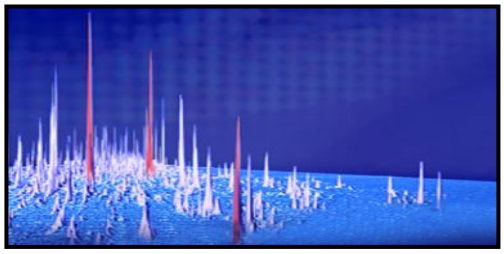
MsCompare™ Univariate and Multivariate Data Analysis Tools:
- A Quick Starting Guide -

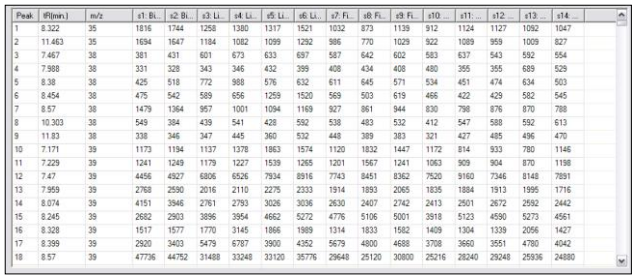


Introduction: this quick starting guide teaches how to find significant and relevant peaks discriminating different groups or classes of samples in your project. The tutorial will focus on results obtained after Peak Picking and/or Peak Matching. See the Quick reference guides on Peak Picking and Peak Matching. This tutorial assumes you are familiar with the basics of MsCompare.

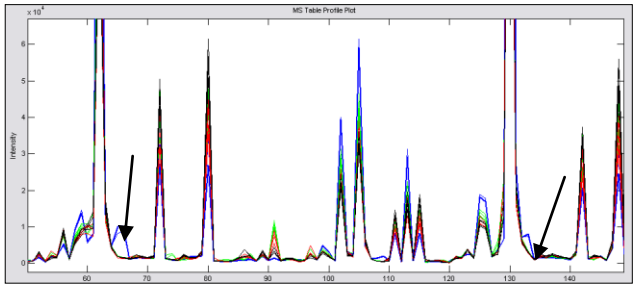
1. **Load or Create a Project:** start MsCompare and load or create a project containing all of your samples. If not done before, you can create classes or groups by selecting the samples from the Sample Listbox and entering the Class Name, next to the Class Label button. To sort the samples, use the Edit Sample List from the Menu. Open the list file (*.blf) and reorder your samples. Save and Exit. You will have to reload the data to see the effects. In MsCompare, class colors are directly related to Trace colors and many of the methods are interactive. The default coloring order of classes is: Blue, Red, Green, Black, Magenta, Cyan, Orange, Purple… , up to 10 classes. The order is related to the alphabetical order of you class names. To set classes with a specific color, the easiest way is to add a number before the Class Name.

2. **Exploratory Data Analysis:** before running Peak Picking or Peak Matching you should have an idea about some specific details of the data. A good starting point is always to explore your samples in the MsCompare. The MsCompare module has many tools to directly interact with your data. Decide if certain artifacts are present, check the alignment of your samples visually, get a feeling of the peak widths, decide at what level peaks are relevant and see if normalization of samples is important, etc. etc.
   Start with PCA (Principal Component Analysis) on the TIC or BPC traces to detect outlying samples. If you already observe nice group separation, you probably have an easy problem in which some of the major peaks are responsible for the differences between the groups.

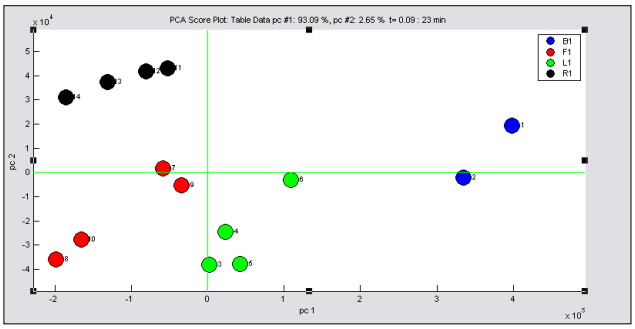
3. **Run Peak Picking or Peak Matching:** See the tutorials on how to perform Peak Picking or Peak Matching. It is assumed that the results (one big table containing all peaks for all samples) is present or can be loaded from disk. When clicked in the table, the EIC's of the selected peak will be plotted to the lower window. You can plot EIC traces or MS spectra at any resolution and automatically zoom in on the peaks of interest.
   MsCompare has many tools to directly explore the table in a graphical interactive way. Table plots are available from the Menu (Table Functions). Example: on the right a small part of the Table Profile Plot is shown, a graphical presentation of all intensities for all samples and all peaks. Unique peaks or peaks responsible for group separation can be directly observed from the color and unique behavior. Clicking in the plot will extract the EIC's for the selected peak.

