

Detecting Novel Samples in Mass Spectral Data: A Clustering Approach

Vladimir Svetnik and Andy I. Liaw
Biometrics Research Department
Merck Research Laboratories
Rahway, NJ

September 17, 2001

Abstract

The problem of identifying novel samples from a library of mass spectral data is seen as the problem of identifying outliers in very high dimensional space. We propose a method of identifying such novel samples based on hierarchical clustering. The method produces a measure of outlyingness for each sample in the library, which can be used to choose candidate samples for follow-up. A bootstrap procedure is also presented for assessing the confidence one can have on the outlying samples identified by the method. A synthetic example and a real data set are used to illustrate the proposed methods.

1 Introduction

In the early phase of the drug discovery process, many tasks involve identification of chemical entities that stand out from a large collection, which may be of interest to scientists. This can often be viewed as the statistical problem of outlier identification. In this paper, we present a technique for finding multivariate outliers that was motivated by a project to identify novel fungal extracts from large “libraries”, based on mass spectrometry data of these extracts. For each fungal extract sample, the mass spectrum consists of hundreds of intensity (or abundance) measurements at a sequence of mass to charge ratios (or channels). We can consider the samples as points in a high dimensional space. While it is well known that outlier identification and clustering in such high dimension is extremely difficult, we argue that the nature of the mass spectral data does not lend itself to dimension reduction.

Recently there has been significant attention in the data mining community to the outlier detection problem in large, high dimensional data sets, see for example (Ramaswamy et al., 2000; Bruenig et al., 2000; Guha et al., 2000). Similar to (Guha et al., 2000), our approach utilizes hierarchical clustering to solve this problem. Advantages of this approach are that it places fewer assumptions on the data than many other methods and it can be used with various (dis)similarity measures. Based on the clustering algorithm, we develop a complete procedure for outlier identification. We also propose a bootstrap procedure to assess the “confidence” one would have in the outliers identified by the method.

Here we briefly describe the biological and chemical experiment that motivated our work. Full details of the experiment are presented in (An et al., 2001). Natural products, such as fungi and other microorganisms, are rich source of potential drugs. However, many fungi are difficult, if not impossible, to grow under lab conditions. (An et al., 2001) described a technique in which the genome of a slow-growing fungus is randomly cut into

about 1,000 pieces, each inserted into the genome of an easy-to-grow “host” fungus. This results in 1,000 transgenic fungi, which are as easy to grow as the original host fungi, but may possess some novel characteristics of the donor fungus. To determine the novelty of the transgenic fungi, the extracts from these transgenic fungi are subjected to mass spectrometry. For each fungal extract, we have one mass spectrum consisting of intensity (abundance) measurements at 800 mass to charge ratios. Given this $1,000 \times 800$ data matrix, the task is then to find fungal extracts with “novel” mass spectral profiles. Before we apply the procedure described subsequently, we do some preprocessing to reduce the dimension: All channels (mass to charge ratios) with maximum intensity accross all samples of less than 1,000 units (a detection limit) are deemed “not informative” and eliminated from the data matrix. This usually reduces the number of columns (channels) by as much as 40%.

The outline of the paper is as follows: The proposed procedure for novelty detection is presented in Section 2. The bootstrap procedure for assessing confidence is presented in Section 3. We show how the procedure is used on a toy data set and a real data set in Section 4. Some concluding remarks are given in Section 5.

2 Outlier Detection by Sequential Clustering

The method we propose is based on hierarchical clustering. One potential advantage of this approach is its ability to deal with the situation where outliers as well as “normal” (i.e., non-novel) samples are grouped in multiple clusters. The basic idea is that novel samples should be in small, isolated clusters (possibly having only one member). Novel samples are defined based on the following assumptions:

1. There are at most a fraction, τ , of outliers in the data.
2. These outliers belong to clusters that each contain at most a fraction, ν , of the data.

In other words, a sample must belong to a cluster that contains no more than $\lceil n\nu \rceil$ samples to be considered a potential outlier, and there can be no more than $\lceil n\tau \rceil$ outliers in all. We refer to item 2 above as the ν -criterion, and clusters that satisfy this criterion as ν -clusters. Clearly we expect both τ and ν to be small. We used $\tau = 0.05$ and $\nu = 0.05$ in our analysis, which was based on the scientists’ educated guess.

