MS data analysis: Preprocessing and Pattern Recognition of the UKCTOCS Data: Pilot Study

CLRC Technical Report
15-11-2005/15-03-2006,
March 2006

A.Gammerman, Z.Luo, I.Nouretdinov, V.Vovk, A.Chervonenkis
Computer Learning Research Centre, Royal Holloway, University of London
and
M.Waterfield, M.Kabir and J.Timms
The Ludwig Insitute for Cancer Research
and
Paul Tempst, John Philip, Josep Villanueva
Memorial Sloan-Kettering Cancer Center, New York
and
U.Menon, A.Rosental, I.Jacobs
University College London and Institute of Women's Health

1 Abstract

Recent advances in analysis of human serum proteome aim to establish novel disease biomarkers that would allow to make early detection of diseases. The current techniques include analysis of the serum data with using mass spectrometry (m/s). The output of m/s work is the large volume of high-dimensional data and it requires modern methods of data analysis. This report describes several machine learning techniques that have been developed in order to establish "proteomic pattern diagnostic". The techniques were applied to a subset of large bank of serum data collected in UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) study over 7 years from 1995-2001. The report first describes the data, and a set of pre-processing techniques that include calibration, subtracting the baseline, normalising, and calibrating the data. Subsequently, peak identifications and peak alignment techniques were applied in the latest stages of pre-processing. Naturally, many factors affect the existence and the position of peaks, but the tube type and transit time taken to store the data are among the most important ones. The report compares different types of the

tubes and the transit time to establish potential problems caused by these factors. The main part of the report devoted to pattern recogniton techniques that have been applied to classify between the cases and control samples. In addition to classical recognition method when the diagnosis is made at certain time, we also experimented with the "dynamic" of the diagnostic process by considering different time slots (0.5, 1, 2, ... 7 years) before the actual diagnosis is made. The results show that certain peaks (and the corresponding peptides) are very important for early diagnostic of OC, and they are much more informative in early stages than CA125.

2 Data

The serum samples used in the pilot study were collected in the UKCTOS project from 1995 to 2001. The data were subsequently analysed using the MALDI-TOF mass spectrometer at Sloan-Kettering Institute. In this pilot study the data were divided into two sets. The **Set 1** has 266 samples, of which 91 case samples taken from 19 women, and 175 control samples. These control samples were selected to match those case samples (usually 2 control samples were selected for each case sample). The women in the study were observed for 7 years (1995-2001) and each of these women samples taken in those years - we shall call them serial samples. Typically, each of the 19 cancer women has 2 to 12 serial samples taken in those years; and each of the healthy women has serial samples of five to nine and most of them have 6 samples. For all cancer women, the last sample was taken at the moment when the diagnosis was fixed. The study also included two different types of tubes that were used to collect samples: 233 with BD red top, and 33 using the Greiner tubes. The transit time taken between points when the samples when obtained and stored were also recorded and used in our analysis. In addtion, the **Set 2** was also obtained that has 305 samples from 50 healthy women. Each of healthy women has serial samples of 5 to 9 and most of them have 6 samples.

Other information such as date of birth, CA125, date of sample taken, date of sample received at the lab and tube type used for serum collection are also available.

Both the Set 1 and Set 2 serum samples were analysed by MALDI-TOF based mass spectrometry (MS). The MS dataset was generated at Memorial Sloan-Kettering Cancer Center (MSKCC), New York, USA in December 2004. For each serum sample, low mass against charge ratio (m/z) range of [700, 4000] and high range m/z of [4000, 15000] were obtained separately. In total, 125,206 data points generated by mass spectometer where mass against charge (m/z) values are ranged between 700 and 15000 with the corresponding intensities. These data points are related to time-of-flight (TOF) or clock tick measurement. An example of raw mass spectrometry (MS) data file is shown in Figure 1.

Overall, 565 serum samples (of possible 571 samples) were successfully analysed and the corresponding MS data files were obtained. These serum samples contain 475 control samples (174 in set 1 and 301 in set 2) and 90 case samples.

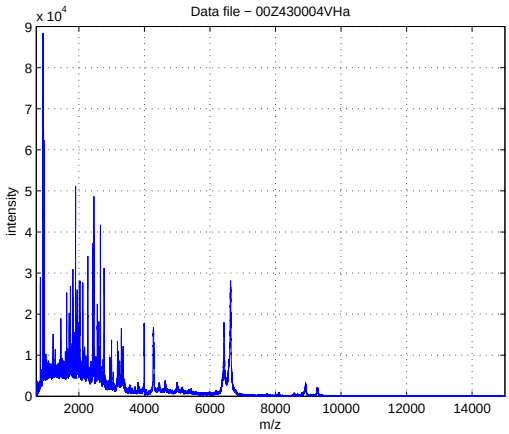

Figure 1: An example of raw MS data

3 Pre-processing

Mass spectrometry instruments are very sensitive and artifacts can be introduced into spectra from physical, electrical or chemical sources in experiments. Pre-processing is an important step to attempt to remove these systematic artifacts and isolate the true protein signal. The goals of pre-processing are to reduce noise, normalise the spectra from different samples and reduce dimensionality of MS data. Our assumption is that each spectrum can be considered as composed of three components: true peak signal, baseline, and random noise.

In this sections we describe our pre-processing of the raw data that includes: calibration, baseline subtraction, smoothing, normalisation and peaks alignment. We start with the raw data and perform the calibration first.

3.1 Calibration

We used the 13 peaks and calibrant file associated with each sample to perform calibration. In addition, we assumed that the relationship between m/z value (M) and time-of-flight (T) for the mass spectrometer used in the experiments can be represented as

$$M = B \times (T - A)^2 \tag{1}$$

where A and B are two constants determined by experiment setup. Our calibration algorithm is presented as follows.

As it is clear from the procedure, we calculated the constants A and B, and re-assigned the m/z (or M) their correct values. Note that calibration is performed separately on low mass range data of [700,4000] and high mass range data of [4000, 15000].

Algorithm 1 Calibration

Require: m/z values of 13 peaks
  fix A_s and B_s in the formula (1) (for example A_s=1.0 and B_s=0.5)
  find out TOF values for these 13 peaks (TOF_s) using the formula (1) and A_s and B_s
  for each sample's calibrant C_i **do**
    find the m/z values of these 13 peaks in C_i
    find out A_i and B_i which optimises $\sum_{k=1}^{13} (TOF_i{}^k - TOF_s{}^k)^2$
  end for
  {we now have optimal A_i and B_i for each sample}
  for each sample's raw data R_i **do**
    for each m/z value M_j in R_i **do**
      $TOF_j = \sqrt{\frac{M_j}{B_i}} + A_i$
      $M_j{}^* = B_s \times (TOF_j - A_s)^2$ {the corresponding intensity value is not changed}
    end for
  end for

3.2 Denoising

Our goal is to remove white noise associated with true peak signals. The undecimated discrete Wavelet approach is used. The basic idea is

- transform from the time domain into the wavelet domain;
- discard any wavelet coefficients below certain threshold
- transform any remaining wavelet coefficients back to the time domain.

The assumption is that white noise should be distributed over most wavelet coefficients at low levels while true signal should be represented in a few of relatively large wavelet coefficients at high level. The undecimated discrete Wavelet is an established tool in image processing and MATLAB code are freely available in the Rice Wavelet Toolbox [2].

3.3 Baseline Subtraction

The goal of baseline subtraction is to remove systematic artifacts, usually due to matrix and chemicals used in the experiments or to the detector overload. Ideally baseline should rest on zero. Chemical and electronic noise produces a background intensity which typically decreases when the mass m/z increases.

Baseline subtraction involves two steps: baseline estimation followed by subtraction of the estimated baseline from the raw mass spectrum. An example of baseline estimation is shown in Figure 2. Our baseline estimation procedure can be described as follows.

Algorithm 2 Baseline estimation

Require: spectrum S
Require: moving window size W
Require: threshold K
  finished = false
  B=S
  while finished == false **do**
    for each moving window B_W of S **do**
      if no of points in $B_W \leq K$ **then**
        finished = true
      else
        find out mean μ and variance σ_2 of those points in B_W
        if all points are in $[\mu - 1.5\sigma, \mu + 1.5\sigma]$ **then**
          finished = true;
        else
          remove those points which are above μ B_W
        end if
      end if
    end for
  end while
  baseline is represented by those points in B

Optionally, the raw data can be binned before baseline estimation procedure is applied. An example is shown below where the raw data was binned by a window size of 10.

3.4 Smoothing

Mass spectra of serum samples also exhibit an additive high frequency noise component. The presence of this noise hampers peak identification and we need to reduce the influence of this high frequency noise. One way is to smear out the high frequency noise signal in the spectra by averaging the intensities within a moving window.

3.5 Normalisation

Due to variation in sample preparation and deposition on the target, matrix crystallisation and ion detection, samples are not directly comparable before normalisation. The goal of normalisation is to make sure that the total amount of ions across different samples are the same. This is done by calculating the sum of all intensity values and then dividing each intensity value by the sum. We then multiple these intensity values with a constant C (for example we set C=2×10^5 in our experiments). The results are shown below.