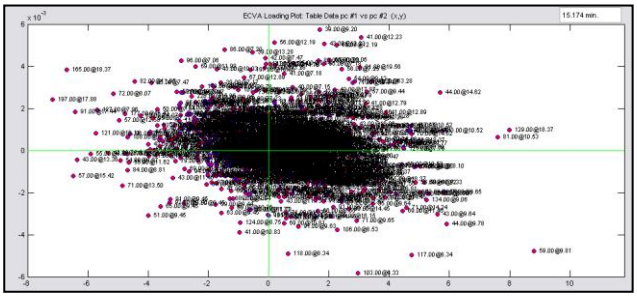


Peak	t[min.]	m/z	s1: Bl..	s2: Bl..	s3: Li..	s4: Li..	s5: Li..	s6: Li..	s7: Fi..	s8: Fi..	s9: Fi..	s10 ..	s11: ..	s12 ..	s13: ..	s14 ..
1	8.322	35	1816	1744	1258	1380	1317	1521	1032	873	1139	912	1124	1127	1092	1047
2	11.463	35	1694	1647	1184	1082	1099	1292	986	770	1029	922	1089	959	1009	827
3	7.467	38	381	431	601	673	633	697	587	642	602	583	637	543	592	554
4	7.988	38	331	328	343	346	432	399	408	434	408	480	355	355	689	529
5	8.38	38	425	518	772	988	576	632	611	645	571	534	451	474	634	503
6	8.454	38	475	542	589	656	1259	1520	569	503	619	466	422	429	582	545
7	8.57	38	1479	1364	957	1001	1094	1169	927	861	944	830	798	876	870	788
8	10.303	38	549	384	439	541	428	592	538	483	532	412	547	588	592	613
9	11.83	38	338	346	347	445	360	532	448	389	383	321	427	485	496	470
10	7.171	39	1173	1194	1137	1378	1863	1574	1120	1832	1447	1172	814	933	780	1146
11	7.229	39	1241	1249	1179	1227	1539	1265	1201	1567	1241	1063	909	904	870	1198
12	7.47	39	4456	4927	6806	6526	7934	8916	7743	8451	8362	7520	9160	7346	8148	7891
13	7.959	39	2768	2590	2016	2110	2275	2333	1914	1893	2065	1835	1884	1913	1995	1716
14	8.074	39	4151	3946	2761	2793	3026	3036	2630	2407	2742	2413	2501	2672	2592	2442
15	8.245	39	2682	2903	3896	3954	4662	5272	4776	5106	5001	3918	5123	4590	5273	4561
16	8.328	39	1517	1577	1770	3145	1866	1989	1314	1833	1582	1409	1304	1339	2056	1427
17	8.399	39	2920	3403	5479	6797	3900	4352	5679	4800	4688	3708	3660	3551	4780	4042
18	8.57	39	47736	44752	31488	33248	33120	35776	29648	25120	30800	25216	28240	29248	25936	24880



4. **Run PCA on the Table:** start exploring the table by running PCA on the peak list (optionally decide on scaling, normalization etc.). PCA is an unsupervised multivariate technique. It does not specifically search for groupings. The score plot on the right already shows a very nice separation, but in more difficult problems this will not be the case. You can check the loading plot to see which peaks are responsible, but is not a very easy task. Often you will only see large peaks sticking out in the loading plot and probably you will check no more than 2 principal components. Even auto-scaling, making all peaks equally important, is often not very easy (see loading plot on the right).





5. **Multivariate Analysis Tools: Two-Class or Multi-Class Problems:**

the Multivariate Tools in MsCompare consist of: PCA, PLS-DA, ECVA and Hierarchical Clustering. **Clustering and PCA** are so-called unsupervised techniques; they do not use class information to find the solution. **PLS and ECVA** are supervised techniques, these explicitly use the class information to find the solution (regression).

