CEPiNS (Conserved Exon Prediction in Novel Species) - ***User Manual/Tutorial***

This project has a homepage which can be accessed through the following URL:

http://cepins.org/

The homepage contains:

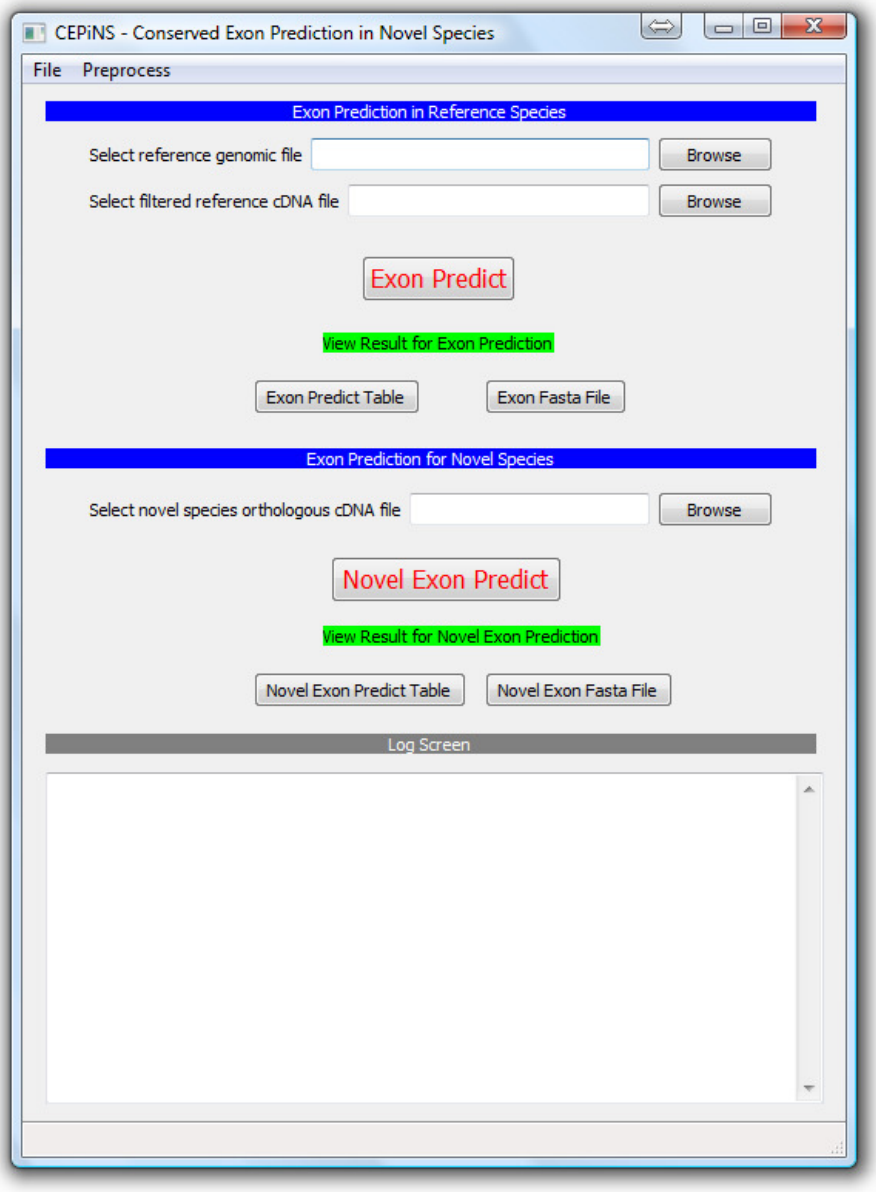
1. Latest version of CEPiNS (Installer which run all versions of Windows)
2. User Manual
3. Example dataset
4. Source code

Additional Information:

This program does not work if your computer has BioEdit installed.

How to use CEPiNS:

After running CEPiNS the user interface is as follows:

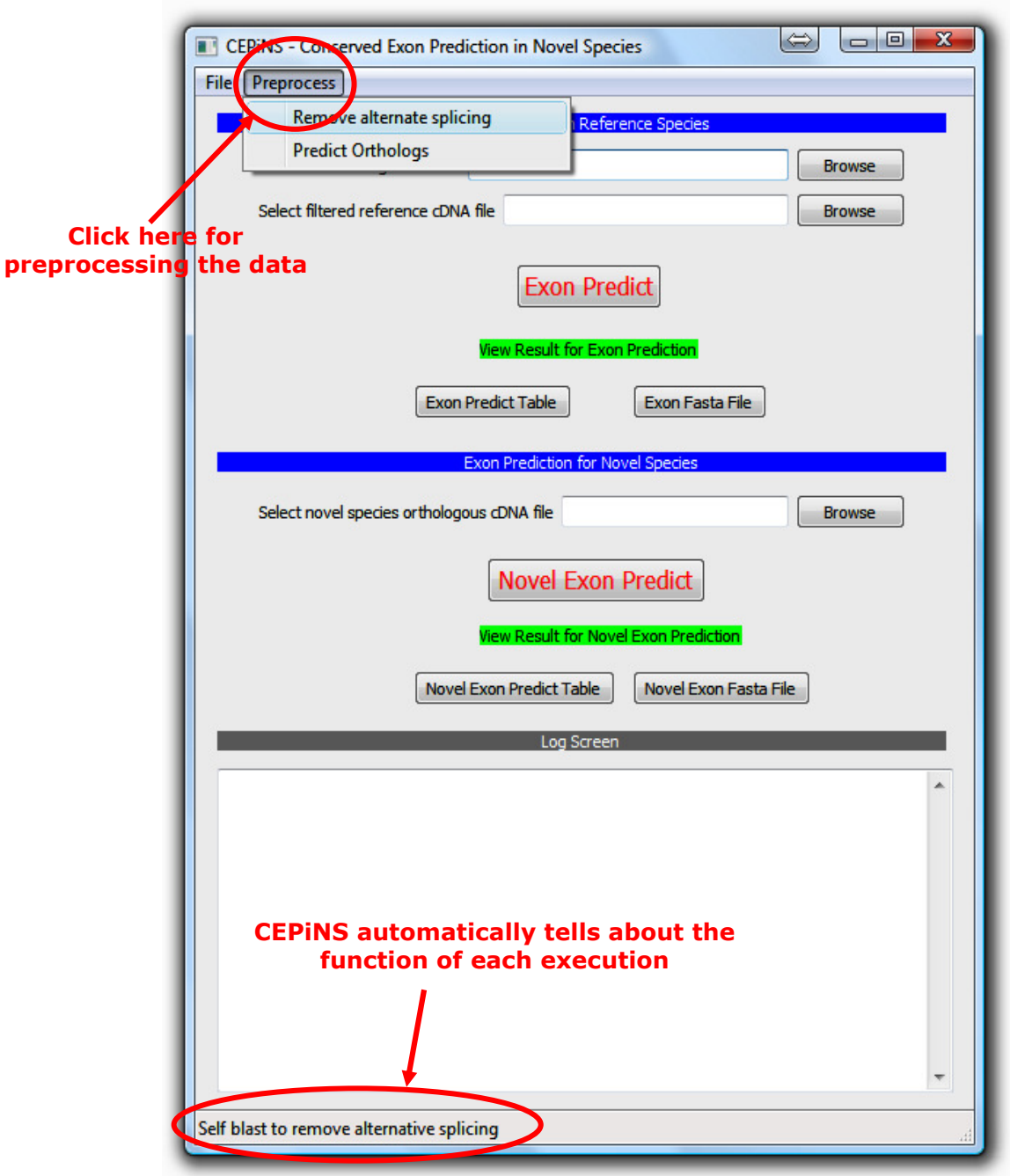


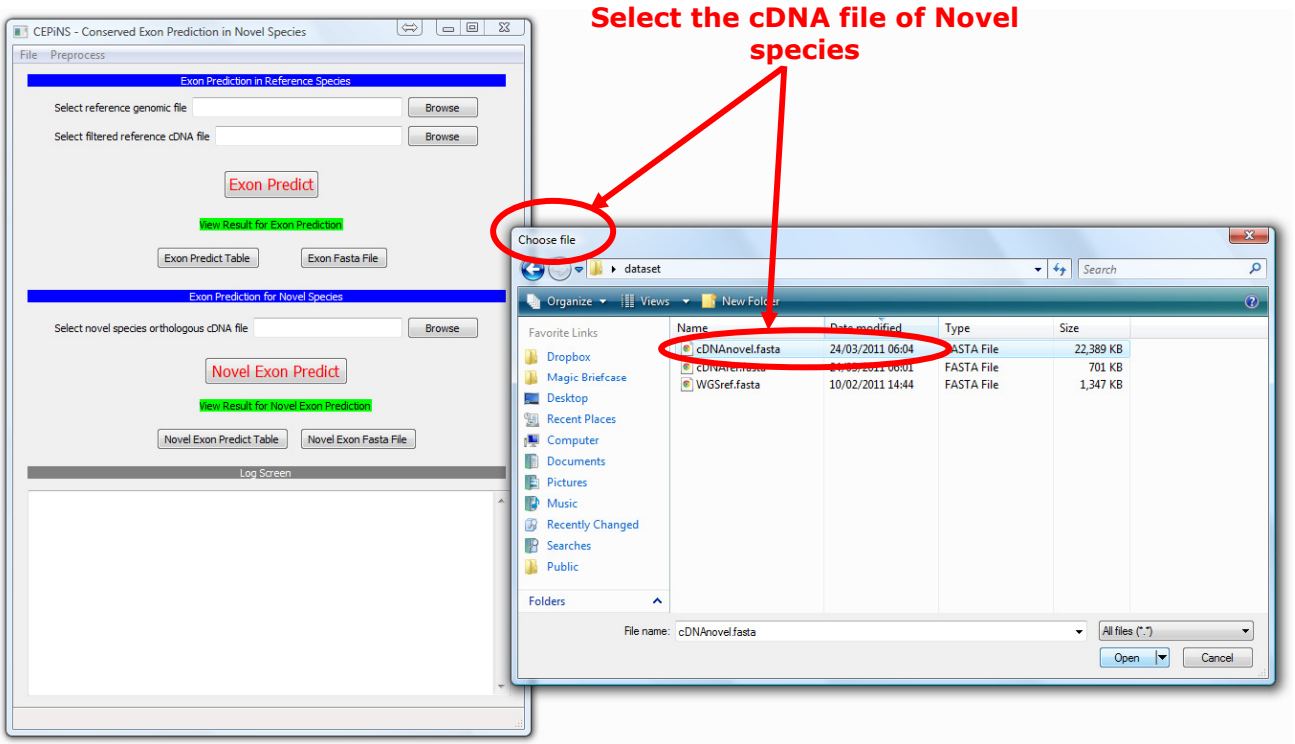
CEPiNS has 3 different parts:
1. Data filtering
    A. Remove alternate splicing
    B. Predict Orthologs
2. Exon Prediction in Reference Species
3. Exon Prediction in Novel Species