The outline of the procedure is as follows:

1. Compute the (dis)similarity matrices (we used three: Euclidean distance, Pearson correlation, and rank correlation).
2. For each distance/similarity matrix, perform the following Sequential Clustering Procedure (SCP):
 - (a) Perform agglomerative hierarchical clustering with average linkage.
 - (b) For each sample spectrum, \mathbf{x}_i , find the smallest number of clusters, k_i^{\min} , such that \mathbf{x}_i belongs to a cluster with fewer than $n\nu$ samples.
3. Take the $\lceil n\tau \rceil$ samples with the smallest k_i^{\min} (if there are ties in k_i^{\min} , take all the ties) as the outlying samples.
4. With the three lists of outliers (one for each (dis)similarity), take the union of the three as the final list.

The rationale for this procedure is as follows. There are two important choices when one performs hierarchical clustering that greatly influence the result: what distance measure is

used and how many clusters to divide the data into. The general rule for the first question is that one should choose a measure that is known to be appropriate for the problem at hand, if possible. After discussions with our scientific colleagues, we decided to use the three mentioned above because each of these measures emphasizes distinctive aspects of similarity. For example, Pearson and rank correlation ignore any constant offset and/or scaling between two spectra, while Euclidean distance takes such factors into account. Because the number of outliers on the overall list is the union of the sets of outliers found by SCP for each metric — and thus usually more than that found by each separate measure — the false positive rate is somewhat increased. Because we are generally more concerned about the false negative rate — missing novel spectra — we believe that this is the best tradeoff to make.

As to the second question, a satisfactory choice of number of clusters is a difficult problem (Gordon, 1999), but not of primary interest here. To circumvent this difficulty, we use step 2b described above, instead of fixing the number of clusters.

This procedure utilizes the following monotonicity property of hierarchical clustering methods: Denote by S_{ik} the size (cardinality) of the cluster a sample \mathbf{x}_i belongs to when the number of clusters is k , and by $S_{i\ell}$ when this number is ℓ . Then for any two numbers of clusters k and ℓ such that $k < \ell$, size $S_{i\ell}$ is always less than or equal to S_{ik} , i.e., the size of a cluster to which a sample belongs to is non-increasing with respect to the number of clusters.

According to this property if the number of clusters is varied from $k = 2$ to n , then once a sample \mathbf{x}_i becomes a member of a v -cluster at some $k = k_i^{\min}$, and hence “appears” as an outlier, it will be a member of some v -cluster at all k such that $k > k_i^{\min}$. Another property of hierarchical clustering is that the more outlying a sample \mathbf{x}_i , the smaller k_i^{\min} is. We can thus use k_i^{\min} , $i = 1, \dots, n$, to rank all n samples according to how outlying they are. Suppose that the ranking of k_i^{\min} is done; i.e., $k_{(1)}^{\min} \leq \dots \leq k_{(n)}^{\min}$. Here $k_{(i)}^{\min}$ denotes the i -th position in this ascending sequence. We denote by $\mathbf{x}_{(i)}$ the sample that corresponds to $k_{(i)}^{\min}$; i.e., having the i -th smallest k^{\min} . Since we expect the proportion of novel samples to be no more than $n\tau$, we select the first $n\tau + m$ samples as outliers. Here m is the number of samples in excess of $\lceil n\tau \rceil$ whose ranks are tied with $k_{\lceil n\tau \rceil}^{\min}$. We refer to this set $L = \{\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(\lceil n\tau \rceil + m)}\}$ as the outlier list. We have one such list for each distance measure, and we combine the three lists (by taking the union of the three lists) to form the overall list. Note that since we are only interested in samples with the $\lceil n\tau \rceil + m$ smallest k^{\min} , we do not need to determine k^{\min} for all samples. We can instead stop the procedure at a much smaller k as soon as the $\lceil n\tau \rceil + m$ -th smallest k^{\min} is determined.

3 Outlier Confidence Values

Once the list of novel samples is determined, confidence values of the samples’ novelty can be estimated. This confidence value is an estimate of the probability that a sample would be named as an outlier had the experiment been repeated over many times; i.e., with respect to the population of possible “non-novel” transgenic samples (produced under identical conditions), would the sample in question be singled out as “outlying”? Outliers with high confidence are thus more likely to represent real novelty. Therefore confidence values can be used to prioritize samples for follow up.

