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 Liu and Wong (2003) to better understand the methodology applied to build the prediction model. Users are also encouraged to read Emanuelsson, Nielsen and Heijne (1999) if they are unfamiliar with protein subcellular localization problem.

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)
- Dataset of cTP sequences. The datasets are available at http://www.cbs.dtu.dk/services/ChloroP/pages/datasets.php

Note: Download the two files – 75 sequences containing cTP (in fasta format) and 75 sequences not containing cTP (in fasta format)

To determine if a protein is a chloroplast transit peptide based on its primary sequence.

Build a prediction model which can predict if input sequences are chloroplast transit peptide, and evaluate the accuracy of the model based on 10-fold cross-validation.

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

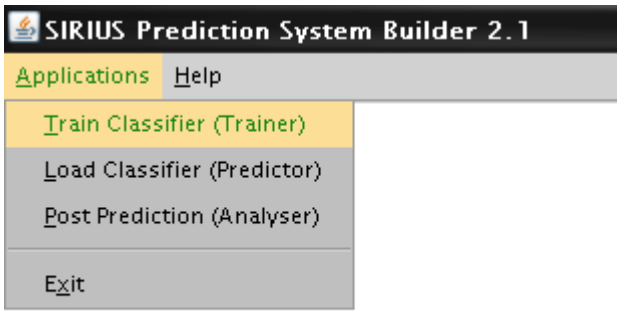


Figure 1. Start up Train Classifier

Step 1: Define Data

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

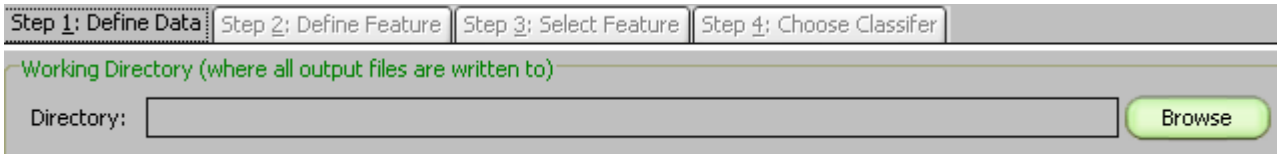


Figure 2. Select location to save files to

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

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

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

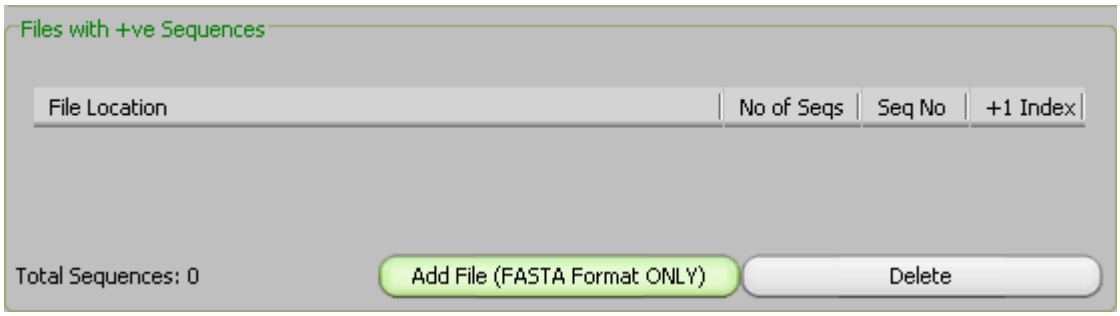


Figure 3. Select file to add

5. a) From the pop-up window, choose cTP.75seq.fasta and click *Open*.

   b) If a "Set +1 Index" window appears, click *OK* (Figure 4).