1. Data filtering

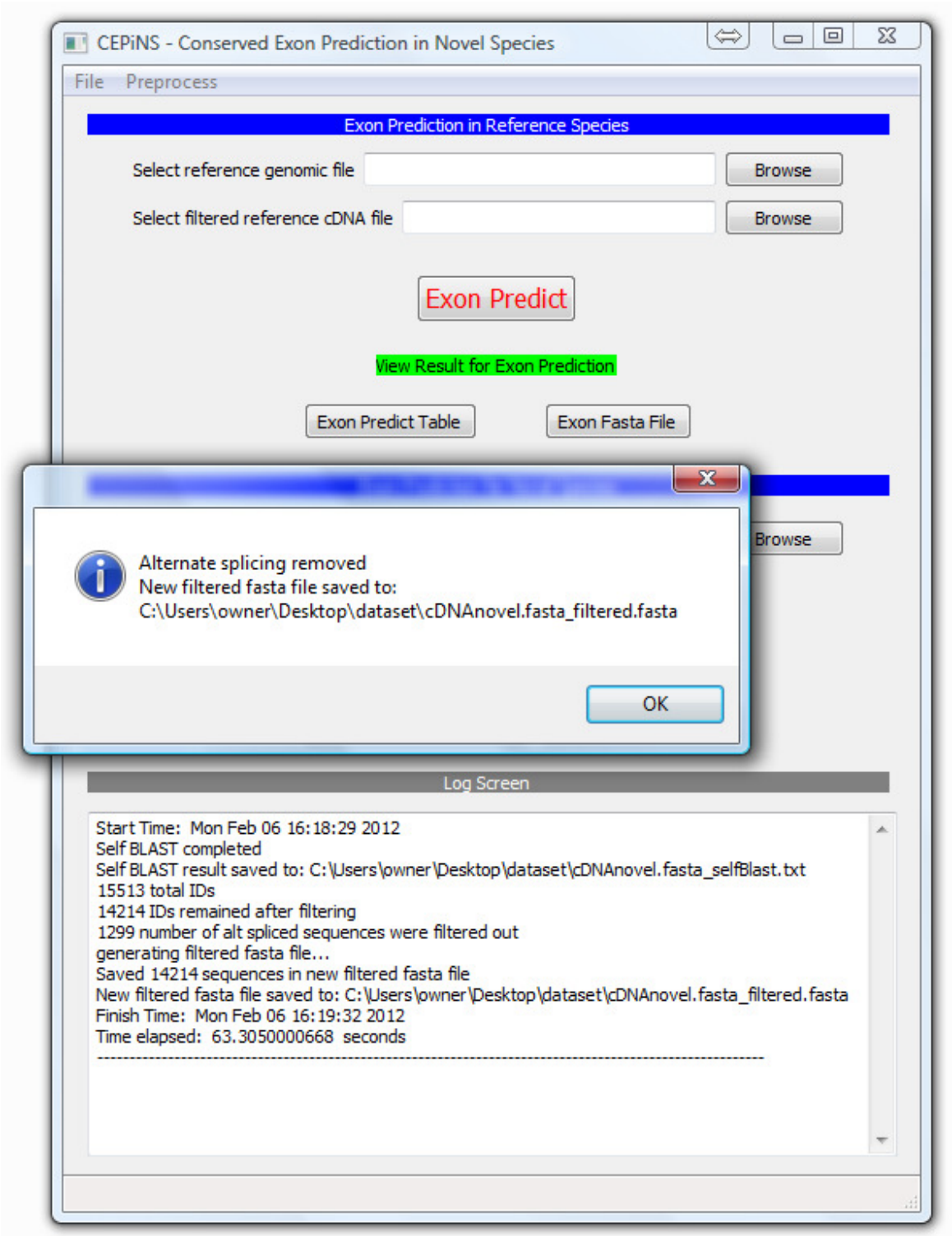
A. Remove alternate splicing:

I. Click on 'Preprocess' tab from menu bar
II. Select 'Remove Alternate Splicing'
III. Select cDNA fasta file
IV. Click on 'Select'

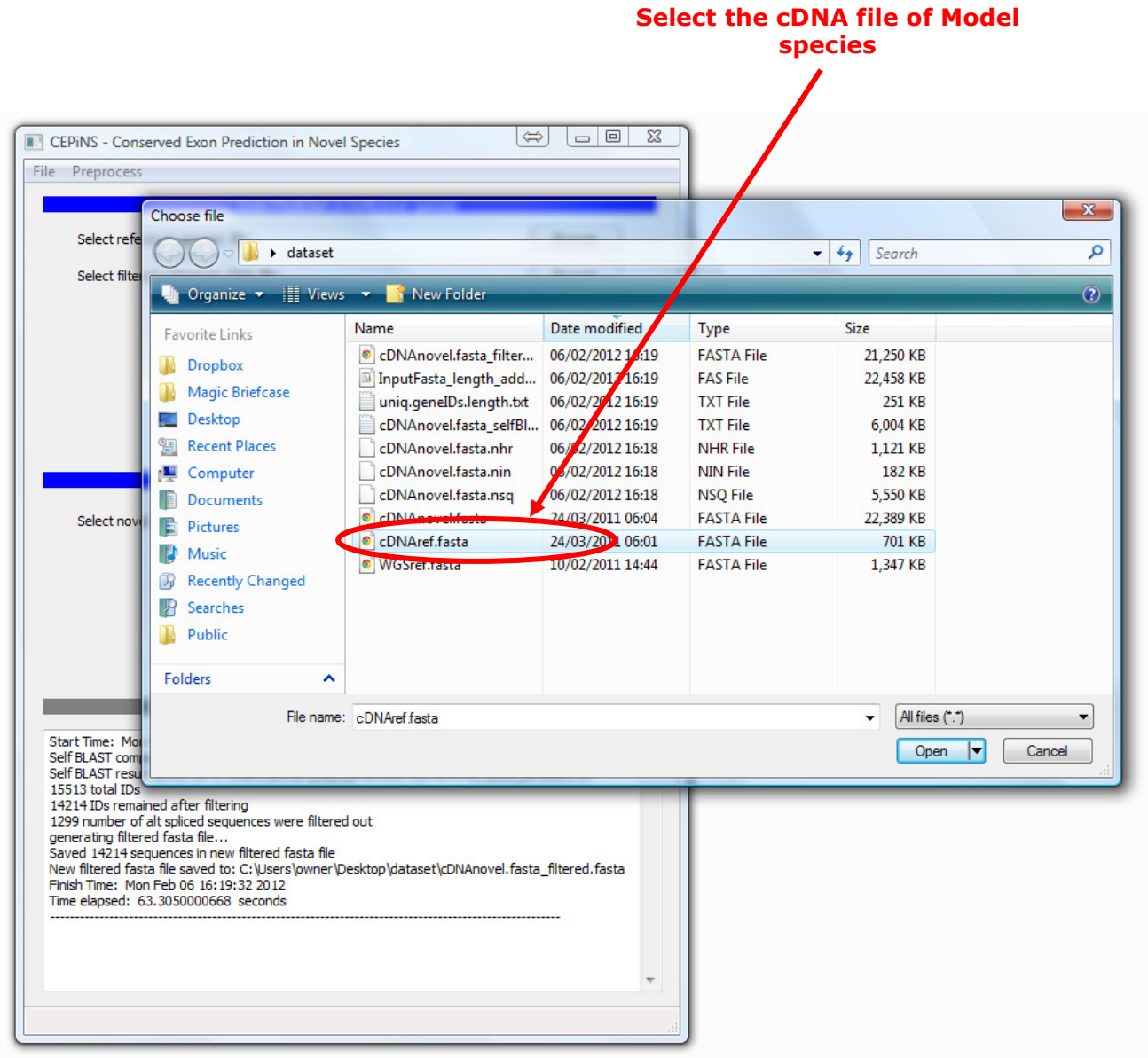


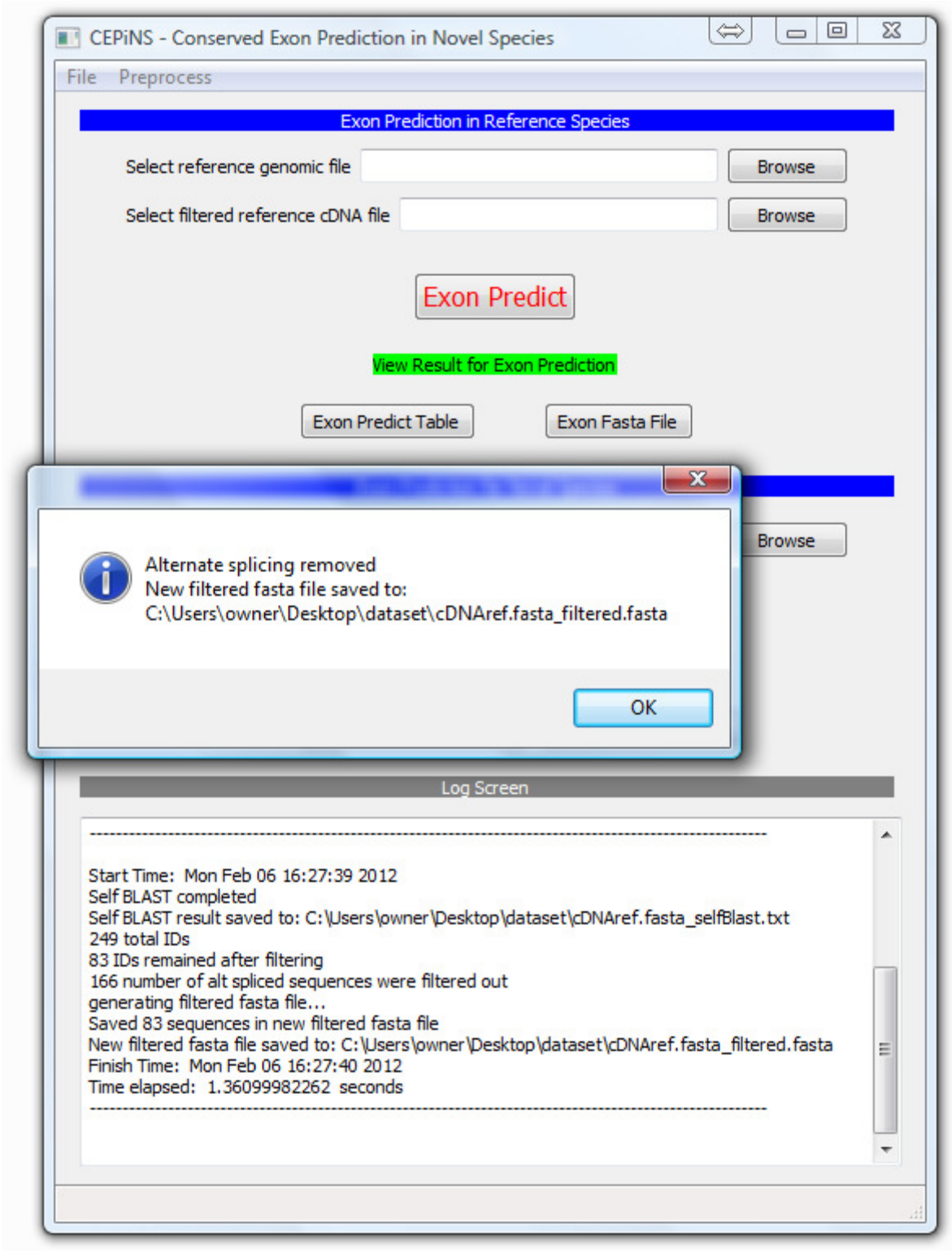


After the execution a new filtered fasta file will be created. The new file name and the saved location will be appeared in a pop up box. The default location is the original dataset location. In the 'Log Screen', all the details will be appeared including the execution time. The 'Log Screen' data can be saved in a text file by clicking 'Save log…' tab from the 'File' menu bar. This log data can be used for keeping records for project's workflow and publication.


