BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–30

Supplement to: A Robust Peak Detection Method for RNA Structure Inference by High-throughput Contact Mapping

Jinkyu Kim , Seunghak Yu , Byonghyo Shim , Hanjoo Kim ,
Hyeyoung Min , Eui-Young Chung , Rhiju Das , and Sungroh Yoon

This document contains the supplemental material mentioned in the main article. Section 1 describes the details of how to train the classifier used in the inter-profile peak analysis and how to normalize and score the features used for better training. Section 2 explains the procedure to install and use the code we developed for this study. Section 3 contains the information on the profile and batch data we used for the experiments presented in the main article. Section 4 presents the enlarged version of the figures in the article.

1 IMPLEMENTATION DETAILS

1.1 Training and validating the classifier

For training the SVM-based binary classifier used in the intra-profile peak detection step, we first examined all potential peaks appearing in the 2002 profiles and extracted 38-dimensional feature vectors from them, as explained in Section 4.3 of the main article. We then sampled 490 profiles randomly in order to train the classifier. After training, we validated the performance of the classifier with the 1512 profiles unused for training. This procedure is depicted in Figure 1. We repeated this process 20 times, generating 20 different training-validation set pairs. The performance of the classifier was not very sensitive to the training sets used: Over the 20 different training-validation set pairs, the obtained standard deviation values of precision and recall were only 0.0091 and 0.0114, respectively.

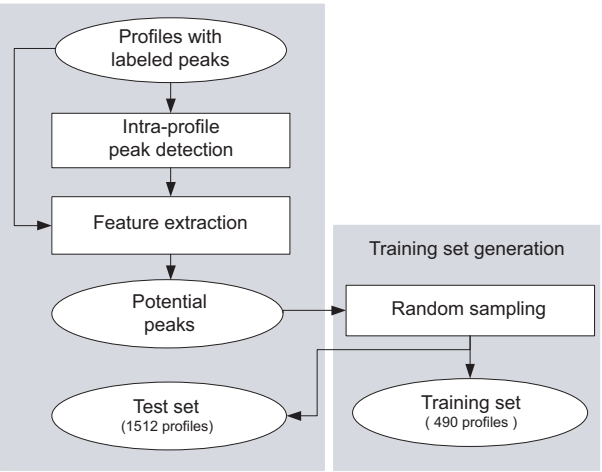


Fig. 1. Overview of generating training sets and test sets.

Table 1. Adjusting feature values

Ranges	Weight	Ranges	Grade	Ranges	Grade
$0.4 < x$	$3 + x$	$6 < y$	-3	$z > 0.9$	-2
$0.3 < x < 0.4$	$2.5 + x$	$4 < y < 6$	-2.5	$0 < z < 0.9$	2
$0.275 < x < 0.3$	$2 + x$	$3 < y < 4$	-1		
$0.25 < x < 0.275$	$1 + x$	$2 < y < 3$	0		
$0.2 < x < 0.25$	$-1 + x$	$0 < y < 2$	1		
$0 < x < 0.2$	$-3 + x$				

Variables x, y, and z represent the adjusted intensity, the number of nearby peaks in other profiles, and the relative peak location, respectively.

1.2 Normalizing and adjusting feature values

For better classification results, we preprocessed the extracted feature values, especially the features such as ‘adjusted intensity’, ‘the relative peak location’, and ‘the number of nearby peaks in other profiles’. These features can have values in a wide range, and using them without normalization could result in a very poor result. The widely-used normalization scheme we used is as follows:

$$\text{normalized value} = \frac{\text{original value} - \text{mean}}{\text{standard deviation}} \quad (1)$$

After this normalization, we further adjusted the values of these three features according to our empirical observation. That is, to see what separates a true peak from a false one, we collected the information on the aforementioned three features over the peaks appearing in the entire profiles. As depicted in Table 1, we then created several ranges of feature values and gave each range a different weight.

The feature ‘adjusted intensity’ usually has values ranging from the threshold value used to approximately two. When the adjusted intensity of a potential peak is larger than 0.7, then this candidate is normally a true peak. (Even a false peak could have adjusted intensity higher than 0.7, but most of these false peaks are removed by other features.) On the other hand, when the adjusted intensity of a peak candidate is smaller than 0.2, then this candidate is typically a false peak. In addition, we observed that when a peak candidate has more than four nearby peaks in other profiles this candidate is typically a false peak. A true peak usually has less than or equal to four nearby peaks in other profiles. In case of the feature ‘the relative peak location’, we divided the entire ranges into two (0∼0.9 and 0.9∼1) in order to penalize those peaks that are located near the end of a profile.

2 SOFTWARE USER MANUAL

2.1 Requirements

- MATLAB **7.6.0 or higher** (for `uitable` support)
- A display with screen resolution of 1280x1024 or higher

2.2 Installation

1. Download the software and data at `http://dna.korea.ac.kr/pub/mohca` (The software and data have been provided for the reviewers as our submission package and will publicly be available at this web site on acceptance of the manuscript.).
2. Unzip the files into a working directory of your choice. In what follows, we assume that the name of this directory is `$HOME`.
3. (On Linux platform only) Move to `$HOME/libsvm` and open `Makefile`. Change `MATLABDIR` into the location where your MATLAB program is installed. Quit the file and compile LIBSVM by typing `make`. GNU g++ (version 3.4 or lower) is recommended. Refer to `http://www.csie.ntu.edu.tw/~cjlin/libsvm` for additional details.

2.3 Directory structure

- `$HOME`: The root directory. Contains MATLAB codes and the following subdirectories.
- `$HOME/Datalist`: Contains the following files:

    1. `data_79batches.txt`: The names of the 79 batches from which the 2002 profiles used for our experimental studies were derived.
    2. `data_19batches.txt`: The names of the 19 batches from which the 494 "noisy" profiles used for the robustness test are drived (this test result is shown in Figure 8 in the main article).
    3. `data_60batches.txt`: The names of the remaining 60 batches.
    4. `label_79batches.txt`: The names of the golden data file corresponding to the 79 batches along with the information on the location of the radical source attachment point, the RNA solution condition, and the type of radical source used in order.
    5. `label_19batches.txt`: The names of the golden data file corresponding to the 19 batches along with the above extra information.
    6. `label_60batches.txt`: The names of the golden data file corresponding to the 60 batches along with the above extra information.

- `$HOME/libsvm`: Contains files needed for running LIBSVM.
- `$HOME/MOHCA_data`: Has 35 subdirectories, each of which contains `*_profile.mat` and `*_peakinfo.mat` files. The former contains the index and y-coordinate values of a profile; the latter the corresponding golden peak information such as the profile index and position of ·OH cleavage.
- `$HOME/trainingset`: Contains the training data we used to generate the result presented in the main article. (It is also possible to generate a training set randomly using the GUI explained below.) The three files included in this directory are as follows:

    1. `traindata_79batches.mat`: The training data for the all 79 batches.
    2. `traindata_19batches.mat`: The training data for the the "noisy" 19 batches.
    3. `traindata_60batches.mat`: The training data for the remaining 60 batches.

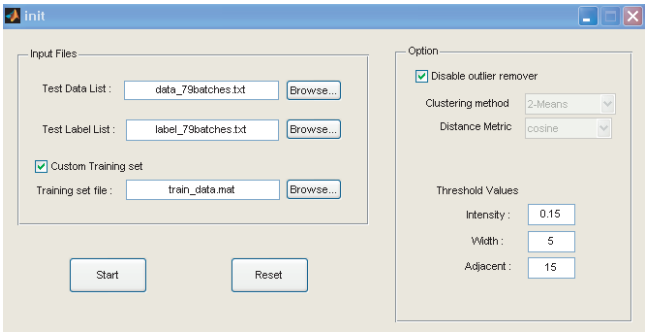


Fig. 2. Main GUI window to set parameters and launch the software.

2.4 GUI support

2.4.1 Main GUI window Figure 2 shows the main GUI window that is used to start the peak detection procedure. In addition, the user can specify the following options using this window:

- Outlier remover: As explained in Section 4.4, we provide an outlier remover that is based upon k-means clustering. We set $k = 2$ and provide optionally four types of distance metrics: cosine, correlation, squared Euclidean distance, and the sum of absolute differences. The default option is not to use the remover, but the user can change it by unchecking the checkbox.
- Threshold values: As explained in Section 4.2 and 4.4, the intra-profile peak detector uses three types of parameters: intensity, width, proximity. The default values 0.15, 5 and 15, respectively.

2.4.2 Result summary window Figure 3 presents a snapshot of the result summary window we developed. The results provided include the following:

- Result table: The table located in the upper right pane in the snap shot.
- Profile curve: The curve located in the lower right pane.
- F-measure (as α varies from 0 to 1): The plot drawn in the middle left pane.
- Box plot of precision, recall, and F-measure(α = 0.25): The plot located in the lower left pane.

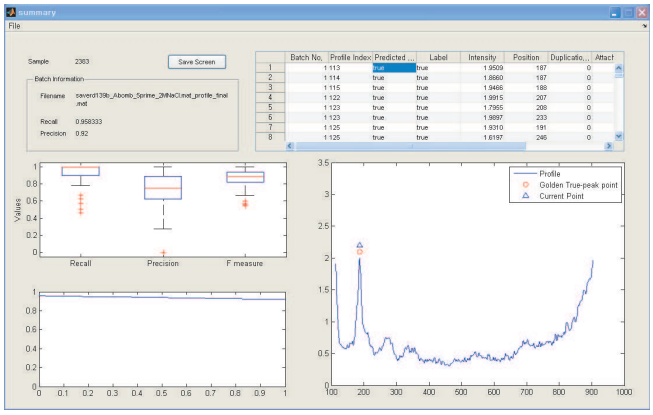


Fig. 3. Result summary window.

The result table contains 10 types of information: batch index, profile index, predicted label, label compared with golden data, adjusted intensity, position of ·OH cleavage, the number of nearby peaks in other profiles, the location of radical source attachment points, RNA solution conditions, and the type of radical source. When you select a certain peak in the result table, you can see the corresponding profile curve in the profile curve pane, in which a golden peak is marked as a triangle and a detected peak as a circle. You can the corresponding batch name and precision and recall values in the left of the result table. To show detection performance, we provide the box plot and the F-measure plot. The former is with respect to the entire results; the latter is with respect to the result for each batch.

2.5 An example run

We provide an example run, which can reproduce the result we report in the main article.

1. Start a MATLAB session.
2. Go to your working directory where the program and data have been installed.
3. Type `init` to launch the main window.
4. Configure your run as follows:
   - In the 'Input Files' section, load the file `$HOME/Datalist/data_79batches.txt` as 'Test Data List'.
   - As 'Test Label List', load the file `$HOME/Datalist/label_79batches.txt`.
   - Enable the checkbox 'Custom Training Set', and load the file `$HOME/trainingset/traindata_79batches.mat`.
   - In the 'Option' section, enable the checkbox 'Disable outlier remover'.
   - Use the default parameter: 0.15, 5 and 15 for 'Intensity', 'Width', and 'Adjacent', respectively.
5. Press the 'Start' button.
6. Wait until the result summary window appears.

2.6 Known issues

- In the summary window, the peak curve often freezes when the graph refresh rate is too fast.
- It was found that the single set overview is sometimes delayed because of using large-scale data.
- In a display with resolution less than 1280x1024, only part of the result window is seen.

3 DATA LIST

This section presents the batches used for our experimental studies. The actual data for these batches can be found in `$HOME/MOHCA_data`. Tables 2 and 3 list the batches used for generating the entire 2002 profiles. Table 4 lists the bathes used for the robustness test presented in Figure 8 of the main article. Table 4 shows the unused batches.

Table 2. The 79 batches from which the 2002 profiles used in the experimental studies are derived (continued in Table 3).

