SIRIUS Prediction System Builder 2.1

Target audience

Users who want to learn using Sirius PSB by example. Users who have a background in data mining and/or have used WEKA previously would find this easier to understand. Users are advised to read Koh and Wong (2007) to better understand this example.

How to use this guide

As this guide is in a teach-by-example format, you would require the following to effectively use the guide:

- SIRIUS Prediction System Builder 2.1(Sirius.jar)
- Arabidopsis genomic sequences from Koh et al (2007). The datasets are available at http://www.comp.nus.edu.sg/~wongls/projects/dnafeatures/giw07-supplement/

note: download "GIW2007-v2.zip" as this version contains the fasta files

To predict the location of Polyadenylation sites in Arabidopsis Genomic Sequences

Build a prediction model (based on the methodology described in Koh et al, 2007) which can predict positions of Polyadenylation sites given Arabidopsis genomic sequences and evaluate the model's accuracy using datasets from Koh et al (2007).

1. Click *Applications*, then *Train Classifier (Trainer)* (Figure 1).

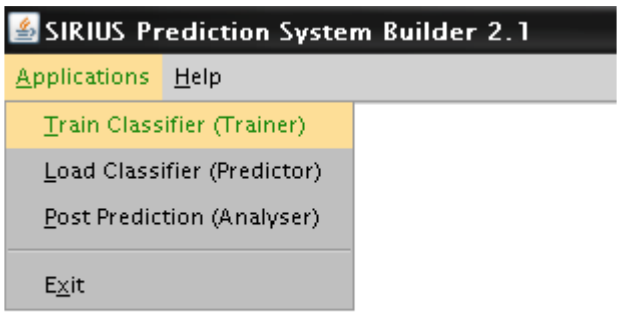


Figure 1. Start up Train Classifier

Step 1: Define Data

1. From **Working Directory**, click *Browse* (Figure 2).

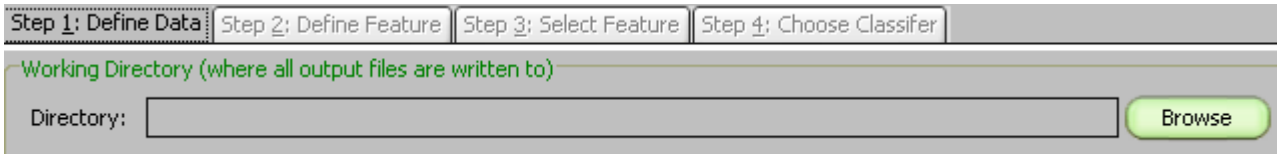


Figure 2. Select location to save output files to

2. From the pop-up window, choose the location where you want the output files saved and click *Open*.

3. From **Sequence Type**, select *DNA*.

4. From **Files with +ve Sequences**, click *Add File (FASTA Format ONLY)* (Figure 3).

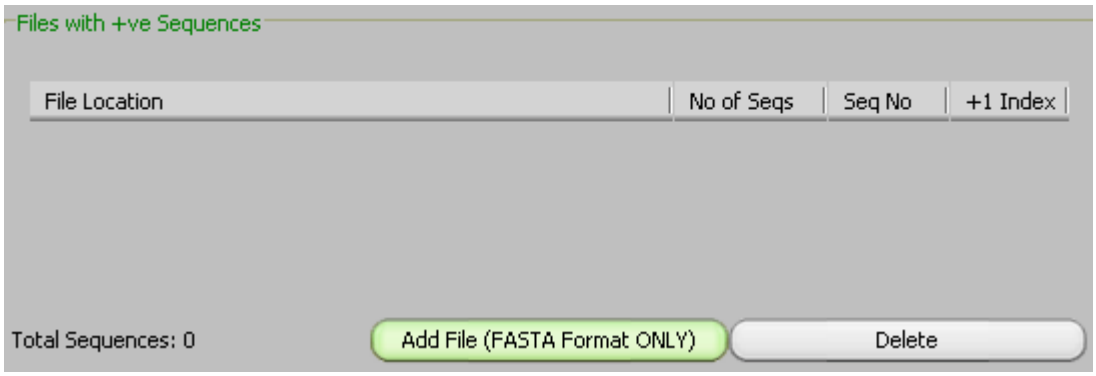


Figure 3. Select file to add

5. From the pop-up window, go to Dataset B (of the GIW datasets), choose EST_Supported.fasta and click *Open*.

6. Repeat step 5 for the same file from Dataset C and Dataset D (Figure 4).

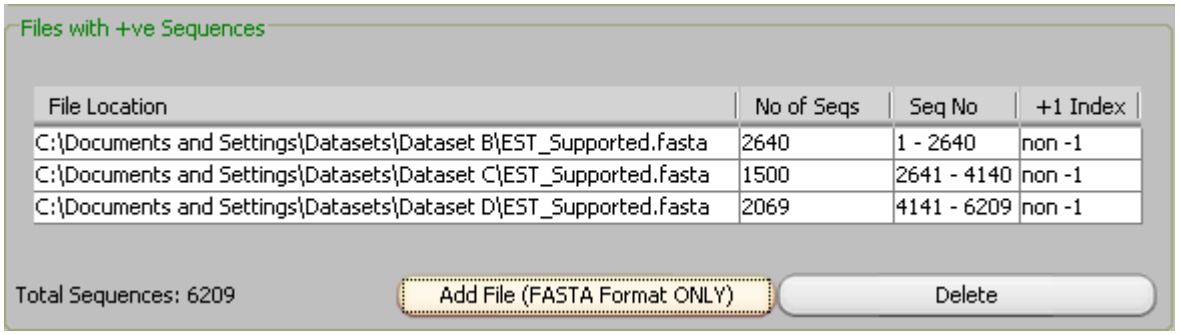


Figure 4. After all 3 EST_Supported.fasta files are added

7. From **Files with -ve Sequences**, click *Add File (FASTA Format ONLY)* (Figure 5).

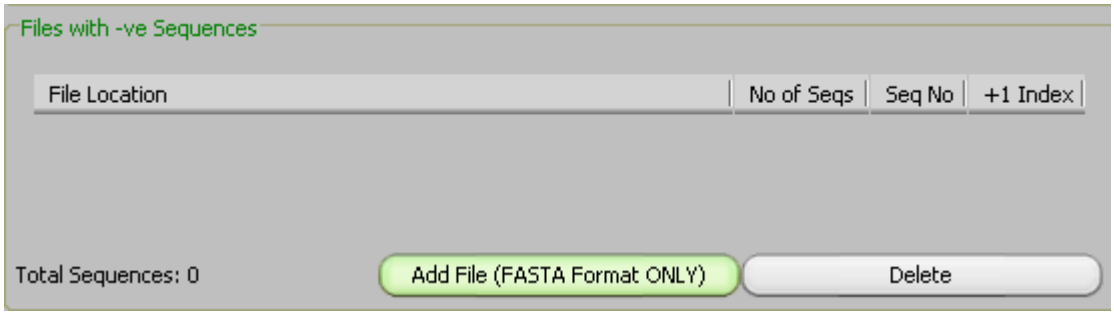


Figure 5. Select file to add

8. From the pop-up window, go to Dataset B, choose 5UTR.fasta and click *Open*.

9. Repeat step 8 to add Coding.fasta and Intronic.fasta from Dataset B.