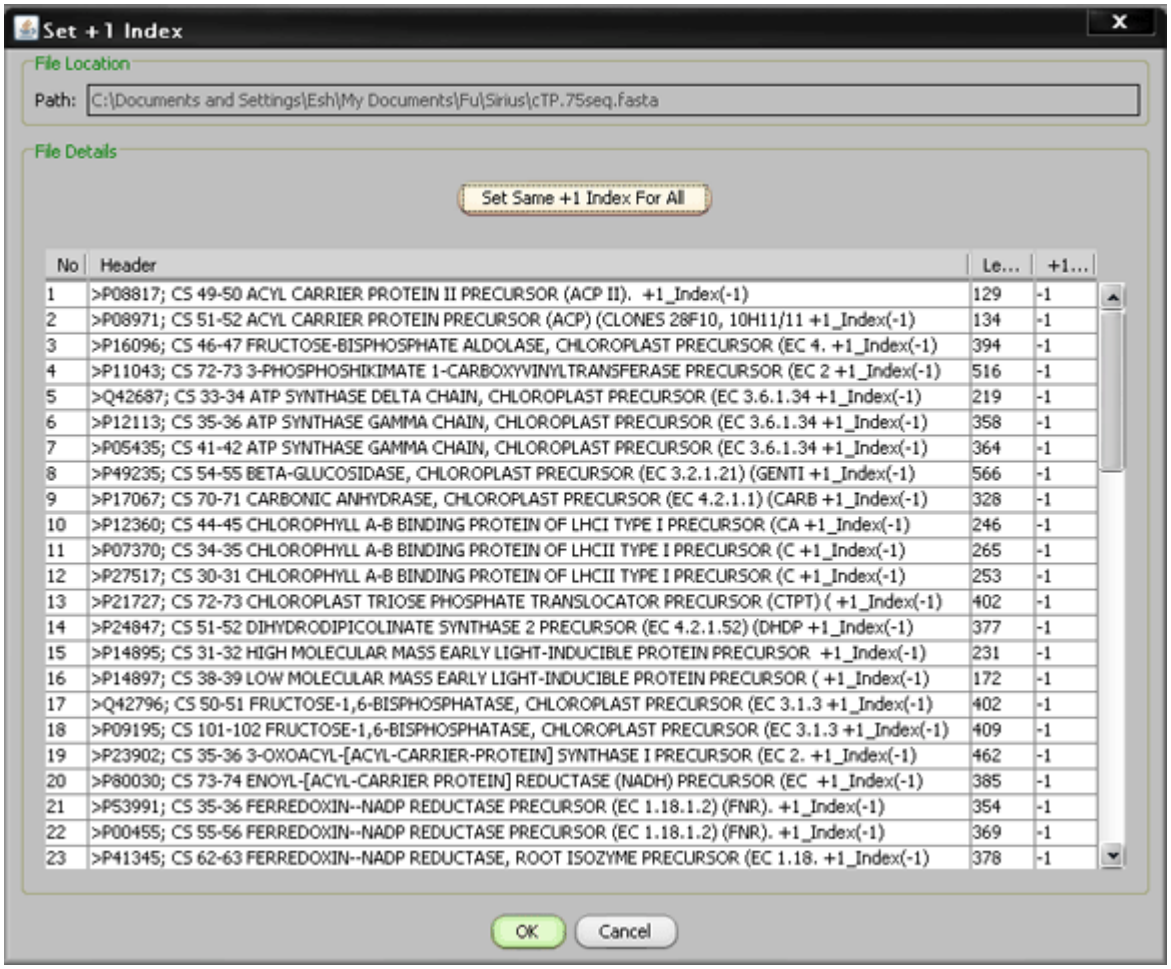


Figure 4. Set +1 Index

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

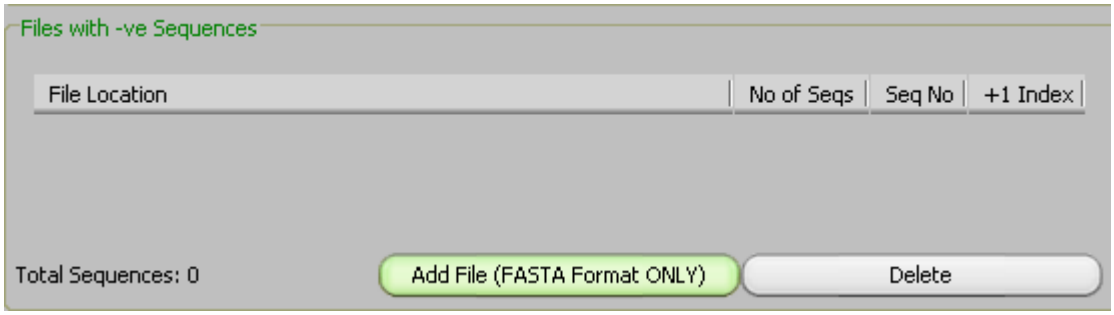


Figure 5. Select file to add

7. a) From the pop-up window, choose non_cTP.75seq.fasta and click *Open*.

   b) Similarly, if the "Set +1 index" window appears, click *OK*.