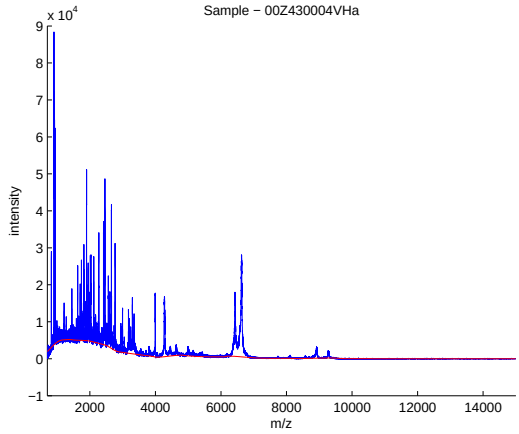

Figure 2: Estimated baseline (shown in red)

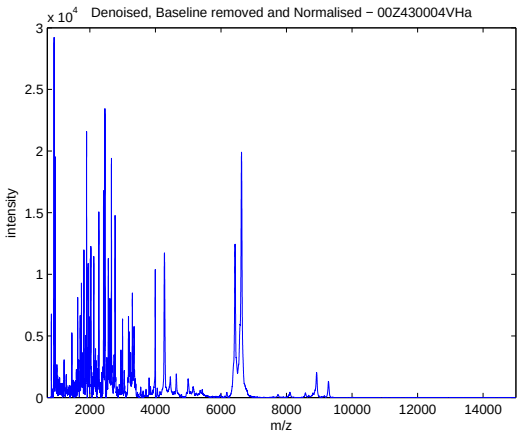

Figure 3: An example preprocess data after denoising, baseline subtraction and normalisation

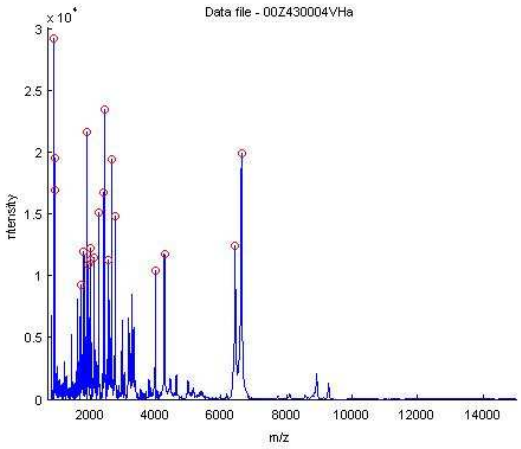

Figure 4: Peak identification

3.6 Peak Identification

A peak in mass spectra indicates the relative abundance of a protein. Peak identification is concerned with identifying peaks within a single mass spectrum. The identification of peaks in a mass spectrum is complicated by the error in measuring the abundance as well as the mass error rate. The goal of peak identification is to identify set of m/z values which comprise peaks which are higher than the noise level of a mass spectrum. The peak identification algorithm finds local maxima with a certain signal-to-noise ratio (eg SNR=4) and choose the local maxima higher than a threshold of the noise level as peaks. For example, local maxima of a mass spectrum are located by finding the m/z ratios with the highest intensity among their N neighbours. The noise level is defined as the average of the intensities at the m/z ratio within a moving window with a fixed size (eg 500). The peaks identified were quantified as height at the local maximum. An example of peaks identified is given in Figure 4, where peaks are represented as circles where the absolute peak height exceeds the threshold of 9000. For the purpose of peak alignment, we use only peaks which have intensity value exceeding the threshod of 9000.

3.7 Peak Alignment

To make inference about trends across a number of spectra, we need to relate the peaks identified in one spectrum to the peaks founded in another spectrum. This process of matching peaks which represent the same protein across several spectra is known as "peak alignment". In peak alignment, the peaks of multiple mass spectra within the mass error rate are grouped together as a "peak group".

Given a number of peak points (or features), we need to find out a unique

correspondence between them. The problem is that not all peaks appear in every sample. Therefore, one-to-one correspondence does not exist between every two samples. A simple approach is to construct a super set of all peaks and use it as the anchor of alignment - every sample is aligned to this super set. For this purpose the superset is splitted to clusters. Cluster definition goes in two steps. First we find all intervals between neighbouring peak positions in the superset exceeding some fixed values d_1, where d_1 is some distance metrics based on mass resolution (eg 1500ppm). These intervals split the m/z axe to clusters of order 1. Then we test if each sample has no more than 1 peak in a cluster. If so, the cluster is considered as final. Otherwise, we look whether there is an interval exceeding $d_2 < d_1$, dividing occurrences of one sample peaks within the cluster. If so we divide the cluster of order 1 to smaller ones. Otherwise we consider it as final.

Now for each sample peak we assign the number of its cluster (in case when there are more than 1 peak of the same cluster for the same sample we take into account only the largest of them).

Now all peaks are aligned to a certain cluster. Every sample is then characterised by a numerical vector, of dimension n (n is the number of final clusters) with the coordinates equal to the height of a peak corresponding to each cluster, zero if there is no such peak. These coordinates are considered as a set of sample features for pattern recognition.

Note that we align peaks from all samples without discriminating among controls and cases.

4 Preliminary results

The preprocessing steps described in section 3 were applied to each of 565 raw MS data files. Figures 5 and 6 illustarte an example of preprocessed serial samples from a cancer woman and a normal woman, respectively.

Heatmap of the preprocessed dataset is shown in Figure 7, where the colours represent intensity value and the while line indicates the two sets. Set 1 is in the top part and set 2 in the bottom part. Heatmap provides an at-a-glance view of alignment of spectra and it shows the data is well calibrated.

In total, 6924 peaks were identified in all the analysed serum samples. After the pre-processing, 173 peak groups were identified after peak alignment. 15 peak groups were found which are present in more than 20% of all samples, see table 1.

These 15 peaks are shown in Figures 8 and 9, where blue lines represent controls and red cases.

The peak variation can be measured by coeffficient of variation (CV) which is a statistical measure of the deviation of a variable from its mean. CV is calculated as follows:

$$CV = \frac{standard\ deviation}{mean}$$

The mean, standard deviation and CV of the 15 peaks is given in table 2 where

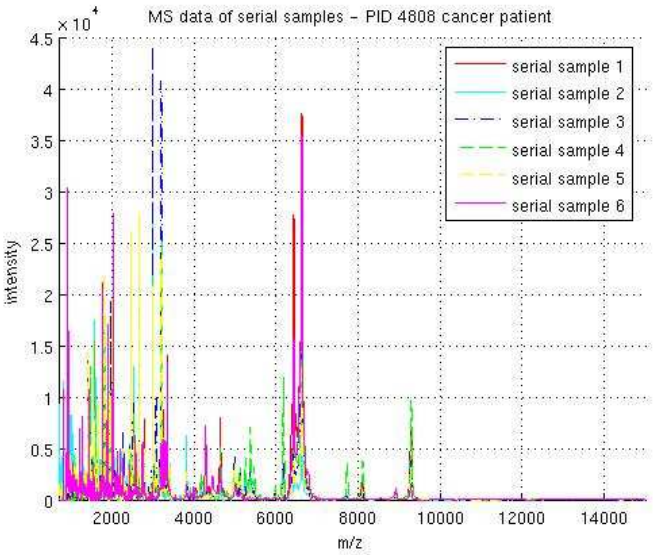

Figure 5: Pre-processed MS data of serial samples - Cancer Patient (ID=4808)

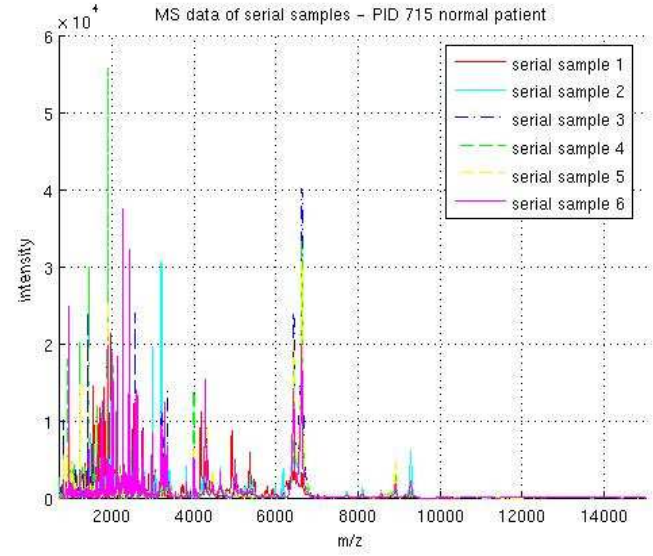

Figure 6: MS data of serial samples - Normal Patient (ID=715)

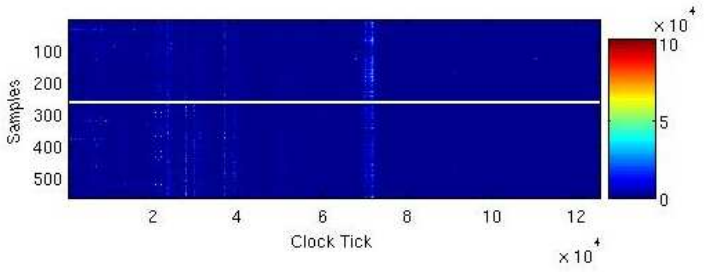

Figure 7: Serial Samples - Cancer

Peak No	Peak ID	m/z value	No of samples having the peak	percentage
1	216	6657.0–6679.1	434	76.00
2	173	3236.2–3244.3	297	52.01
3	207	6463.1–6479.3	293	51.31
4	160	3028.5–3035.6	238	41.68
5	074	1796.4–1819.2	200	35.02
6	091	1992.9–2012.5	197	34.50
7	119	2302.9–2308.4	192	33.62
8	086	1926.8–1933.3	191	33.45
9	095	2043.0–2055.3	187	32.74
10	098	2054.9–2067.6	161	28.19
11	128	2447.1–2457.8	149	26.09
12	083	1894.2–1901.6	132	23.11
13	132	2482.8–2495.1	131	22.94
14	146	2687.3–2693.6	129	22.59
15	079	1843.7–1849.5	108	18.91