10. Repeat steps 8 and 9 to add the same 3 files from Dataset C and Dataset D (Figure 6).

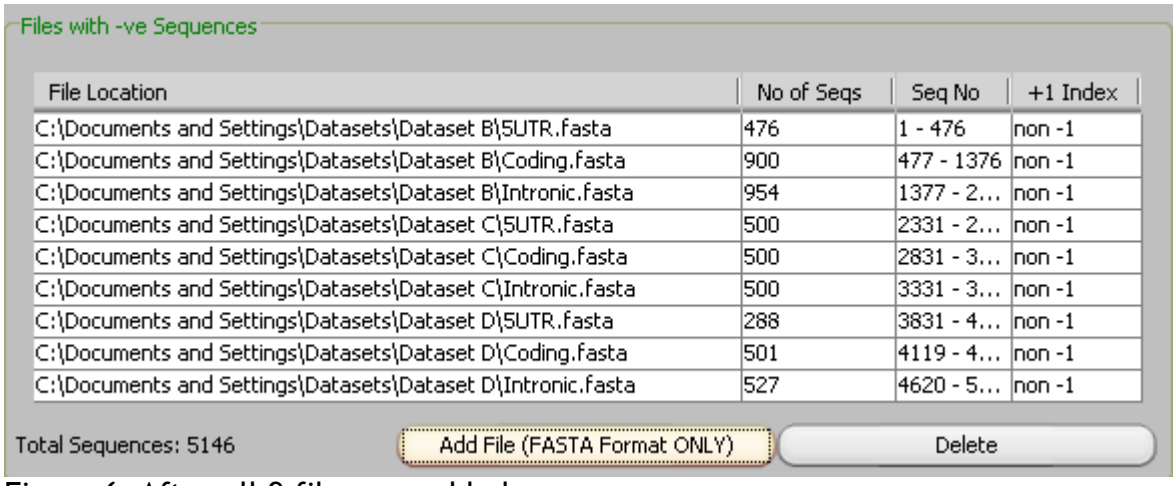


Figure 6. After all 9 files are added

11. From **Split all +ve Sequences into 3 Dataset**, Dataset 1, type 1 and 2640 into the boxes (Figure 7).

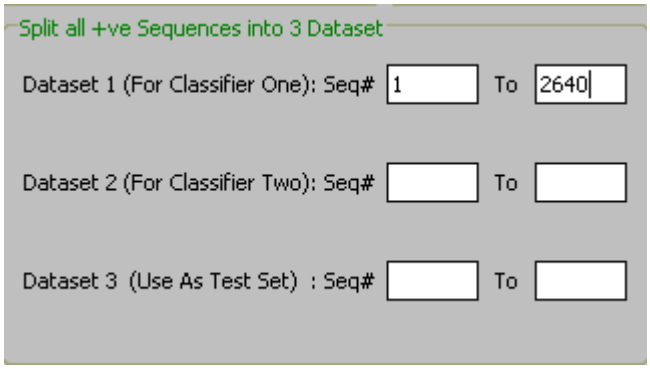


Figure 7. Enter sequence numbers for +ve set

12. For Dataset 2, type 2641 and 4140.

13. For Dataset 3, type 4141 and 6209.

14. Similarly, from **Split all -ve Sequences into 3 Dataset**, Dataset 1, type 1 and 2330 into the boxes (Figure 8).

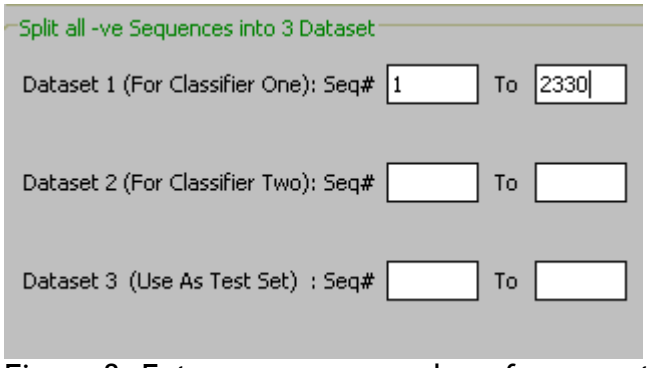


Figure 8. Enter sequence numbers for –ve set

15. For Dataset 2, type 2331 and 3830.

16. For Dataset 3, type 3831 and 5146.

17. Click *Next*.

Step 2: Define Feature

1. From **Choose a feature type to add**, click *K-grams with X mistakes* (Figure 9).



Figure 9. Select K-grams with X mistakes

2. From the pop-up window, click on the *All Permutations* tab.

3. a) From **Window Size and Location**, type From -110 to 5.

   b) From **K-gram with X mistakes allowed**, type *K: 1* and *X: 0* (Figure 10).