No	Batch ID
1	rd139b-Abomb-5prime-2MNaCl
2	rd139b-Abomb-5prime-Native
3	rd139c-Cbomb-5prime-Native
4	rd135c-Ubomb-5prime-Native
5	rd135b-Ubomb-5prime-Native
6	rd135d-Ubomb-5prime-Unfold
7	rd135e-Ubomb-5prime-Native
8	rd135e-Ubomb-5prime-Unfold
9	rd150d-p4p6-ABOMB-3prime-Unfolded-MKbottomC
10	rd150d-p4p6-ABOMB-5prime-Unfolded-MKtopB
11	rd150d-p4p6-ABOMB-5prime-Unfolded-RDtopC
12	rd150g-p4p6-ABOMB-3prime-2MNaCl-RDbottomB
13	rd150g-p4p6-ABOMB-5prime-2MNaCl-MKtopA
14	rd150g-p4p6-ABOMB-5prime-2MNaCl-MKtopB
15	rd150n-p4p6-Abomb-5prime-Native-MKtopC
16	rd150n-p4p6-Abomb-5prime-Native-RDbottomC
17	rd150b-Cbomb-5prime-Unfolded-MKbottomA
18	rd150b-Cbomb-5prime-Unfolded-RDtopC
19	rd150h-p4p6-CBOMB-3prime-2MNaCl-RDbottomA
20	rd150h-p4p6-CBOMB-5prime-2MNaCl-MKtopA
21	rd150h-p4p6-CBOMB-5prime-2MNaCl-MKtopB
22	rd150h-p4p6-CBOMB-5prime-2MNaCl-RDtopC
23	rd150i-p4p6-ABOMB-5prime-Native-MKbottomC
24	rd150k-p4p6-CBOMB-3prime-Native-RDbottomA
25	rd150k-p4p6-CBOMB-5prime-Native-MKbottomB
26	rd150k-p4p6-CBOMB-5prime-Native-MKtopA
27	rd150k-p4p6-CBOMB-5prime-Native-RDbottomC
28	rd150l-p4p6-CBOMB-5prime-Unfolded-LowIncorp-MKtopC
29	rd150c-Ubomb-5prime-Unfolded-MKbottomB
30	rd150c-Ubomb-5prime-Unfolded-MKtopA

Table 3. (Continued.)

No	Batch ID
31	rd150c-p4p6-Ubomb-3prime-Unfolded-RDtopC
32	rd150e-p4p6-UBOMB-5prime-2MNaCl-RDbottomA
33	rd150e-p4p6-UBOMB-5prime-2MNaCl-RDbottomB
34	rd150e-p4p6-UBOMB-5prime-2MNaCl-RDtopB
35	rd150j-p4p6-UBOMB-3prime-2MNaCl-RDbottomA
36	rd150j-p4p6-UBOMB-5prime-Native-MKtopA
37	rd150j-p4p6-UBOMB-5prime-Native-MKtopB
38	rd150n-p4p6-Abomb-5prime-Native-MKtopC
39	rd150n-p4p6-Abomb-5prime-Unfolded-MKtopA
40	rd150o-p4p6-UBOMB-5prime-Unfolded-MKbottomB
41	rd151c-p4p6-ABOMB-3prime-2MNaCl-MKtopA
42	rd151d-p4p6-CBOMB-3prime-Native-MKbottomC
43	rd151d-p4p6-CBOMB-3prime-Native-RDtopA
44	rd151d-p4p6-CBOMB-3prime-Unfolded-MKbottomA
45	mk80-Abomb-3prime-Native
46	mk80-Abomb-3prime-Unfold
47	mk81-Abomb-3prime-Native
48	mk81-Cbomb-3prime-Native
49	mk81-Abomb-3prime-Unfold
50	mk82-Ubomb-3prime-2MNaCl
51	mk82-Ubomb-3prime-Native
52	mk82-Ubomb-5prime-Native
53	mk83-Abomb-3prime-Unfold
54	mk83-Cbomb-5prime-Native
55	mk83-Cbomb-5prime-Unfold
56	mk83-Ubomb-3prime-2MNaCl
57	mk84-Abomb-5prime-2MNaCl
58	mk84-Abomb-5prime-Native
59	mk84-Abomb-5prime-Unfold
60	mk85-Abomb-3prime-Native
61	mk85-Abomb-3prime-Unfold
62	mk86-Abomb-5prime-2MNaCl
63	mk86-Abomb-5prime-Native
64	mk86-Abomb-5prime-Unfolded
65	rd138c-Abomb-5prime-Unfold
66	rd148b-p4p6-CBOMB-5prime-unfolded-RDtop
67	rd148b-p4p6-CBOMB-5prime-unfoldedOLDQUENCH-RDbottom
68	rd148c-secondround-p4p6-CBOMB-5prime-unfolded-RDtop
69	rd148c-thirdround-p4p6-CBOMB-5prime-unfolded-MKtop
70	rd148d-p4p6-CBOMB-5prime-2MNaCl-30mincleave-MKbottom-secondround
71	rd148d-p4p6-CBOMB-5prime-2MNaCl-30mincleave-MKtop-thirdround
72	rd148d-p4p6-CBOMB-5prime-2MNaCl-3mincleave-RDtop
73	rd148d-p4p6-CBOMB-5prime-2MNaCl-3mincleave-RDtop-secondround
74	rd148d-p4p6-CBOMB-5prime-2MNaCl-30mincleave-MKbottom-firstround
75	rd148e-p4p6-CBOMB-5prime-Native-3mincleave-MKbottom
76	rd148e-p4p6-CBOMB-5prime-Native-3mincleave-MKbottom-secondround
77	rd148e-p4p6-CBOMB-5prime-Native-3mincleave-RDbottom-thirdround
78	rd148e-p4p6-CBOMB-5prime-Native-30mincleave-MKtop-thirdround
79	rd148e-p4p6-CBOMB-5prime-Native-30mincleave-RDtop-secondround

Table 4. The 19 batches from which the "noisy" 494 profiles are derived. These batches are part of the 79 batches listed in the previous two tables.

No	Batch ID
1	rd135c-Ubomb-5prime-Native
2	rd135d-Ubomb-5prime-Unfold
3	rd150d-p4p6-ABOMB-5prime-Unfolded-MKtopB
4	rd150g-p4p6-ABOMB-3prime-2MNaCl-RDbottomB
5	rd150g-p4p6-ABOMB-5prime-2MNaCl-MKtopA
6	rd150g-p4p6-ABOMB-5prime-2MNaCl-MKtopB
7	rd150i-p4p6-ABOMB-5prime-Native-MKbottomC
8	rd150n-p4p6-Abomb-5prime-Unfolded-MKtopA
9	rd150b-Cbomb-5prime-Unfolded-MKbottomA
10	rd150c-p4p6-Ubomb-3prime-Unfolded-RDtopC
11	rd150o-p4p6-UBOMB-5prime-Unfolded-MKbottomB2
12	mk81-Abomb-3prime-Unfold
13	mk83-Ubomb-3prime-2MNaCl
14	mk85-Abomb-3prime-Unfold
15	mk86-Abomb-5prime-Native
16	mk86-Abomb-5prime-Unfolded
17	rd138c-Abomb-5prime-Unfold
18	rd148d-p4p6-CBOMB-5prime-2MNaCl-30mincleave-MKtop-thirdround
19	rd148e-p4p6-CBOMB-5prime-Native-3mincleave-MKbottom

Table 5. The 19 batches that were not used for the experiments.

No	Batch ID
1	rd150d-p4p6-ABOMB-3prime-Unfolded-RDbottomB
2	rd150d-p4p6-ABOMB-3prime-Unfolded-RDtopA
3	rd150n-p4p6-Abomb-5prime-2MNaCl-MKtopB
4	rd150n-p4p6-Abomb-5prime-2MNaCl-RDbottomB
5	rd150b-Cbomb-3prime-Unfolded-RDbottomA
6	rd150h-p4p6-CBOMB-3prime-2MNaCl-MKbottomC
7	rd150h-p4p6-CBOMB-3prime-2MNaCl-RDbottomB
8	rd150l-p4p6-CBOMB-5prime-Unfolded-HighIncorp-RDbottomC
9	rd150c-p4p6-Ubomb-3prime-Unfolded-RDbottomA
10	rd150c-p4p6-Ubomb-3prime-Unfolded-RDtopB
11	rd150e-p4p6-UBOMB-3prime-2MNaCl-MKtopA
12	rd150e-p4p6-UBOMB-3prime-2MNaCl-RDbottomC
13	rd150j-p4p6-UBOMB-3prime-2MNaCl-RDbottomB
14	rd150j-p4p6-UBOMB-3prime-2MNaCl-RDtopC
15	rd151c-p4p6-ABOMB-3prime-2MNaCl-MKtopB
16	rd151c-p4p6-ABOMB-3prime-Native-MKtopC
17	rd151c-p4p6-ABOMB-3prime-Native-RDbottomC
18	rd138a-Ubomb-3prime-2MNaCl
19	rd138a-Ubomb-3prime-Native

4 ENLARGED FIGURE IMAGES

FIGURE 1(A)

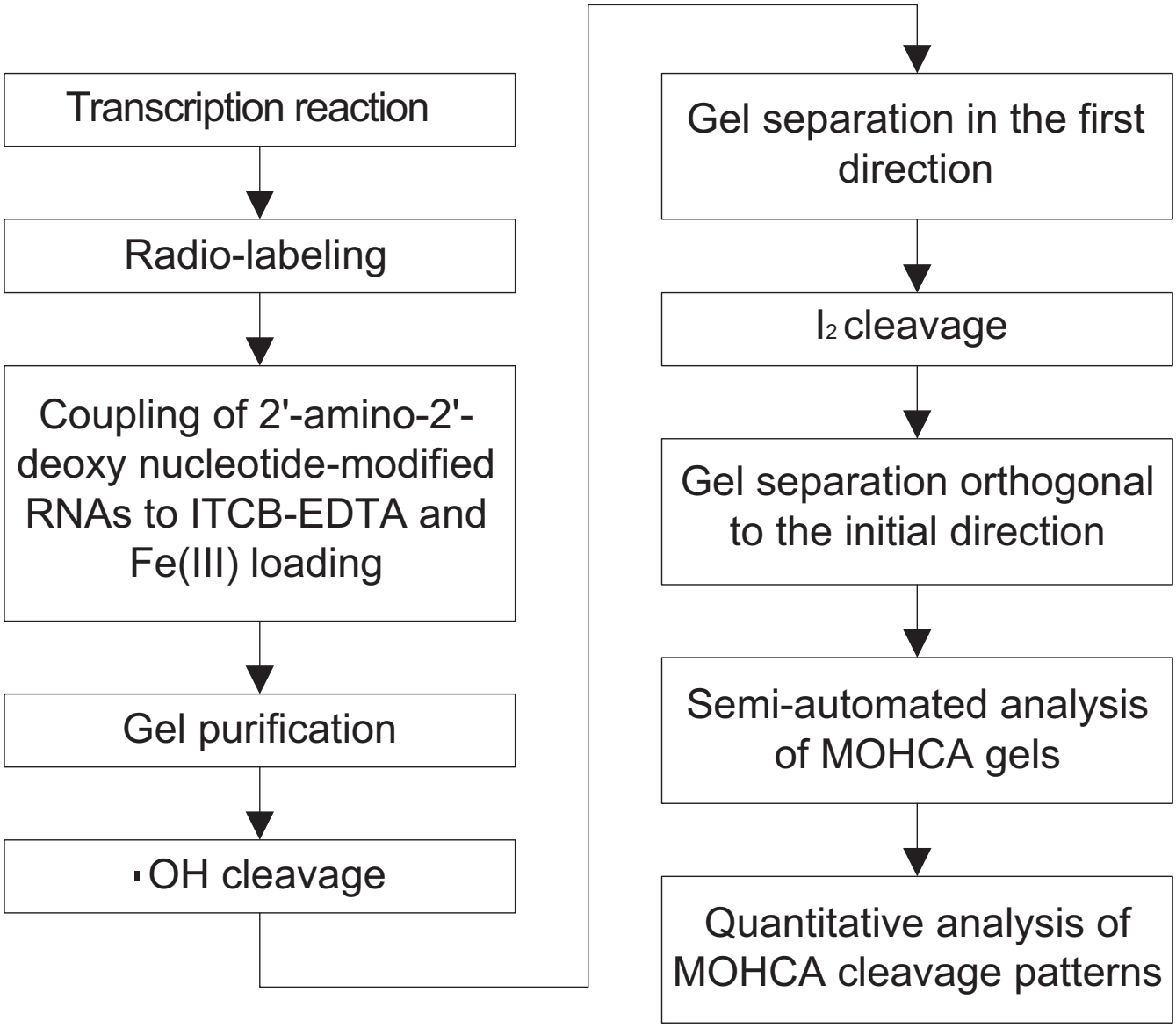


FIGURE 1(B)