MsCompare distinguishes two type of problems, related to the setup of the study: **2-Classes:** you can use the supervised technique PLS-DA (Partial Least Squares Regression) or ECVA for problems consisting of two groups.
Multi Classes: use ECVA (Extended Canonical Variate Analysis), a powerful new technique combining PLS and Linear Discriminant Analysis. From the score plot try to find directions that separate the classes. Then look at the loading plot in the same directions to find the discriminating peaks. Again, often the large peaks stick out in the loading plot.

6. **Univariate Analysis Tools for finding Discriminating Peaks:**

Multivariate techniques in general are variance based, which means that the focus is on the large peaks in your data. Furthermore, it is expected that peaks are highly correlated. In many LC/MS and GC/MS studies the interesting peaks will be very small and the correlation structure with other peaks in your data is missing. In these situations, almost all multivariate techniques will fail, or the interpretation will be very difficult.
We have seen in many studies, that univariate techniques often outperform the multivariate techniques because of the reasons mentioned above. MsCompare has powerful univariate statistics to find your discriminating peaks. We make a distinction between 2-class projects and multi-class projects.

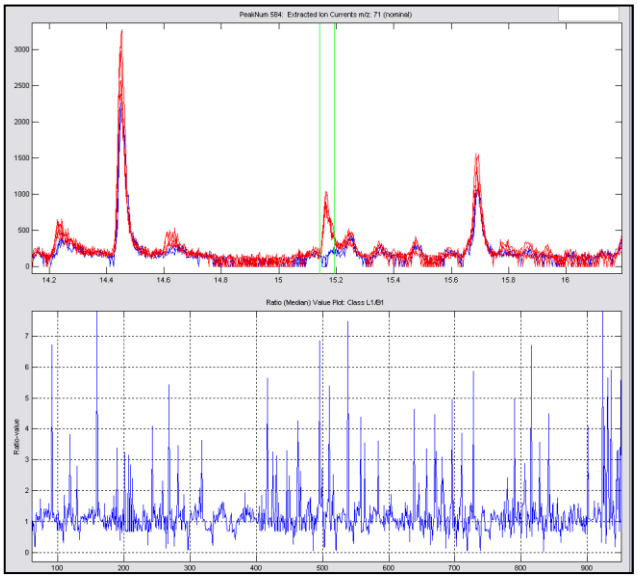
Univariate Statistics – 2 Classes: in MsCompare select from the Menu: Biomarker/Stats > Set Selectivity Rules. You will have to decide which group is expected to contain the up-regulated peaks (some statistics use ratios). Select the option according to the class setup. MsCompare has 7 different statistics for finding discriminating peaks: **ratio, t-test, p-test, uniqueness, full selectivity, % up-regulated and Fisher Discriminant Score.** You can create plots for any of the selected statistics. The plots are interactive, click on a peak and the EIC's or the Profile plots will be generated.

- *Ratio Test:* will calculate the ratio's between the group means or group medians. In one plot you can see both up- and down-regulated peaks.

- *Uniqueness Test:* calculates a value between 100 and -100. The value 100 means unique and up-regulated, -100 means unique and down-regulated. A value of zero means that the group means are equal.

- *Full Selectivity Test:* checks which peaks are larger in one group compared to the other group (must be true for all samples

in the group). The Full Selectivity has a value of 1 or 0.

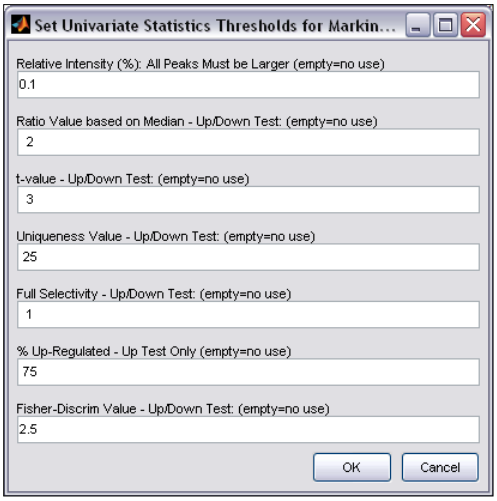


- *Percent Up-Regulated Test:* counts the number of up-regulated samples in your active group, compared to the other group , e.g. for 10 samples in group A, a value of 80% means that for a certain peak, 8 samples are larger and 2 smaller compared to the other group.