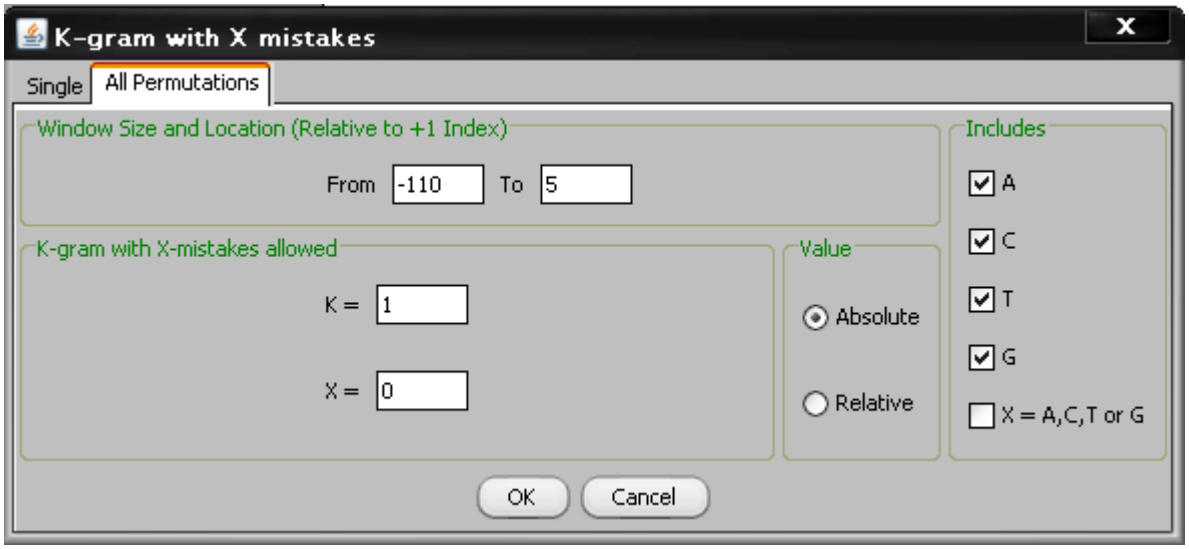


Figure 10. Input K and X values

   c) Click *OK*.

4. Repeat steps 1 to 3, with the following 8 sets of values:

Window Size and Location	K-gram with X mistakes allowed
From -110 to 5	K: 2 , X: 0
	K: 3 . X: 0
From -35 to 15	K: 1 , X: 0
	K: 2 , X: 0
	K: 3 , X: 0
From -50 to 30	K: 1 , X: 0

	K: 2 , X: 0
	K: 3 , X: 0



Note
Total number of features should be 252

5. Click *Generate Features*.

Step 3: Select Feature

1. From **Filter Features**, click *Choose* (Figure 11).

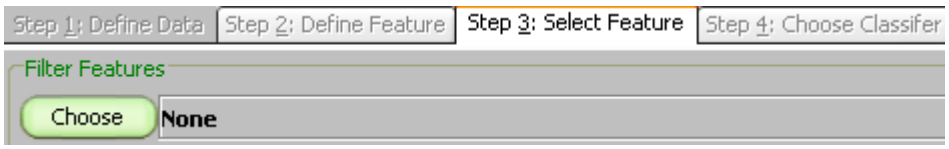


Figure 11. Choose filter

2. From the drop-down window, double-click *filters*, *supervised* then *attributes* to select *AttributeSelection* (Figure 12).

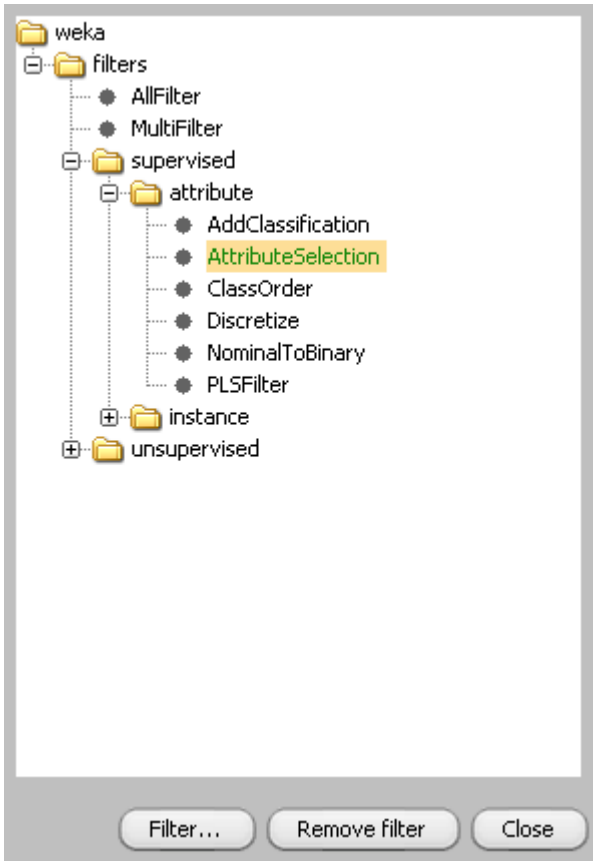


Figure 12. Select filter, AttributeSelection

3. From **Filter Features**, click on *AttributeSelection* (Figure 13).

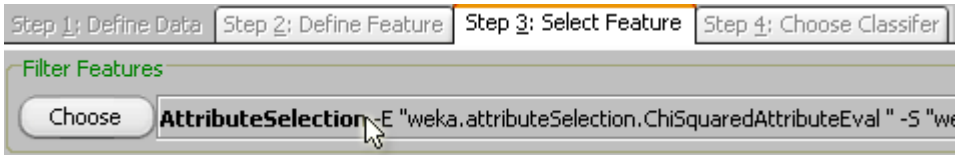


Figure 13. To edit properties of *AttributeSelection*

4. a) From the pop-up window, for **evaluator**, click *Choose* and select *ChiSquaredAttributeEval*.

b) For **search**, click Choose and select *Ranker*.

c) Click on *Ranker* (Figure 14).