FIGURE 1(C)

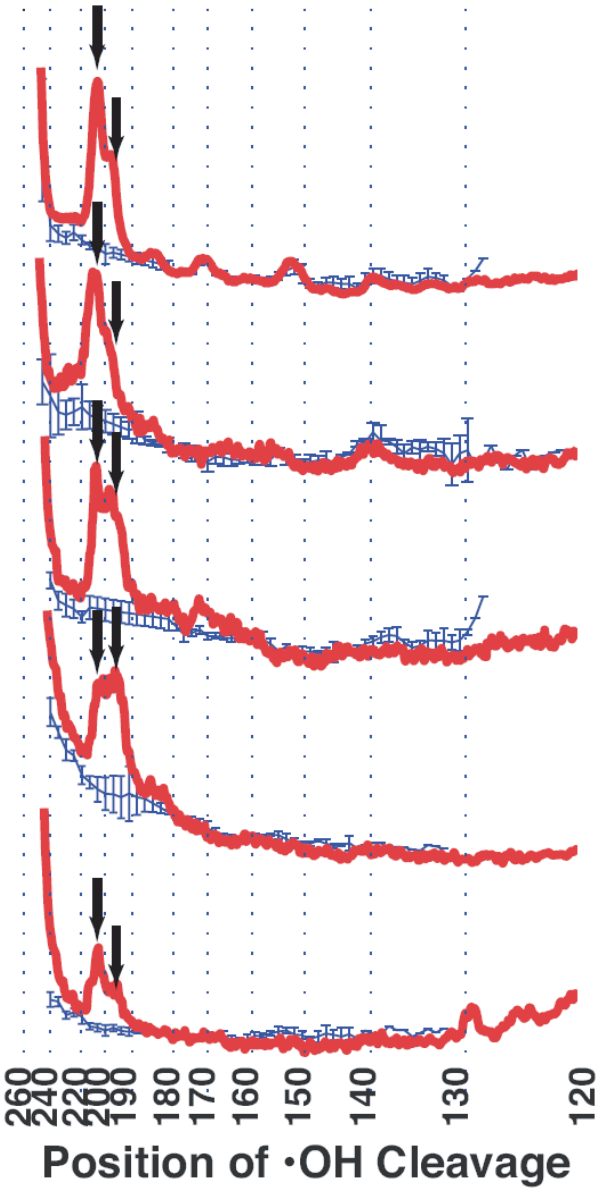

260
240
220
200
190
180
170
160
150
140
130
120
Position of •OH Cleavage

FIGURE 1(D)

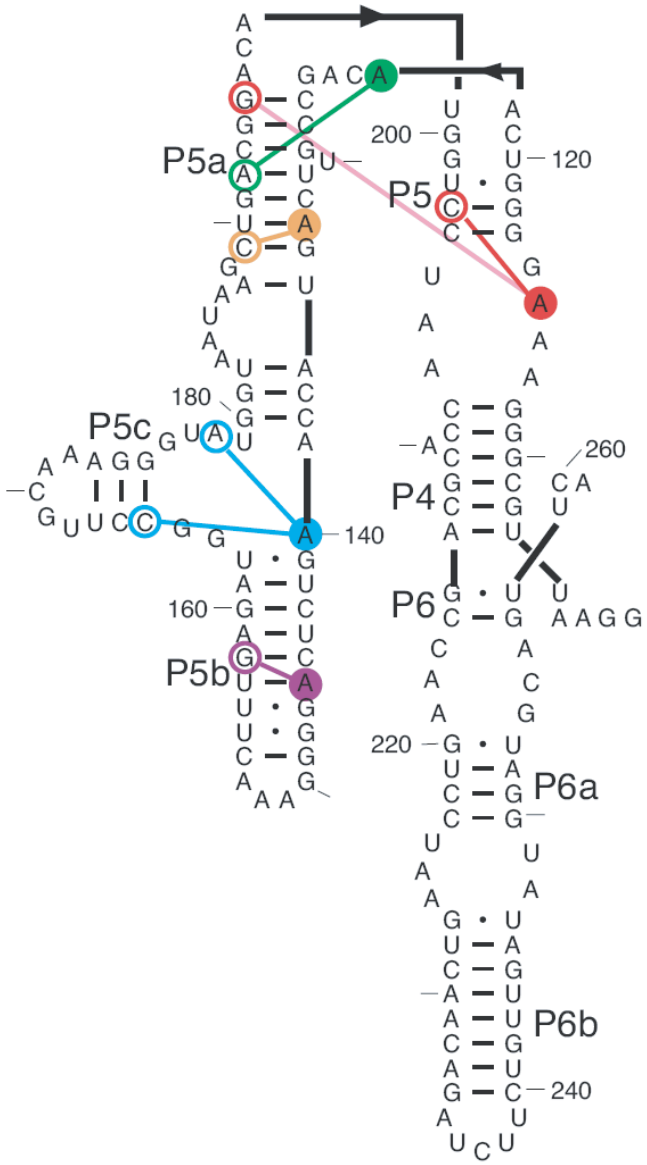


FIGURE 1(E)

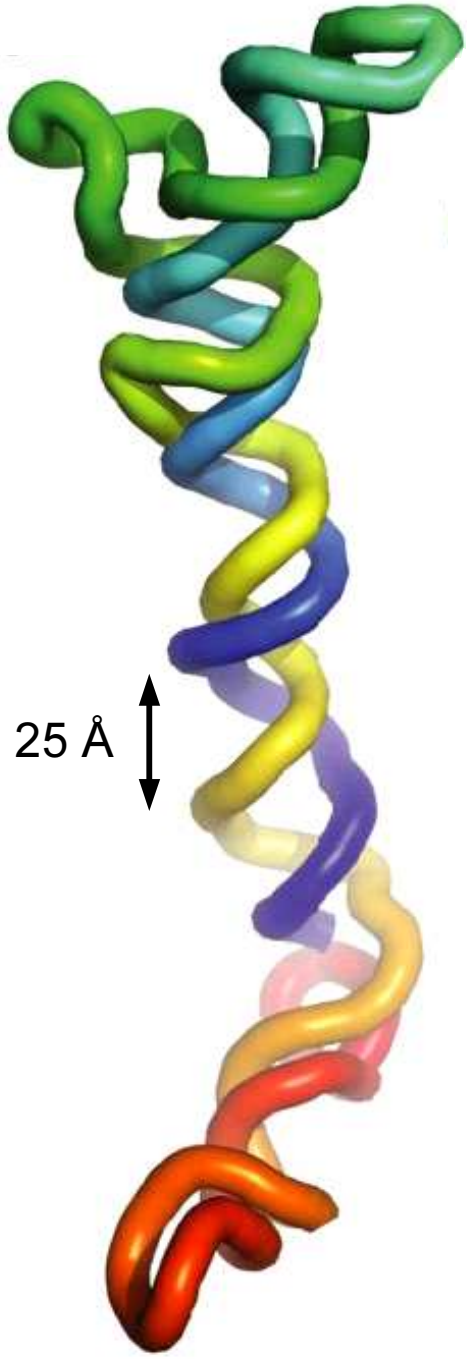


FIGURE 2

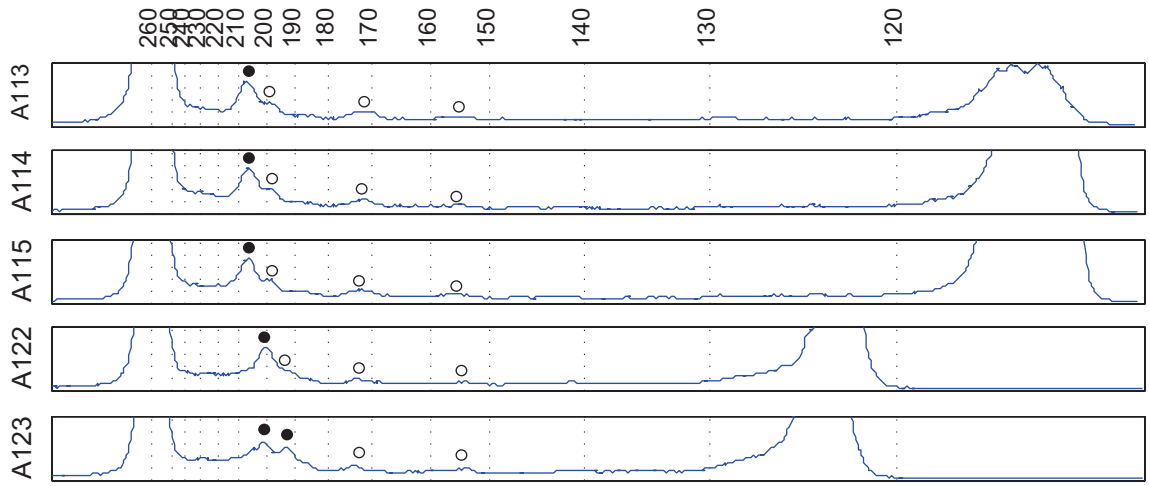

260
250
240
230
220
210
200
190
180
170
160
150
140
130
120
A113
A114
A115
A122
A123

FIGURE 3

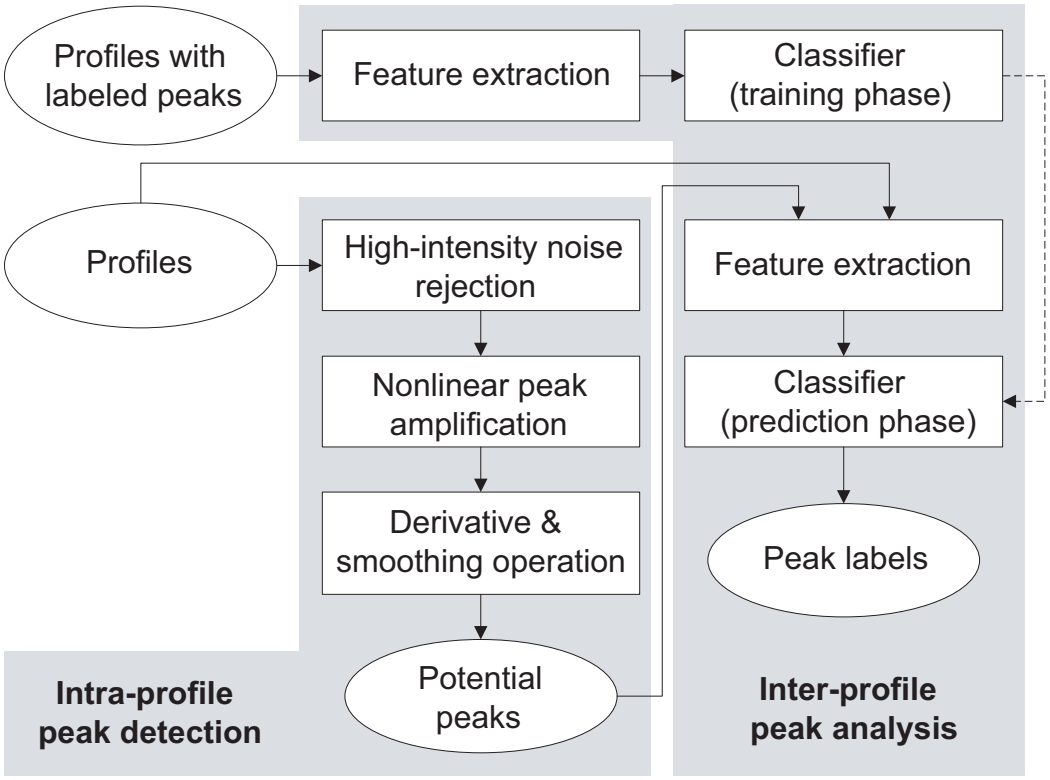

Profiles with labeled peaks
Feature extraction
Classifier (training phase)
Profiles
High-intensity noise rejection
Nonlinear peak amplification
Derivative & smoothing operation
Potential peaks
Feature extraction
Classifier (prediction phase)
Peak labels
Intra-profile peak detection
Inter-profile peak analysis