CEPiNS - Conserved Exon Prediction in Novel Species
File
Preprocess
Exon Prediction in Reference Species
Select reference genomic file
Browse
Select filtered reference cDNA file
Browse
Exon Predict
View Result for Exon Prediction
Exon Predict Table
Exon Fasta File
Alternate splicing removed
New filtered fasta file saved to:
C:\Users\owner\Desktop\dataset\cDNAnovel.fasta_filtered.fasta
OK
Browse
Log Screen
Start Time: Mon Feb 06 16:18:29 2012
Self BLAST completed
Self BLAST result saved to: C:\Users\owner\Desktop\dataset\cDNAnovel.fasta_selfBlast.txt
15513 total IDs
14214 IDs remained after filtering
1299 number of alt spliced sequences were filtered out
generating filtered fasta file...
Saved 14214 sequences in new filtered fasta file
New filtered fasta file saved to: C:\Users\owner\Desktop\dataset\cDNAnovel.fasta_filtered.fasta
Finish Time: Mon Feb 06 16:19:32 2012
Time elapsed: 63.3050000668 seconds

Perform the same steps as above for cDNA/gene sequences fasta file of the reference species (model species).

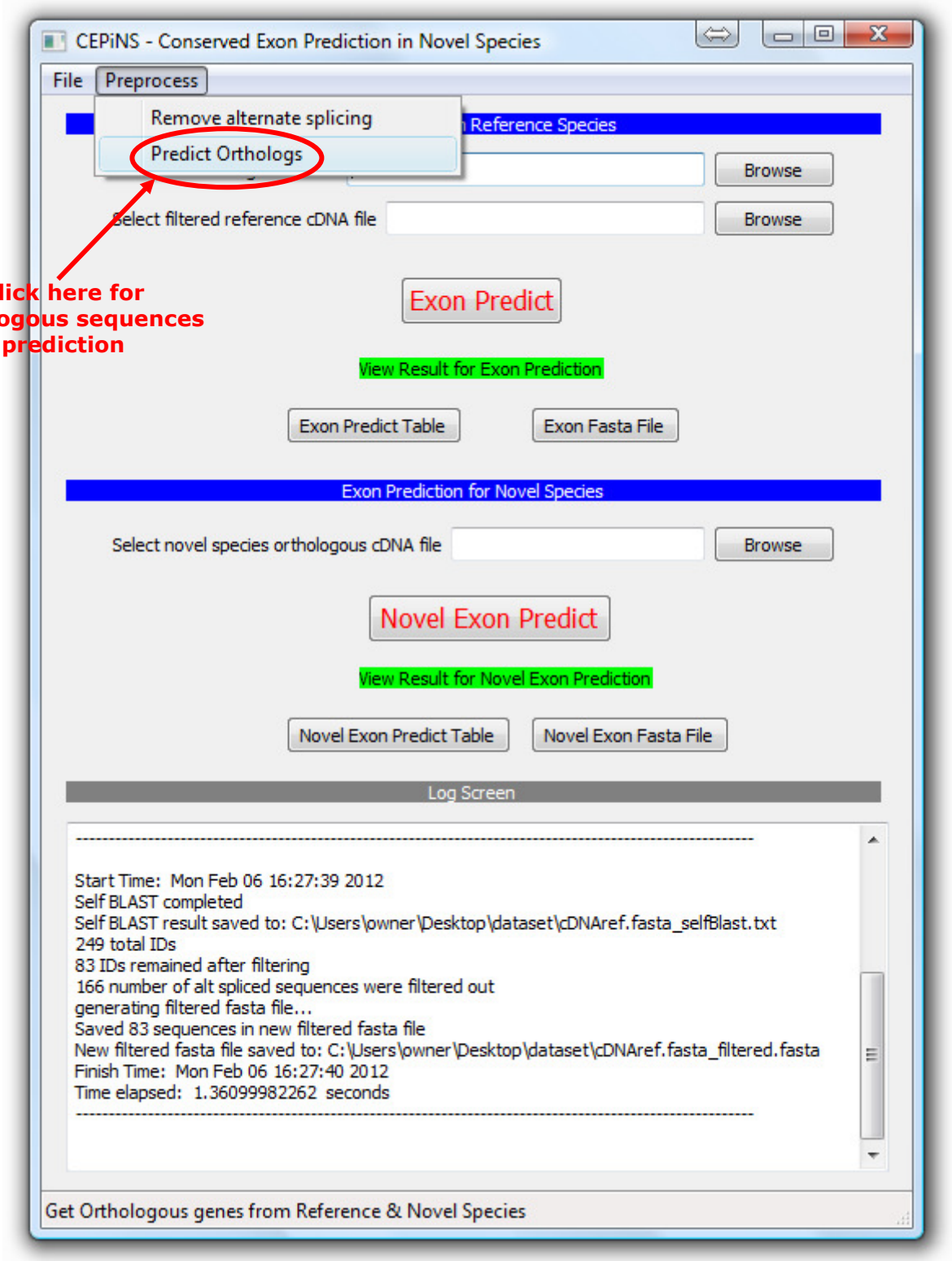




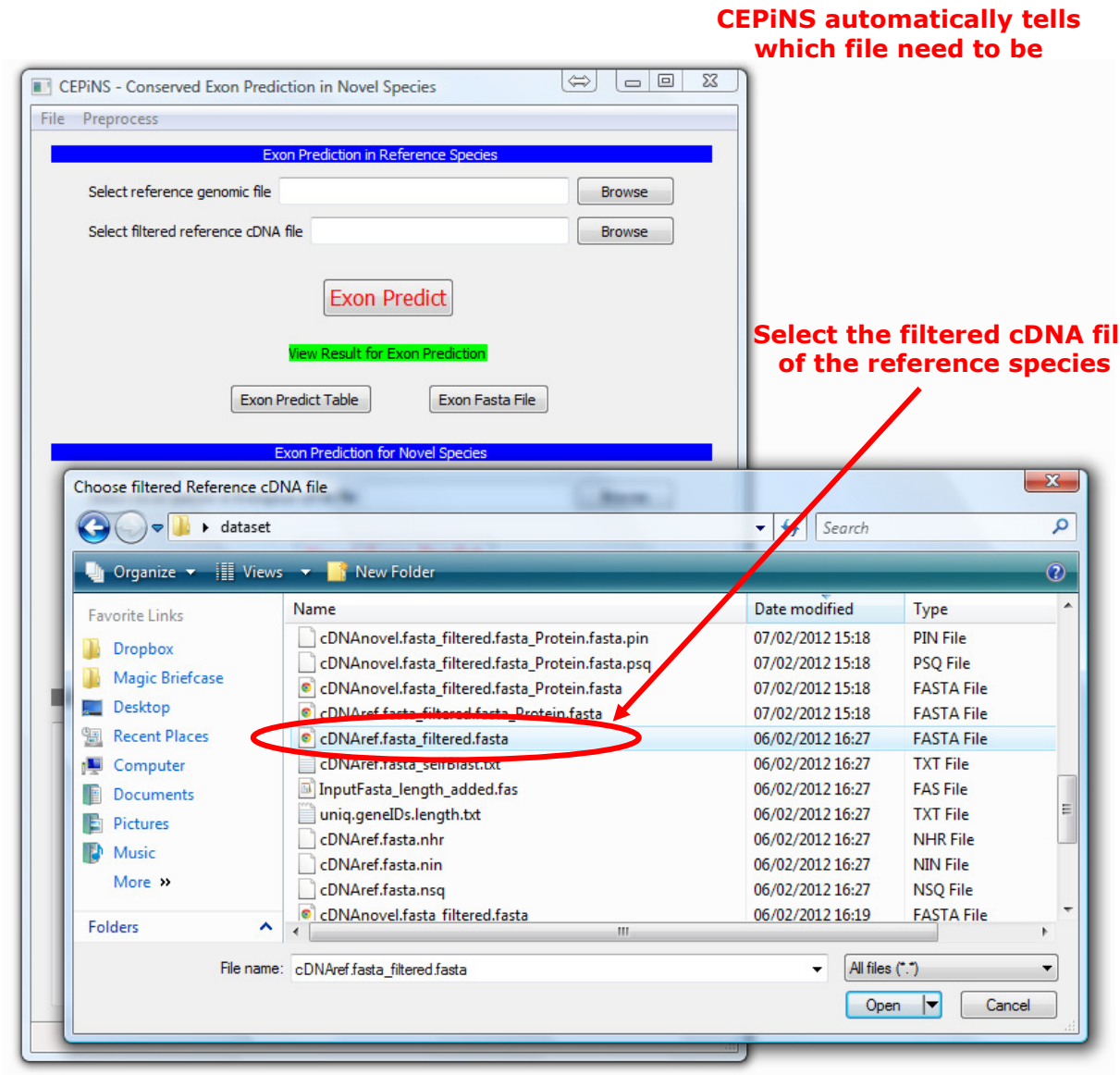
After this step, we will get unique cDNA/gene sequences for both Novel and Reference species where multiple gene copies have been removed. These two files will be used for getting the orthologous sequences.

B. Predict Orthologs:

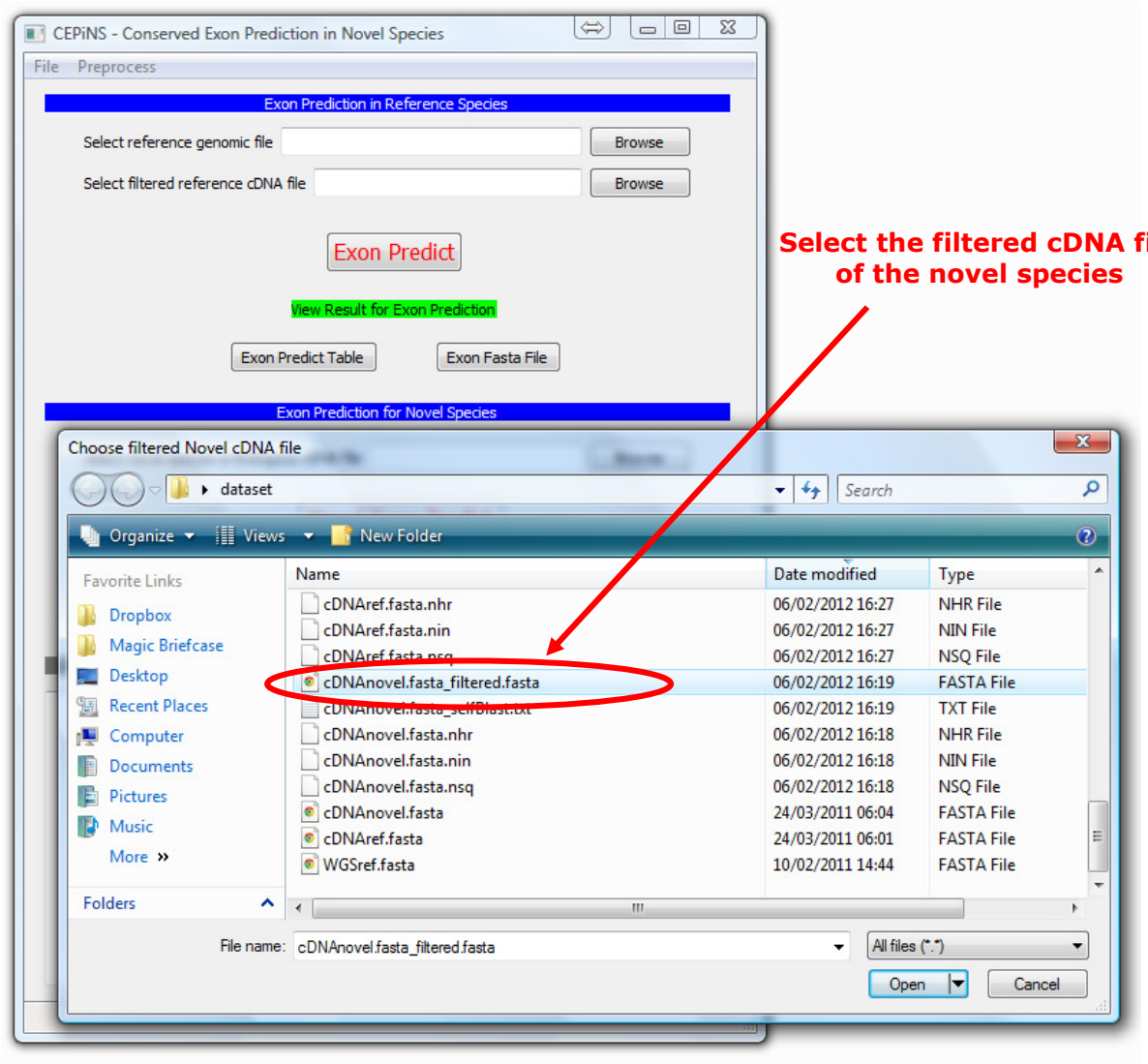
I. Click on 'Preprocess' tab from menu bar
II. Select 'Predict Orthologs'
III. Select filtered novel cDNA fasta file (cDNAnovel.fasta_filtered.fasta)
IV. Click on 'Select'
V. Select filtered reference cDNA fasta file (cDNAref.fasta_filtered.fasta)
VI. Click on 'Select'