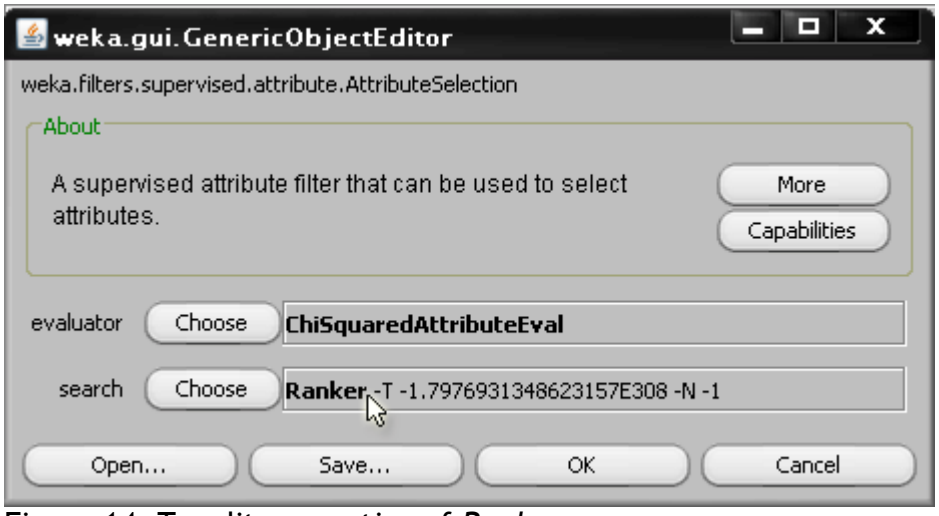


Figure 14. To edit properties of *Ranker*

d) Set the value of *threshold* to 0.

e) Click *OK* (Figure 15).

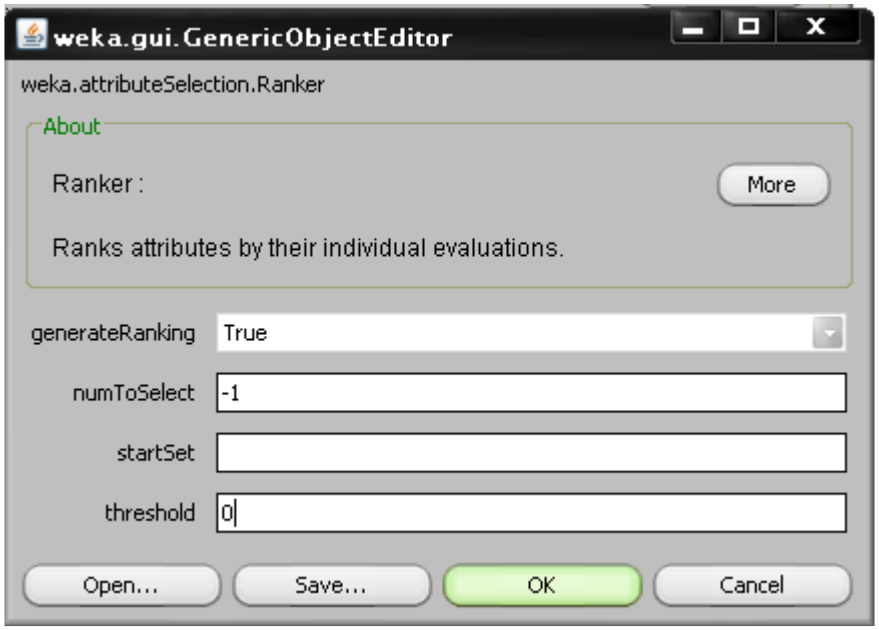


Figure 15. Confirm properties of *Ranker*

f) Click *OK* (Figure 16).