FIGURE 4

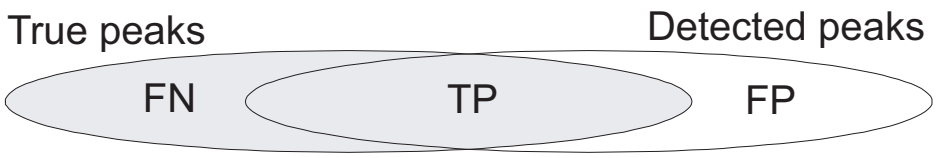


FIGURE 5(A)

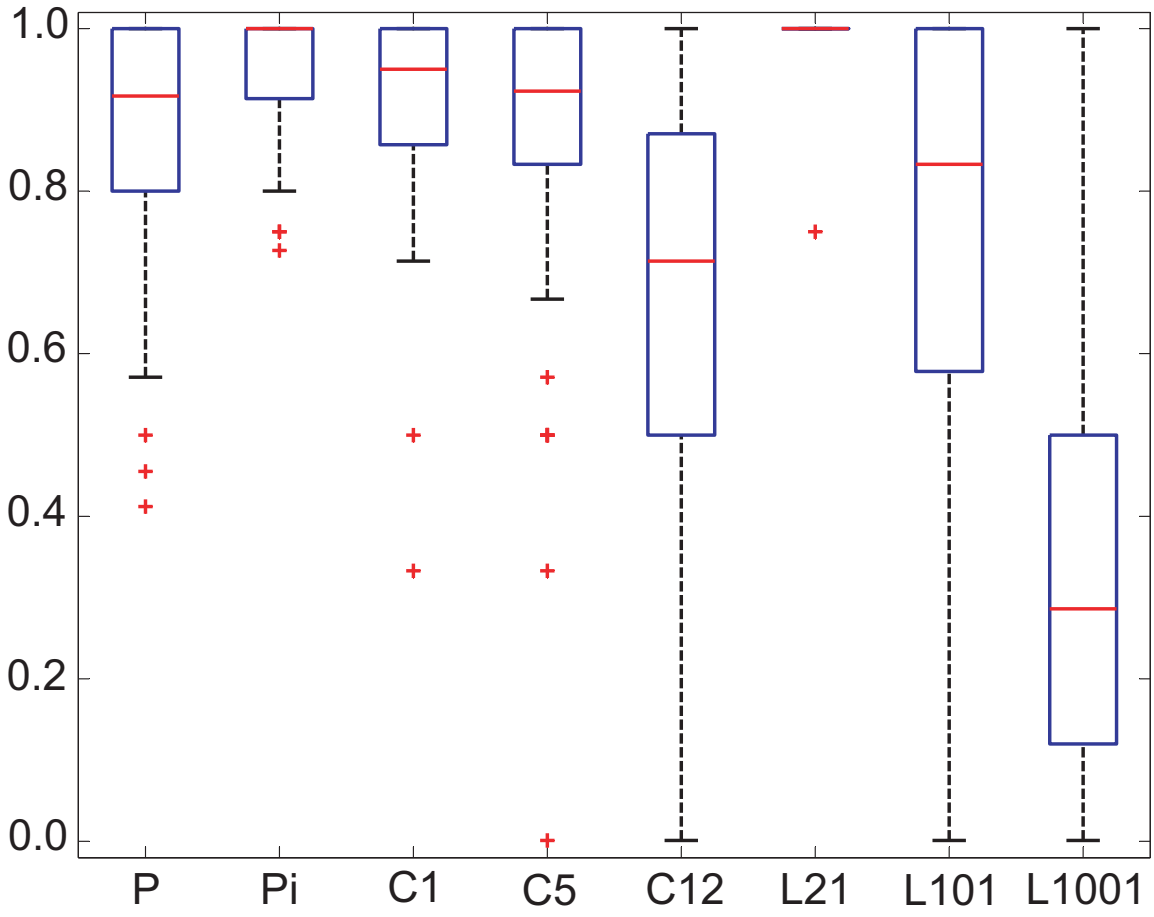


FIGURE 5(B)

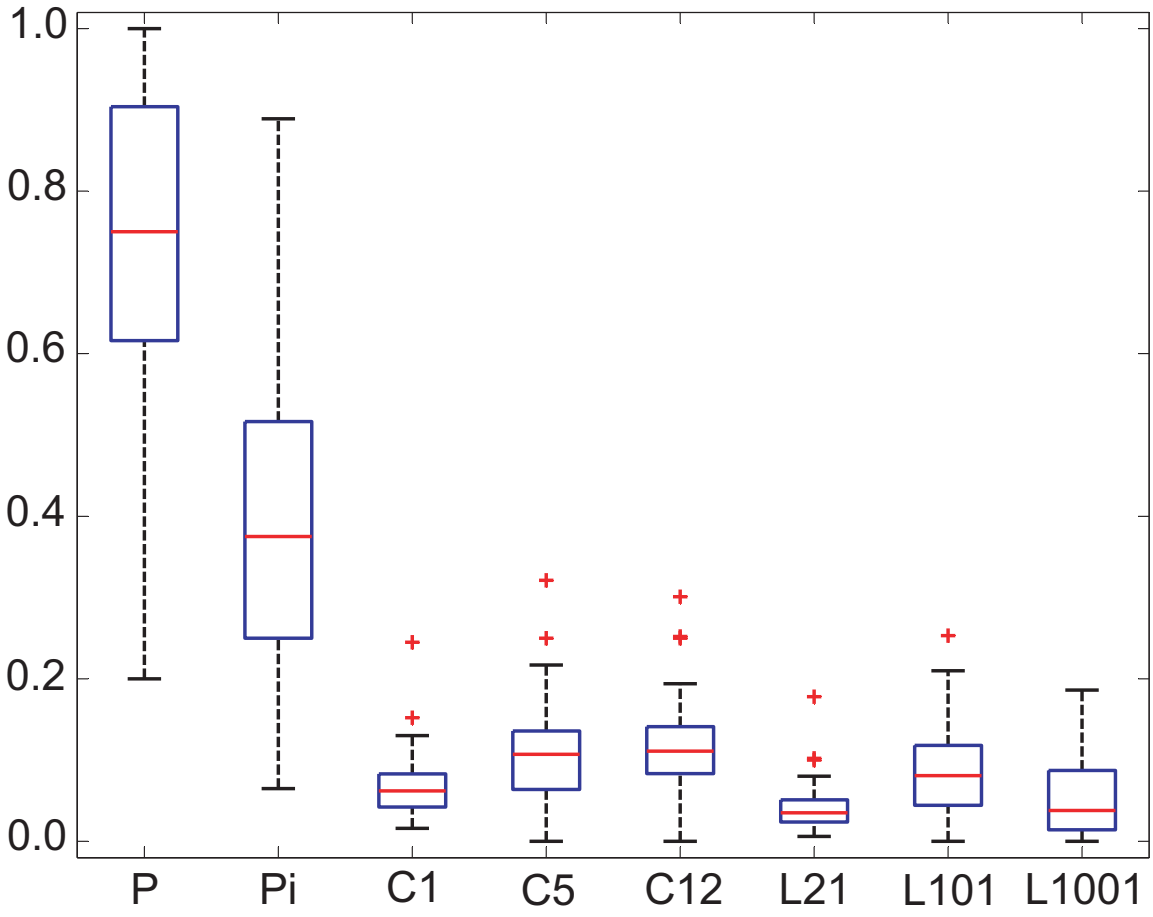


FIGURE 5(C)