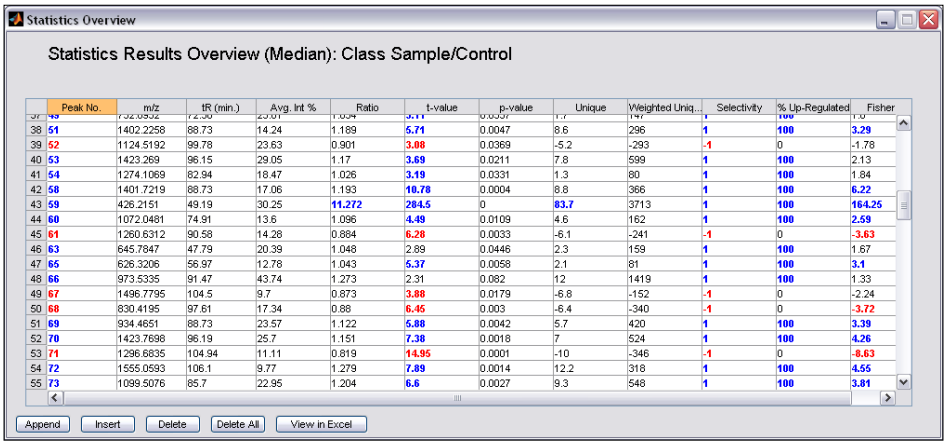
- *Fisher Discriminant Score:* this statistics calculates a value which expresses the difference between the group mean and at the same time takes into account the standard deviation within each group. High values (> 5) have clear separation power and not much spread.

The plot on the right displays part of the ratio graph (lower). Clicking on a peak number will extract the EIC's (top). You can filter (and sort) the full table based on any of the calculated statistics, e.g. keep all peaks in the table having a ratio value larger than 3.0. The above procedure is applicable to multi-group problems too. However, you then should build data sets containing only two groups from the full data set, which is more work.

All Statistics Overview: by selecting this option, you can calculate all the statistics directly. The output will be a table with the calculated statistics for all peaks. Up- and down-regulated peaks will be marked in color. Blue means up-regulated, red means down-regulated. You will have the option to view all peaks or only peaks that are up- or down-regulated. Optionally the full table can be filtered on these peaks. The Overview Table is interactive; clicking an entry will plot the EIC's in MsCompare.
Before the table is generated you will have to decide what is a relevant threshold regarding each of the statistical tests. If any of the tests is positive the peak will pop-up in the table. To use only one type of statistics, clear all other thresholds.



Attention: the combined results include peaks that pass the test for **any** of the individual tests. The test color (up/down) is based on the threshold. To only view peaks that have e.g. a Fisher value > 5, clear all other thresholds.