Table 1: 15 most popular peaks among the samples

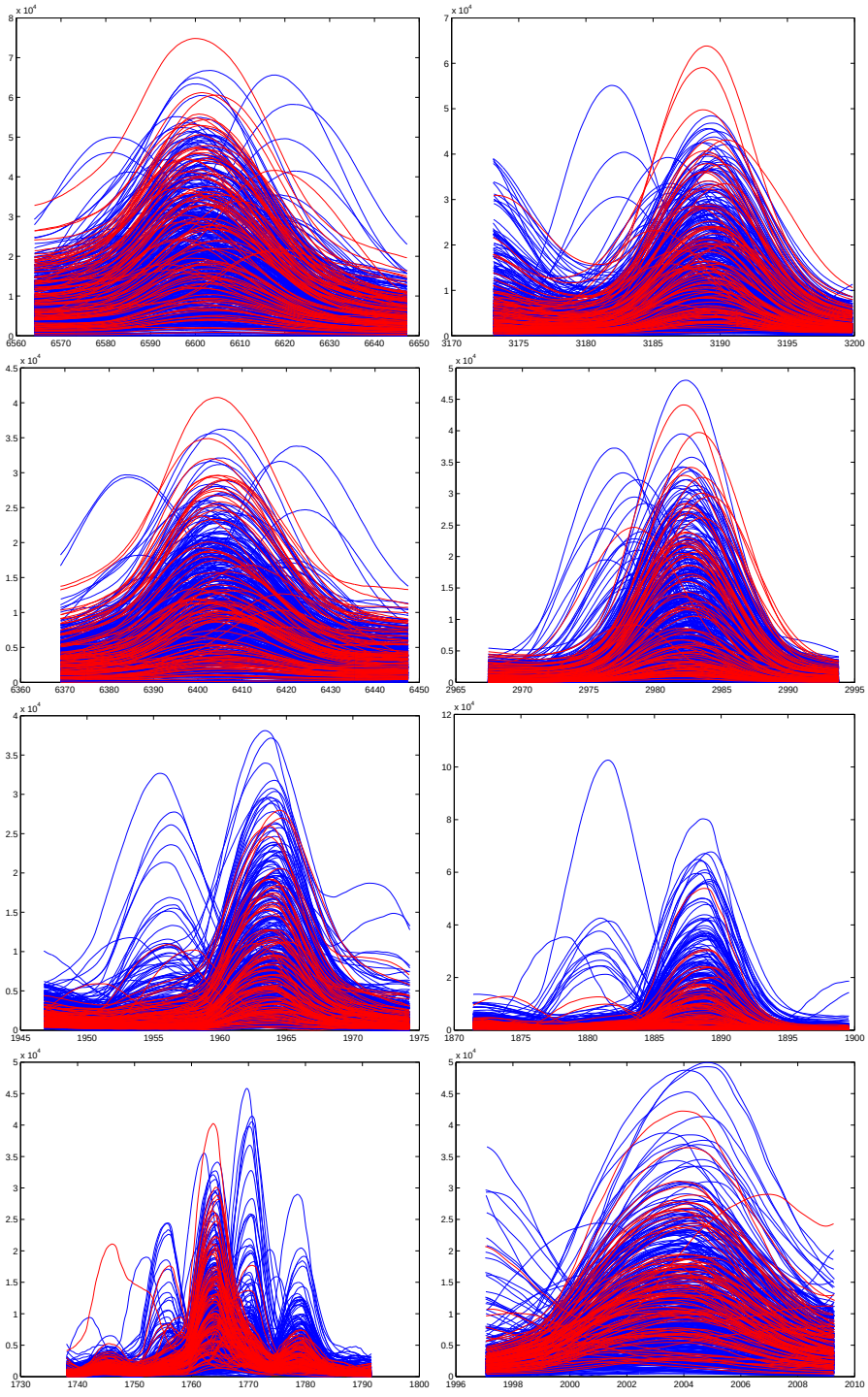

Figure 8: The top 15 peaks (1-8 peaks)

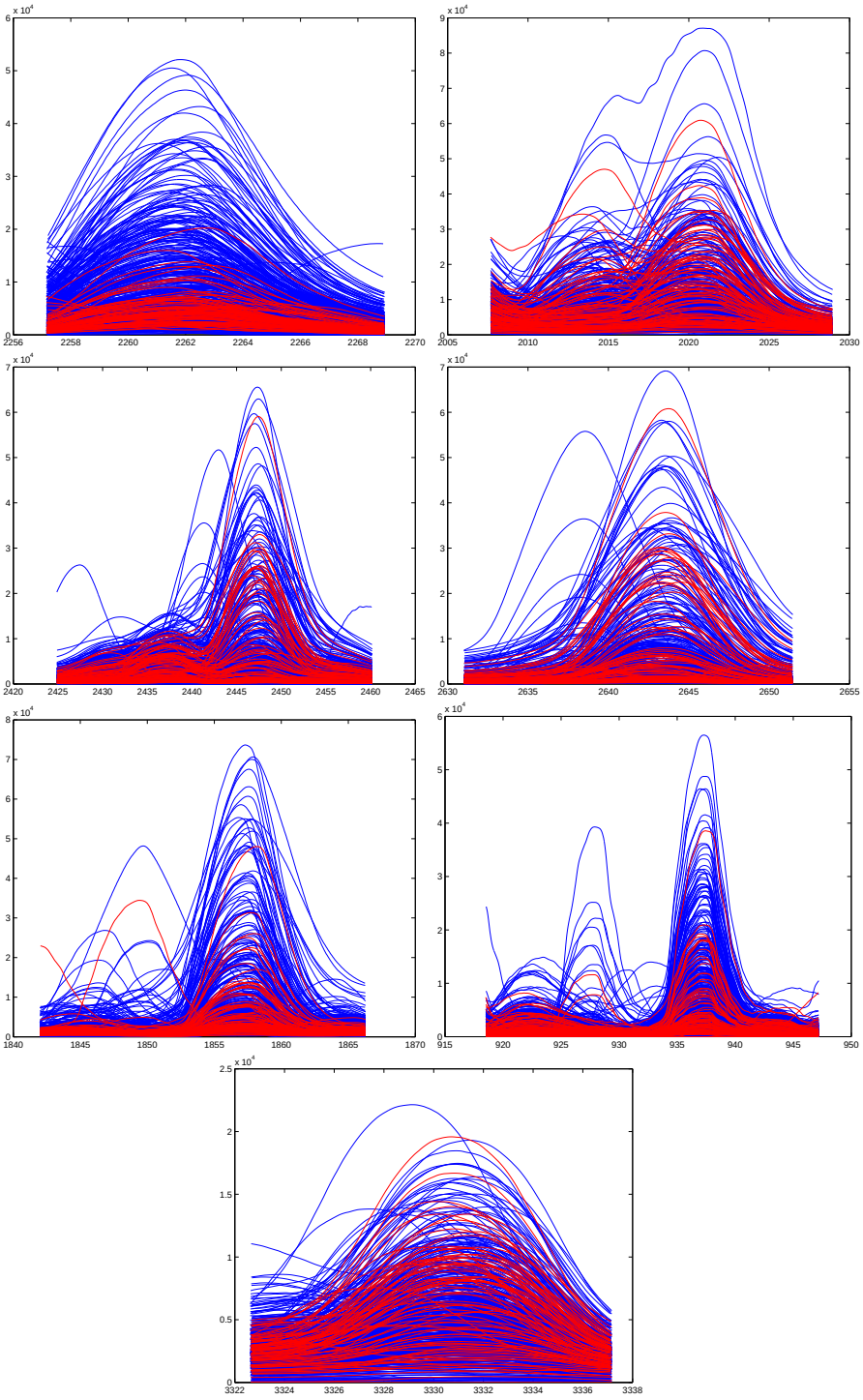

Figure 9: The top 15 peaks (9-15 peaks)

Peak No	Peak ID	m/z value	Mean	standard deviation	CV
1	216	6657.0–6679.1	26592	12788	0.48
2	173	3236.2–3244.3	21366	10033	0.47
3	207	6463.1–6479.3	16272	6169	0.38
4	160	3028.5–3035.6	18566	7472	0.40
5	074	1796.4–1819.2	16091	6788	0.42
6	091	1992.9–2012.5	16386	6034	0.37
7	119	2302.9–2308.4	19136	8942	0.47
8	086	1926.8–1933.3	23049	14418	0.63
9	095	2043.0–2055.3	18029	8311	0.46
10	098	2054.9–2067.6	21287	12195	0.57
11	128	2447.1–2457.8	19664	9069	0.46
12	083	1894.2–1901.6	25205	15468	0.61
13	132	2482.8–2495.1	23006	12365	0.54
14	146	2687.3–2693.6	23633	12971	0.55
15	079	1843.7–1849.5	17869	6710	0.38

Table 2: CV of 15 most popular peaks

Set 1	BD red top	Greiner-Bio-One
Controls	152	22
Cases	79	11
Set 2	251	50

Table 3: Distribution of BD red top and Greiner-Bio-One serum gel tubes

we only considered the peaks which have intensity values exceeding the threshold of 9000.

We are interested in peak changes in the samples and want to investigate how tube type and transit time could impact peak changes.

Note that we have found a interval (m/z) for each peak in "peak alignment" procedure. These intervals are then fixed and the highest intensity values within these intervals are used to represent the peaks for classification procedure.