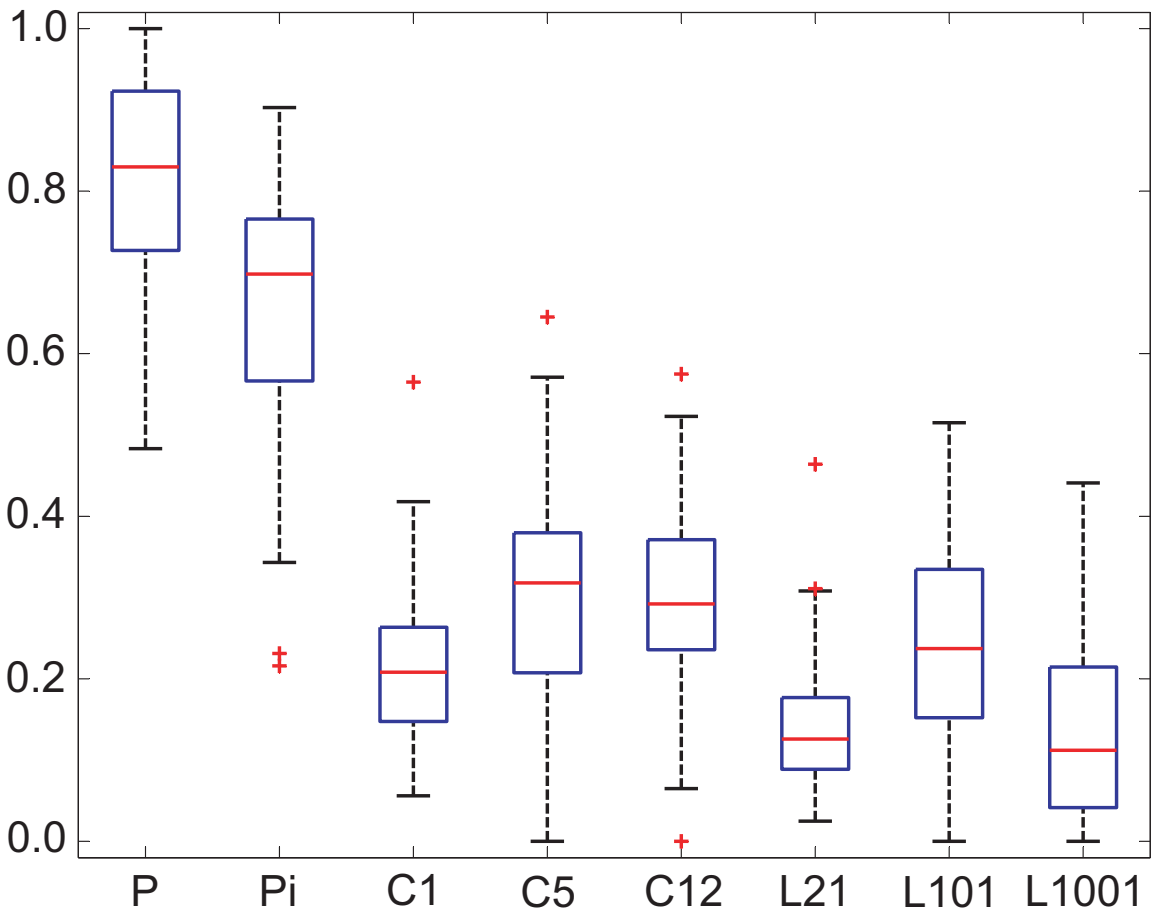


FIGURE 6

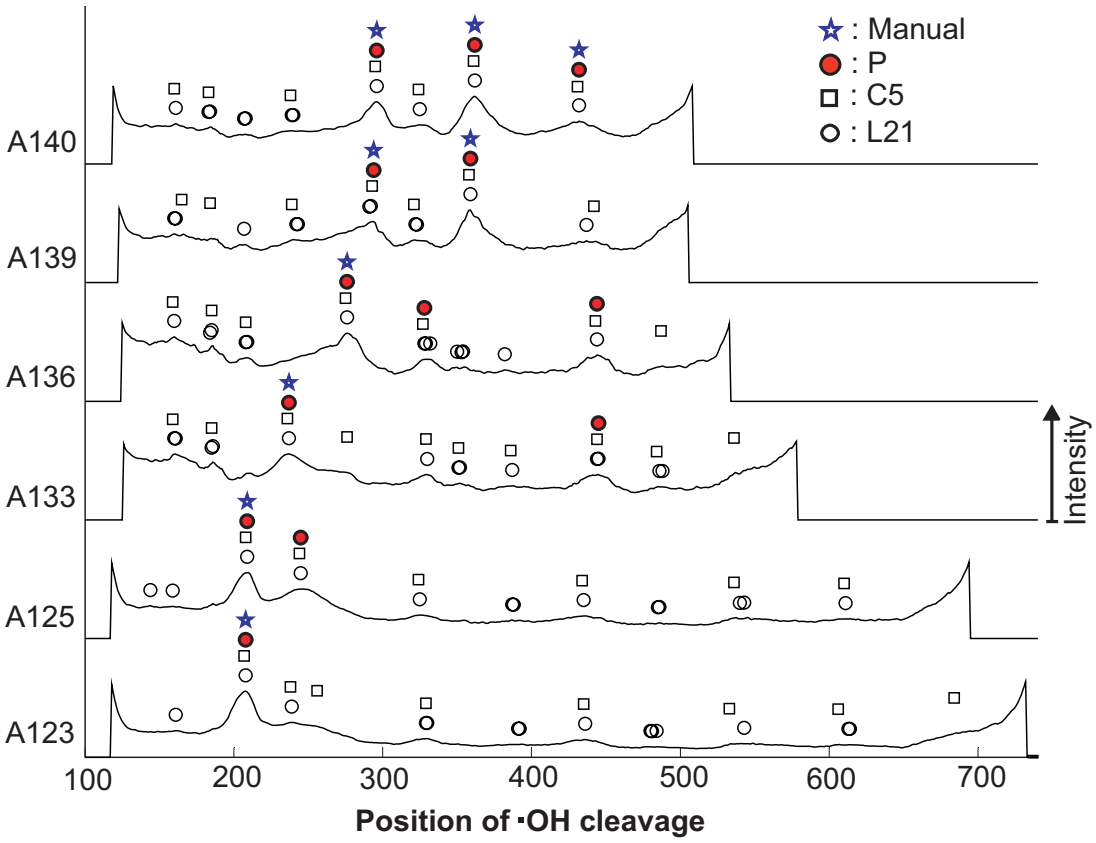

: Manual
: P
: C5
: L21
A140
A139
A136
A133
A125
A123
Intensity
100
200
300
400
500
600
700
Position of ·OH cleavage

FIGURE 7

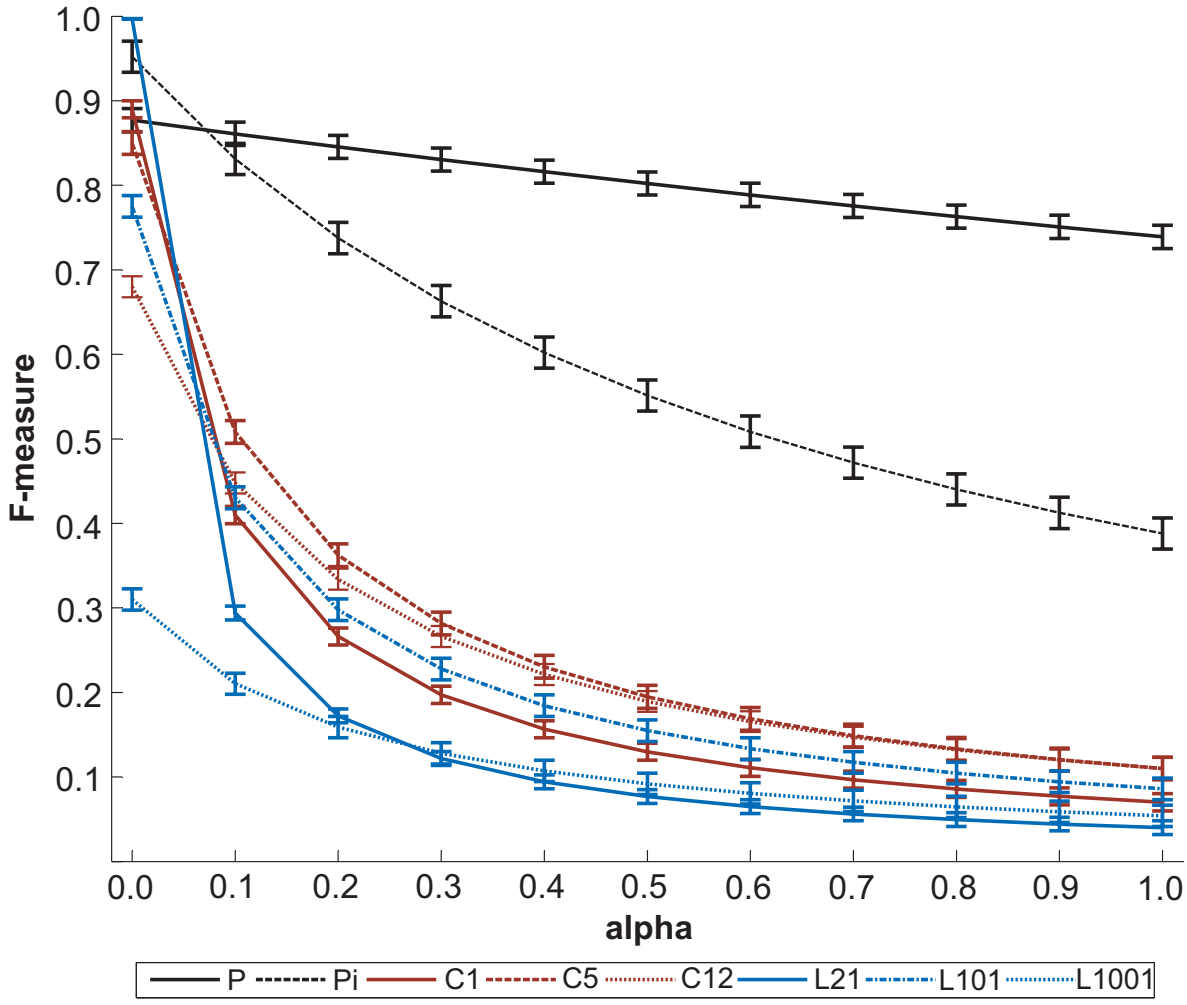

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
F-measure
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
alpha
P
Pi
C1
C5
C12
L21
L101
L1001

FIGURE 8(A)

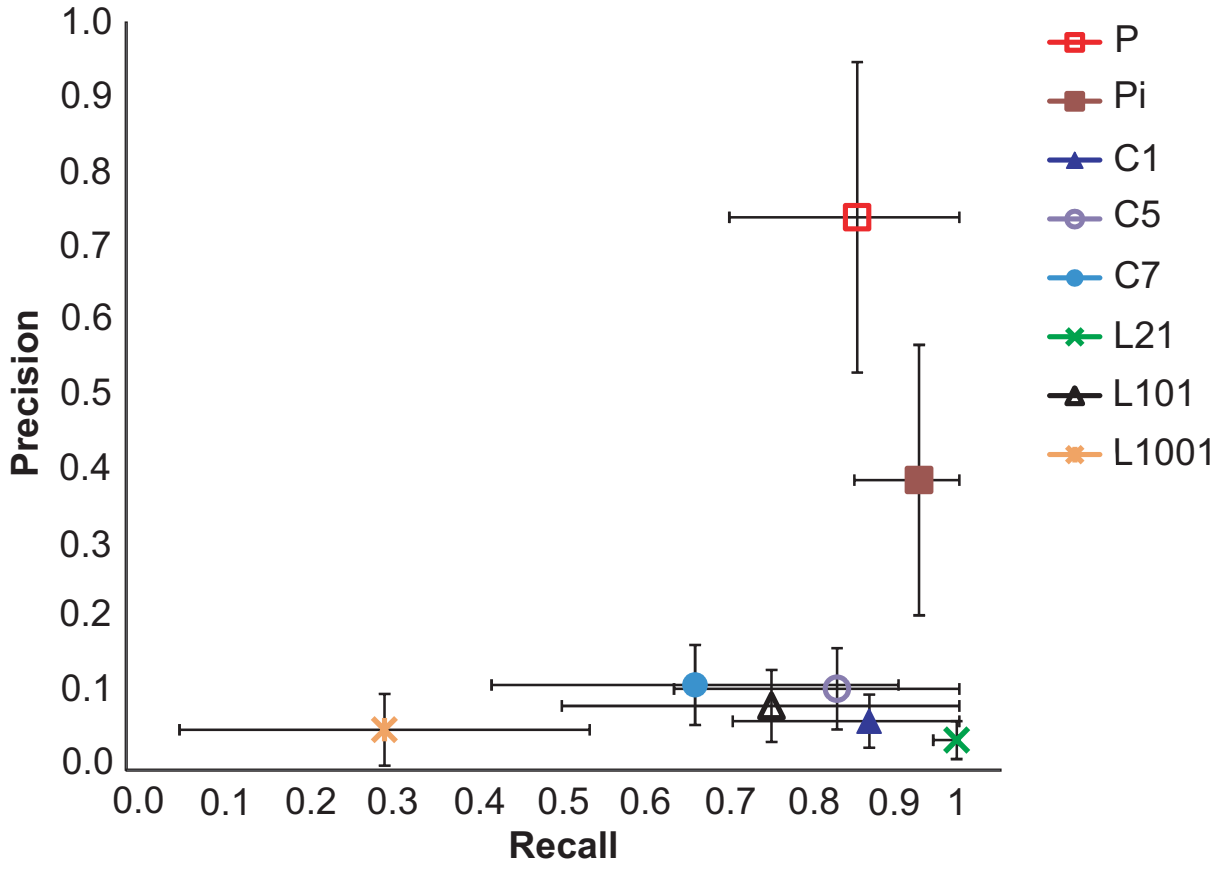

Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall
P
Pi
C1
C5
C7
L21
L101
L1001

FIGURE 8(B)

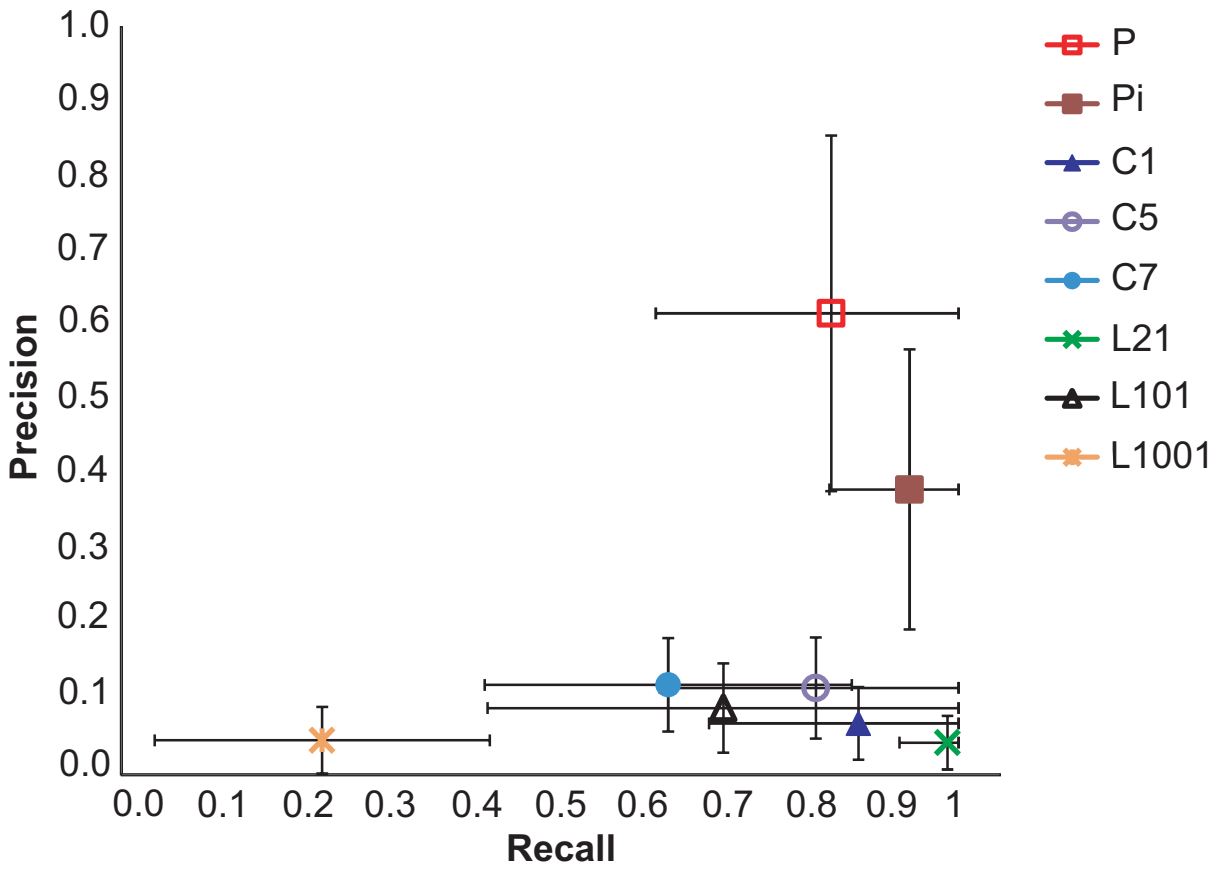

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Precision
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall
P
Pi
C1
C5
C7
L21
L101
L1001

