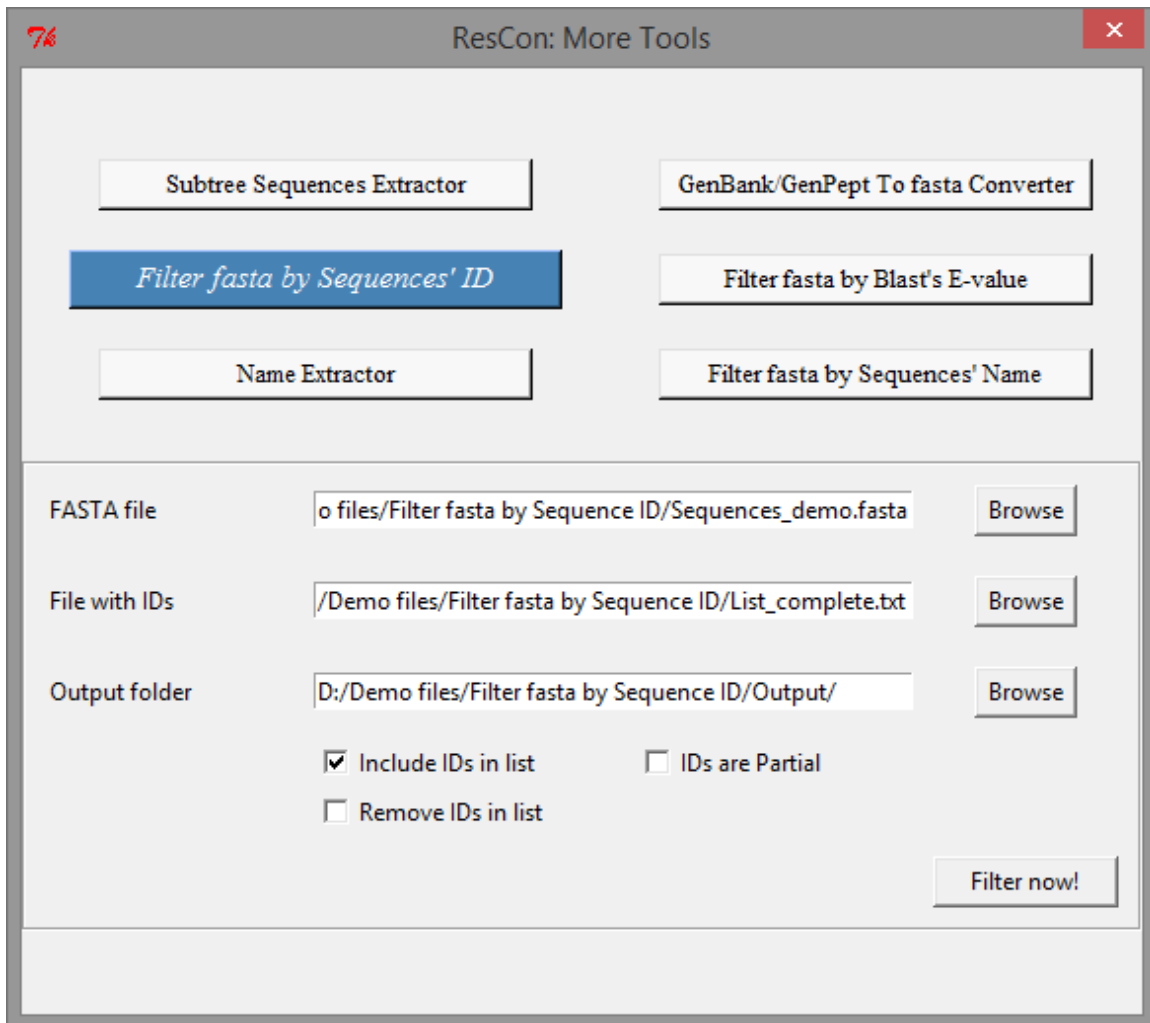
Check one of these boxes depending on if you would like to include or exclude sequences in fasta file based on sequence IDs present in 'File with IDs'.

IDs are partial:

Check this box if 'File with IDs' has partial IDs instead of complete ID. If partial IDs are to be used, it is a good idea to choose partial IDs that are unique (i.e. one partial ID corresponds to only one sequence in FASTA file).

Demo - 1 (Using *Complete* sequence IDs to filter fasta sequences):

Choose demo files 'Sequences_demo.fasta' and 'List_complete.txt' from folder 'Demo files\ 3. Filter fasta by Sequence ID\' to get a feel of how the tool "ResCon - Filter fasta by Sequences' ID" works. Make sure 'IDs are Partial' checkbox is unchecked.



Click 'Filter now' after populating the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder.

Demo – 2 (Using *Partial* sequence IDs to filter fasta sequences):

Now choose demo files '*Sequences_demo.fasta*' and '*List_partial.txt*' from folder 'Demo files\3. Filter fasta by Sequence ID\' and check 'IDs are Partial' checkbox.

The screenshot shows a software window titled "ResCon: More Tools". Inside, there are several buttons for different tools: "Subtree Sequences Extractor", "GenBank/GenPept To fasta Converter", "Filter fasta by Sequences' ID" (highlighted in blue), "Filter fasta by Blast's E-value", "Name Extractor", and "Filter fasta by Sequences' Name". Below these buttons is a configuration section for the "Filter fasta by Sequences' ID" tool. It includes three text input fields: "FASTA file" with the path "o files/Filter fasta by Sequence ID/Sequences_demo.fasta", "File with IDs" with the path "D:/Demo files/Filter fasta by Sequence ID/List_partial.txt", and "Output folder" with the path "D:/Demo files/Filter fasta by Sequence ID/Output/". Each field has a "Browse" button to its right. Below the fields are two checked checkboxes: "Include IDs in list" and "IDs are Partial", and one unchecked checkbox: "Remove IDs in list". A blue "Filter now!" button is located at the bottom right of the configuration section.






Click 'Filter now' after populating the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder.

Note: You may extract sequences using either complete sequence IDs or partial sequence IDs when 'IDs are partial' checkbox is selected. But be aware that if you are extracting sequences using complete sequence IDs, keeping 'IDs are partial' checkbox unchecked will work faster than having it checked.

4. Filter fasta by Blast's E-value:

This tool reads a file of sequences in fasta format and a XML file containing BLAST results and then extract the sequences based on user provided E-value.