4.1 Tube type

Two types of tubes used for serum sample collected: BD red top and Greiner-Bio-One serum gel. The distribution of two is given in the table 3.

Peak changes of different tube types in controls and cases are presented in Figures 10 and 11.

4.2 Transit time

The serum samples were collected from various collection centres distributed across the country and then posted to the lab for storage. All the serum samples

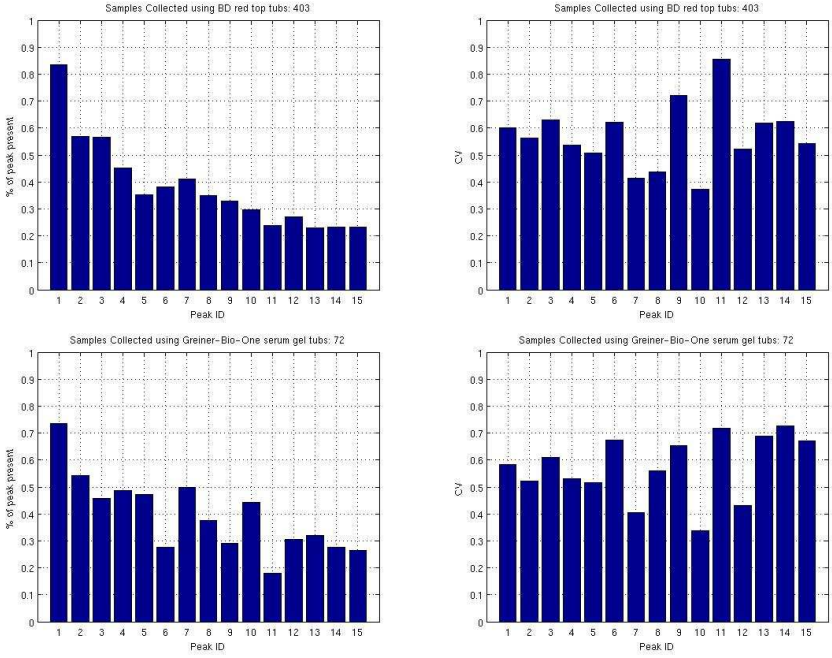

Figure 10: Effect of tube types on all control samples (top row: BD red top tubes; bottom row: Greiner-Bio-One gel tubes)

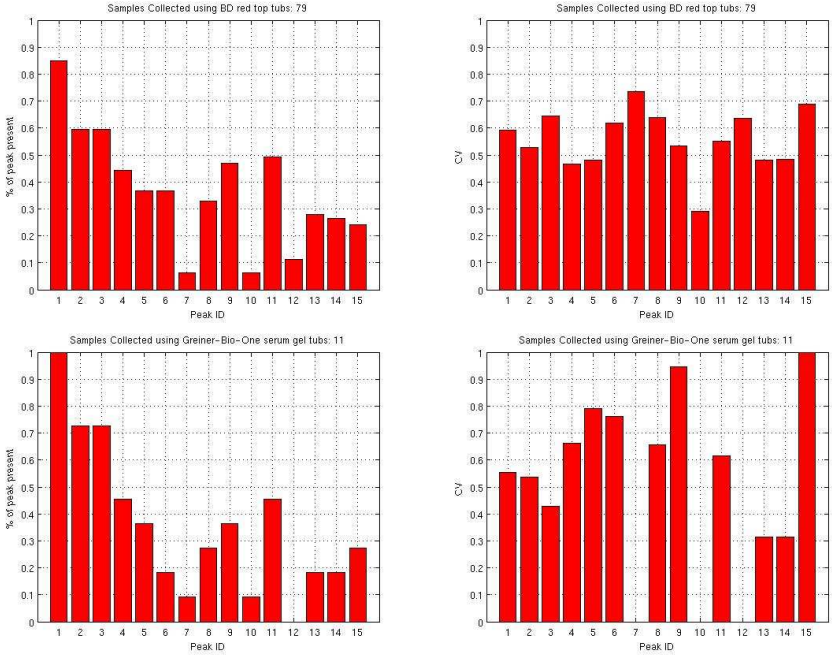

Figure 11: Effect of tube types on all case samples (top row: BD red top tubes; bottom row: Greiner-Bio-One gel tubes)

Set 1	O day	1 day	2 day
Controls	16	194	91
Cases	12	53	25
Set 2	11	134	29

Table 4: Distribution of transit time in the serum samples

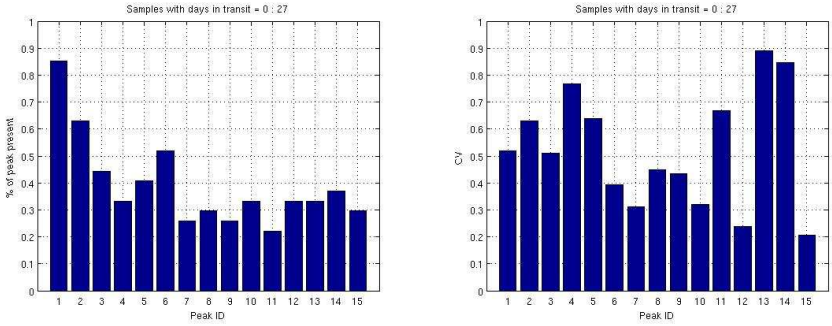

Figure 12: Effect of transit time on all control samples (top row: transit time=0 day; middle row: transit time=1 day; bottom row: transit time=2 day)

analysed were 48 hours or less in transit. We divided the serum samples into three groups based on the transit time: 0 day (samples arrived on the same day in the lab), 1 day (samples arrived next day in the lab) or 2 day (samples arrived in the lab two days after the collection). their distribution given in table 4.

5 Discussion

Peak variation is huge even with normal samples. Transit time can influence peak presence and variation in mass spectra. One day transit time and two day transit time have similar peak characterises Zero day transit time has a different peak characterises from one day or two day transit time How to best pre-process MS raw data is still an open problem.

6 Pattern recognition techniques

6.1 The techniques

There are many methods of pattern recognition that can be applied to medical diagnostics. Usually every object presented for pattern recognition is described by a set of features (quantitive or qualitive). A subset of objects with known classification forms the training set. Pattern recognition algorithms construct

decision rules on the basis of this training set. Another subset with known classification can be used as a test (or validation) set. The quality of a decision rule is estimated by the number of errors in the test set. In the case of two classes there are two type of errors: the first type is when an object of the first class is recognised by the decision rule as an element of the second one. The second type is when on the contrary an object of the second class is recognised as an element of the first one. An algorithm of pattern recognition often has a parameter regulating the proportion of the two types of errors.

In applications to medical diagnostics a patient serves as an object for pattern recognition. Patients are grouped in classes according to their diseases or their absense. And the features are symptoms, some general information such as age, sex and quite often additional objective measurements (for example, the intensities of peaks).

In the current report we used two pattern recognition methods: the nearest neighbour algorithm and support vector machine. Their main advantage is that they are fast and accurate.

The first, the nearest neighbour method, starts from presenting all objects of the training set as the points in multidimensional space with coordinates corresponding to the list of features. A new object is also presented as a point in the space and the nearest (in a certain metrics) point of the training set is searched. The new object is referred to the same class as the nearest point of the training set has. We can weight, if required, the distances from the points of different classes and in this way to regulate the proportion of different error types.

Another approach is that of the support vector machine. In this case any two objects are compared using a certain positively defined kernel. In the simplest case, when dot product is used as the kernel, a linear decision rule is constructed in the space of features. If more complicated kernels are used, then a nonlinear separating surface is constructed by support vector machine on the base of the training set.

6.2 Significance level of selected features

The basis of feature space in our data is formed by the intensities of the selected peaks on each MS-spectrogram. That means that for each sample, after preliminary processing and peak alignment, 15 values of the selected 15 peaks intensities were fixed as a vector characterizing the sample. These values were fixed independent on whether a peak was large or small as a maximum of the spectrogram values within an interval corresponding to the peak. Before using pattern recognition procedures it would be useful to check if the probability distribution of the peaks intensities is different for cancer samples versus controls. To estimate the difference for a certain significance level we used Mann-Whitney test. The test returns P value of whether two sample sequences could be generated by the same probability distribution, assuming that the samples are independent.

Peak No	m/z	Norm.int.	Med.ratio	p-value
1	6656.98–6679.08	21867.5	1.64063	0.0715372
2	3236.2–3244.32	7980.09	1.86326	0.0612298
3	6463.12–6479.3	9804.44	1.21398	0.226691
4	3028.51–3035.62	4964	0.992385	0.731099
5	1796.44–1819.21	10390	0.910727	0.959652
6	1992.91–2012.49	5025.18	1.04361	0.843728
7	2302.92–2308.44	5532.25	0.74315	0.081379
8	1926.81–1933.27	3409.02	0.847595	0.621526
9	2043.03–2055.27	14192	0.713101	0.260507
10	2054.92–2067.64	12860.5	0.873336	0.504046
11	2447.08–2457.78	3183.31	0.798343	0.138589
12	1894.17–1901.56	4049.99	0.594754	0.0328834
13	2482.78–2495.11	2725.27	0.735022	0.611702
14	2687.25–2693.65	2482.69	0.428217	0.415259
15	1843.67–1849.51	1896.73	0.961602	0.962433
	log(CA125)	2.5337	1.79882	1.34912e-006

Table 5: All samples from 19 women with cancer and 218 healthy women

First we used the serum sample analyses from all 19 women with OC diagnosis, versus 219 healthy women. The samples were taken over a period of 7 years - (serial samples). To find the level of significance we consider only the last time-point for each patient in these serial data since for the patients it was the point, when the diagnosis was made and the disease was found. Then for each of the 15 peaks intensities for each serum sample were calculated and the test was applied. In the Table 5 the results are shown. In the column 2 m/z interval corresponding to a peak is fixed, column 3 shows median intensities of each m/z peak in the control set, column 4 corresponds to peak intensity ratio, which was calculated by dividing median value of a peak in the cancer group by the median value in the control set, and column 5 gives corresponding P values. For comparison, a line corresponding to CA125 analysis was added.