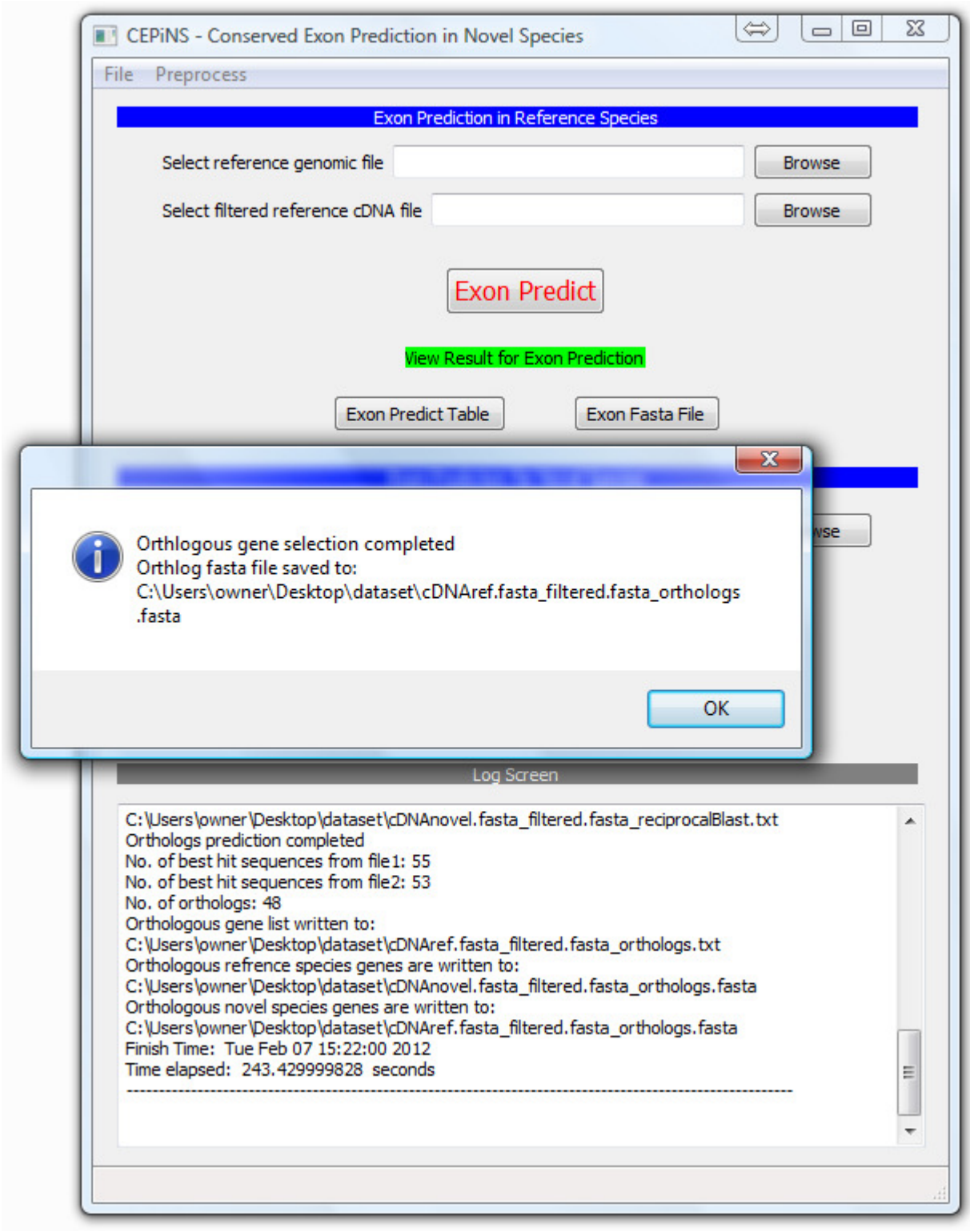


After selecting the 'Predict Orthologs' tab from 'Preprocess' menu bar CEPiNS will be directed to browse option where the filtered cDNA/gene fasta file of reference species is required to be selected.



CEPiNS will be directed to open the filtered cDNA/gene fasta file of novel species.





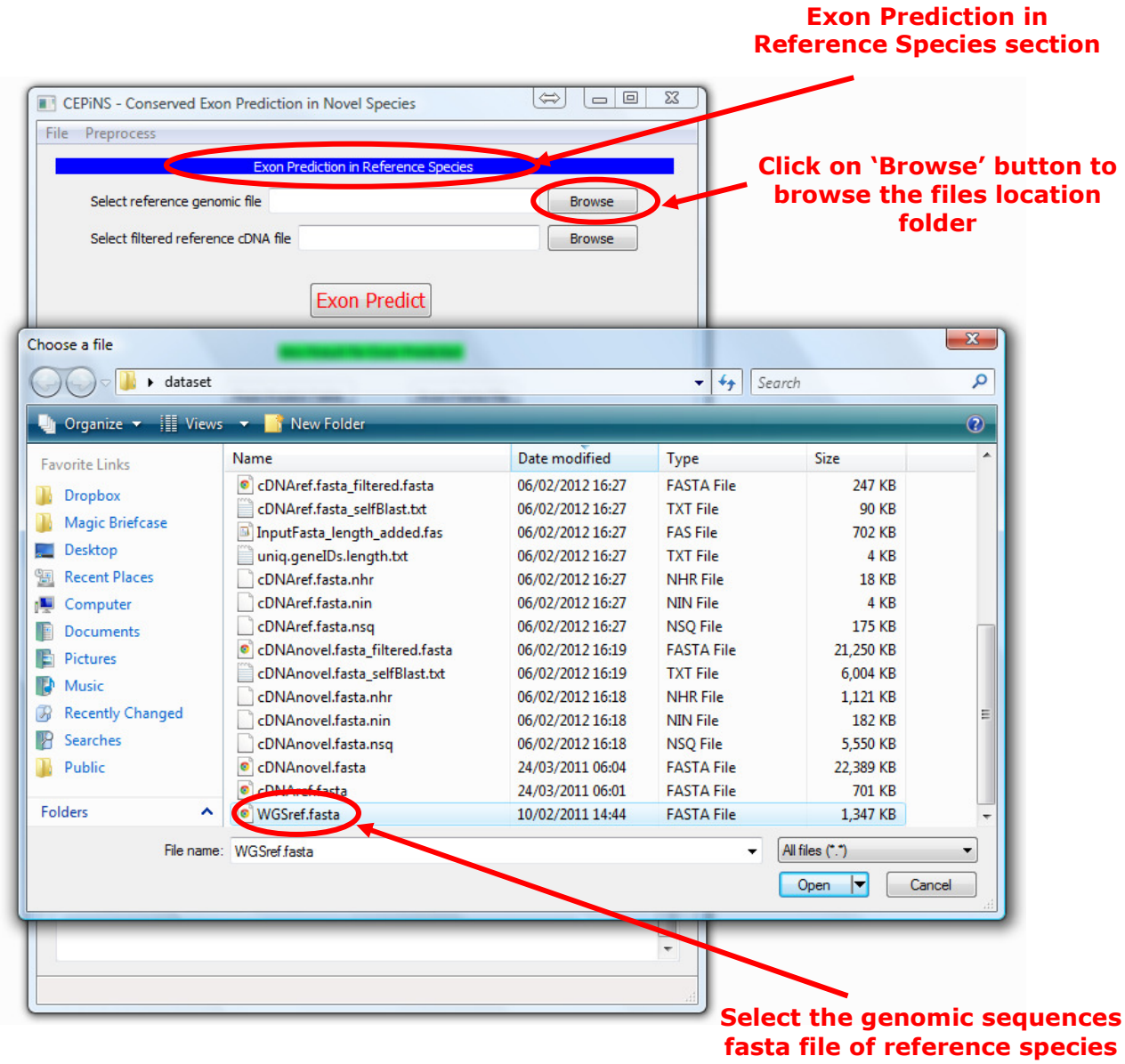
After this step, you will get the orthologous sequences for both reference and novel species. The ouput files name will be appended according to the original file names where '_orthologs.fasta' will be added at the end of original file names.