8. From **Split all +ve Sequences into 3 Dataset**, Dataset 1, type 1 and 75 into the boxes (Figure 6).

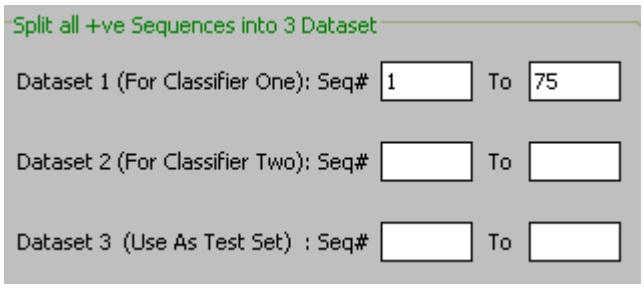


Figure 6. Enter sequence numbers for +ve set

9. Similarly, from **Split all -ve Sequences into 3 Dataset**, Dataset1, type 1 and 75 into the boxes (Figure 7).

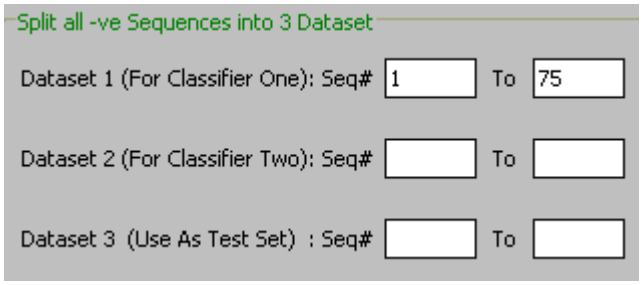


Figure 7. Enter sequence numbers for -ve set

10. Click *Next*.

Step 2: Define Feature

1. From **Choose a feature type to add**, click *K-grams with X mistakes* (Figure 8).



Figure 8. 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 0 to 100.

   b) From **K-gram with X mistakes allowed**, type *K: 1* and *X: 0* (Figure 9).

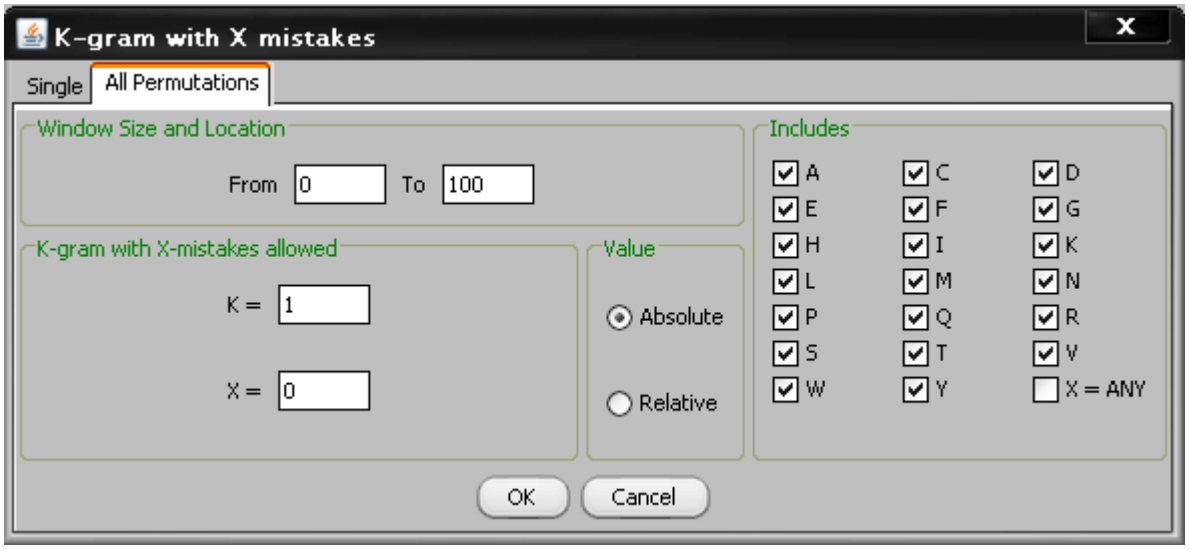


Figure 9. Input K and X values

   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 0 to 100	K: 2 , X: 0
	K: 3 . X: 0
From 101 to 150	K: 1 , X: 0
	K: 2 , X: 0
	K: 3 , X: 0
From 151 to 200	K: 1 , X: 0
	K: 2 , X: 0
	K: 3 , X: 0