Since it is infeasible to repeat the experiment to find out what proportion of the time SCP identifies a sample as an outlier, it is necessary to develop a statistical model on which to base the estimate. We have used the bootstrap method (Efron and Tibshirani, 1993) to do this. The details are as follows. Bootstrap matrices \mathbf{X}_b^* , $b = 1, \dots, B$, are simulated by randomly selecting samples (rows) from the original matrix \mathbf{X} with replacement. The SCP

is applied to each matrix \mathbf{X}_b^* , giving rise to B overall bootstrap outlier lists L_1, \dots, L_B . The bootstrap confidence value estimate λ_i is the ratio $\lambda_i = m_i/M_i$, where m_i is the number of times the sample \mathbf{x}_i is selected as outlier in the B bootstrap SCP, and M_i is the number of bootstrap matrices \mathbf{X}_b^* that contains the sample \mathbf{x}_i . In terms of prioritization, one could choose to follow up on those samples whose confidence values are greater than 0.9, for example.

4 Examples

In this section we illustrate the proposed Sequential Clustering Procedure (SCP) and the bootstrap confidence values with one toy example and one real dataset.

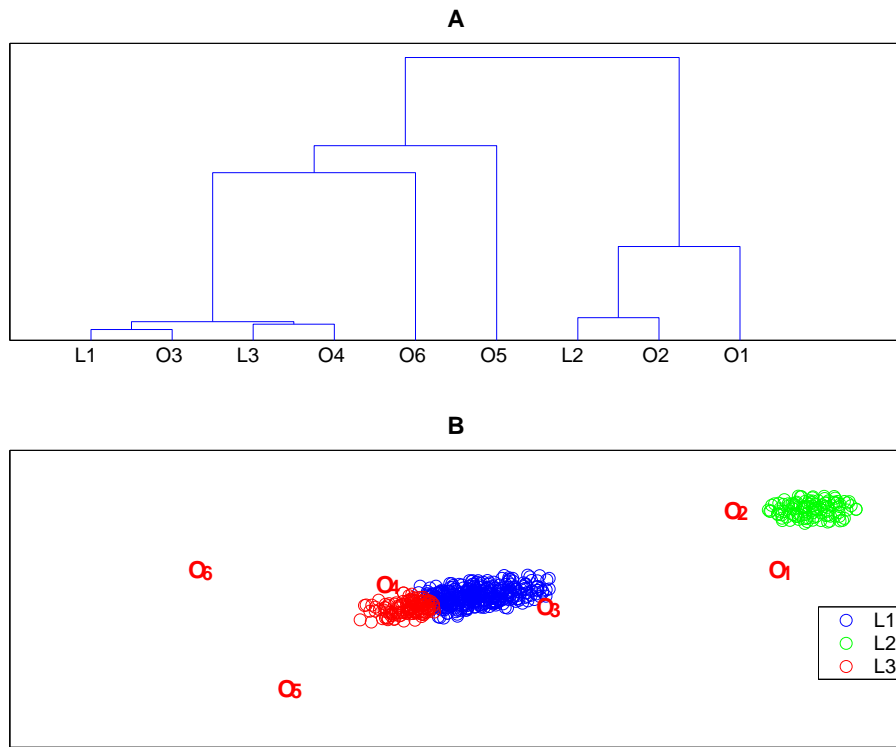


Figure 1: Toy example data with one “large” cluster, one “small” cluster and six outliers. (A) shows the dendrogram (cut at nine clusters) from average linkage hierarchical clustering. (B) shows the data color coded by the cluster membership. L_1 , L_2 and L_3 denotes the clusters containing non-outlying samples.