Consider this scenario – Sequences in file '*Sequences_demo.fasta*' was BLASTed against *E. coli*'s IPPase and results are shown in the figure below. Now let's say you want to filter sequences whose E-value is below or above certain threshold value. Bingo! This is where this tool is helpful.

| Alignments  Download  Graphics  Distance tree of results  Multiple alignment  | | | | | | | Max score | Total score | Query cover | E value | Ident | Accession |
|--|---|--|--|--|--|--|-----------|-------------|-------------|---------|-------|--------------|
| Description | | | | | | | | | | | | |
| <input type="checkbox"/> | YP_004282029 Bacteria Aquificae Desulfurobacteriales Desulfurobacterium thermolithothrophum DSM 11699 inorganic pyrophosphatase [Desulfurobacte | | | | | | 184 | 184 | 100% | 9e-64 | 59% | Query_105684 |
| <input type="checkbox"/> | YP_004366927 Bacteria Deinococcus-Thermus Deinococcus Marinithermus hydrothermalis DSM 14884 inorganic pyrophosphatase [Marinithermus hydrot | | | | | | 122 | 122 | 99% | 8e-40 | 41% | Query_105691 |
| <input type="checkbox"/> | YP_005257157 Bacteria Firmicutes Clostridia Sulfobacillus acidophilus DSM 10332 inorganic pyrophosphatase [Sulfobacillus acidophilus DSM 10332] | | | | | | 111 | 122 | 89% | 9e-36 | 46% | Query_105688 |
| <input type="checkbox"/> | YP_004268229 Bacteria Planctomycetes Planctomycetia Planctomycetes brasiliensis DSM 5305 inorganic pyrophosphatase [Planctomycetes brasiliensis D | | | | | | 111 | 111 | 97% | 2e-35 | 40% | Query_105693 |
| <input type="checkbox"/> | YP_004238254 Bacteria Bacteroidetes Flavobacteriia Weeksella virosa DSM 16922 inorganic pyrophosphatase [Weeksella virosa DSM 16922] | | | | | | 106 | 120 | 87% | 6e-34 | 45% | Query_105692 |
| <input type="checkbox"/> | YP_005009634 Bacteria Bacteroidetes Sphingobacteriia Niastella koreensis GR20-10 inorganic pyrophosphatase [Niastella koreensis GR20-10] | | | | | | 105 | 105 | 96% | 3e-33 | 39% | Query_105689 |
| <input type="checkbox"/> | YP_004788406 Bacteria Bacteroidetes Flavobacteriia Muricauda rustringensis DSM 13258 inorganic pyrophosphatase [Muricauda rustringensis DSM 1 | | | | | | 99.0 | 99.0 | 88% | 7e-31 | 47% | Query_105680 |
| <input type="checkbox"/> | YP_004445607 Bacteria Bacteroidetes Sphingobacteriia Haliscomenobacter hydroxialis DSM 1100 inorganic pyrophosphatase [Haliscomenobacter hydro | | | | | | 98.6 | 111 | 97% | 1e-30 | 38% | Query_105690 |
| <input type="checkbox"/> | YP_007049604 Bacteria Cyanobacteria Nostocales Nostoc sp. PCC 7107 inorganic pyrophosphatase [Nostoc sp. PCC 7107] | | | | | | 95.1 | 95.1 | 100% | 2e-29 | 39% | Query_105683 |
| <input type="checkbox"/> | YP_007162693 Bacteria Cyanobacteria Oscillatoriothrix deaei Cyanobacterium laponinum PCC 10605 inorganic pyrophosphatase [Cyanobacterium aponi | | | | | | 94.7 | 94.7 | 89% | 3e-29 | 40% | Query_105677 |
| <input type="checkbox"/> | YP_007139175 Bacteria Cyanobacteria Nostocales Calothrix sp. PCC 6303 inorganic pyrophosphatase [Calothrix sp. PCC 6303] | | | | | | 94.4 | 94.4 | 100% | 4e-29 | 39% | Query_105687 |
| <input type="checkbox"/> | YP_004164947 Bacteria Bacteroidetes Flavobacteriia Cellulophaga algaicola DSM 14237 inorganic pyrophosphatase [Cellulophaga algaicola DSM 14237] | | | | | | 94.0 | 94.0 | 88% | 7e-29 | 46% | Query_105679 |
| <input type="checkbox"/> | YP_004262651 Bacteria Bacteroidetes Flavobacteriia Cellulophaga lytica DSM 7489 inorganic pyrophosphatase [Cellulophaga lytica DSM 7489] | | | | | | 93.6 | 93.6 | 88% | 8e-29 | 41% | Query_105678 |
| <input type="checkbox"/> | YP_007070257 Bacteria Cyanobacteria Oscillatoriothrix deaei Leptolyngbya sp. PCC 7376 inorganic pyrophosphatase [Leptolyngbya sp. PCC 7376] | | | | | | 90.5 | 90.5 | 100% | 1e-27 | 36% | Query_105681 |
| <input type="checkbox"/> | YP_007132473 Bacteria Cyanobacteria Pleurocapsales Stanieria cyanosphaera PCC 7437 inorganic pyrophosphatase [Stanieria cyanosphaera PCC 74 | | | | | | 88.6 | 88.6 | 100% | 5e-27 | 38% | Query_105685 |
| <input type="checkbox"/> | YP_007127661 Bacteria Cyanobacteria Oscillatoriothrix deaei Gloeocapsa sp. PCC 7428 inorganic pyrophosphatase [Gloeocapsa sp. PCC 7428] | | | | | | 87.0 | 87.0 | 91% | 3e-26 | 39% | Query_105686 |
| <input type="checkbox"/> | YP_007065375 Bacteria Cyanobacteria Nostocales Calothrix sp. PCC 7507 inorganic pyrophosphatase [Calothrix sp. PCC 7507] | | | | | | 86.3 | 86.3 | 100% | 4e-26 | 38% | Query_105682 |
| <input type="checkbox"/> | NP_066952 Eukaryota Metazoa Chordata Homo sapiens inorganic pyrophosphatase [Homo sapiens] | | | | | | 40.4 | 40.4 | 87% | 3e-09 | 28% | Query_105676 |

XML BLAST file:

This field should be filled with BLAST data in XML file format. If you use NCBI for BLAST, there is an option to download data in XML format.

FASTA file:

This field should be filled with a fasta file preferably the same fasta file used to obtain BLAST data.

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'XML BLAST file' is located. You may choose to change this to different folder, though.

E-value Threshold:

The threshold E-value you would like to use can be specified as an integer value or as a decimal value. If it is a decimal number, for example 0.0023, it could be entered as 0.0023 or as 2.3E-4 or as 2.3e-4. See [Tips and Tricks / Troubleshooting](#) to avoid potential mistakes.