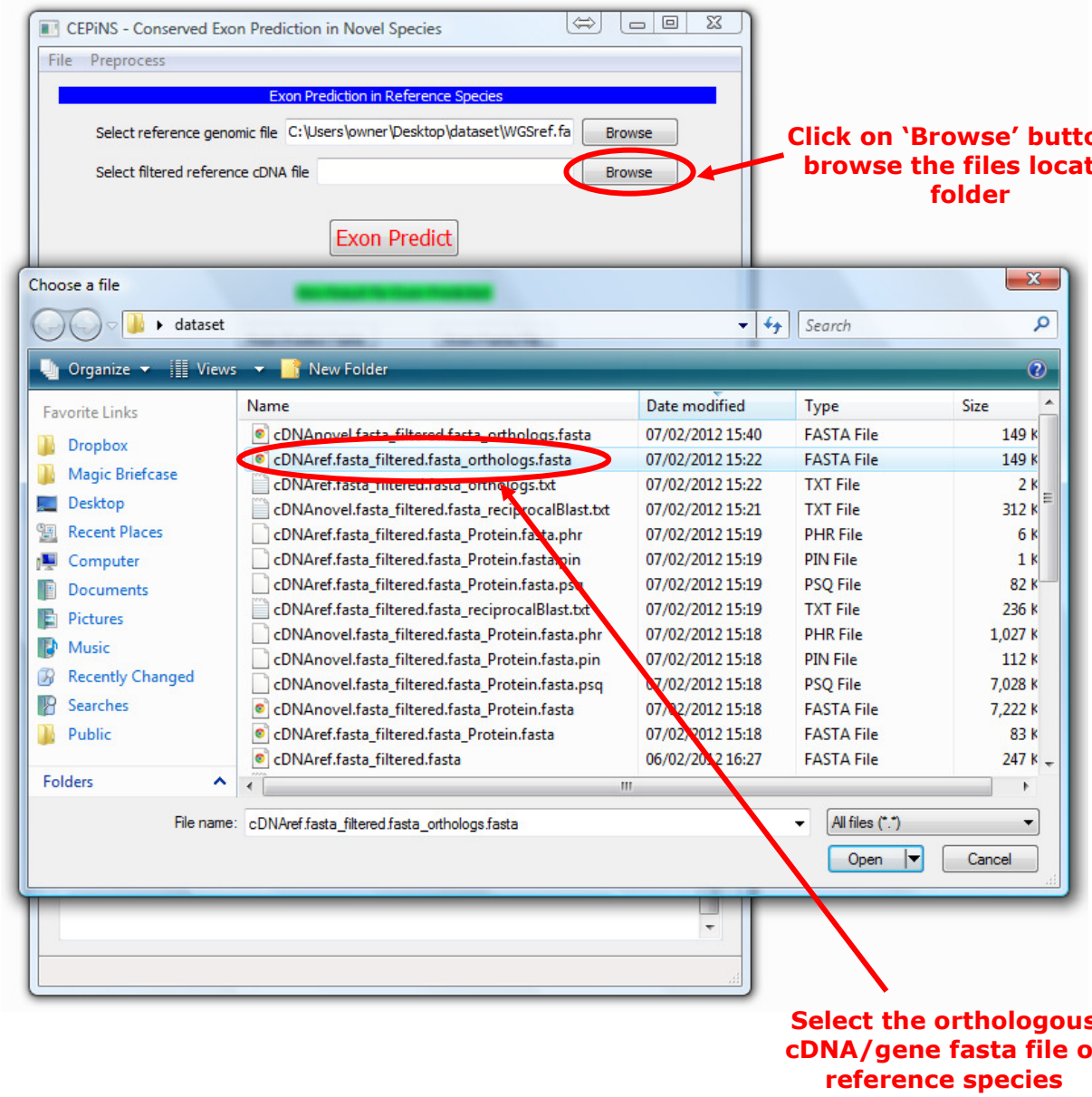
2. Exon Prediction in Reference Species:

I. Select reference genomic file ('WGSref.fasta')

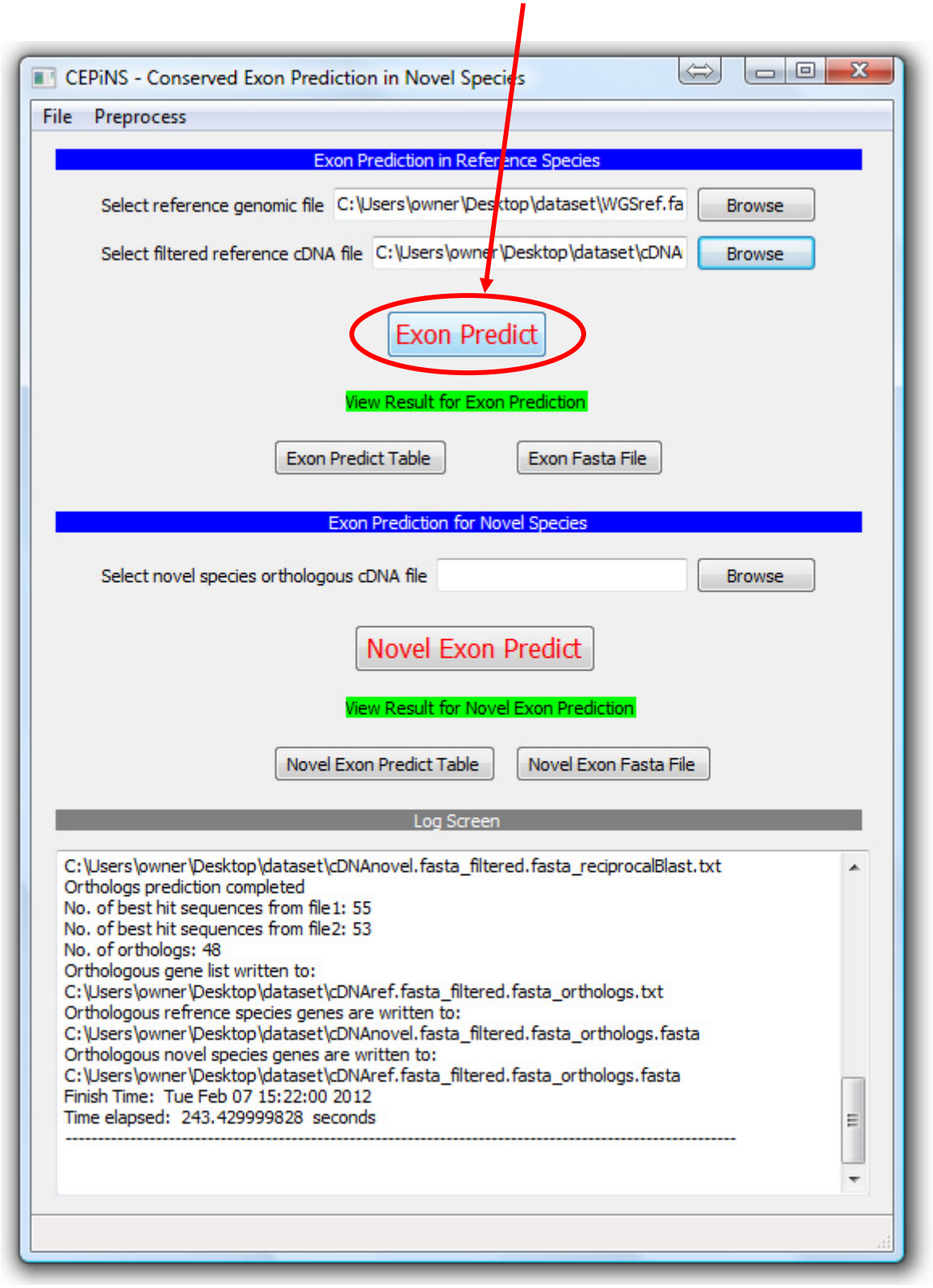
II. Select filtered reference cDNA file ('cDNAref.fasta_filtered.fasta' or 'cDNAref.fasta_filtered.fasta_orthologs.fasta')