The data for the toy example is shown in Figure 1B, which consists of two variables. There are two clusters, one about 70% and the second, smaller cluster about 30% of the data. There are six outliers added to the data, labelled as O_1, \dots, O_6 . Figure 1A shows the dendrogram of the data clustered with average linkage and $k = 9$ clusters. Note that the last two merges involve O_5 and O_6 , which are the most outlying points in the data. Note also that in this case, $k_{O_5}^{\min} = 3$, since if we set the number of clusters to 3, O_5 would be in a cluster by itself (and it remains that way for all $k > 3$). Likewise O_6 has $k^{\min} = 4$, since it is in a cluster by itself for all numbers of clusters $k \geq 4$. In this example, we used $\tau = 0.02$

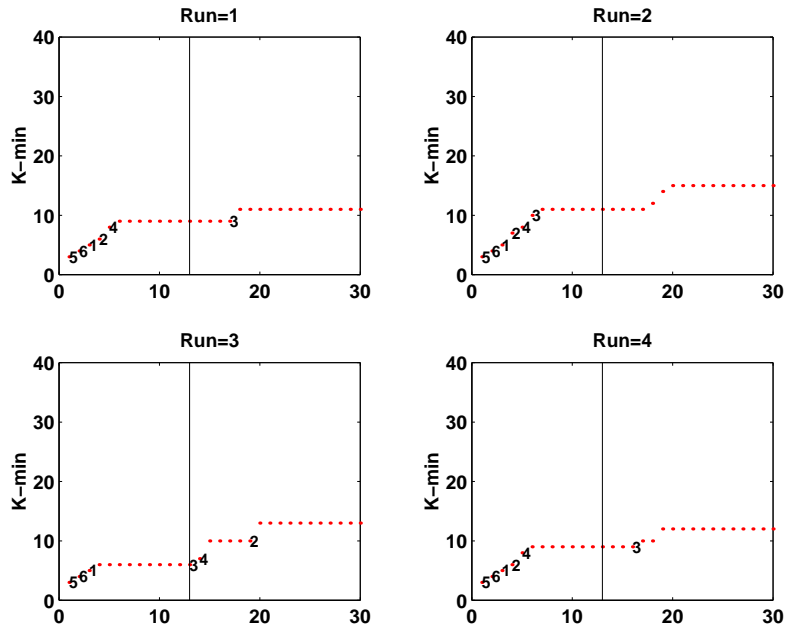


Figure 2: The 30 smallest k^{\min} values of four simulation runs of the toy example, where the six outliers are kept and the two clusters are simulated from bivariate gaussians. Note that the extreme outliers consistently have the lowest k^{\min} values.

and $v = 0.02$.

To illustrate the utility of the bootstrap confidence values, we generated four more sets of data that are similar to the toy example, with the same six outliers, but different realizations of bivariate normal data for the two clusters. Figure 2 shows the 30 smallest k^{\min} values for the four datasets. We see that the more extreme outliers, O_1 , O_5 and O_6 consistently have the smallest k^{\min} , while O_2 , O_3 and O_4 would occasionally have k^{\min} so large that they do not make the list of outliers. We take the original toy example and calculate the bootstrap confidence values (using $B = 100$ bootstrap samples), as described in the previous section. The result is shown in Figure 3. We see that the six outliers all have very high confidence values, but those for O_1 , O_5 and O_6 clearly larger than those of O_2 , O_3 and O_4 . Furthermore, other samples in the dataset all have very low confidence values (near zero).

Next, we show the result of applying SCP to a real dataset. The mass spectra of a library of 764 transgenic fungal extracts are analyzed. The data contain 764 rows and 468 columns (after preprocessing, as described in Section 1). We used three dis/similarity measures: Euclidean distance, Pearson correlation and rank correlation. We used threshold of $\tau = 5\%$ and $v = 5\%$. The combination of the three outlier lists gives a total of 54 outliers. The largest 150 confidence values, based on 500 bootstrap samples, are shown in Figure 4 with the outliers identified by SCP coded by filled circle. As can be seen in the figure, all 54 samples identified have confidence values higher than 90%. Note that there are a handful of samples that were not identified by SCP as outliers, but also have bootstrap confidence values higher than 0.9. These may very well be border-line cases and are worth closer examination. Figure 5 shows the mass spectra of the 8 of the 54 outliers identified. Also shown is the “median” spectrum, formed by taking the median of each column (channel) across all rows (samples). The median spectrum can be seen as a “typical” sample. As

can be seen, most of the samples identified do have distinctly different spectra.

5 Conclusion

In this paper we have presented a procedure for identifying outlying samples in a high dimensional data set. The method utilizes hierarchical clustering, because we wanted the method to work in the situation where “normal” as well as “novel” samples can form clusters, and the novelty of a sample is judged by the number of neighbors it has. A bootstrap procedure is also given for assessing the confidence one can have in the outliers identified.