Lower than E-threshold:

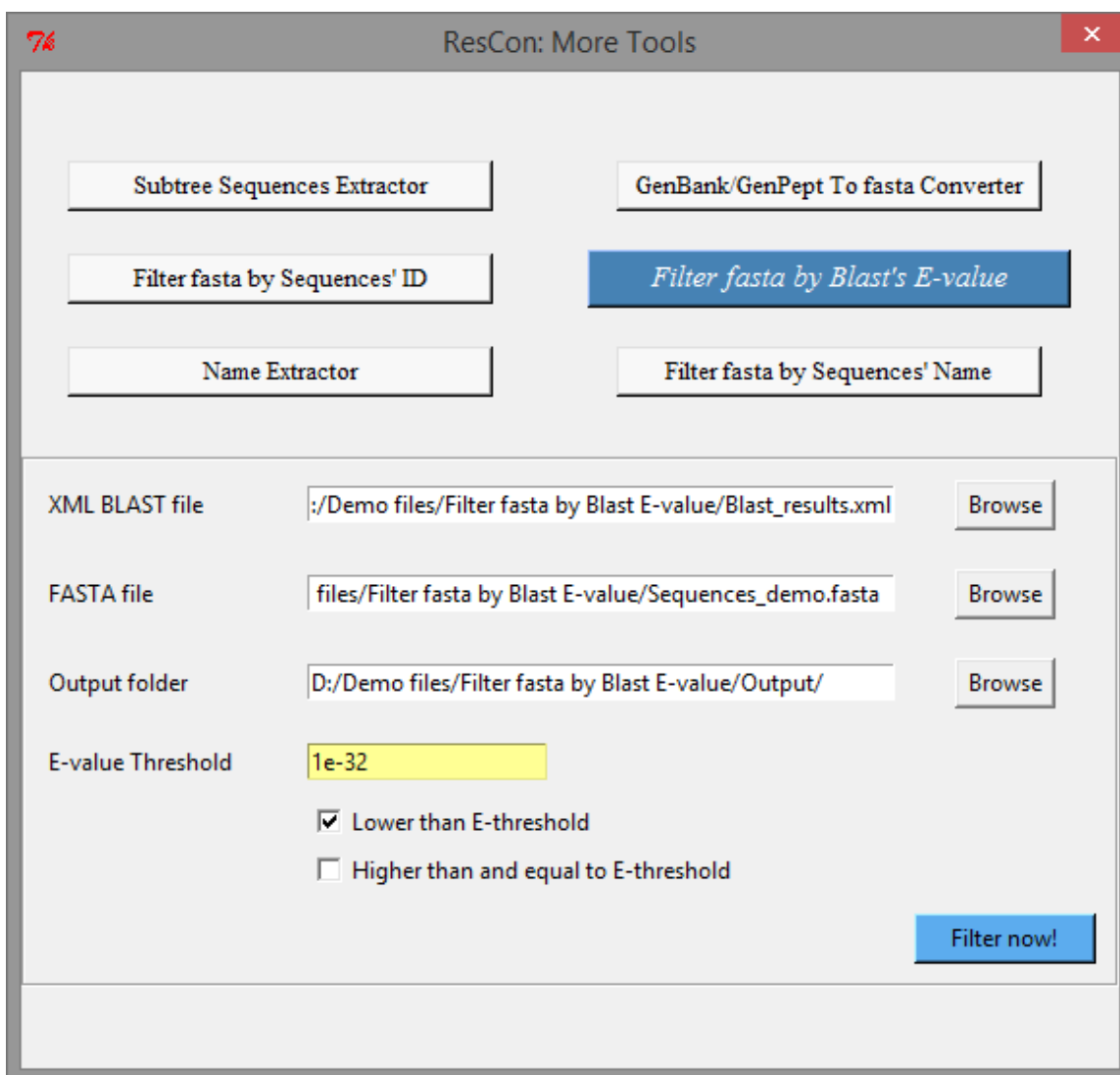
Check this box if you need to extract sequences whose E-values are less than the E-value threshold entered.

Higher than and equal to E-threshold:

Check this if you need to extract sequences whose E-values are higher than and equal to the provided E-value threshold. Note that this will result in extracting sequences that are not in blast XML file but are present in provided fasta file that do not have any E-value listed in XML file. Such instances arise when a sequence does not share any significant similarity to the query sequence.

Demo:

Use the demo files from folder 'Demo files\4. Filter fasta by Blast E-value\' to get a feel of how the tool 'ResCon - Filter fasta by Blast's E-value' works. Here, we will extract sequences whose E-value is less than 1e-32. Click 'Convert now' after filling in the required fields as shown in figure below. When the job is done, files will be written in to the selected output folder.



Tips and Tricks / Troubleshooting:

1. Error: Number of sequences filtered is not same as expected number of sequences:

Possibly this is due to E-value threshold entered. Even though E-values appear as numbers with only one decimal place in NCBI blast table, they actually are not. They are actually numbers rounded off to only one decimal place. For example, be it 2.3, 2.30116 or 2.2844, all these will show up as just 2.3.

This is important to know because if E-value threshold is provided as 2.3 and if certain sequence's E-value is 2.30116, then this sequence record would not be written into output file when 'lower than E value' is selected.

5. Filter fasta by Sequences' Description:

This tool reads a file of fasta sequences and then extracts the sequences based on sequence descriptions provided by you. You may provide either complete sequence description or part of the sequence description. See [Appendix A](#) to learn more about fasta sequence description.

FASTA file:

This must be a fasta file. Number of sequences in it are not limited. Of course, sequences must have descriptions in their header.

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called 'Output' where 'FASTA file' is located. You may choose to change this to a different folder, though.

Get Descriptions from text file - Checkbox:

- ***If unchecked:***

This indicates that you need to provide descriptions (or partial descriptions) as a list, not from a file.

List of Descriptions:

This is the field where you will provide descriptions as double comma separated terms. For example: "*DNA polymerase,,Helicase*". Double commas are used as delimiters, instead of single commas, because single comma may sometimes inherently be a part of the description. Note that you may not add additional space character if they are not part of the sequence description itself, or else you may not extract the intended sequence records.

- ***If checked:***

This indicates that you need to provide descriptions (complete or partial) in a text file.

File with Descriptions:

This should be a text file containing one sequence description per line. If 'Filter using partial Descriptions' checkbox is selected, each line in text file should have a partial description. Number of descriptions allowed in this file is not restricted. Note that you may not include additional space character if they are not part of the sequence description itself, or else you may not extract the intended sequence records.

Filter using Partial Descriptions:

Check this box if you need to match only certain part of the sequence descriptions from fasta sequence headers. Assuming you have certain regularity in the structure of descriptions of all sequences present in your fasta file, you may use regular expression so as to filter fasta sequences on the basis of partial descriptions.