Note
Total number of features after filtering is 25260

5. Click *Generate Features*.

Step 3: Select Feature

1. Click *Next* (Figure 10).

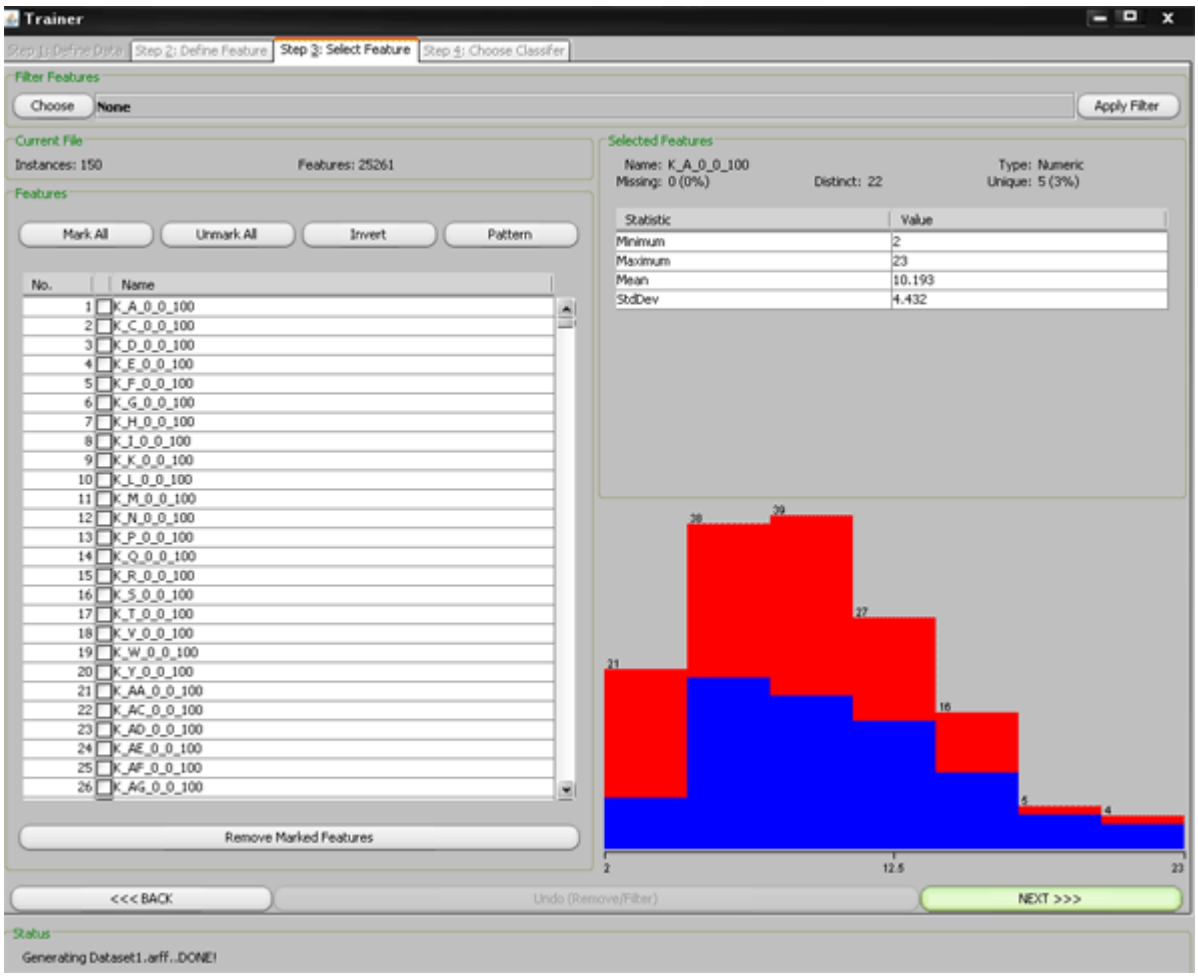


Figure 10. Proceed to next screen



Note

Observant users may have noticed that the number of features is 25261, this is because a class feature has been added (scroll down the list to see it)

Step 4: Choose Classifier

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

2. From the drop-down window, double-click *classifiers*, then *meta* then select *AttributeSelectedClassifier* (Figure 11).