We see that P values for almost all peaks are very high in comparison with that of CA125. This means that difference in distributions for cases and controls is not reliable; that is it is unlikely that the control and OC are coming from the same probablity distribution. But according to UKCTOCS program serial blood serum samples were taken from the same people. It allows, first, to average the results of MS-spectrum analysis over the serial samples, increasing the accuracy, and, second, to take in account changes in an organism, preceding the moment when the diagnosis was fixed. So it was reasonable to test averaged over serial samples peak intensities, whether their probability distributions are different for normal and cancer cases. But then only those people, who had serial blood sampling, could be used. We agreed that the number of samples in averaging should be equal 4, and so only those women that had not less than 4 serial samples could be tested. After this restriction there remained only 11 patients

Peak No	m/z	Norm.int.	Med.ratio	p-value
1	6656.98–6679.08	16894.7	1.79531	0.146522
2	3236.2–3244.32	8336.85	1.83058	0.00287992
3	6463.12–6479.3	11772.8	0.958328	0.969521
4	3028.51–3035.62	9580.59	0.579158	0.276178
5	1796.44–1819.21	11221.3	0.851011	0.702397
6	1992.91–2012.49	5350.47	1.00478	0.456231
7	2302.92–2308.44	18027.5	0.204231	8.3028e-005
8	1926.81–1933.27	1452.5	1.21277	0.702397
9	2043.03–2055.27	13793.7	0.697387	0.0890776
10	2054.92–2067.64	8145.34	1.32386	0.131237
11	2447.08–2457.78	4106.13	0.46941	0.0266847
12	1894.17–1901.56	3195.9	0.747088	0.0585808
13	2482.78–2495.11	2635.57	0.760037	0.803859
14	2687.25–2693.65	3419.46	0.303162	0.369238
15	1843.67–1849.51	1463.76	1.40435	0.276178
	log(CA125)	2.48491	1.83414	3.38873e-005

Table 6: The last samples of 11 cases and 49 controls

with cancer diagnosis and 50 healthy women remained.

To make the results comparable we first applied the test only to the last sample for these 11 cases and 50 controls. The results are shown in the Table 6. The results are not significantly different from those presented in the Table 5.

Then we applied the test to peak intensities averaged over the 4 last serum samples (for cancer cases it means the last 4 blood samples preceding the moment when the diagnosis was made).

We see that P values for some of the peaks becomes neglectively small, and it means that averaged intensities for cases and controls are really different for those peaks.

It was also interesting to see if distribution changes precede the moment, when the diagnosis was fixed. For this purpose we excluded from averaging the last moment, averaging intensities only over 3 samples preceding the last one. The test results are shown in the Table 8.

Again we see that P values for some of the peaks (peaks 7, 8, 10, 11) are very small. Still the difference in distributions does not mean that a feature can be used for reliable diagnosis. It is due to the fact that distributions may be different, but overlapping, and the overlapping distribution inevitably implies errors.

7 Feature space and recognition field definition

We want to mention again: the basis of the feature space is formed by the intensities of the selected 15 peaks on each MS- spectrogram.

Peak No	m/z	Norm.int.	Med.ratio	p-value
1	6656.98–6679.08	19368.6	1.11224	0.214318
2	3236.2–3244.32	9269.6	1.71457	0.00184568
3	6463.12–6479.3	9801.5	0.851109	0.67427
4	3028.51–3035.62	6714.35	1.82998	0.0110576
5	1796.44–1819.21	9658.65	1.04334	0.491607
6	1992.91–2012.49	6277.1	1.70491	0.0230013
7	2302.92–2308.44	11776.6	0.386615	7.13864e-006
8	1926.81–1933.27	12182	0.408536	8.98873e-005
9	2043.03–2055.27	11990.8	1.01132	0.349216
10	2054.92–2067.64	8475.77	1.85873	0.00162039
11	2447.08–2457.78	10066.2	0.370796	7.13864e-006
12	1894.17–1901.56	9581.67	0.881836	0.0373087
13	2482.78–2495.11	4151	2.53842	0.00142066
14	2687.25–2693.65	3891.43	2.85622	0.0066715
15	1843.67–1849.51	2601.92	2.35004	0.000226796
	log(CA125)	2.3302	1.52829	0.000123051

Table 7: Peak intensities averaged over the 4 last serum samples

Peak No	m/z	Norm.int.	Med.ratio	p-value
1	6656.98–6679.08	20913.6	1.07215	0.566557
2	3236.2–3244.32	9726.48	1.61889	0.00791912
3	6463.12–6479.3	9797.79	0.791477	0.503718
4	3028.51–3035.62	6093.17	1.97543	0.00196881
5	1796.44–1819.21	8525.61	1.14811	0.411363
6	1992.91–2012.49	6871.61	1.77838	0.0469392
7	2302.92–2308.44	8965.76	0.549563	0.000328641
8	1926.81–1933.27	14211.6	0.363002	0.000105243
9	2043.03–2055.27	9994.74	1.28807	0.0123262
10	2054.92–2067.64	7805.26	1.7933	0.000380284
11	2447.08–2457.78	11266.5	0.231518	5.96232e-006
12	1894.17–1901.56	11658.4	0.888413	0.0756175
13	2482.78–2495.11	4463.06	2.3992	0.00306475
14	2687.25–2693.65	3858.96	2.94571	0.00560348
15	1843.67–1849.51	2615.57	2.89076	0.00142066
	log(CA125)	2.32955	1.38926	0.00162039

Table 8: Averaging peak intensities only over 3 samples preceding the last one

According to the classical scheme of pattern recognition we had to consider as ill only those women, for whom the disease was found at the moment of the last sampling, and to use for diagnosis only the data of this last sample. All the other cases should be considered as control: i.e. at that moment a woman was healthy. But,

1. from the medical point of view it is vitally important to predict the disease for some time before it was actually diagnosed,

2. experience of diagnostics on the basis of CA125 protein has shown that not only the present value of a protein concentration is significant for diagnosis, but also its comparison with prehistory [3].

Given this, we decided to include in the number of input parameters a value, characterising its prehistory.

We had to take in account that there was only very small number of ill women presented for training. To increase the number of we defined as an **object for recognition** not a woman, but **a moment of sampling**. (It is necessary to remind, that the last moment of sampling for women in the set cases was the time, when the disease was found.) We included in the only those cases (moments), which had at least k preceding measurements from the same person. (Value k was a fixed constant equal to 2 or 3).

All such cases from the control group formed the first class (healthy). So it included all moments of healthy women sampling, which had at least k predecessors. Then we fixed some time interval T and defined the second class (cancer) in the following way. We included in this class those sampling moments (from the group cases), after which the disease was found not later than in time T and not less than k samples preceded it, see figure 13. So if $T = 0$, we refer to the class II only the moments, when the disease was detected by other means. If $T = 2$ years, we refer to the second class all the sampling moments, which are not more than 2 years before the disease was detected. In the case $T = \infty$ all sampling moments of a woman who got ill are included in the second class. All the other sampling moments were not included in the massive for training and recognition due to the following reasons. The k preceding samples were necessary to take in account prehistory. Then, if the disease was found later than in time T, both answers (healthy, ill) can not be considered as errors, and it is dangerous to include these cases in the training set, as far as it is unknown what time before the disease detection changes in protein concentration start.

We remind once more, that as the objects of recognition the moments of sampling, satisfying the above conditions, were taken. The input parameters were the intensities of the selected peaks, and to take in consideration the prehistory, they were added by the average value over prehistory (really average over k preceding samples to put all the objects in equal conditions). Formally

$$X_i{}^*(t_j) = 1/k \sum X_i(t_{j-p}),$$

where $X_i(t_j)$ is the i-th peak intensity at the current moment t_j, $p = 1, ..., k$ and $X_i{}^*(t_j)$ is the additional feature, responsible for prehistory. Of course it

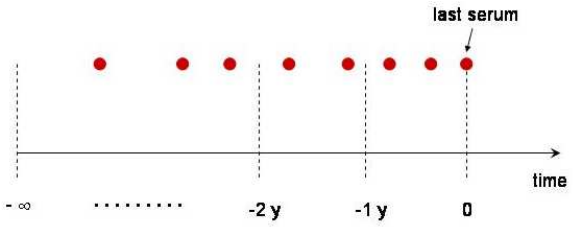

Figure 13: A moment of sampling for cancer women

would be more accurate to normalise this difference by individual mean square deviation in the prehistory. But very low number of preceding samples for a fixed person makes impossible satisfactory estimation of the mean square deviation.

8 Feature Selection

As the result of these restrictions we had in the case $k = 3$ and $T = 0$, 151 objects of the class 1 and 11 objects of the class 2, in the case $k = 3$ and $T = 2$ years, 151 objects of the class 1 and 31 objects of the class 2, in the case $k = 3$ and $T = \infty$, 151 objects of the class 1 and 37 objects of the class 2.