Regular Expression:

This field should contain the regular expression that matches to your intended part of sequence descriptions. By default, the tool uses `(.+)\[.+` as the regular expression. This means that the part of sequence description from its beginning up to character before space followed by a open square brace will be used for matching. For example, if sequence description is

inorganic pyrophosphatase [Escherichia coli]

then use of default regular expression will result in

inorganic pyrophosphatase

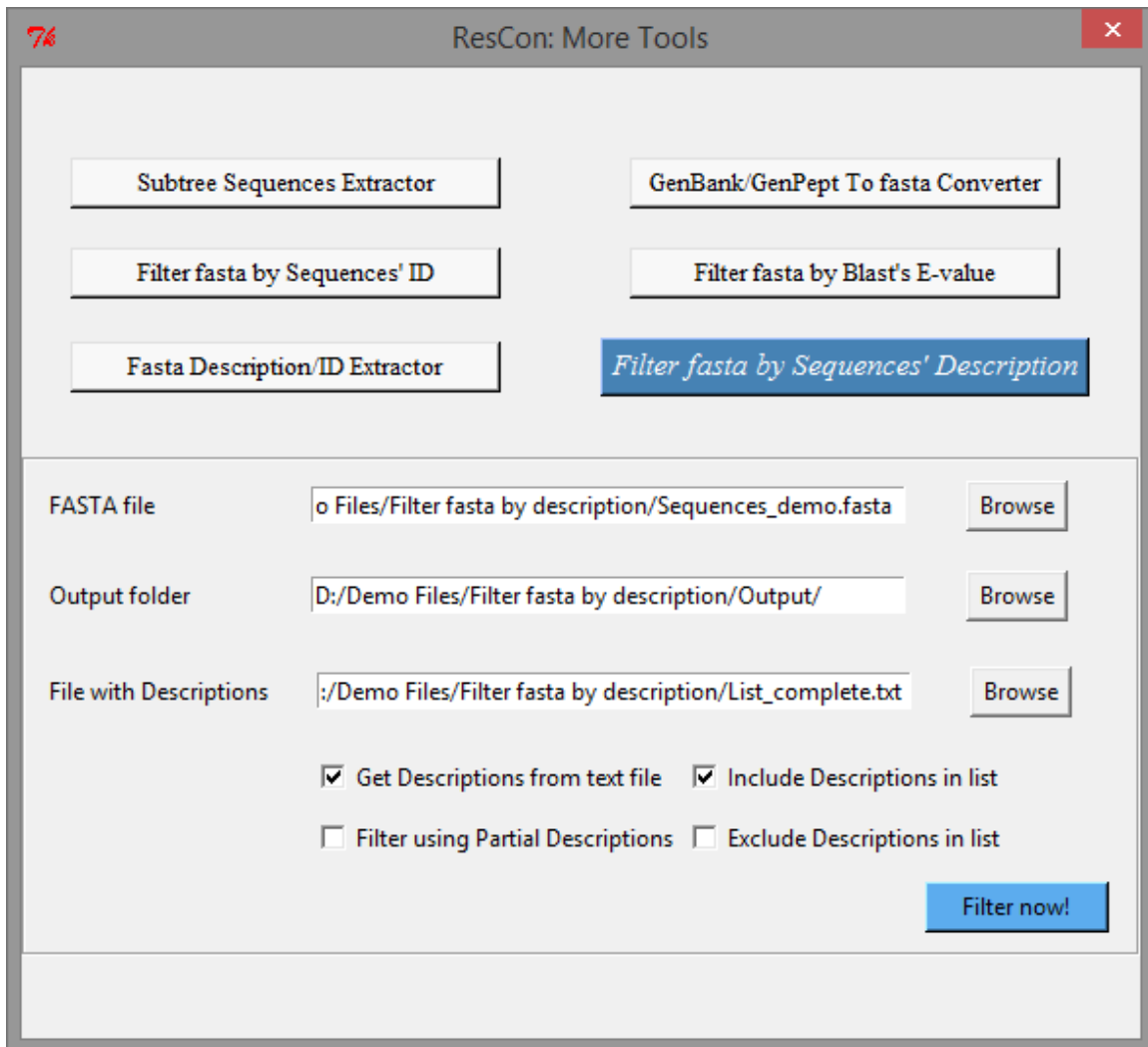
While regular expression seems intimidating at first, trust us it does not take long to learn using it. This web site could be a good start- <http://www.proftpd.org/docs/howto/Regex.html>. You may use any text editor, which allows use of regular expressions or wild card as part of their inbuilt search (ctrl + F) feature, to test and practice them. For example, notepad++ (not notepad), text wrangler, gedit, Microsoft word, etc. have this feature.

Include / Exclude Descriptions in list:

Check one of these boxes depending on if you need to include or exclude the sequences that match to the provided complete or partial descriptions.

Demo 1 - Using Complete Descriptions from a text file:

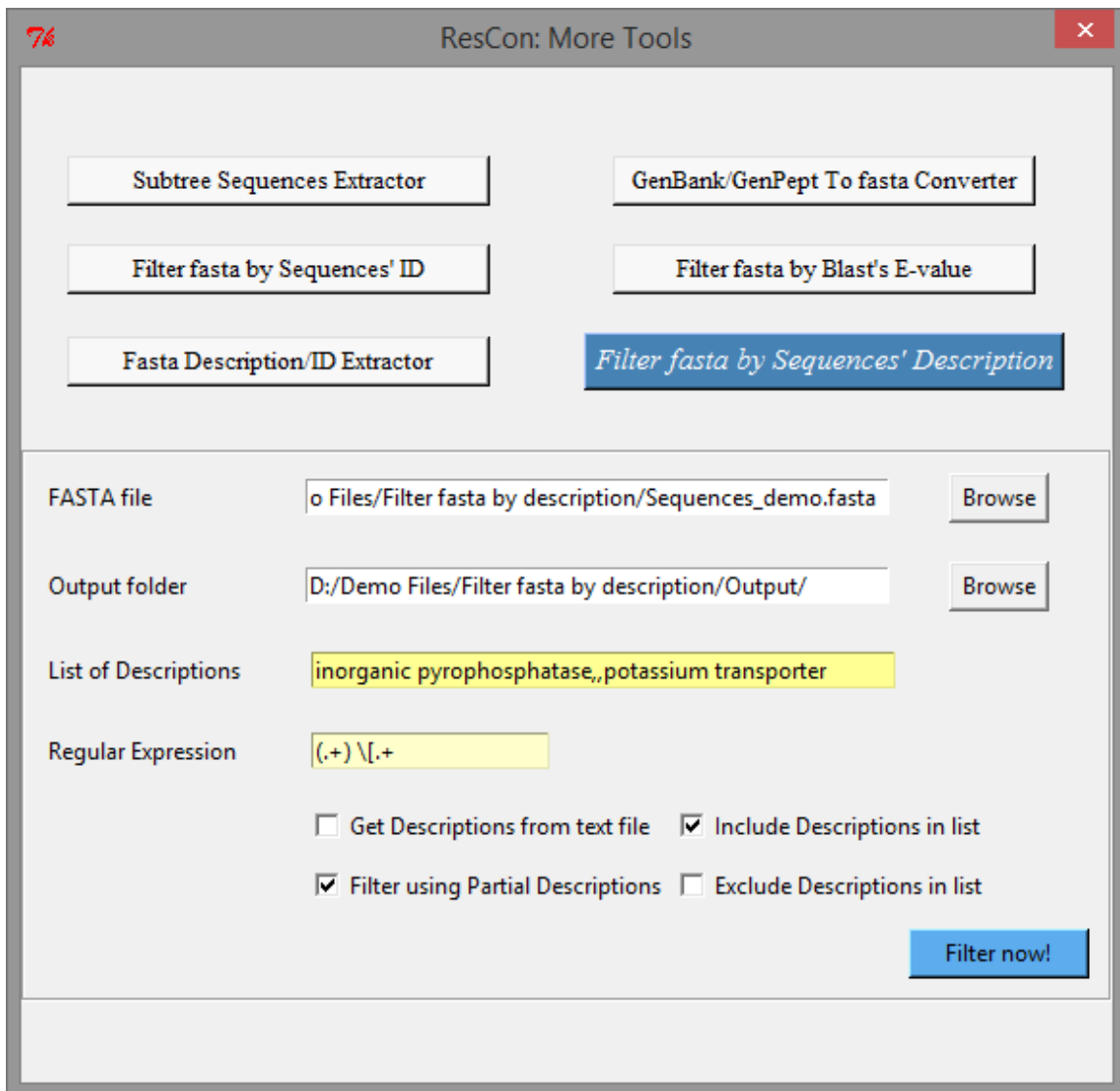
Use the demo files 'Sequences_demo.fasta' and 'List_complete.txt' from folder 'Demo Files\5. Filter fasta by description\' to get a feel of how the tool 'ResCon - Filter fasta by Sequences' Description' works. Here, we will extract fasta sequences that have matching descriptions provided in a text file.