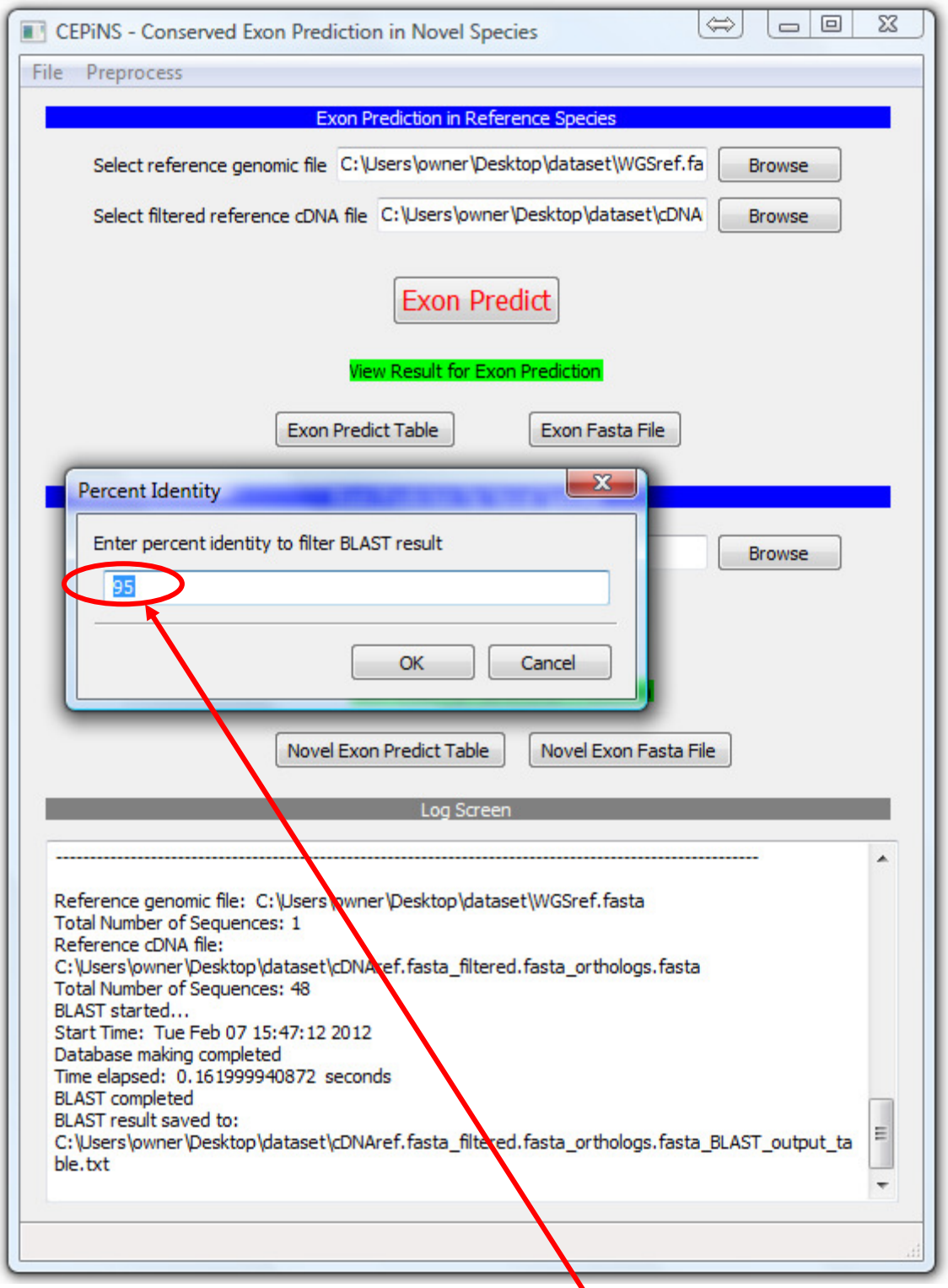
III. Click on 'Exon Predict' button



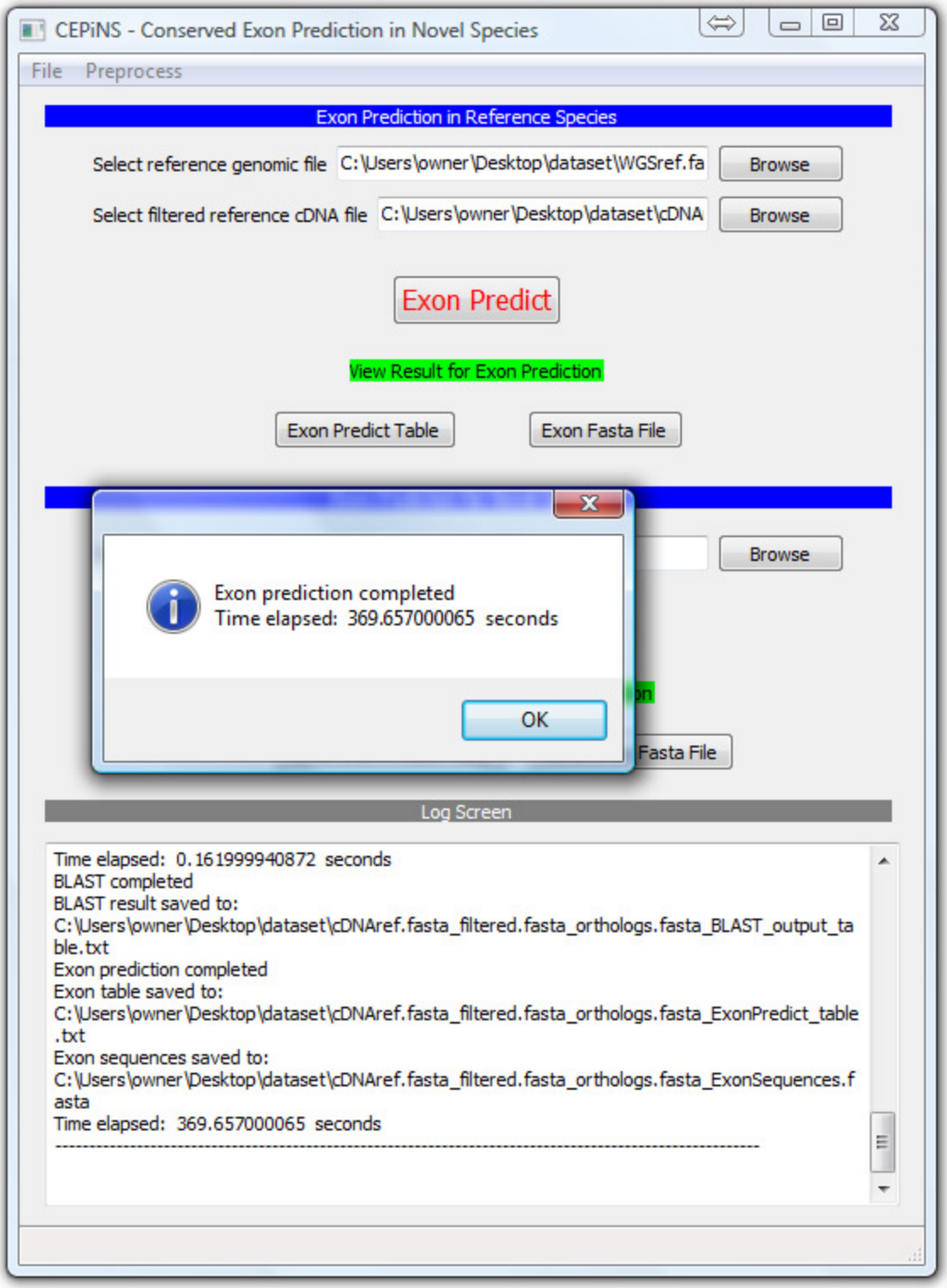

CEPiNS - Conserved Exon Prediction in Novel Species
File Preprocess
Exon Prediction in Reference Species
Select reference genomic file
C:\Users\owner\Desktop\dataset\WGSref.fa
Browse
Select filtered reference cDNA file
Browse
Exon Predict
Click on 'Browse' button to browse the files location folder
Choose a file
dataset
Search
Organize
Views
New Folder
Favorite Links
Dropbox
Magic Briefcase
Desktop
Recent Places
Computer
Documents
Pictures
Music
Recently Changed
Searches
Public
Folders
Name
Date modified
Type
Size
cDNAnovel.fasta_filtered.fasta_orthologs.fasta 07/02/2012 15:40 FASTA File 149 k
cDNAref.fasta_filtered.fasta_orthologs.fasta 07/02/2012 15:22 FASTA File 149 k
cDNAref.fasta_filtered.fasta_orthologs.txt 07/02/2012 15:22 TXT File 2 k
cDNAnovel.fasta_filtered.fasta_reciprocalBlast.txt 07/02/2012 15:21 TXT File 312 k
cDNAref.fasta_filtered.fasta_Protein.fasta.phr 07/02/2012 15:19 PHR File 6 k
cDNAref.fasta_filtered.fasta_Protein.fasta.pin 07/02/2012 15:19 PIN File 1 k
cDNAref.fasta_filtered.fasta_Protein.fasta.psq 07/02/2012 15:19 PSQ File 82 k
cDNAref.fasta_filtered.fasta_reciprocalBlast.txt 07/02/2012 15:19 TXT File 236 k
cDNAnovel.fasta_filtered.fasta_Protein.fasta.phr 07/02/2012 15:18 PHR File 1,027 k
cDNAnovel.fasta_filtered.fasta_Protein.fasta.pin 07/02/2012 15:18 PIN File 112 k
cDNAnovel.fasta_filtered.fasta_Protein.fasta.psq 07/02/2012 15:18 PSQ File 7,028 k
cDNAnovel.fasta_filtered.fasta_Protein.fasta 07/02/2012 15:18 FASTA File 7,222 k
cDNAref.fasta_filtered.fasta_Protein.fasta 07/02/2012 15:18 FASTA File 83 k
cDNAref.fasta_filtered.fasta 06/02/2012 16:27 FASTA File 247 k
File name: cDNAref.fasta_filtered.fasta_orthologs.fasta
All files (*.*)
Open
Cancel
Select the orthologous cDNA/gene fasta file of reference species

Click on 'Exon Predict' button



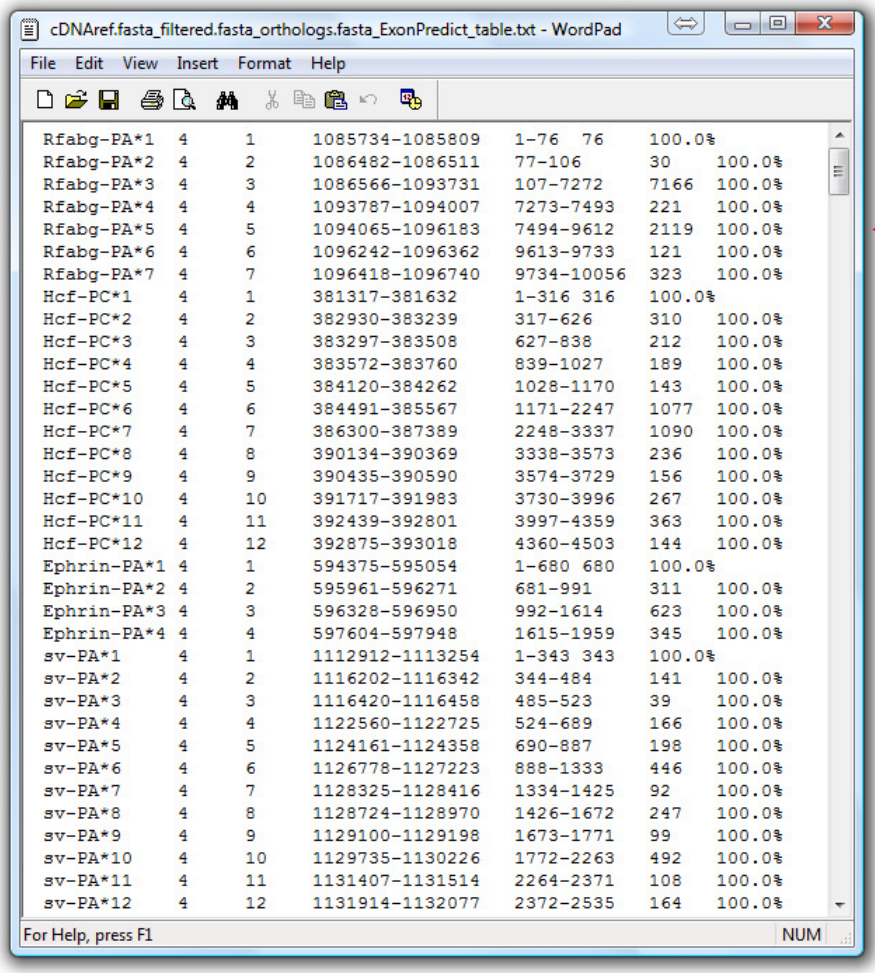


CEPiNS asks for the percent identity to filter BLAST result which is 95 by default. But user can change the value according to diversity of two species.

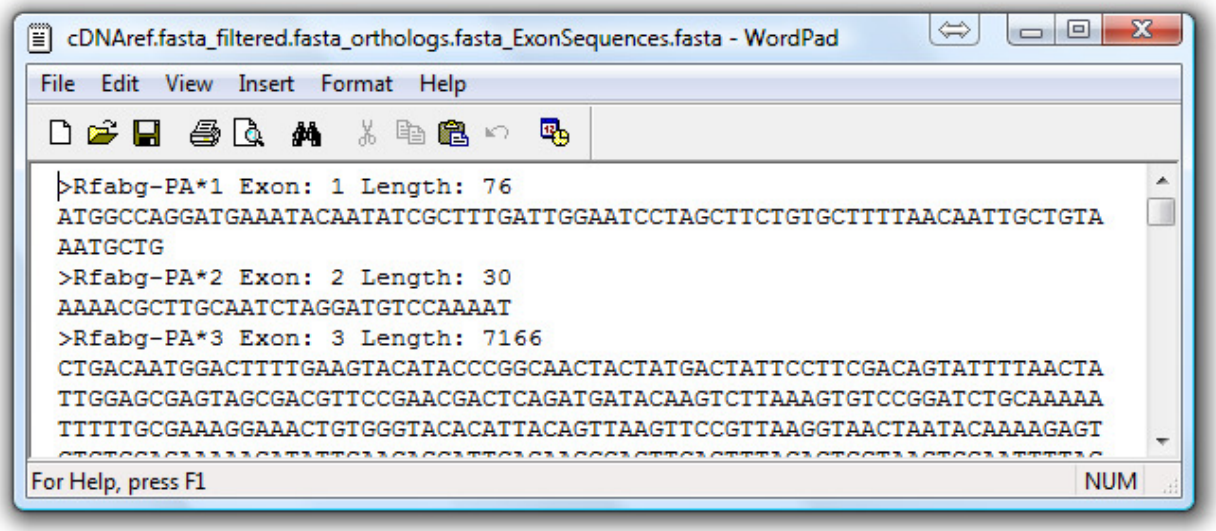


After this step, you will get the exon boundaries table text file and a fasta file with the exon sequences which can be opened by clicking 'Exon Predict Table' and 'Exon Fasta File' respectively.

In the exon boundaries table text file, the columns represent cDNA Sequence ID, Location, Exon number, Genomic region location, Location in cDNA, Exon length and Percent Identity respectively.