In any case the number of objects in the class 2 (cancer) was very low, so it was impossible to divide to training and control sets without sufficient losses in the training set. Then to have fair estimation of recognition result we had to use the method called leave one out. It is to leave one object out, execute the training procedure using all other objects and test the object left out. Then the procedure is repeated leaving out all objects one by one, the number of errors and correct answers is calculated and serves as fair estimation of recognition quality. In our case, if a sample of some person was left out, then all samples of the same person was also left out. This was made in order not use any sample of a person in the training, if a sample of the same person was used for control.

The other moment is that having low number of examples and rather large number of features one can easily find a decision rule correctly recognising all the samples in the training set, but making a lot of errors in the control set. To find really good decision rule in this case it is necessary to decrease the number of features severely. It is possible to look through different combinations of rather small feature number from initial set.

The other moment is that, if we use leave one out procedure to evaluate a certain combination of features and then choose the best, the result can be wrong. The gained estimation may deviate from the real value, and this deviation becomes large if the number of combinations is grate. If we put the choice of the best combination within leave-one-out procedure, then possibly we would have as the best different combinations for different objects left out, which is unacceptable from biological point of view.

L	M	S
x_t	$x_t - \bar{x}$	$\frac{x_t - \bar{x}}{\delta}$

These considerations forced us look for the least number of features, and the least number is 1. On the other hand if we find one feature, discriminating healthy and cancer groups, it could serve as biomarker for the disease.

Still experiments has shown that there was no single feature (one of the 15 peak intensities), which can satisfactory discriminate the classes. Then we turned to the idea of using prehistory: try to discriminate classes in two dimensional space with coordinates ${X_i}^*(t_j)$ and $X_i(t_j)$, as defined above. Alternatively, we can define other measurements, see the table below:

Now we gained acceptable results.

The procedure was as follows. For each feature pair (${X_i}^*(t_j)$ and $X_i(t_j)$) we applied leave-one-out method with the nearest neighbour recognition rule. Then we compared the results and looked for the best pairs.

9 Recognition results

We illustrate recognition results for two cases: $T = 0$, which means that we consider as cancer only those sampling moments, when the disease was found by other means immediately, and $T = 1, 2, 3, 4, 5$ years, which means that we consider as cancer also those sampling moments, when the disease was found not later than in 1, 2, 3, 4, 5 years after sampling. Not less than 3 samples should precede a sample moment presented for classification. After these restrictions we had 11 ill women and 50 healthy ones.

For the case $T = 0$ we had for recognition 11 sampling moments in the class **cancer** (because for every woman only one sampling moment was taken) and 151 sampling moments in the class **healthy** (because several samples from one woman was included).

For the case $T = 2$ we had for recognition 31 sampling moments in the class cancer (because now one patient could be sampled several times within time interval T) and again 151 sampling moments in the class healthy.

The nearest neighbour recognition procedure has a parameter which regulates the error proportion of different kinds: **type 1 -cancer sample classified as healthy, and type 2 - healthy samples classified as cancer.** This parameter consists of two parts: number of neighbours and the voting threshold. For example, parameter 2/5 means that 5 neighbours neighbours are considered, and the classification is 'case' if 2 or more of 5 neighbours are 'cases'.

Below we show the results gained using the best selected features listing several values for different proportion two type errors in Table 9.

Figures 14 to 37 show the distributions of cases (crosses) and controls (zeros). The x-axis represents peak intensity in the moment of measurement and the y-

Samples	Interval	Peak 1	Peak 2	Error rate
37	-∞, 0	10 (L,M,S)	6 (M)	2.7%
33	-3, 0	10 (L, M, S)	6 (M)	3.0%
31	-2, 0	10 (L, M, S)	6 (M)	3.2%
27	-1, 0	10 (L, M)	CA125 (M)	3.7%
25	-8/12, 0	10 (L, M)	CA125 (M)	2%
23	-6/12, 0	10 (L, M)	CA125 (L, M)	0%
16	-2/12, 0	10 (L, M, S)	6 (S)	0%
11	0, 0	10 (L, M, S)		0%

axis is average value of the peak intensity over prehistory (really over k preceding samples (k=3)).

Peak 10. m/z =[2054.9 2067.6]				
Error	*Case T = 0.*			
Param.	1/5	1/4	1/3	2/5
Type 1	2(18%)	3 (27%)	5 (45%)	6 (54%)
Type 2	26(17%)	21 (14%)	16 (10%)	9 (5.8%)
	Case T = 1.			
Param.	1/5	1/4	1/2	2/5
Type 1	6(22%)	7 (26%)	9 (33%)	12 (44%)
Type 2	52(34%)	43 (28%)	23 (15%)	14 (9.1%)
	Case T = 2.			
Param.	1/5	1/4	1/2	2/5
Type 1	6 (19%)	8 (25%)	10 (32%)	11 (35%)
Type 2	52 (34%)	45 (29%)	24 (16%)	15 (9.7%)
	Case T = 3.			
Param.	1/5	1/4	1/3	2/5
Type 1	7 (21%)	10 (30%)	11 (33%)	13 (39%)
Type 2	60 (39%)	52 (34%)	33 (21%)	14 (9.1%)
	Case T = 4.			
Param.	1/5	1/4	1/2	2/5
Type 1	6 (17%)	9 (25%)	12 (34%)	13 (37%)
Type 2	60 (39%)	52 (34%)	30 (19%)	18 (12%)
	Case T = 5.			
Param.	1/5	1/3	1/2	2/5
Type 1	6 (16%)	8 (22%)	11 (30%)	12 (32%)
Type 2	60 (38%)	34 (22%)	29 (19%)	19 (12%)

Table 9: Classification results using the best selected features

Peak 7. m/z=[2302.9 2308.4]				
Error	*Case T = 0.*			
Param.	2/5	1/1	2/3	2/2
Type 1	4 (36%)	5 (45%)	6 (54%)	8 (72%)
Type 2	11 (7.1%)	5 (3.2%)	7 (4.6%)	2 (1.3%)
	Case T = 1.			
Param.	2/5	2/4	3/5	4/5
Type 1	4 (15%)	6 (22%)	10 (37%)	16 (59%)
Type 2	18 (12%)	17 (11%)	12 (7.8%)	7 (4.5%)
	Case T = 2.			
Param.	1/5	1/4	2/5	2/4
Type 1	4 (13%)	5 (16%)	6 (19%)	8 (26%)
Type 2	40 (26%)	38 (25%)	21 (13%)	19 (12%)
	Case T = 3.			
Param.	1/5	1/4	2/5	2/4
Type 1	4 (12%)	5 (15%)	6 (18%)	7 (21%)
Type 2	40 (26%)	38 (25%)	22 (14%)	19 (12%)
	Case T = 4.			
Param.	1/5	1/4	2/5	2/4
Type 1	5 (14%)	7 (20%)	8 (23%)	9 (26%)
Type 2	42 (27%)	41 (27%)	29 (19%)	25 (16%)
	Case T = 5.			
Param.	1/5	1/4	2/5	2/4
Type 1	5 (14%)	7 (19%)	8 (22%)	9 (24%)
Type 2	42 (27%)	41 (27%)	29 (19%)	25 (16%)

Table 10: Classification results using the best selected features

Peak 11. m/z =[2447.1 2457.8]				
Error	*Case T = 0.*			
Param.	1/3	2/4	2/3	2/2
Type 1	0 (0%)	2 (18%)	3 (27%)	6 (54%)
Type 2	7 (4.6%)	6 (3.9%)	4 (2.6%)	1 (0.7%)
	Case T = 1.			
Param.	1/5	3/5	1/2	2/3
Type 1	0 (0%)	1 (3.7%)	4 (15%)	7 (26%)
Type 2	12 (7.8%)	6 (3.9%)	7 (4.6%)	5 (3.2%)
	Case T = 2.			
Param.	1/5	3/5	2/3	2/2
Type 1	0 (0%)	1 (3.2%)	4 (13%)	12 (38%)
Type 2	12 (7.8%)	6 (3.9%)	5 (3.2%)	3 (2%)
	Case T = 3.			
Param.	1/5	3/5	3/4	3/3
Type 1	0 (0%)	1 (3%)	7 (21%)	16 (48%)
Type 2	12 (7.8%)	6 (3.9%)	5 (3.2%)	2 (1.3%)
	Case T = 4.			
Param.	1/5	2/4	3/5	4/5
Type 1	0 (0%)	1 (2.9%)	2 (5.7%)	11 (31%)
Type 2	13 (8.5%)	7 (4.6%)	6 (3.9%)	5 (3.2%)
	Case T = 5.			
Param.	1/5	2/4	3/5	4/5
Type 1	0 (0%)	1 (2.7%)	2 (5.4%)	12 (32.4%)
Type 2	13 (8.4%)	7 (4.5%)	6 (3.9%)	5 (3.2%)