Click 'Filter now' after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder. The output file should have 35 sequences written in to it.

Demo 2 - Using Partial Descriptions as a list:

Use file 'Sequences_demo.fasta' from folder 'Demo Files\5. Filter fasta by description\' for this demo. Here we will extract sequences whose partial descriptions (based on regular expression) match with partial descriptions we provide as a list. For 'List of Descriptions' box, enter 'inorganic pyrophosphatase,,potassium transporter'. Note that partial descriptions are separated by double comma and no extra space in the end of partial descriptions.



The screenshot shows a software window titled "ResCon: More Tools" with a red close button in the top right corner. The window contains several tool buttons in a grid: "Subtree Sequences Extractor", "GenBank/GenPept To fasta Converter", "Filter fasta by Sequences' ID", "Filter fasta by Blast's E-value", "Fasta Description/ID Extractor", and "Filter fasta by Sequences' Description" (which is highlighted in blue). Below these buttons are input fields and checkboxes. The "FASTA file" field contains "o Files/Filter fasta by description/Sequences_demo.fasta" with a "Browse" button. The "Output folder" field contains "D:/Demo Files/Filter fasta by description/Output/" with a "Browse" button. The "List of Descriptions" field contains "inorganic pyrophosphatase,,potassium transporter" and is highlighted in yellow. The "Regular Expression" field contains "(.+)\[.+" and is also highlighted in yellow. There are four checkboxes: "Get Descriptions from text file" (unchecked), "Include Descriptions in list" (checked), "Filter using Partial Descriptions" (checked), and "Exclude Descriptions in list" (unchecked). A blue "Filter now!" button is located at the bottom right of the configuration area.

Click 'Filter now' after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder. The output file should have 326 sequences written in to it.

6. Fasta Description/ID Extractor:

This tool enables you to extract descriptions or identifiers, complete or partial, from sequence headers in a fasta file. See [Appendix A](#) to learn more about fasta sequence description and identifier. You may choose between **1. Description Extractor** and **2. Identifier Extractor**. Here, we explain only the features of ‘Description Extractor’ as ‘Identifier Extractor’ has the same features except that the latter will extract sequence identifiers.

FASTA file:

This must be a fasta file. Number of sequences in it are not limited. Of course, sequences must have descriptions in their header.

Output folder:

This indicates where output files will be saved. By default, ResCon will select to save output files in to a folder called ‘Output’ where ‘FASTA file’ is located. You may choose to change this to a different folder, though.

Extract Partial Description:

Check this box if you need to extract only certain part of the sequence descriptions from fasta sequence headers. Assuming you have certain regularity in the structure of descriptions of all sequences present in your fasta file, you may use regular expression so as to extract only the partial descriptions.

Regular Expression:

This field should contain the regular expression that matches to your intended part of sequence descriptions. By default, the tool uses ‘(.+) \[.+\’ as the regular expression. This means that the part of sequence description from its beginning up to character before space followed by a open square brace will be extracted. For example, if sequence description is

inorganic pyrophosphatase [Escherichia coli]

then use of default regular expression will result in

inorganic pyrophosphatase

While regular expression seems intimidating at first, trust us it does not take long to learn using it. This web site could be a good start- <http://www.proftpd.org/docs/howto/Regex.html>. You may use any text editor, which allows use of regular expressions or wild card as part of their inbuilt search (ctrl + F) feature, to test and practice them. For example, notepad++ (not notepad), text wrangler, gedit, Microsoft word, etc. have this feature.

Note for ‘Identifier Extractor’:

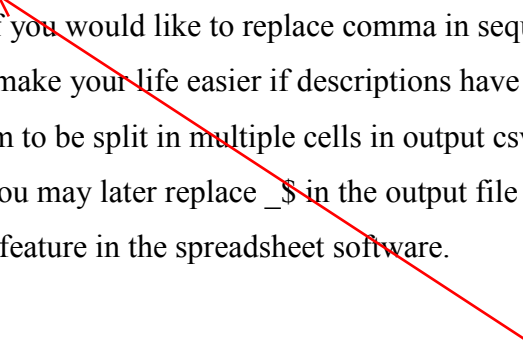
The default regular expression used in Identifier Extractor is ‘(\w+)\|.+’. This will extract the part of the identifier up to the character before the first pipe ‘|’ character. For example, for sequence identifier

WP_123890|Bacteria|Escherichia|coli

the part that will be extracted is ‘WP_123890’.

Replace comma with _\$:

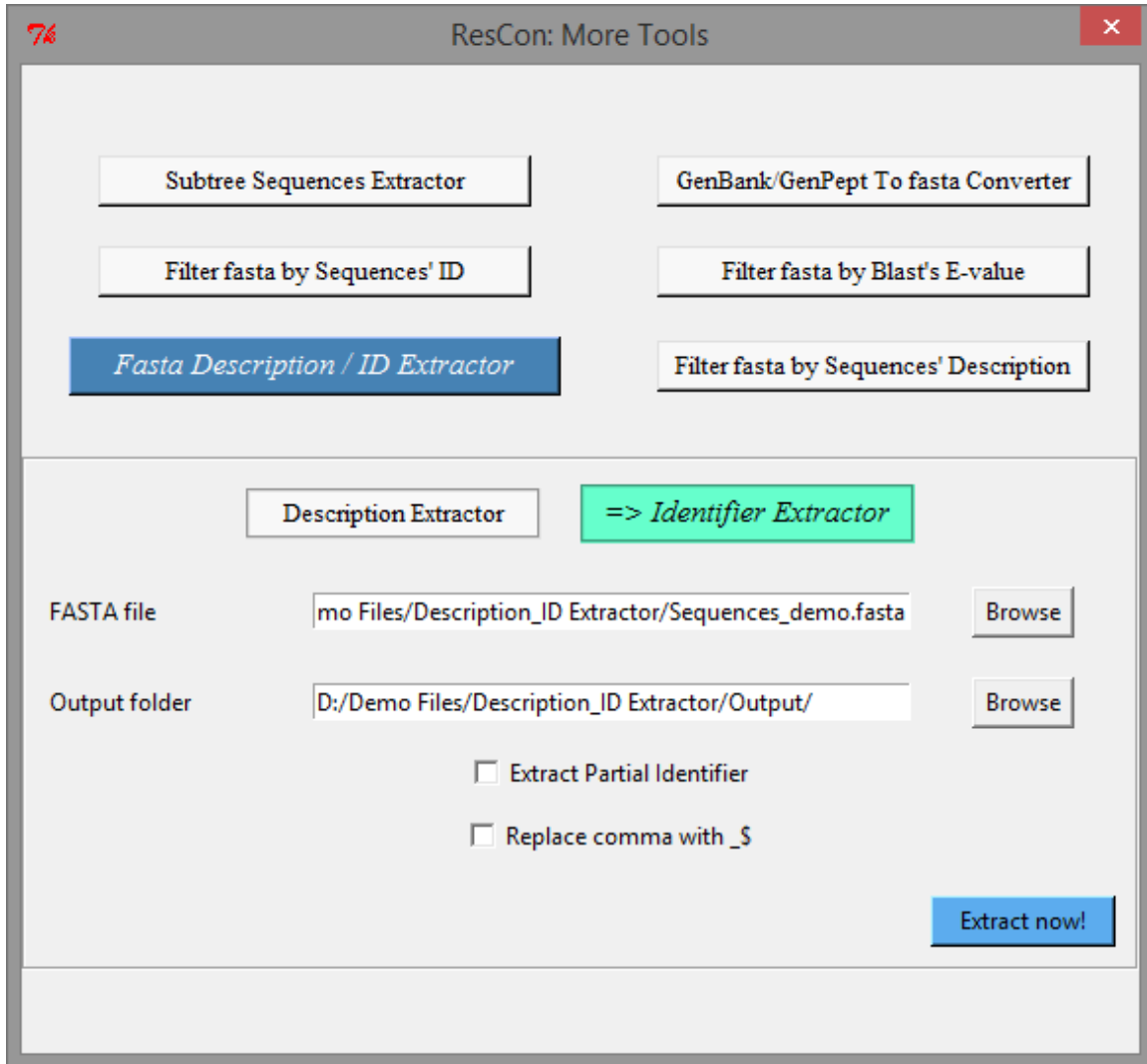
Check this box if you would like to replace comma in sequence descriptions with symbols ‘_’\$. This will make your life easier if descriptions have comma symbol in them but you do not want them to be split in multiple cells in output csv (comma separated values) file. If needed, you may later replace _\$ in the output file with a comma symbol using ‘find and replace’ feature in the spreadsheet software.



Now, you can choose which symbol(s) to use to replace comma. Go to File -> Edit Settings

Demo 1 – Extracting complete Identifiers:

Use demo file from folder ‘Demo Files\6. Description_ID Extractor\’ to get a feel of how the tool ‘ResCon – Fasta Description/ID Extractor’ works. Here, we will extract complete identifiers from a fasta file.



Click ‘Extract now’ after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder. The output csv file will show the list of complete identifiers and will also show their count.

Demo 2 – Extracting Partial Descriptions:

Use demo file from folder ‘Demo Files\6. Description_ID Extractor\’ to get a feel of how the tool ‘ResCon – Fasta Description / ID Extractor’ works. Here, we will extract partial descriptions from a fasta file by using the default regular expression. Also, we will choose to replace commas in the description with ‘_’.

The screenshot shows the 'ResCon: More Tools' window. It contains several buttons for different tools: 'Subtree Sequences Extractor', 'GenBank/GenPept To fasta Converter', 'Filter fasta by Sequences' ID', 'Filter fasta by Blast's E-value', 'Fasta Description / ID Extractor' (highlighted in blue), and 'Filter fasta by Sequences' Description'. Below these is a section with two buttons: '=> Description Extractor' (highlighted in green) and 'Identifier Extractor'. The 'Description Extractor' section includes fields for 'FASTA file' (containing 'mo Files/Description_ID Extractor/Sequences_demo.fasta'), 'Output folder' (containing 'D:/Demo Files/Description_ID Extractor/Output/'), and 'Regular Expression' (containing '(.+) \[.+' and highlighted in yellow). There are 'Browse' buttons for the file and folder fields. Below the regular expression field are two checked checkboxes: 'Extract Partial Description' and 'Replace comma with _\$'. An 'Extract now!' button is located at the bottom right of the configuration section.

Click ‘Extract now’ after filling in the required fields as shown in above figure. When the job is done, files will be written in to the selected output folder. The output csv file will show the list of partial descriptions and will also show their count.

Appendix A: Understanding fasta format

Complete understanding of fasta format will enable easy understanding and proper use of ResCon. Fasta sequence has two sections: a header line and one or multiple lines of sequence data. More than one sequence records may be present in fasta file.

| 1 | 2 Identifier | 3 Description |
|---|--|---------------|
| | <pre>>WP_0202604i Bacteria Escherichia coli inorganic pyrophosphatase MSLLNVPAGKDLPEDIYVVIEIPANADPIKYEIDKESGALFVDRFMSTAMFYP CNYGYINHTLSLDGDPVDVLVPTPYPLQPGSVIRCRPVGVLKMTDEAGEDAKL IAVAVPHSKLSKEYDHIKDVNDLPELLKAQIAHFFEHYKDLE</pre> | |

First line is the header line and it has three parts.

1. First character in header line is always “>” (greater than symbol). This signifies the beginning of a new sequence record. This symbol “>” should always be present for each sequence record for a valid fasta format.
2. The section that follows “>” until the first empty space is called as *identifier* or *ID*. Obviously identifier cannot have empty space in them. In the figure above, red highlighted part is the identifier. This part is optional in fasta format. However, this is the part showing id of sequences and hence *identifier* is necessary to use ResCon.

Certain programs may limit the length of identifier to certain number of characters. For example, Clustal omega allows up to 127 characters and hence ResCon recommends maximum length of identifier to 127.