The proposed method does have its short-comings. It depends quite heavily on the clustering procedure being able to put the data into clusters that are not too different from the “true” clusters. If the clustering algorithm fails to produce grouping that is “close” to the “truth”, the proposed method will likely fail. Also, it requires choice of threshold parameters τ and v , which, if possible, should be derived from existing knowledge of the problem at hand.

In our work we used three (dis)similarity measures in the hope to capture as many aspects of “novelty” as possible. In other applications, one should choose a measure that is known to make sense, if available.

References

- An, Z., Harris, G., Zink, D., Giacobbe, R., Sangari, R., Lu, P., Greene, J., Gerald, B., Meyers, C., Armbruster, J., Smith, S., Svetnik, V., Gunter, B., Liaw, A., Masurarekar, P., Liesch, J., Steven, G., and Strohl, W. (2001). Expression of cosmid-size DNA of slow-growing fungi in *Aspergillus nidulans* for secondary metabolite screening. *Nature Biotechnology* (submitted).
- Bruenig, M. M., Kriegel, H. P., Ng, R. T., and Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceeding of the ACM SIGMOD Conference on Management of Data*.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.
- Gordon, A. D. (1999). *Classification*. Chapman & Hall/CRC.
- Guha, S., Rastogi, R., and Shim, K. (2000). Cure: An efficient clustering algorithm for large databases. In *Proceeding of the ACM SIGMOD Conference on Management of Data*.
- Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceeding of the ACM SIGMOD Conference on Management of Data*.

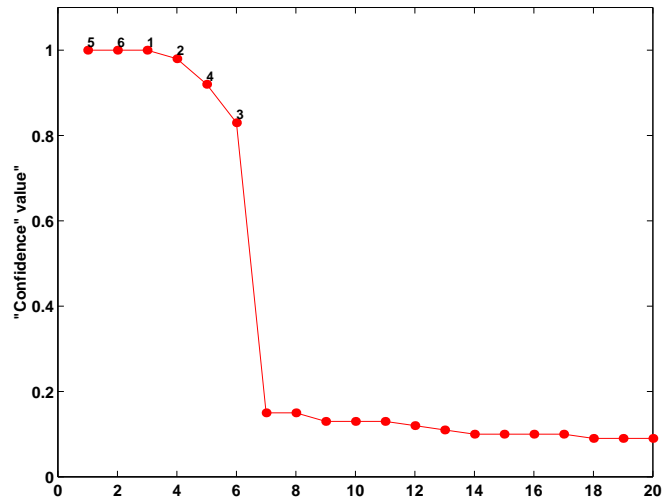


Figure 3: The 20 largest bootstrap confidence values in the toy example. Note the six outliers have the highest confidence values.

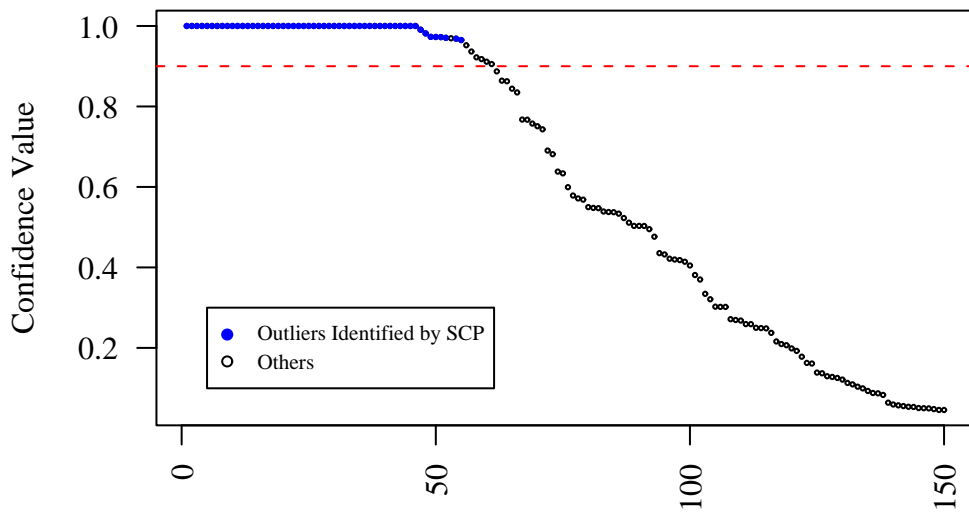


Figure 4: The 150 largest bootstrap confidence values in the transgenic fungal library.