Table 11: Classification results using the best selected features

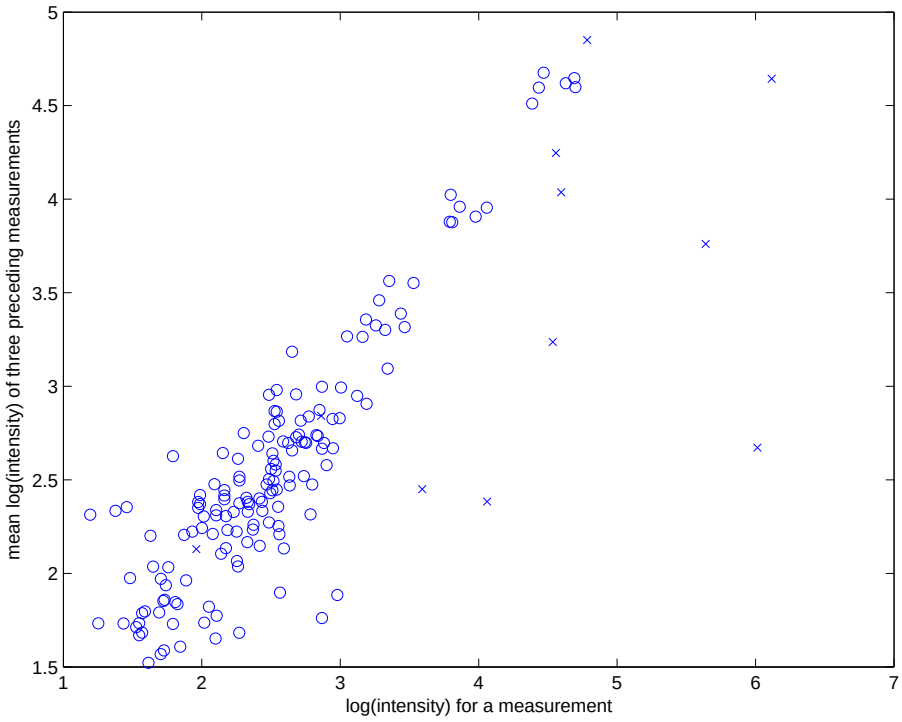

Figure 14: CA125 - last moment cases vs. all controls

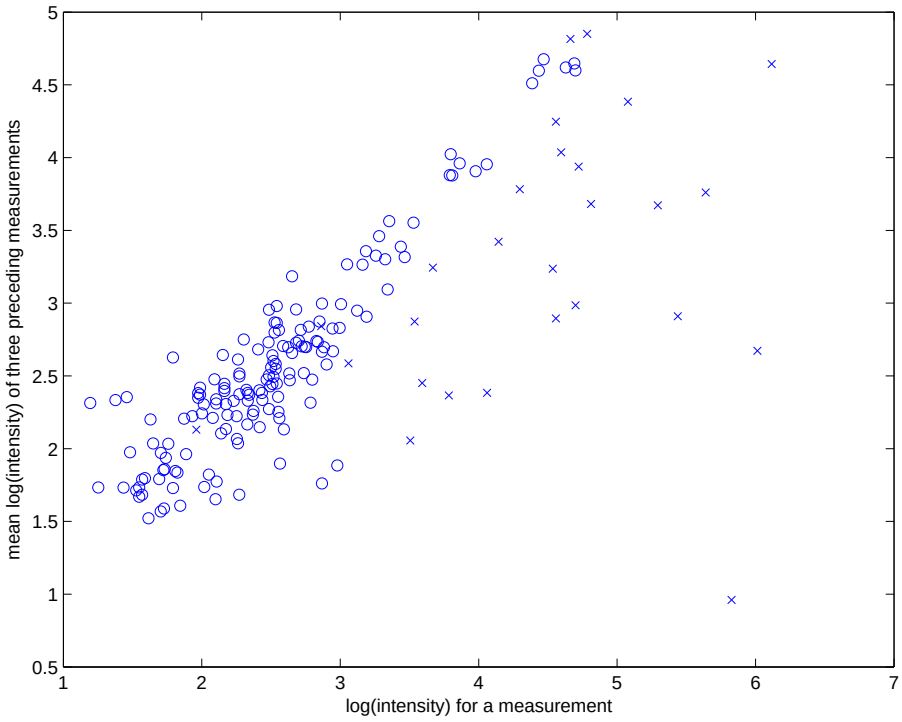

Figure 15: CA125 - cases for 1 year vs. all controls

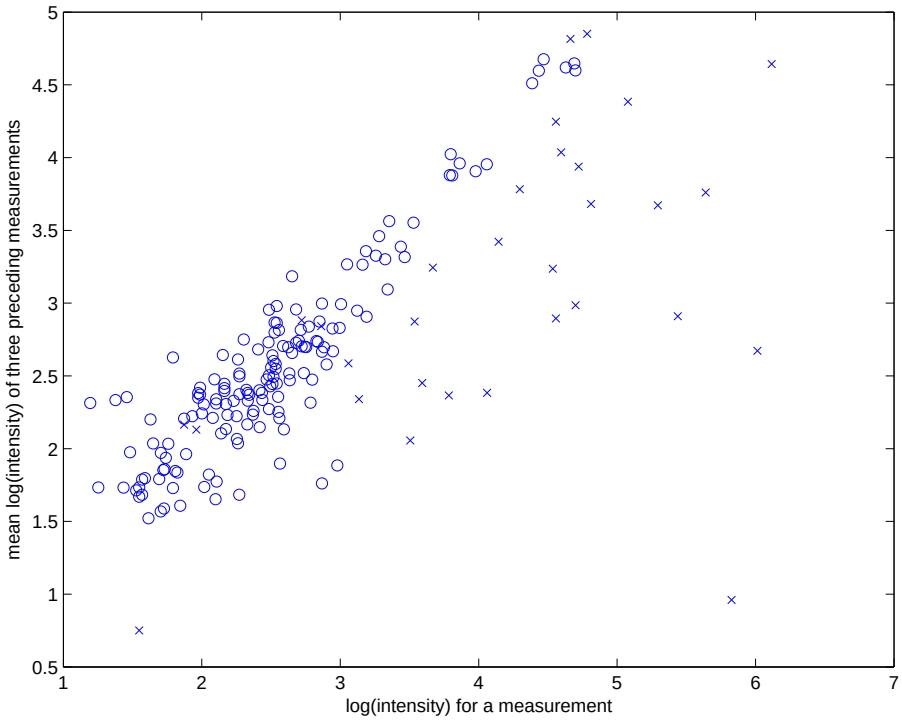

Figure 16: CA125 - cases for 2 years vs. all controls

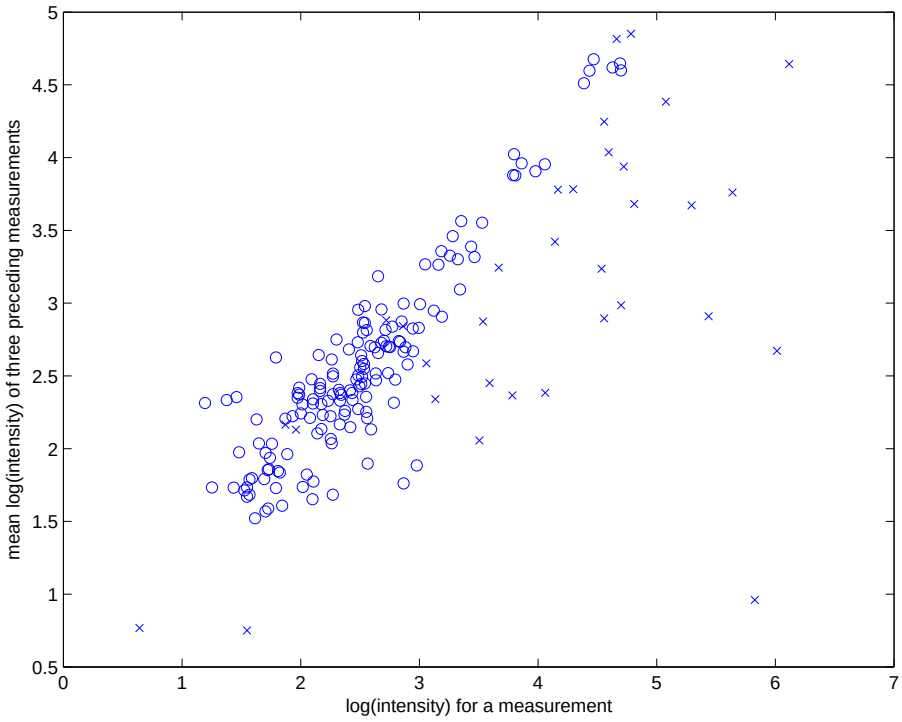

Figure 17: CA125 - cases for 3 years vs. all controls

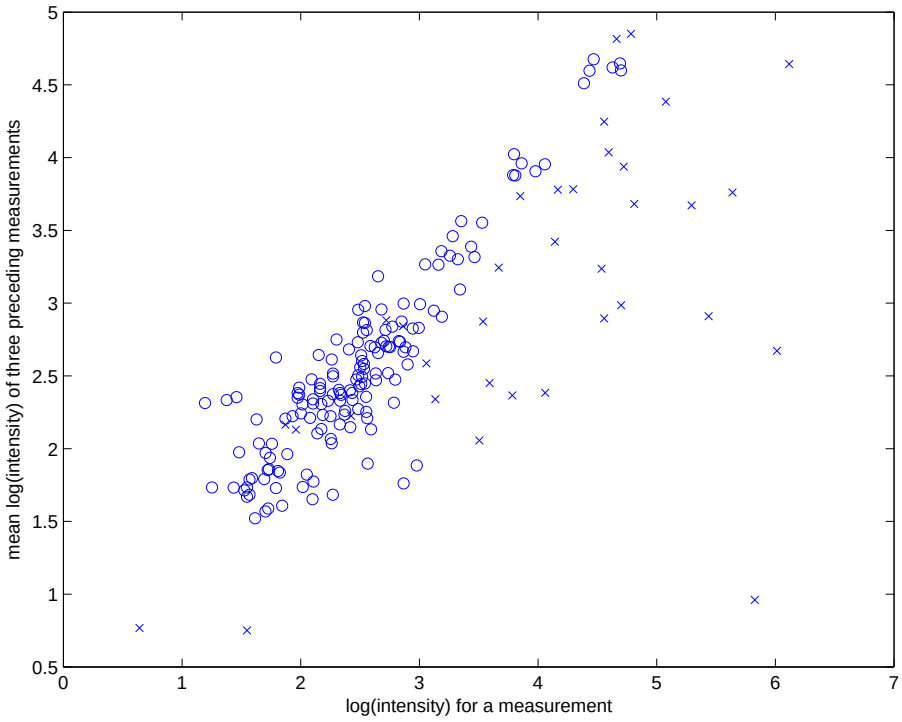