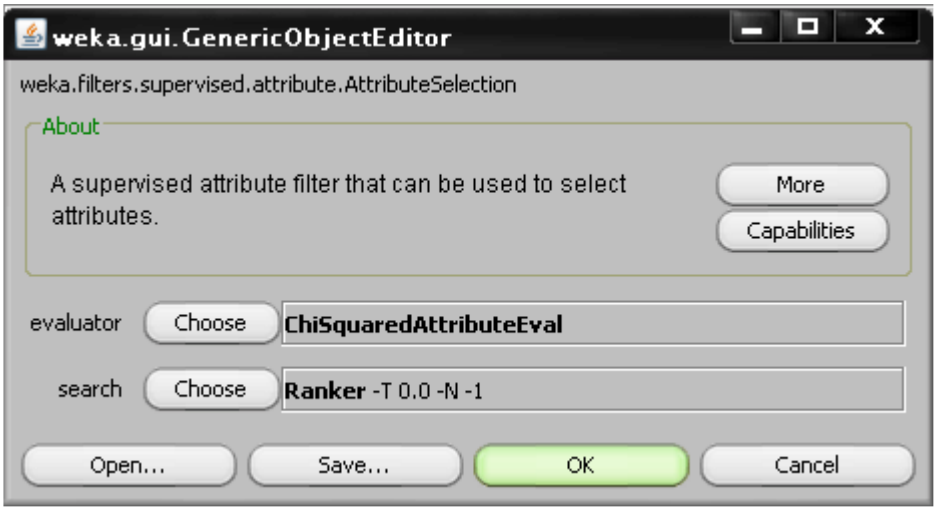


Figure 16. Confirm properties of filter

5) Click *Apply Filter*.



Note
Total number of features after filtering is 221

6) Click *Next*.

Step 4: Choose Classifier

1. From **Level One Classifier**, click *Choose*.

2. From the drop-down window, double-click *classifiers*, *functions* then select *SMO* (Figure 17).

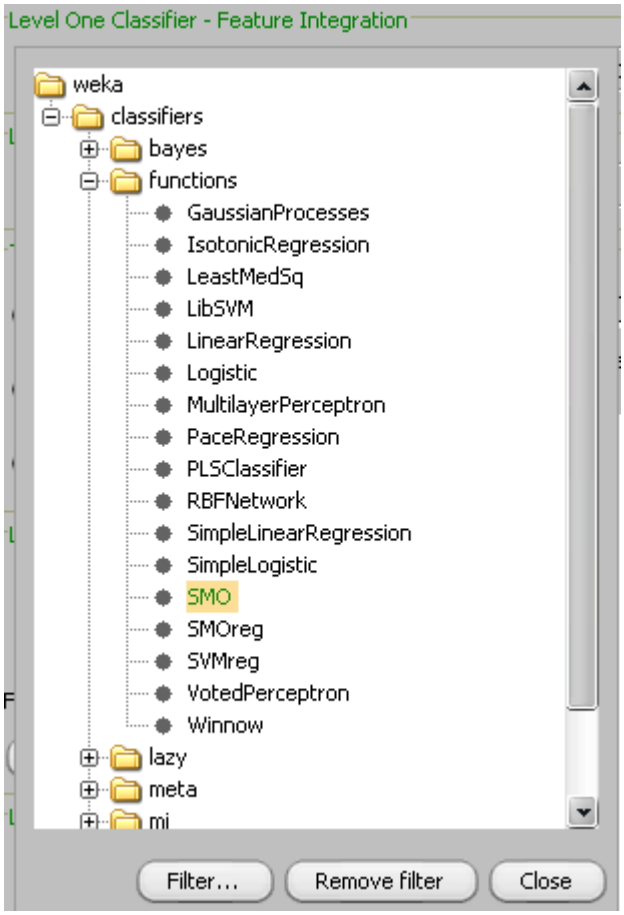


Figure 17. Select function, *SMO*

3. From **Level One Classifier**, click on *SMO* (Figure 18).

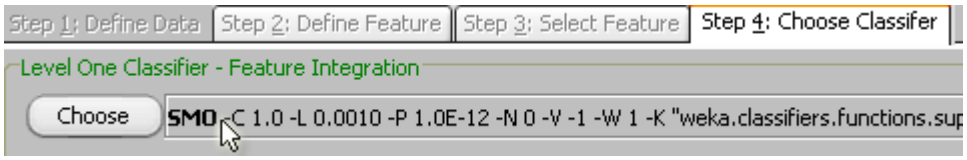


Figure 18. To edit properties of *SMO*

4. a) From the pop-up window, for **buildLogisticModels**, select *True*.

   b) Select *OK* (Figure 19).