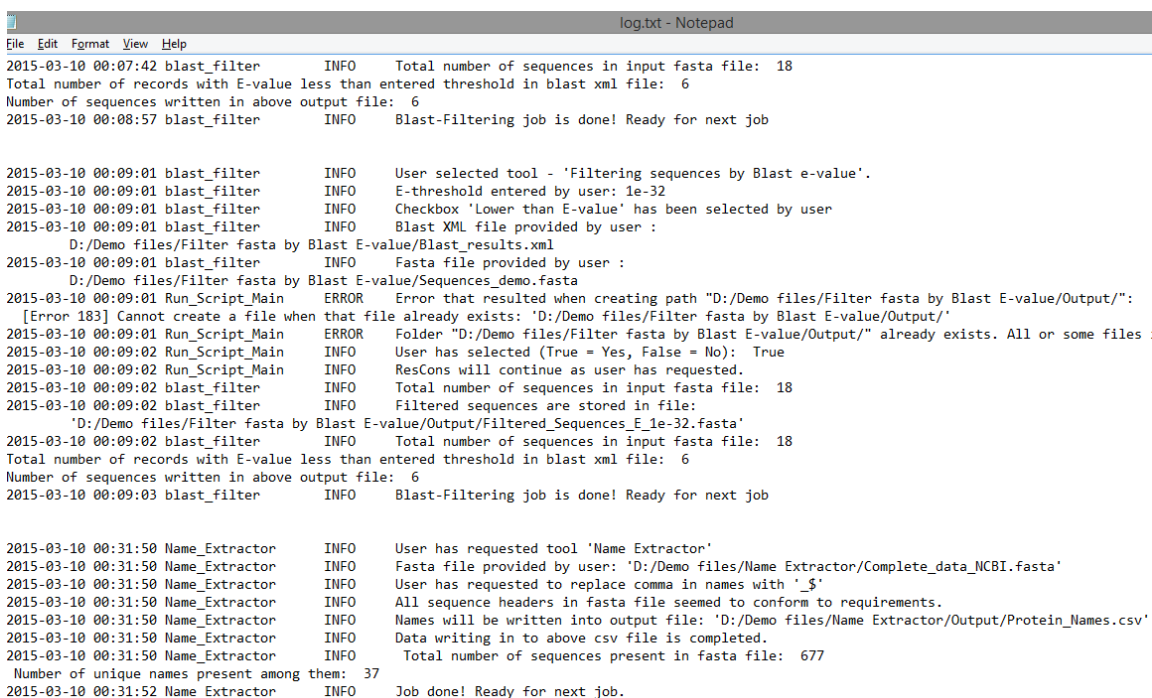
3. Rest of the header that follows identifier is called *description*. This part is optional and space character is allowed here.

Appendix B: Log file

Now 'Native Log file' can be accessed easily: File menu -> Open Log file

ResCon writes log files along with output files and they serve two purposes:

- 1) It serves as a notebook (history) storing all the input details you have used.
- 2) It tracks the processes of ResCon and also stores any error that may come up.



```
log.txt - Notepad
File Edit Format View Help
2015-03-10 00:07:42 blast_filter INFO Total number of sequences in input fasta file: 18
Total number of records with E-value less than entered threshold in blast xml file: 6
Number of sequences written in above output file: 6
2015-03-10 00:08:57 blast_filter INFO Blast-Filtering job is done! Ready for next job

2015-03-10 00:09:01 blast_filter INFO User selected tool - 'Filtering sequences by Blast e-value'.
2015-03-10 00:09:01 blast_filter INFO E-threshold entered by user: 1e-32
2015-03-10 00:09:01 blast_filter INFO Checkbox 'Lower than E-value' has been selected by user
2015-03-10 00:09:01 blast_filter INFO Blast XML file provided by user :
D:/Demo files/Filter fasta by Blast E-value/Blast_results.xml
2015-03-10 00:09:01 blast_filter INFO Fasta file provided by user :
D:/Demo files/Filter fasta by Blast E-value/Sequences_demo.fasta
2015-03-10 00:09:01 Run_Script_Main ERROR Error that resulted when creating path "D:/Demo files/Filter fasta by Blast E-value/Output/":
[Error 183] Cannot create a file when that file already exists: 'D:/Demo files/Filter fasta by Blast E-value/Output/'
2015-03-10 00:09:01 Run_Script_Main ERROR Folder "D:/Demo files/Filter fasta by Blast E-value/Output/" already exists. All or some files :
2015-03-10 00:09:02 Run_Script_Main INFO User has selected (True = Yes, False = No): True
2015-03-10 00:09:02 Run_Script_Main INFO ResCons will continue as user has requested.
2015-03-10 00:09:02 blast_filter INFO Total number of sequences in input fasta file: 18
2015-03-10 00:09:02 blast_filter INFO Filtered sequences are stored in file:
'D:/Demo files/Filter fasta by Blast E-value/Output/Filtered_Sequences_E_1e-32.fasta'
2015-03-10 00:09:02 blast_filter INFO Total number of sequences in input fasta file: 18
Total number of records with E-value less than entered threshold in blast xml file: 6
Number of sequences written in above output file: 6
2015-03-10 00:09:03 blast_filter INFO Blast-Filtering job is done! Ready for next job

2015-03-10 00:31:50 Name_Extractor INFO User has requested tool 'Name Extractor'
2015-03-10 00:31:50 Name_Extractor INFO Fasta file provided by user: 'D:/Demo files/Name Extractor/Complete_data_NCBI.fasta'
2015-03-10 00:31:50 Name_Extractor INFO User has requested to replace comma in names with '_'
2015-03-10 00:31:50 Name_Extractor INFO All sequence headers in fasta file seemed to conform to requirements.
2015-03-10 00:31:50 Name_Extractor INFO Names will be written into output file: 'D:/Demo files/Name Extractor/Output/Protein_Names.csv'
2015-03-10 00:31:50 Name_Extractor INFO Data writing in to above csv file is completed.
2015-03-10 00:31:50 Name_Extractor INFO Total number of sequences present in fasta file: 677
Number of unique names present among them: 37
2015-03-10 00:31:52 Name_Extractor INFO Job done! Ready for next job.
```

In the screenshot above showing a log file, you will notice multiple blocks and each block corresponds to a process request. Final block of data in the log file corresponds to the most recent process request. *So, in short, to find logs for a particular process request, open log file from corresponding output folder, scroll all the way down and the final block of logs is what you are looking for.*

If you happen to run into any error (hopefully you will not!) using ResCon, you will want to see 'native log file' to know the nature of error. Location of this native log file in your computer differs depending on your operating system. See the instructions below to find this log file:

Now 'Native Log file' can be accessed easily: File menu -> Open Log file

For windows:

Located at 'C:\temp' under title 'Logs_ResCon.txt'. If your computer does not have C drive, look for temp folder in any other drive it has.

For Mac:

Right click on ResCon.app and click 'Show Package Contents'. You will find log file titled "Logs_ResCon.txt" under 'Contents\Resources\'.

For Ubuntu:

Log file titled "Logs_ResCon.txt" is present in the same folder that contains ResCon executable program.