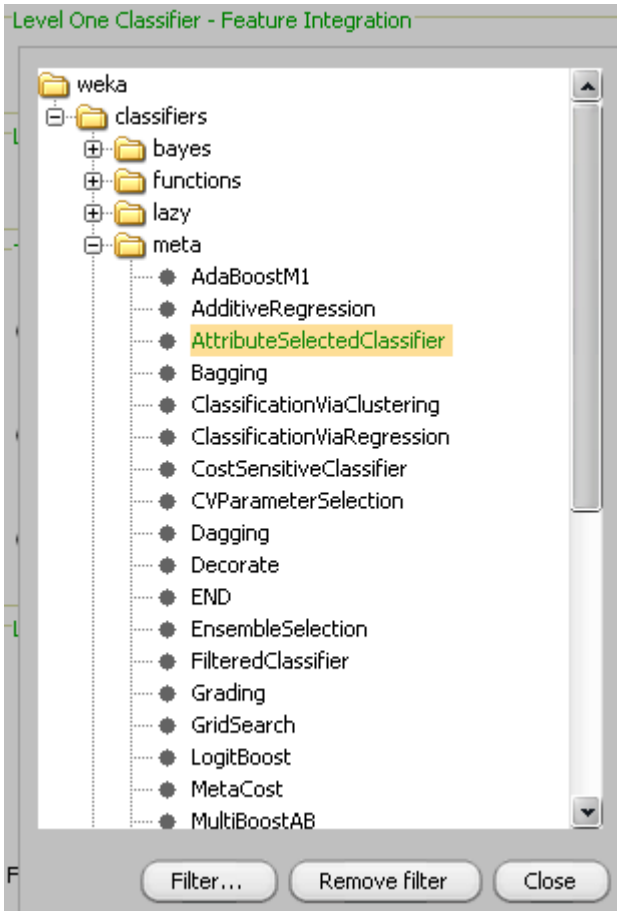


Figure 11. Select classifier, AttributeSelectedClassifier

3. From **Level One Classifier**, click on *AttributeSelectedClassifier* (Figure 12).



Figure 12. To edit properties of *AttributeSelectedClassifier*

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 13).



Figure 13. To edit properties of *Ranker*

   d) Set **threshold** to 0.

e) Click *OK* (Figure 14).

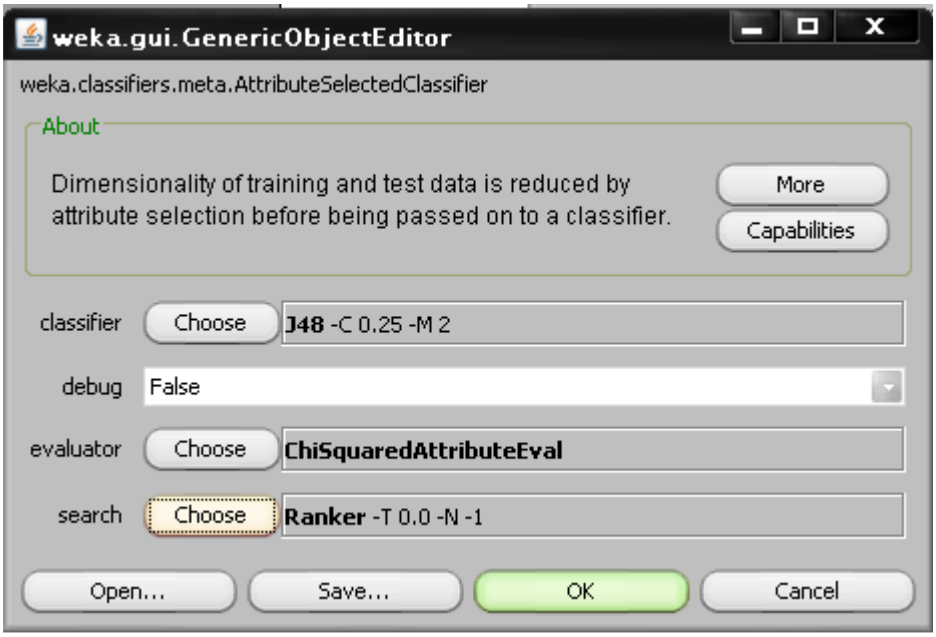


Figure 14. Confirm properties of classifier

5. From **Test Options**, select *Cross-Validation*.



Note
Ensure that it is 10 fold cross-validation

6. From **Level One Classifier Information**, click *Start*. The results will then be shown in the **Output** section (Figure 15).