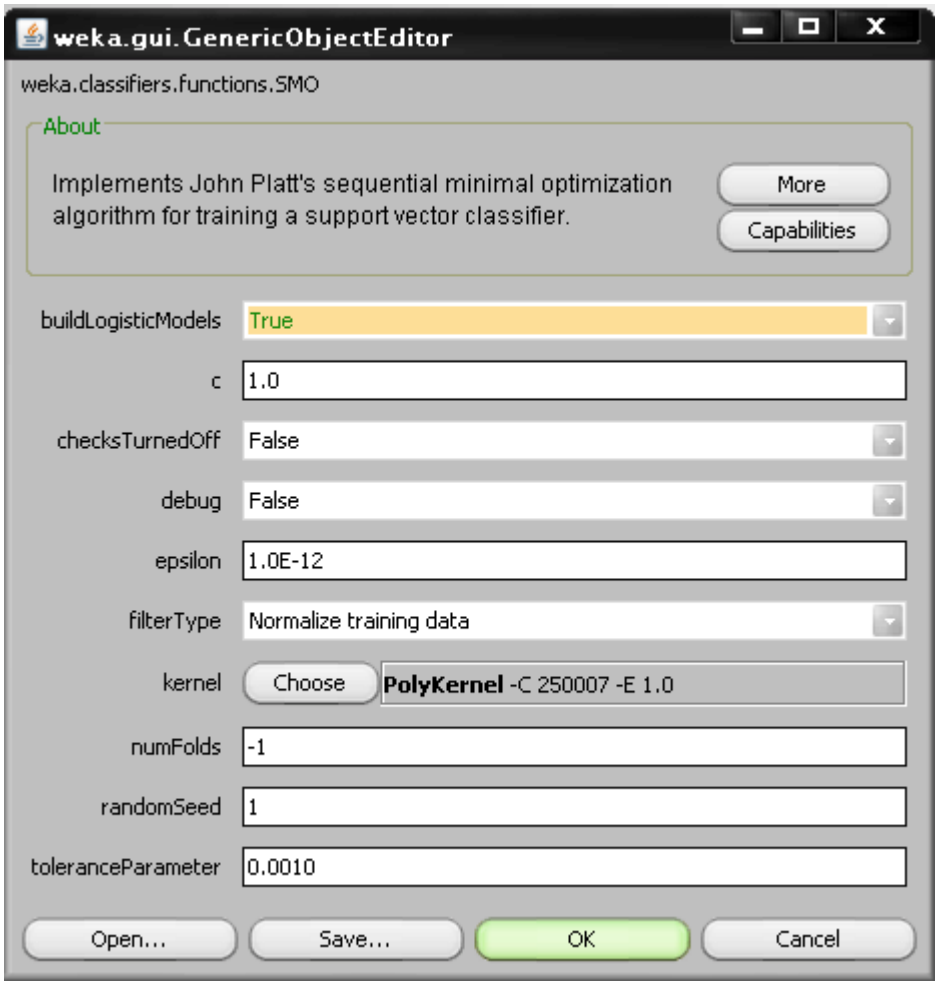


Figure 19. Confirm properties of classifier

5. From **Level One Classifier Information**, click *Start*.

6. Repeat steps 1 to 4 for **Level Two Classifier** (Figure 20).

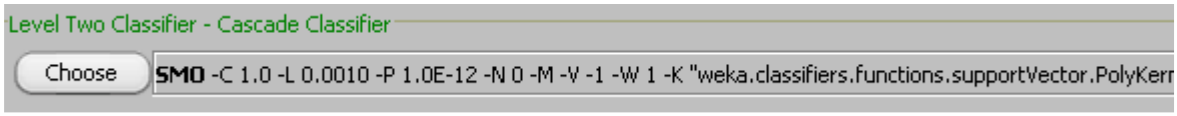


Figure 20. Settings for Level Two Classifier completed

7. From **Test Options**, select Use *Dataset 3* (Figure 21).

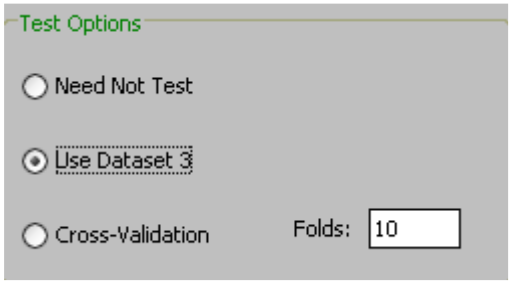


Figure 21. Select *Dataset 3* from Test Options



Note

Also ensure 'Set Upstream' and 'Set Downstream' values are set at -40 and 41 respectively

8. From **Level Two Classifier Settings**, click *Start*. The results will then be shown in the **Output** section (Figure 22).