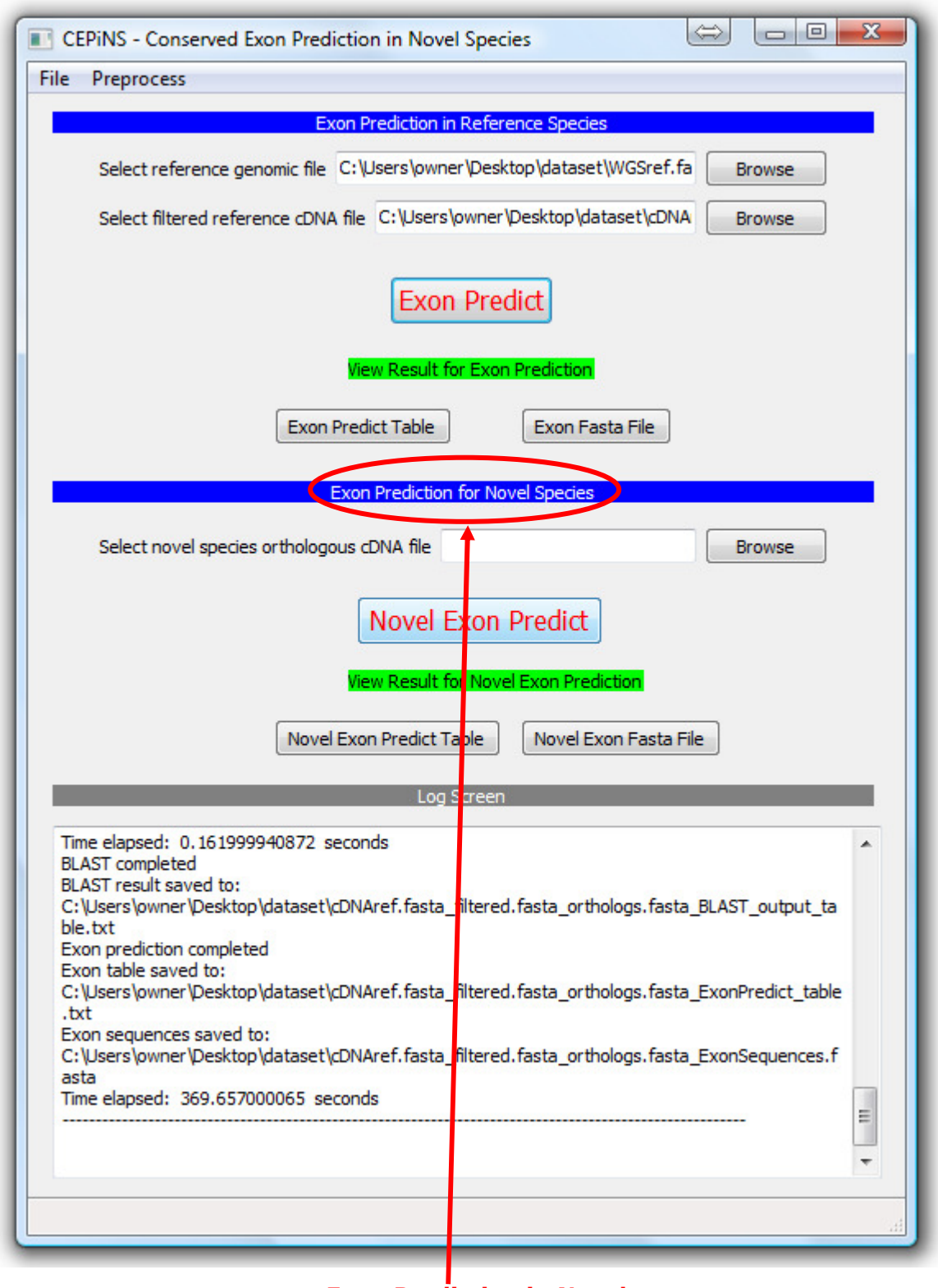
In the exon sequences fasta file, the header contains cDNA Sequence ID, Exon number and Length.



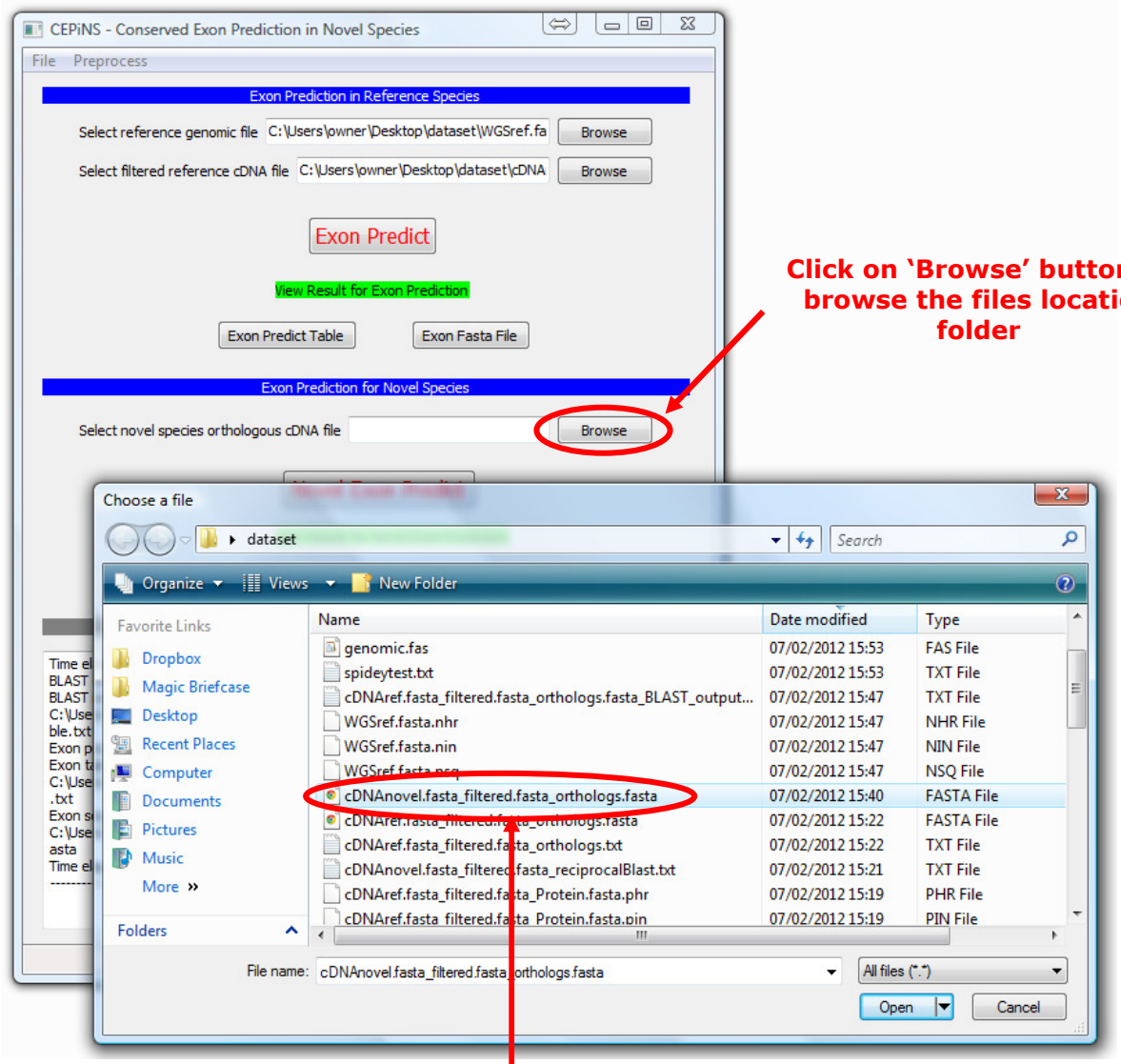
3. Exon Prediction in Novel Species:

I. Select novel species orthologous cDNA file (cDNAnovel.fasta_filtered.fasta_orthologs.fasta')

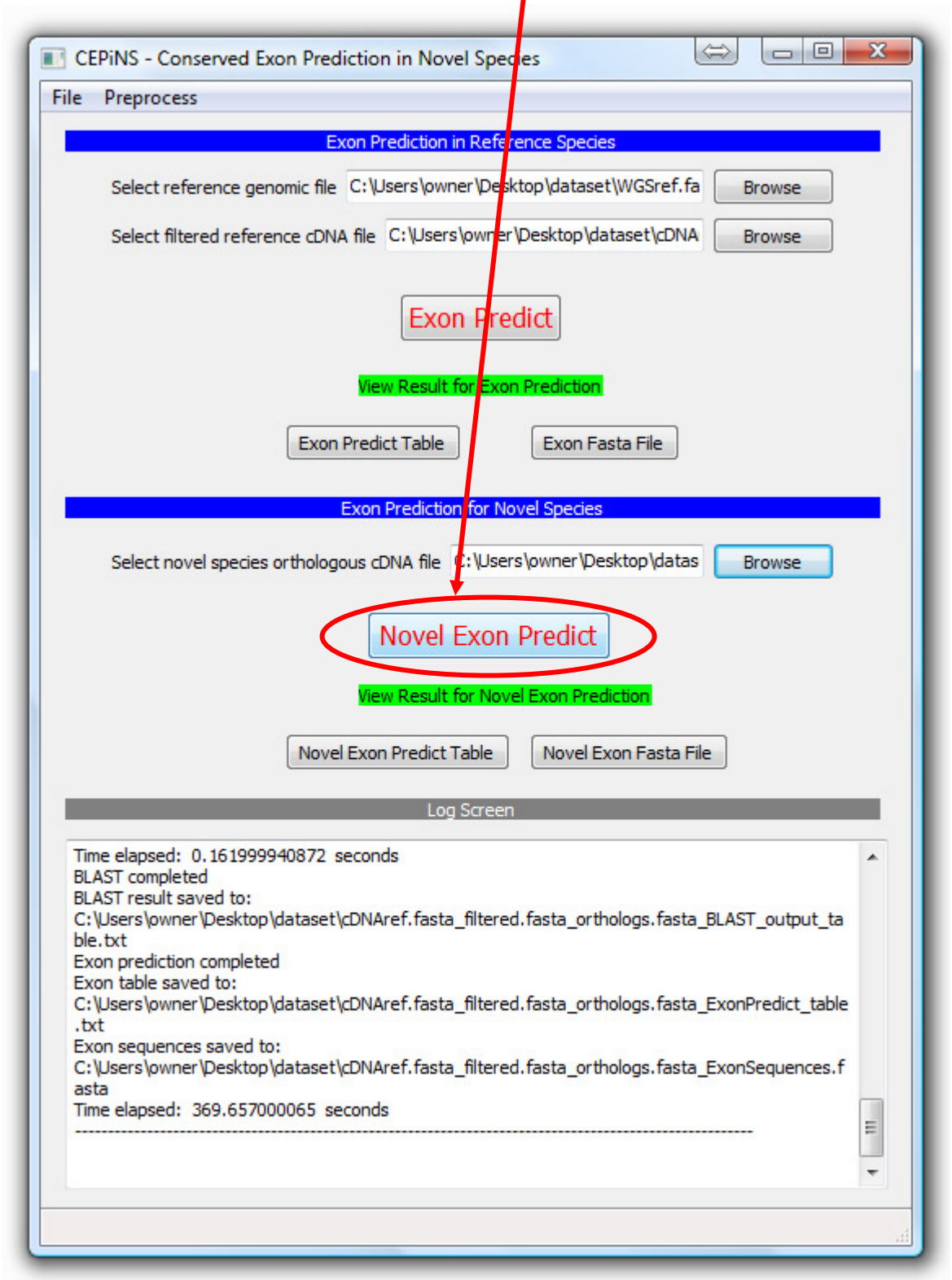
II. Click on 'Novel Exon Predict' button