Appendix C: Clustal Omega Parameters

```
# Sequence Input
(["-i", "--in", "--infile", "infile"],
        "Multiple sequence input file",
        filename=True,
        equate=False),
(["--hmm-in", "HMM input", "hmm_input"],
        "HMM input files",
        filename=True,
        equate=False),
_Switch(["--dealign", "dealign"],
        "Dealign input sequences"),
(["--profile1", "--p1", "profile1"],
        "Pre-aligned multiple sequence file (aligned columns will be kept fix).",
        filename=True,
        equate=False),
(["--profile2", "--p2", "profile2"],
        "Pre-aligned multiple sequence file (aligned columns will be kept fix).",
        filename=True,
        equate=False),
(["-t", "--seqtype", "seqtype"],
        "{Protein, RNA, DNA} Force a sequence type (default: auto).",
        equate=False,
        checker_function=lambda x: x in ["protein", "rna", "dna",
                                         "Protein", "RNA", "DNA", "PROTEIN"]),
_Switch(["--is-profile", "isprofile"],
        "disable check if profile, force profile (default no)",
        equate=False),
(["--infmt", "infmt"],
        """"Forced sequence input file format (default: auto)
Allowed values: a2m, fa[sta], clu[stal], msf, phy[lip], selex,
st[ockholm], vie[nna]""",
        equate=False,
        checker_function=lambda x: x in ["a2m", "fa", "fasta",
                                         "clu", "clustal",
                                         "msf",
                                         "phy", "phylip",
                                         "selex",
                                         "st", "stockholm", "vie", "vienna"]]),
```

```

# Clustering


```

```

# Alignment Output


```

```

        checker_function=lambda x: isinstance(x, int)),

# Limits (will exit early, if exceeded):


```

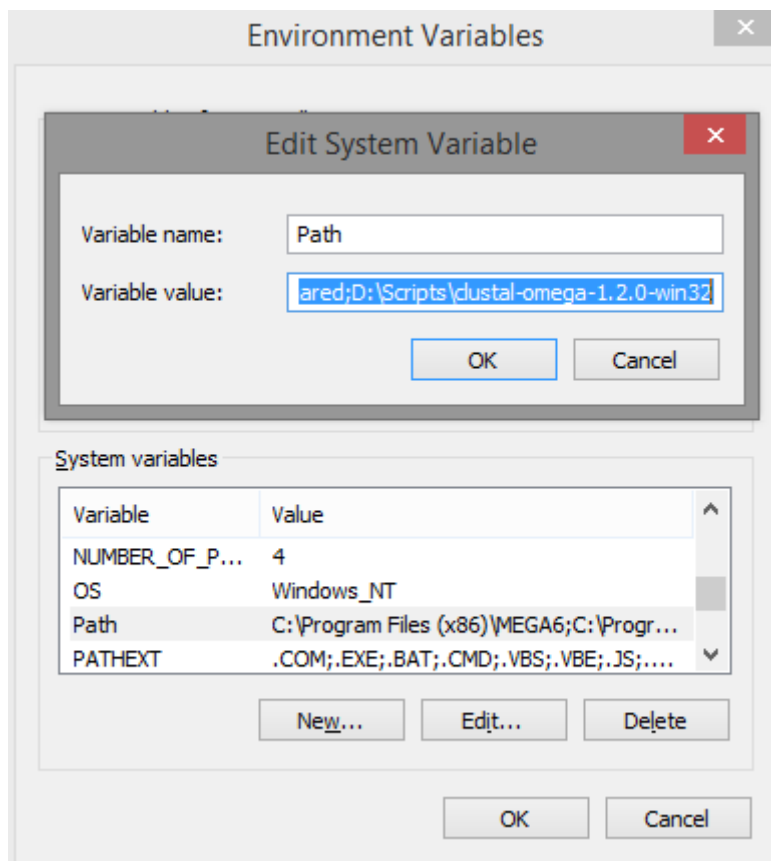
Appendix D: Installing Clustal Omega

Clustal omega is available for download from <http://www.clustal.org/omega/>. Here we provide basic guide on how to get Clustal Omega working on your computer. If this doesn't work for you, you may have to refer to installation instructions available from Clustal Omega website.

For Windows:

Download precompiled binary file for Windows from above link and then extract and save it anywhere in your computer. Next you have to set PATH so that Windows will always recognize where Clustal Omega is installed (stored). Follow the instructions below to set path (you will need administrator password). If this doesn't work for your windows version, google 'how to set path in windows (your version)'.

1. Open 'My Computer' or 'This PC'.
Right click on an empty space and click 'Properties'.
2. Click 'Advanced system settings' and then 'Environment variables'.
3. Select 'Path' under System variables.
4. Now click 'Edit' and add file path of the Clustal omega folder as shown in the image here. Use semi-colon to separate it from other file paths.
5. That's it. Now open command prompt, type 'clustalo' and press Enter. It should say something like this - 'FATAL: No sequence input was provided'. This means installation has been successful.



For UNIX based OS (Mac & Ubuntu):

You may download and use pre-compiled standalone binary file available for your OS or compile from the source code available from <http://www.clustal.org/omega/>. Either way, chances are you need to install 'argtable2' to install Clustal omega successfully. Follow these steps:

1. Download 'argtable2' source code (file name: argtable2-13.tar.gz, at the time of writing) from <http://argtable.sourceforge.net/>.
2. Extract or decompress the downloaded file.
3. Open 'Terminal' and execute these commands.

```
cd argtable2-13      (this is to take you inside the extracted argtable2
                     folder. If this doesn't work, type 'cd ' and drag and
                     drop extracted argtable2 folder into the terminal)
```

```
./configure
```

```
make
```

```
make check          (this step is optional)
```

```
sudo make install   (you will need password)
```

```
make clean
```

4. Now 'argtable2' is installed and next is to install Clustal Omega. Follow these steps to install using its source code.
5. Download file 'Source code .tar.gz (1.2.1)' from <http://www.clustal.org/omega/> and then extract it.
6. Open Terminal and change directory to extracted folder using 'cd' command. (See step 3 for more details on this).
7. Once you are inside the extracted folder (or directory), execute these commands:

```
./configure
```

```
make
```

```
sudo make install   (you will need password)
```

```
make clean
```


8. Now open a new Terminal, type 'clustalo' and press Enter. It should say something like this - 'FATAL: No sequence input was provided'. This means installation has been successful.

If you rather install using precompiled standalone library, download the file for your OS under section called precompiled binaries from this link - <http://www.clustal.org/omega/#Download>. Move this downloaded file to location of your choice and then rename it to 'clustalo'. Next, use following command in Terminal to make it executable:

```
chmod u+x clustalo
```

Above command assumes you are already in the directory/folder that contains clustalo file. If you don't know how to do this, type 'chmod u+x ' in the terminal and then drag and drop 'clustalo' file to that terminal and then press Enter. Installation should be successful now. Follow step 8 to verify Clustal Omega behaves properly.

Alternate method for Ubuntu:

Since Clustal Omega is available through Ubuntu's Advanced Packaging Tool (APT), it might be just easy to install it using apt-get. This has added advantage of installing dependancies (including argtable2) automatically. In such case, use the following command:

```
sudo apt-get install clustalo
```

முற்றும்

(The End)