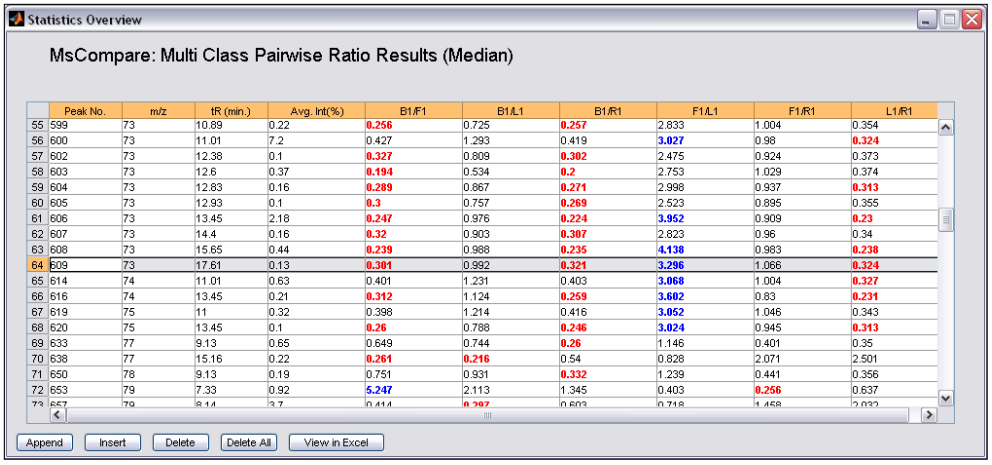


	Peak No.	m/z	tR (min.)	Avg. Int %	Ratio	t-value	p-value	Unique	Weighted Uniq...	Selectivity	% Up-Regulated	Fisher
38	51	1402.2258	88.73	14.24	1.189	5.71	0.0047	8.6	296	1	100	3.29
39	52	1124.5192	99.78	23.63	0.901	3.08	0.0369	-5.2	-293	-1	0	-1.78
40	53	1423.269	96.15	29.05	1.17	3.69	0.0211	7.8	599	1	100	2.13
41	54	1274.1069	82.94	18.47	1.026	3.19	0.0331	1.3	80	1	100	1.84
42	58	1401.7219	88.73	17.06	1.193	10.78	0.0004	8.8	366	1	100	6.22
43	59	426.2151	49.19	30.25	11.272	284.5	0	83.7	3713	1	100	164.26
44	60	1072.0481	74.91	13.6	1.096	4.49	0.0109	4.6	162	1	100	2.59
45	61	1260.6312	90.58	14.28	0.884	6.28	0.0033	-6.1	-241	-1	0	-3.63
46	63	645.7847	47.79	20.39	1.048	2.89	0.0446	2.3	159	1	100	1.67
47	65	626.3206	56.97	12.78	1.043	5.37	0.0058	2.1	81	1	100	3.1
48	66	973.5335	91.47	43.74	1.273	2.31	0.082	12	1419	1	100	1.33
49	67	1496.7795	104.5	9.7	0.873	3.88	0.0179	-6.8	-152	-1	0	-2.24
50	68	830.4195	97.61	17.34	0.88	6.45	0.003	-6.4	-340	-1	0	-3.72
51	69	934.4651	88.73	23.57	1.122	5.88	0.0042	5.7	420	1	100	3.39
52	70	1423.7698	96.19	25.7	1.151	7.38	0.0018	7	524	1	100	4.26
53	71	1296.6835	104.94	11.11	0.819	14.95	0.0001	-10	-346	-1	0	-8.63
54	72	1555.0593	106.1	9.77	1.279	7.89	0.0014	12.2	318	1	100	4.55
55	73	1099.5076	85.7	22.95	1.204	6.6	0.0027	9.3	548	1	100	3.81

Univariate Statistics – More than 2 Classes: MsCompare contains three types of overviews in situations that you have more than 2 classes in your project. One of the tests is the **PairWise Ratio Test**. It will calculate the ratios between all combinations of classes, e.g. for 4 classes A,B,C and D it will calculate ratios between classes A-B, A-C, A-D, B-C, B-D and C-D.

There is no restriction to the number of classes, but the output grows fast. The same test is available for Fisher Discriminant Scores; the **PairWise Fisher Test**. Again, the output is a interactive table containing the test result for the different groups. You can filter the original table so that only up-or down regulated peaks will be left.

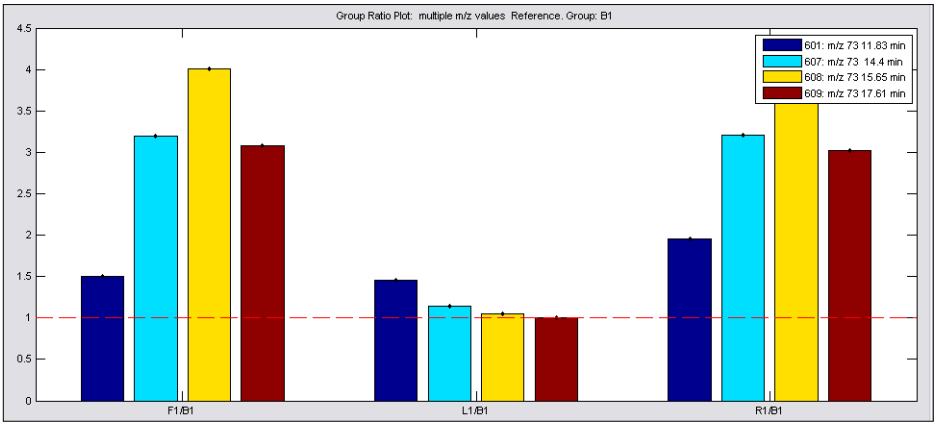
Attention: Please don't use long class names to keep the output compact. Below and example is given for 4 classes: B1, L1, F1 and R1. The peak with number 609 is explored in more detail. For Class F1/L1 this peak is up-regulated and down-regulated when comparing the classes: B1/F1, B1/R1 and L1/R1.