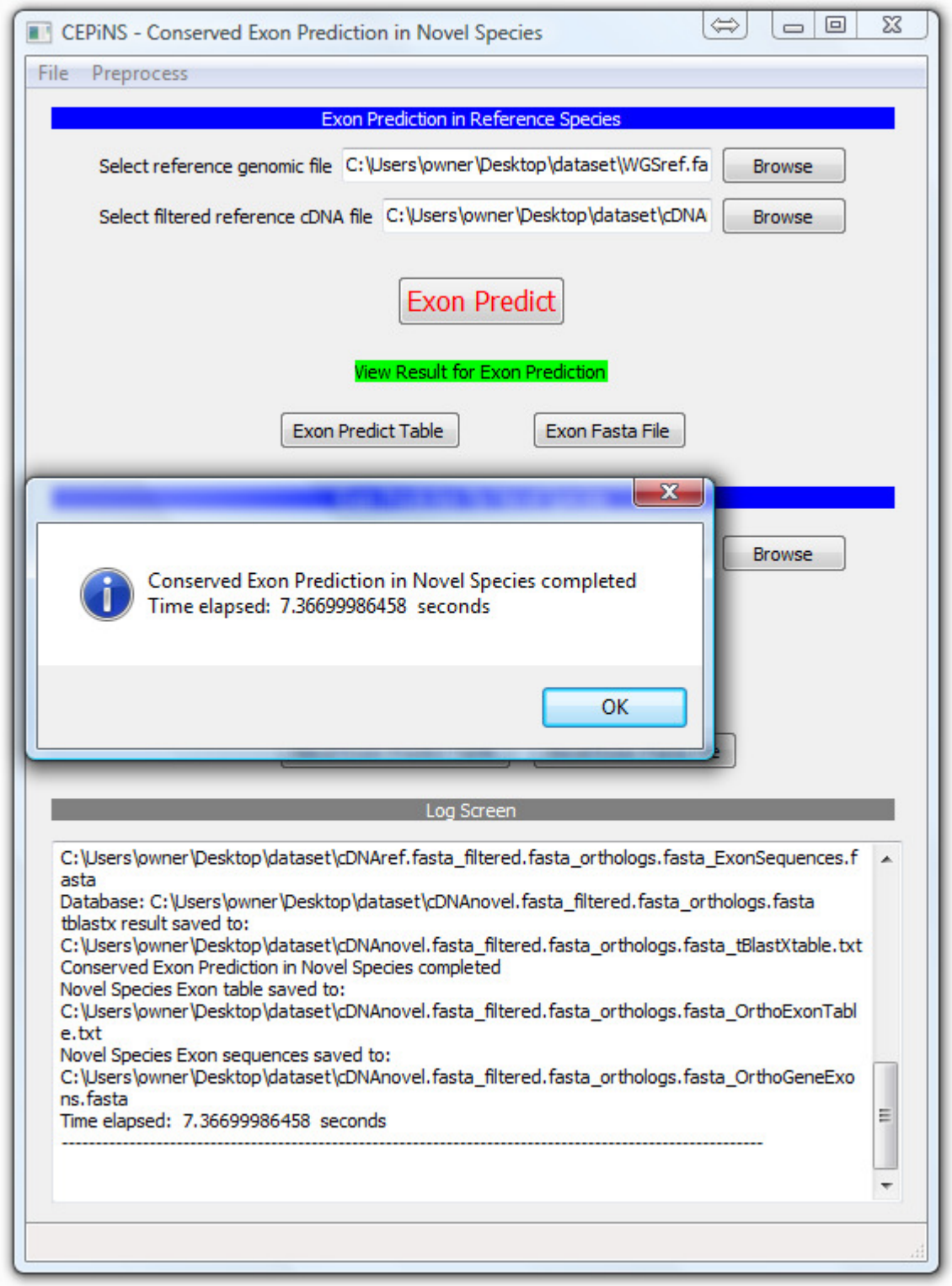


Exon Prediction in Novel Species section


CEPiNS - Conserved Exon Prediction in Novel Species
File Preprocess
Exon Prediction in Reference Species
Select reference genomic file C:\Users\owner\Desktop\dataset\WGSref.fa Browse
Select filtered reference cDNA file C:\Users\owner\Desktop\dataset\cDNA Browse
Exon Predict
View Result for Exon Prediction
Exon Predict Table
Exon Fasta File
Exon Prediction for Novel Species
Select novel species orthologous cDNA file Browse
Choose a file
dataset
Search
Organize
Views
New Folder
Favorite Links
Dropbox
Magic Briefcase
Desktop
Recent Places
Computer
Documents
Pictures
Music
More
Folders
Name
Date modified
Type
genomic.fas 07/02/2012 15:53 FAS File
spideytest.txt 07/02/2012 15:53 TXT File
cDNAref.fasta_filtered.fasta_orthologs.fasta_BLAST_output... 07/02/2012 15:47 TXT File
WGSref.fasta.nhr 07/02/2012 15:47 NHR File
WGSref.fasta.nin 07/02/2012 15:47 NIN File
WGSref.fasta.nsq 07/02/2012 15:47 NSQ File
cDNAnovel.fasta_filtered.fasta_orthologs.fasta 07/02/2012 15:40 FASTA File
cDNAref.fasta_filtered.fasta_orthologs.fasta 07/02/2012 15:22 FASTA File
cDNAref.fasta_filtered.fasta_orthologs.txt 07/02/2012 15:22 TXT File
cDNAnovel.fasta_filtered.fasta_reciprocalBlast.txt 07/02/2012 15:21 TXT File
cDNAref.fasta_filtered.fasta_Protein.fasta.phr 07/02/2012 15:19 PHR File
cDNAref.fasta_filtered.fasta_Protein.fasta.pin 07/02/2012 15:19 PIN File
File name: cDNAnovel.fasta_filtered.fasta_orthologs.fasta
All files (*.*)
Open
Cancel
Click on 'Browse' button to browse the files location folder
Select the orthologous cDNA sequences fasta file of novel species

Click on 'Novel Exon Predict' button





CEPiNS writes exon boundaries of the novel species into a tabular text file format and the exon sequences into a fasta file format which can be opened by clicking 'Novel Exon Predict Table' and 'Novel Exon Fasta File' respectively.

After opening the novel exon table text file in Excel, the Excel sheet looks as following.

Microsoft Excel - cDNAnovel.fasta_filtered.fasta_orthologs.fasta_OrthoExonTable.txt

File Edit View Insert Format Tools Data Window Help

U1

	A	B	C	D	E	F	G	H	I
1	Reference_ID	Exon_Number	cDNA_coordinate	Length	Novel_ID	Exon_Number	cDNA_coordinate	Length	Identity
2	Rfabg-PA*3	3	107-7272	7166	GK13707-PA	3	107-unk	unk	73.32
3	Rfabg-PA*4	4	7273-7493	221	GK13707-PA	4	7291-7511	221	44.29
4	Rfabg-PA*5	5	7494-9612	2119	GK13707-PA	5	7512-9630	2119	67.99
5	Rfabg-PA*6	6	9613-9733	121	GK13707-PA	6	9631-9751	121	82.5
6	Rfabg-PA*7	7	9734-10056	323	GK13707-PA	7	9752-10074	323	83.18
7	Hcf-PC*1	1	1-316	316	GK13582-PA	1	unk-301	unk	98
8	Hcf-PC*2	2	317-626	310	GK13582-PA	2	302-611	310	98.04
9	Hcf-PC*3	3	627-838	212	GK13582-PA	3	612-823	212	88.57
10	Hcf-PC*4	4	839-1027	189	GK13582-PA	4	824-1012	189	88.71
11	Hcf-PC*5	5	1028-1170	143	GK13582-PA	5	1013-1155	143	91.49
12	Hcf-PC*6	6	1171-2247	1077	GK13582-PA	6	1162-unk	unk	78.26
13	Hcf-PC*7	7	2248-3337	1090	GK13582-PA	7	unk-unk	unk	77.23
14	Hcf-PC*10	10	3730-3996	267	GK13582-PA	10	unk-3975	unk	87.1
15	Hcf-PC*11	11	3997-4359	363	GK13582-PA	11	3976-4338	363	95.87
16	Ephrin-PA*1	1	1-680	680	GK13690-PA	1	unk-842	unk	68.66
17	Ephrin-PA*2	2	681-991	311	GK13690-PA	2	843-1153	311	100
18	Ephrin-PA*3	3	992-1614	623	GK13690-PA	3	unk-unk	unk	76.47
19	Ephrin-PA*4	4	1615-1959	345	GK13690-PA	4	1834-unk	unk	74.19
20	sv-PA*4	4	524-689	166	GK14101-PA	4	41-206	166	81.48
21	sv-PA*5	5	690-887	198	GK14101-PA	5	207-404	198	63.08
22	CG11093-PB*2	2	94-1561	1468	GK13606-PA	2	253-unk	unk	95.19
23	bt-PF*1	1	1-313	313	GK13703-PA	1	1-313	313	93.27
24	bt-PF*3	3	336-498	163	GK13703-PA	3	336-498	163	85.19
25	bt-PF*4	4	499-640	142	GK13703-PA	4	499-640	142	91.49

The exon sequences fasta file of the novel species looks as follows.

cDNAnovel.fasta_filtered.fasta_orthologs.fasta_OrthoGeneExons.fasta - WordPad

File Edit View Insert Format Help

```
TGTGGAGCAGATTATCACCGAAAAGCGACTTTTGGACATCTACAATGTCCTGGAGAGGCTCGTGAAGAG
CGTGGTCCCTGGATCAGATGAGAAGGCGTTCCATCTAGCCCTCGACTATCCATTCCGCGCTGGCGCAGC
TAAGAGCATCGTTGGTGTTCGCAGCGATTCGCTTGAGTACAAGAACTGG
>GK13707-PA_Exon-6 Length:121
TGGAAGTTTGTGCGGGCGCAAATAACAGGAACTGTCACCAAATTCGATGGTGCCCTGCTTCATCTGATT
GCTCCAGTGAAAGGACTGTCCTTGGAGGGAGTGCCAGCCGAAAAATTAGTTG
>GK13707-PA_Exon-7 Length:323
GCTTTAACTCTCGACTTGTCGCCACCACTGAAAACAAGGATAACAAGAAGCGCACCAAACTCCAGTTCG
AGAACGATATGGGCATTGACTTTGTTCTCAACAATGGCGGATGGGTATTCGATACACAGAACTTCGACA
AGTTGAAGGCTCCGGAGCAGAAGAAGGTACTCAACCAAGTTACTTCCTCGATTGCCGATACTTTGTTCA
AGACCGAAATCGTCAGTGACTGTCGCTGCTGGCCAGCCCATGGCCTGCATGGTCAGCACAAGTGTGCCA
TTAAGTCGTCCACCTTCGTACCAAACAAGAAACCCAAGGCCGCGTAG
>GK13582-PA_Exon-1 Length:unk
CGGGATTTCGATGGAAGCGCGTGCTGAACCCAACAGGACCTCAGCCTCGTCCTCGTCATGGACATCGCG
CCATCAACATTAAGGAACTGATGGTTGTATTTGGTGGCGGTAACGAGGGAATTGTTGATGAACTTCATG
TCTACAACACAG
>GK13582-PA_Exon-2 Length:310
TTACCAATCAGTGGTATGTGCCGGTGTTGAAGGGCGATGTGCCCAATGGATGCGCTGCTTACGGATTTG
TTGTCGAGGGCACTCGCATGTTCGTATTTGGCGGAATGATTGAGTATGGAAAATACTCTAATGAACTTT
ATGAGCTCCAGGCCACCAAGTGGGAGTGGCGTAAAATGTATCCAGAGTCCCCTGACAACGGGTTGTCAC
CCTGTCCCCGCCTGGGTCATAGTTTCACCATGGTCGGCGACAAGATATTCCTTTTTGGTGGCTTGGCCA
ATGAATCGGATGATCCGAAAAACAATATTCCGAA
>GK13582-PA_Exon-3 Length:212
ATATCTGAACGATTTGTACATCTTGGACACACGCGGAGTGCATAGCCACAACGGCAAGTGGATAATACC
GAAAACGTATGGCGATAGTCCGCCGCCTAGAGAATCGCATACTGGAATCTCGTTTACAAGCAAAACAAC
TGGCAAGCTAAATCTCTTGGTTTATGGTGGGATGAGTGGCTGCCGTTTGGGAGATTTGTGGTTGCTGGA
```

For Help, press F1 NUM