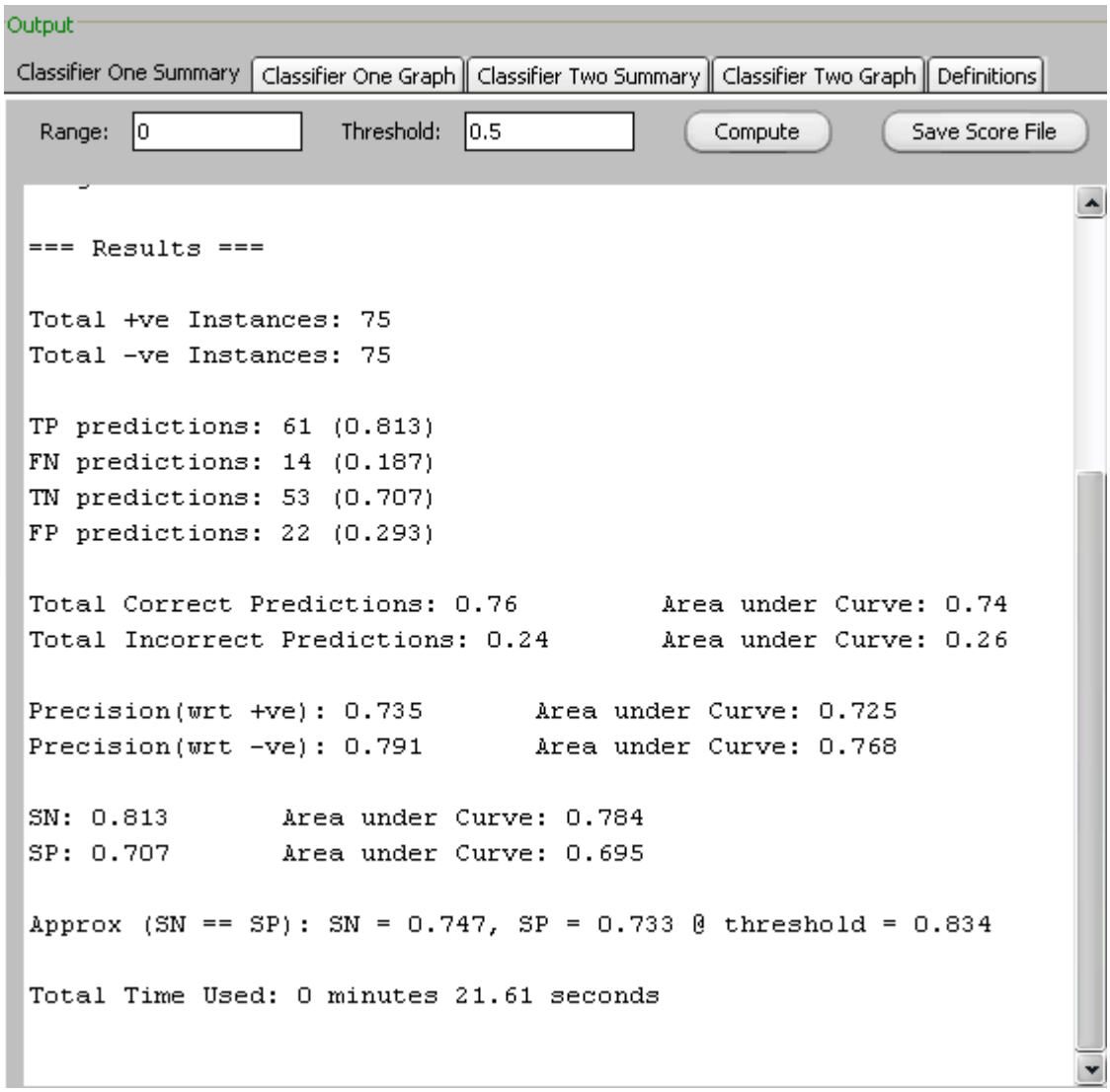


Figure 15. Results spread out over various tabs

REFERENCES

Liu, H. and Wong, L. (2003). Data Mining Tools for Biological Sequences. Journal of Bioinformatics and Computational Biology, Vol.1, No.1, April 2003, pp. 139-167.

Emanuelsson, O., Nielsen, H. and Heijne G.V. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Science, Vol.8, pp. 978-984.