	Peak No.	m/z	tR (min.)	Avg. Int(%)	B1/F1	B1/L1	B1/R1	F1/L1	F1/R1	L1/R1
55	599	73	10.89	0.22	**0.256**	0.725	**0.257**	2.833	1.004	0.354
56	600	73	11.01	7.2	0.427	1.293	0.419	**3.027**	0.98	**0.324**
57	602	73	12.38	0.1	**0.327**	0.809	**0.302**	2.475	0.924	0.373
58	603	73	12.6	0.37	**0.194**	0.534	**0.2**	2.753	1.029	0.374
59	604	73	12.83	0.16	**0.289**	0.867	**0.271**	2.998	0.937	**0.313**
60	605	73	12.93	0.1	**0.3**	0.757	**0.269**	2.523	0.895	0.355
61	606	73	13.45	2.18	**0.247**	0.976	**0.224**	**3.952**	0.909	**0.23**
62	607	73	14.4	0.16	**0.32**	0.903	**0.307**	2.823	0.96	0.34
63	608	73	15.65	0.44	**0.239**	0.988	**0.235**	**4.138**	0.983	**0.238**
64	609	73	17.61	0.13	**0.301**	0.992	**0.321**	**3.296**	1.066	**0.324**
65	614	74	11.01	0.63	0.401	1.231	0.403	**3.068**	1.004	**0.327**
66	616	74	13.45	0.21	**0.312**	1.124	**0.259**	**3.602**	0.83	**0.231**
67	619	75	11	0.32	0.398	1.214	0.416	**3.052**	1.046	0.343
68	620	75	13.45	0.1	**0.26**	0.788	**0.246**	**3.024**	0.945	**0.313**
69	633	77	9.13	0.65	0.649	0.744	**0.26**	1.146	0.401	0.35
70	638	77	15.16	0.22	**0.261**	**0.216**	0.54	0.828	2.071	2.501
71	650	78	9.13	0.19	0.751	0.931	**0.332**	1.239	0.441	0.356
72	653	79	7.33	0.92	**5.247**	2.113	1.345	0.403	**0.256**	0.637
73	657	79	9.14	3.7	0.414	**0.287**	0.603	0.718	1.458	2.032

The last Multi Class Overview Statistic is a so-called **Multi Class Ratio Plot using a Reference Class.** It will calculate the ratios for selected peaks for all groups against a fixed reference group. In the case of 4 classes and class B as the reference class, the following ratios will be calculated. A/B, C/B and D/B. You will be able to specify the reference group. The output will be an interactive table listing the group ratios and a graph of the group ratios for the selected peaks. See the example below.

Four peaks were selected. Class B1 was the Reference Group. The plot shows ratios for the selected peaks between classes F1/B1, L1/B1 and R1/B1. This plot is very useful in situations where you have multiple classes and one reference group e.g. a group of controls.

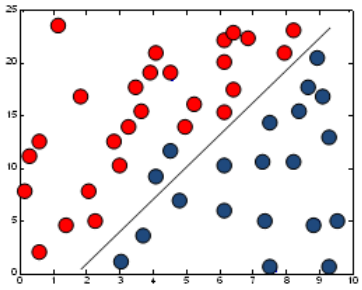


7. If the Problem is Really Multivariate and Peaks are small:

If the solution lies in small peak that have no correlation with larger peaks and these small peaks are not unique or up- or down-regulated, than the multivariate techniques will probably fail, but univariate methods will fail too.

In the example below a scatter plot of two small peaks P1 and P2 is shown for a two class situation. No single peak is able to discriminate between the classes, but **together** they are very discriminative. This is the "real multivariate" power, combining more than a single peak. However, the two peaks are not correlated with the majority of the large peaks in the data set, so they will probably not be detected by PCA or PLS-DA, at least not in the main principal components.

P2



P1

How to proceed? In these cases, use the new **Genetic Optimization Algorithms** to solve the problem. It will search for combinations (2-10) of peaks able to differentiate between the classes. For many peaks, it will be slow, but guaranteed to find the solution!!

Document References:

1. MsX MsCompare - High Resolution Peak Matching QuickRef2013
2. MsX MsCompare - High Resolution Peak Picking QuickRef 2013
3. MsX User Manual