FIGURE 9(A)

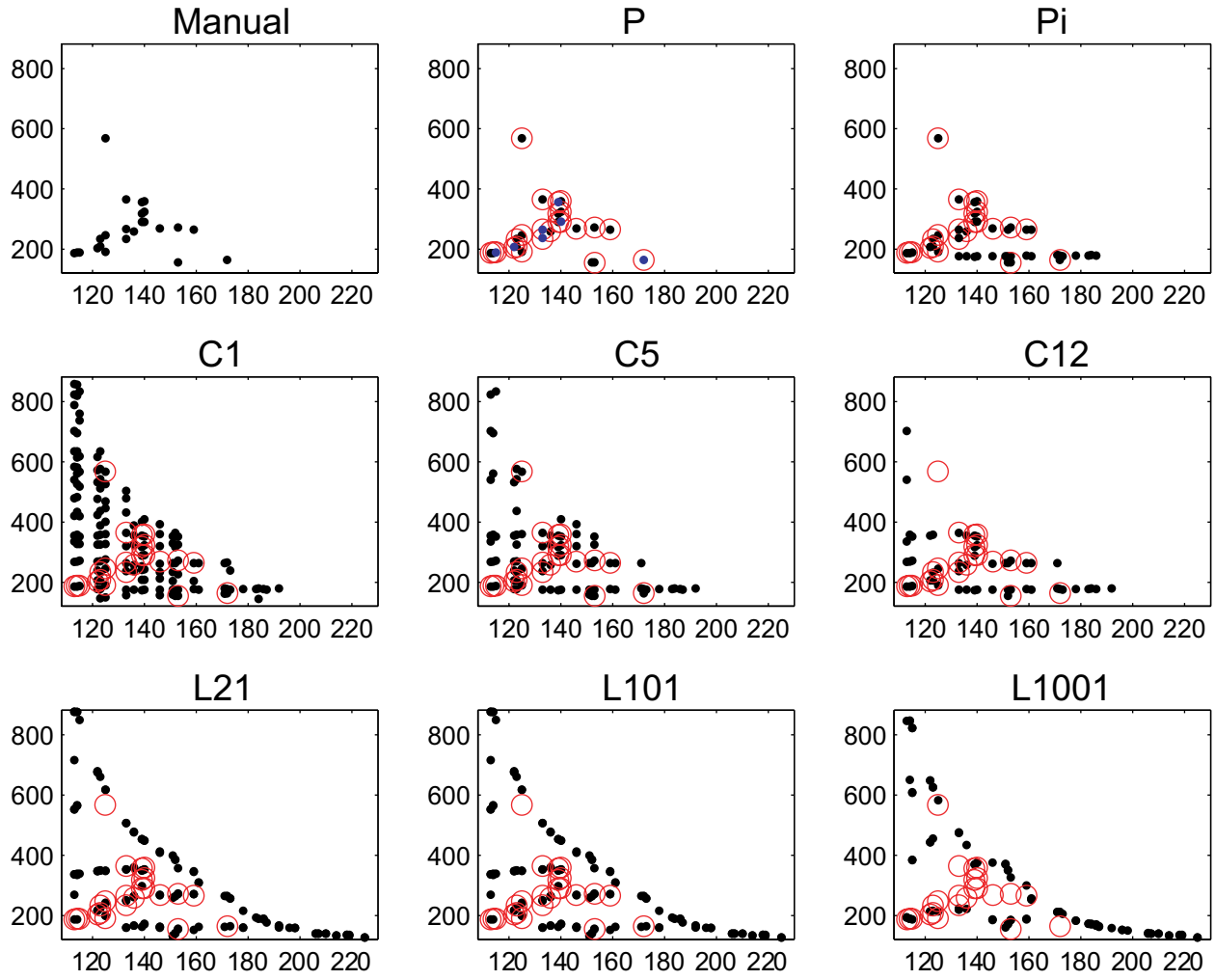

Manual
P
Pi
C1
C5
C12
L21
L101
L1001
800
600
400
200
120 140 160 180 200 220

FIGURE 9(B)

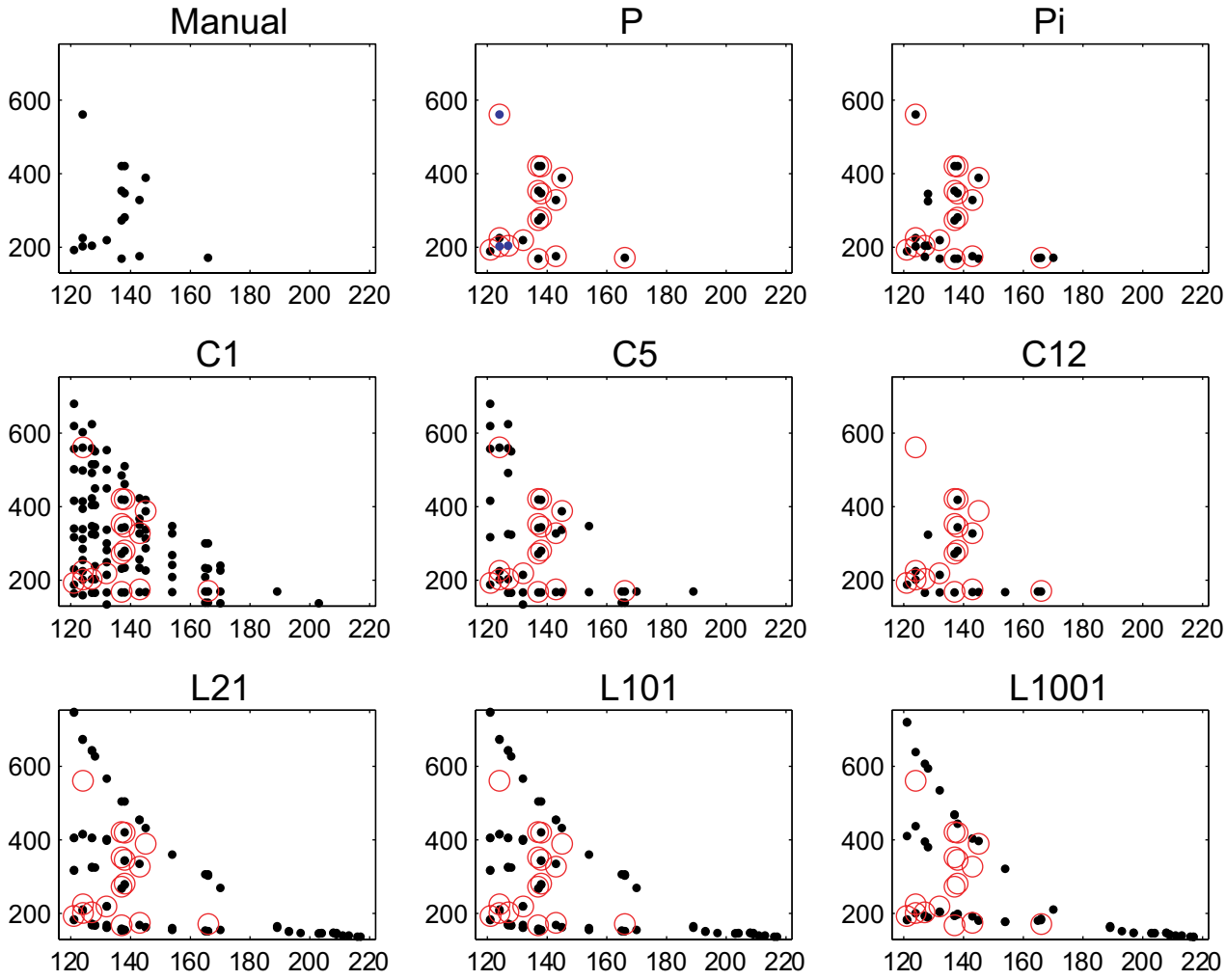

Manual
P
Pi
C1
C5
C12
L21
L101
L1001
600
400
200
120 140 160 180 200 220

FIGURE 10(A)

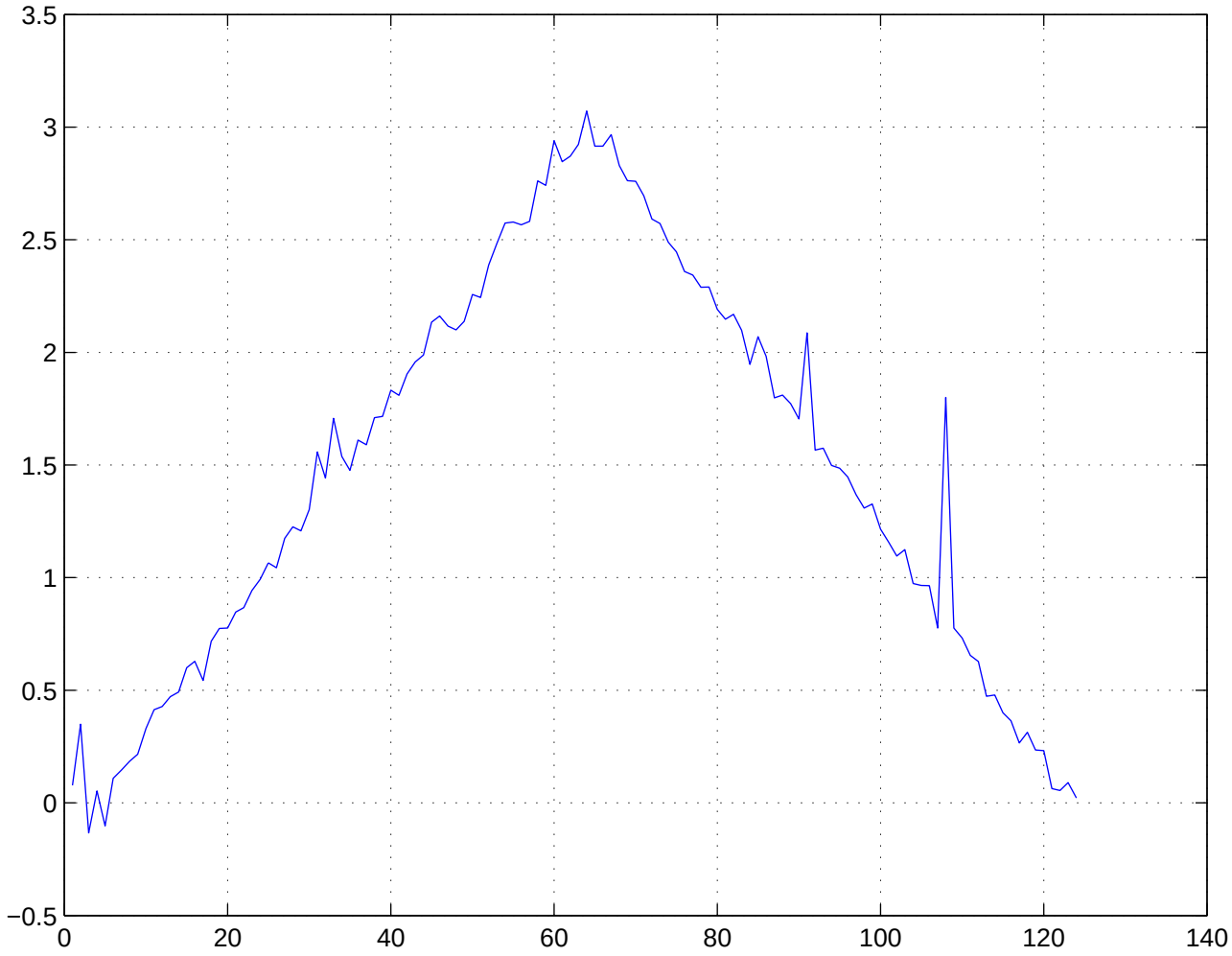


FIGURE 10(B)