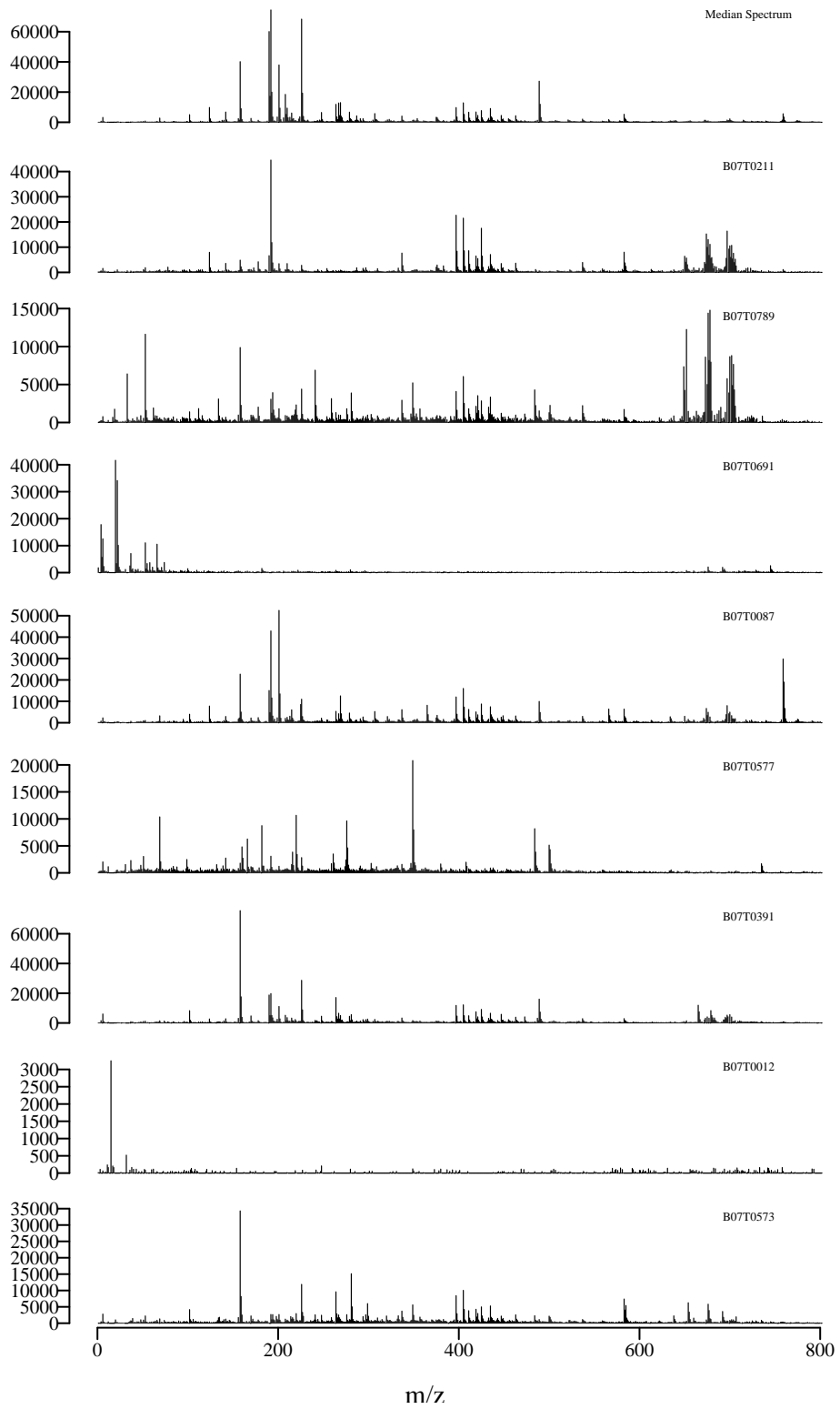


Figure 5: Median spectrum and the spectra of the 8 of the outliers identified by SCP in the transgenic fungal library.