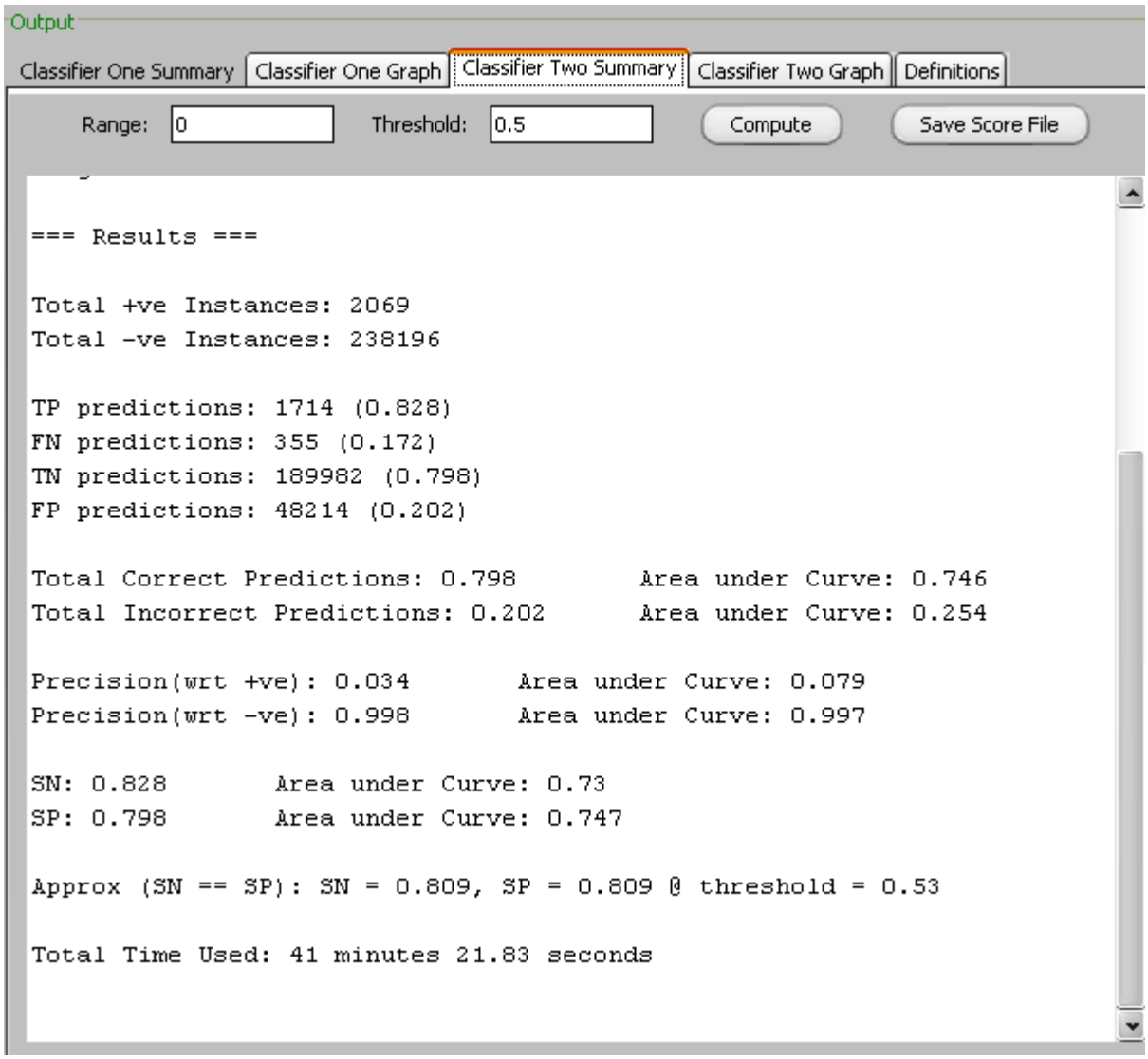


Figure 22. Results spread out over various tabs

REFERENCES

Koh, C.H. and Wong, L. (2007). Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences. In Proceedings of 18th International Conference on Genome Informatics, (Singapore, December 3 – 5, 2007.), pp. 73-82.