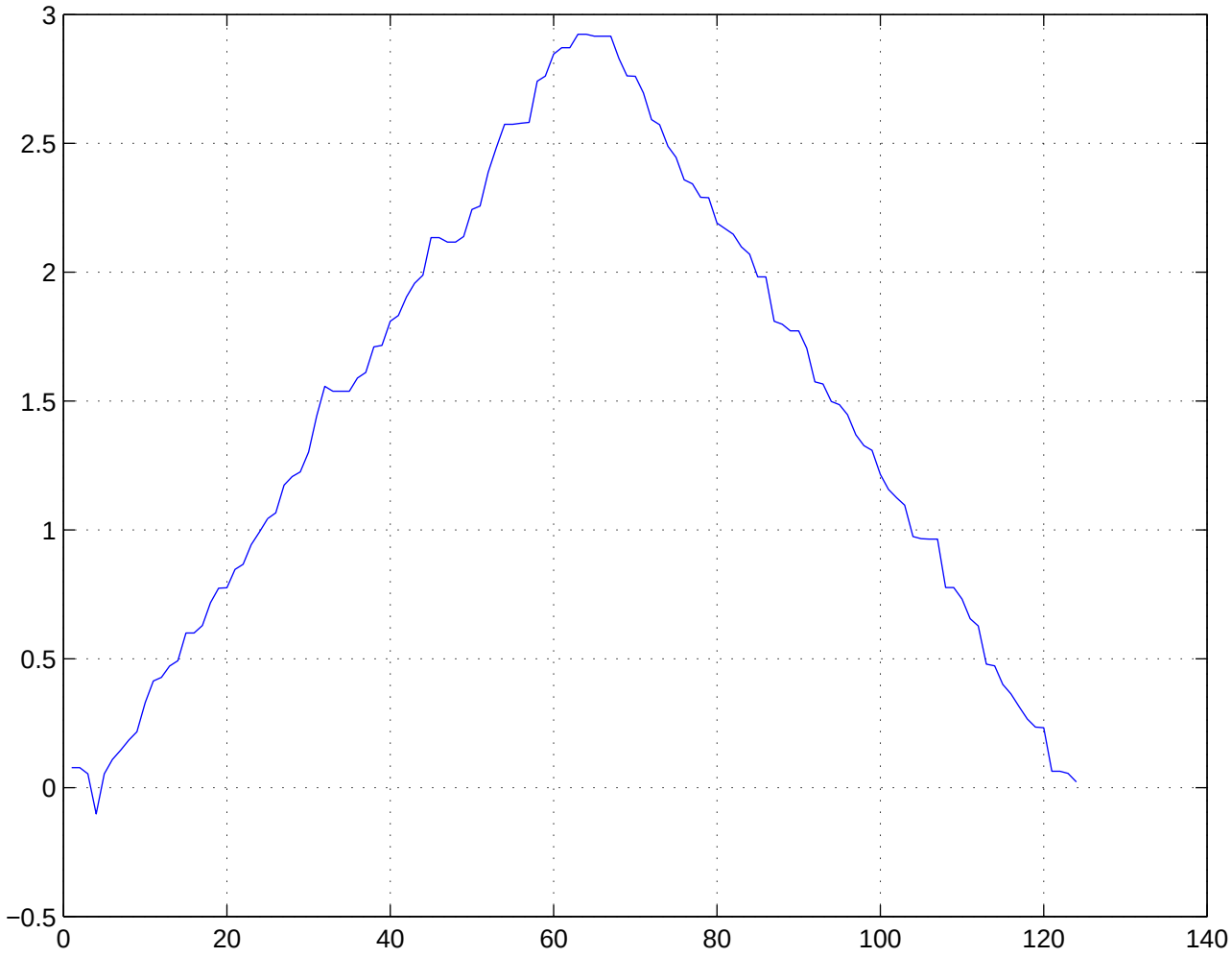


FIGURE 10(C)

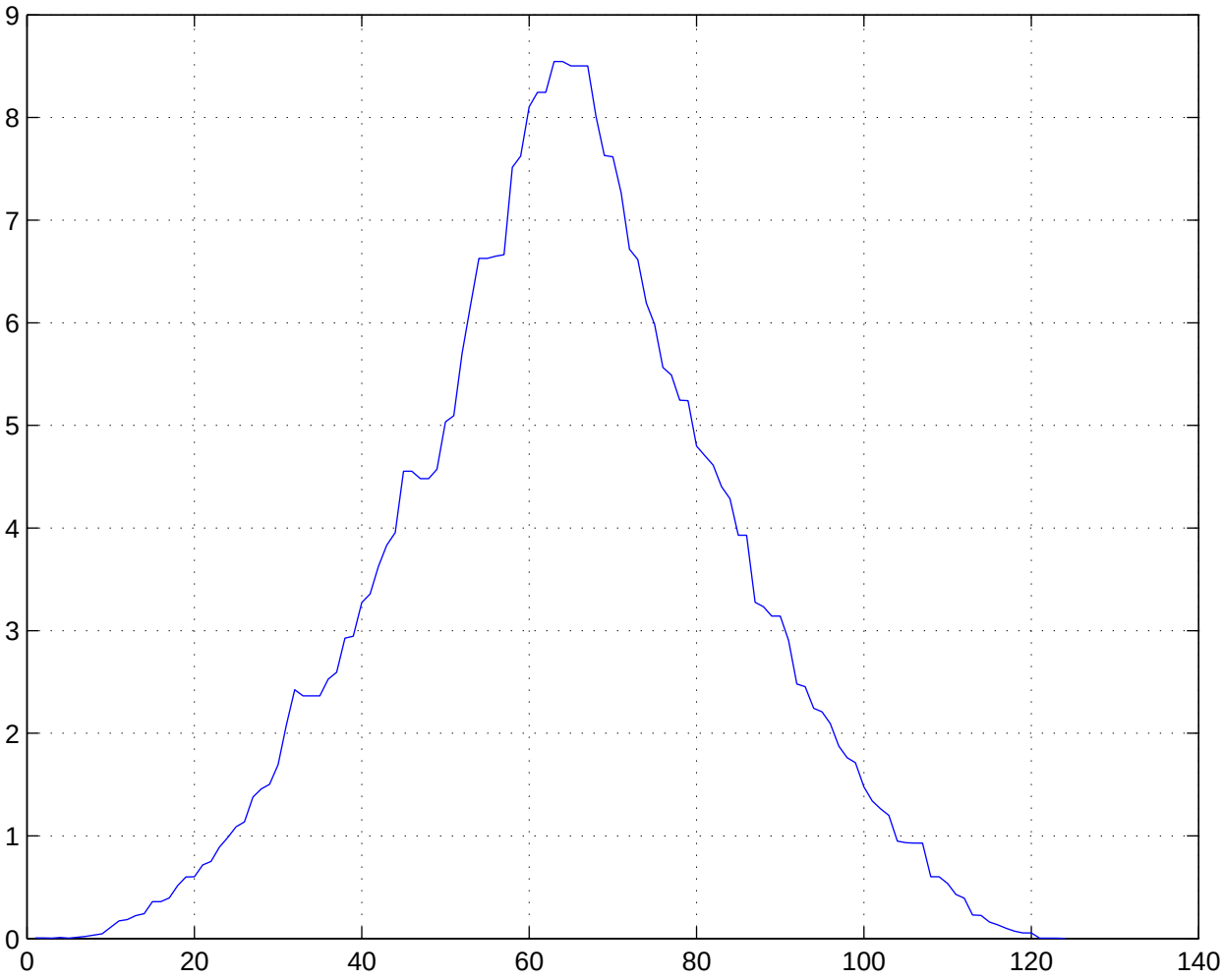


FIGURE 10(D)

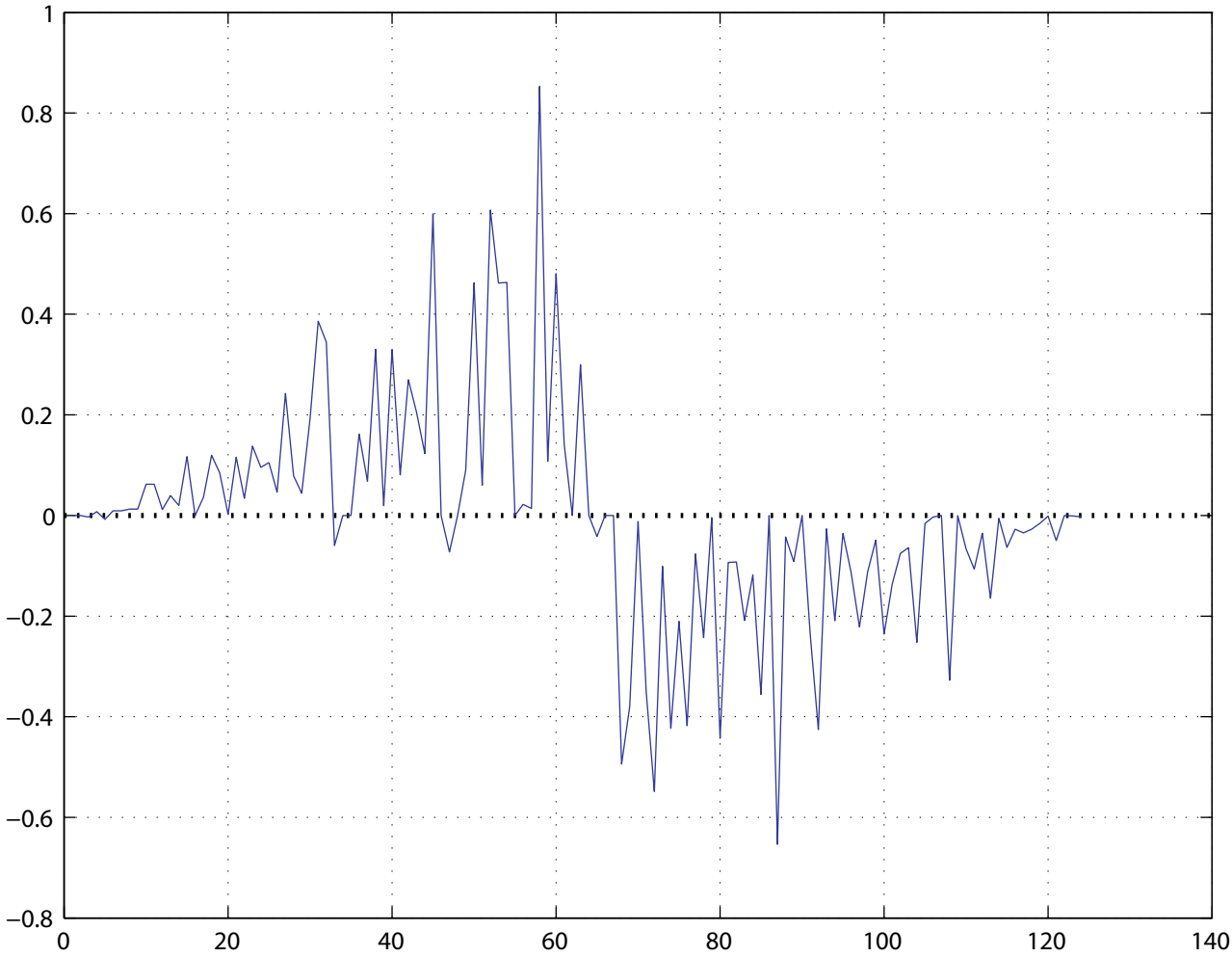


FIGURE 10(E)

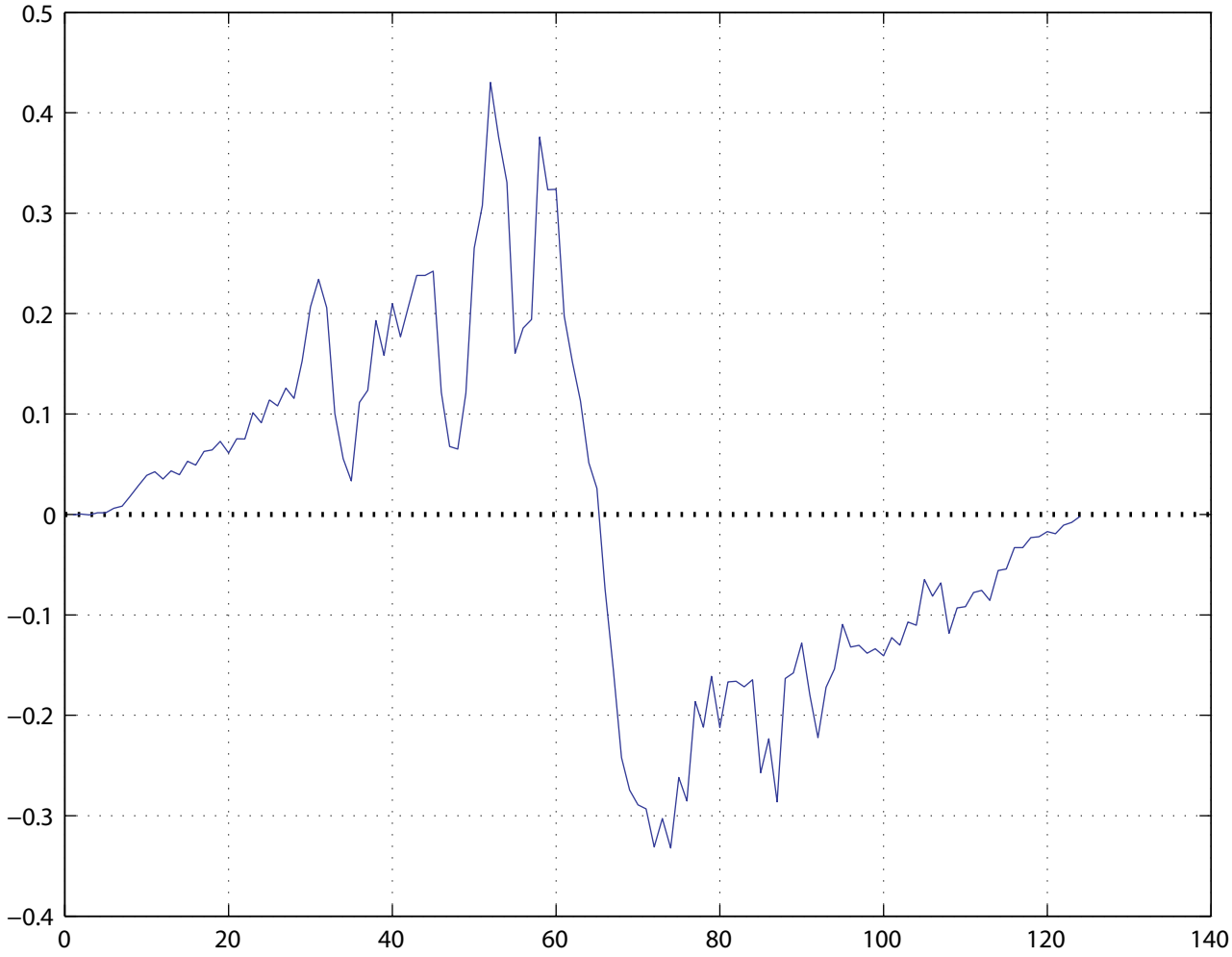


FIGURE 11

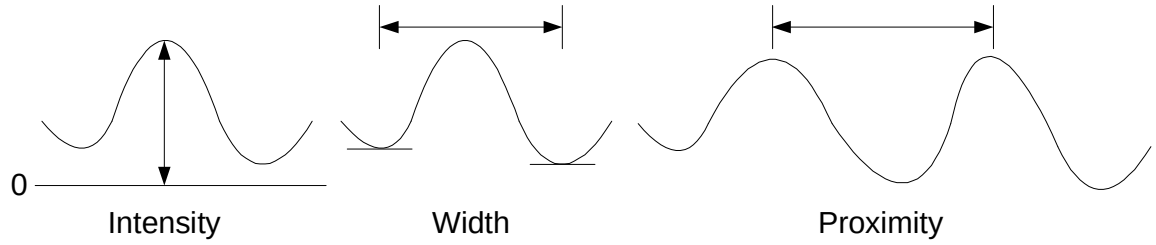
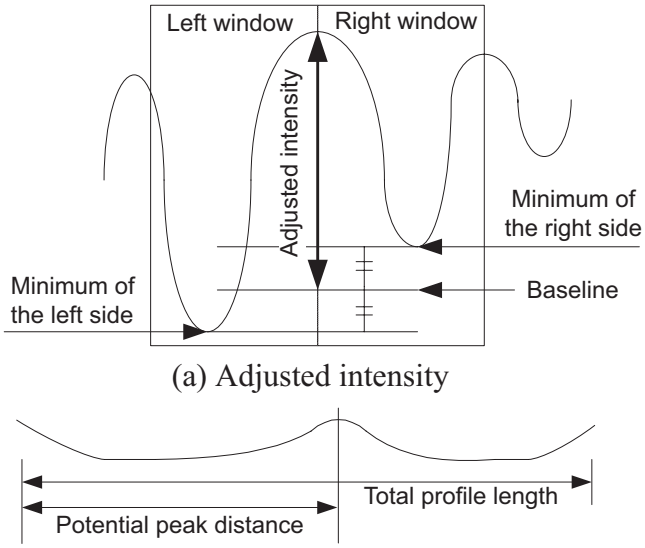
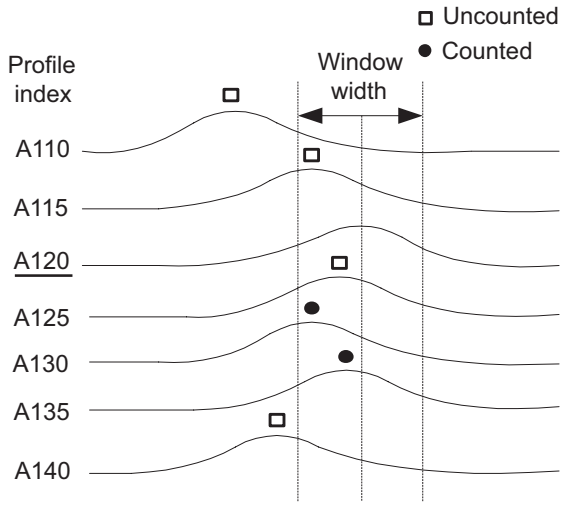


FIGURE 12



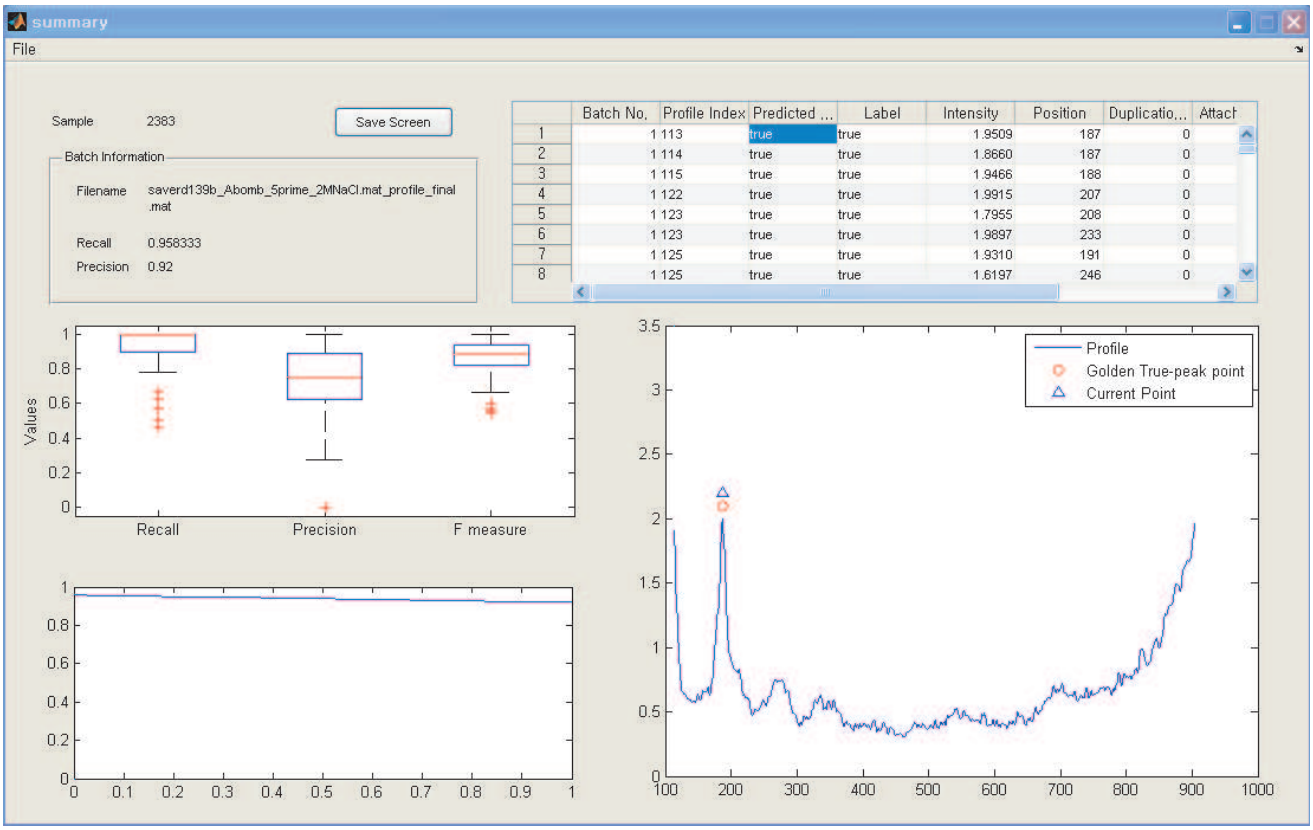
(a) Adjusted intensity

(b) Relative peak location



(c) Number of nearby peaks in other profiles

FIGURE 13


summary
File
Sample 2383
Save Screen
Batch Information
Filename saverd139b_Abomb_5prime_2MNaCl.mat_profile_final.mat
Recall 0.958333
Precision 0.92
Batch No. Profile Index Predicted ... Label Intensity Position Duplicatio... Attact
1 1 113 true true 1.9509 187 0
2 1 114 true true 1.8660 187 0
3 1 115 true true 1.9466 188 0
4 1 122 true true 1.9915 207 0
5 1 123 true true 1.7955 208 0
6 1 123 true true 1.9897 233 0
7 1 125 true true 1.9310 191 0
8 1 125 true true 1.6197 246 0
Values
Recall
Precision
F measure
Profile
Golden True-peak point
Current Point