Figure 18: CA125 - cases for 4 years vs. all controls

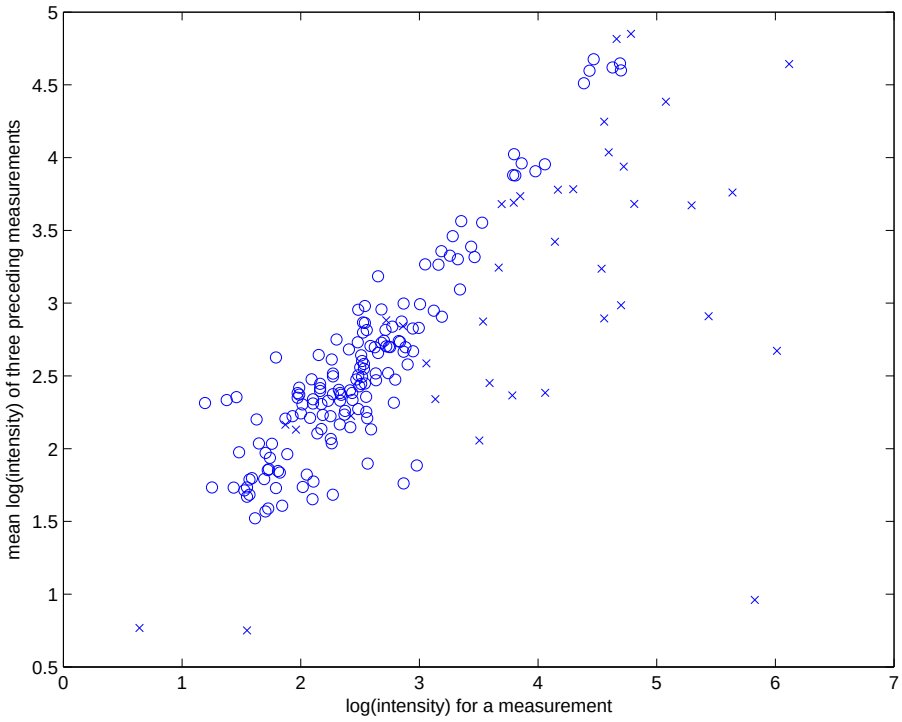

Figure 19: CA125 - cases for 5 years vs. all controls

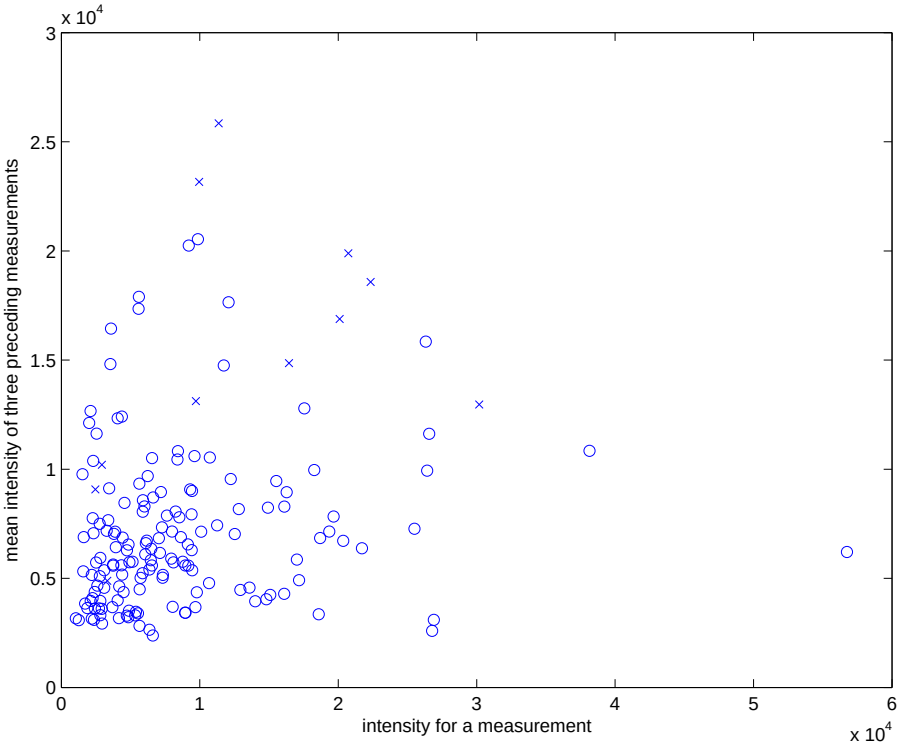

Figure 20: Peak 10: [2054.9 2067.6] - last moment cases vs. all controls

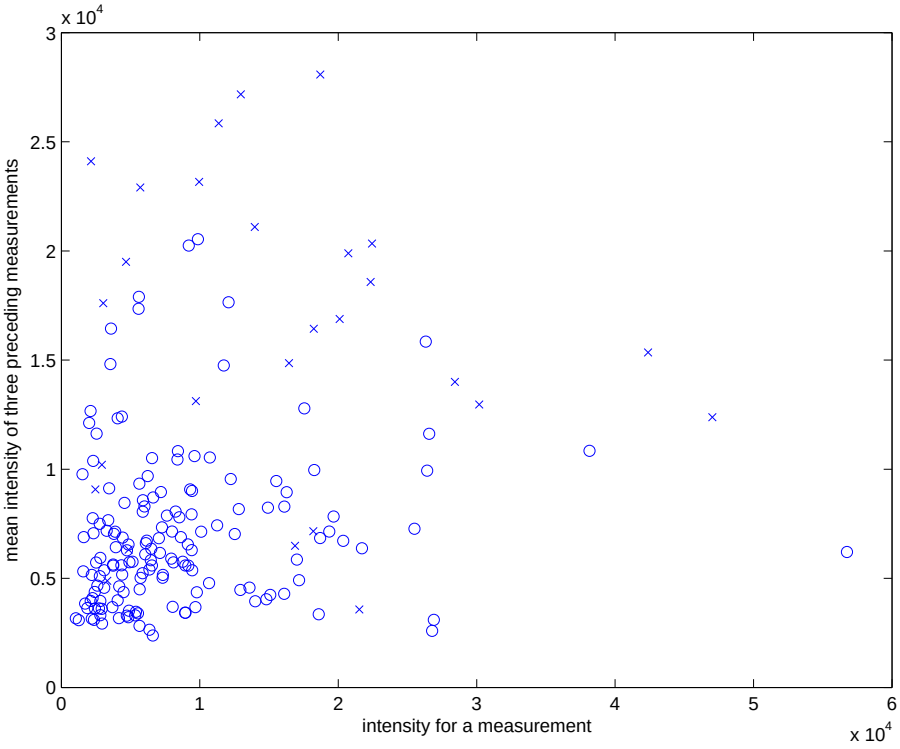

Figure 21: Peak 10: [2054.9 2067.6] - cases for 1 year vs. all controls

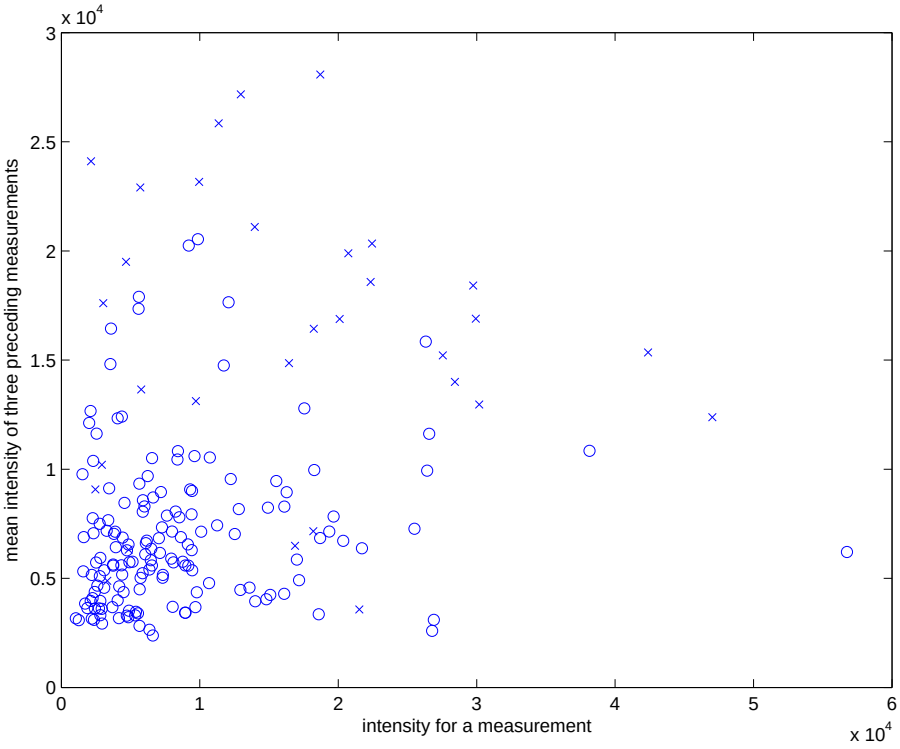

Figure 22: Peak 10: [2054.9 2067.6] - cases for 2 years vs. all controls

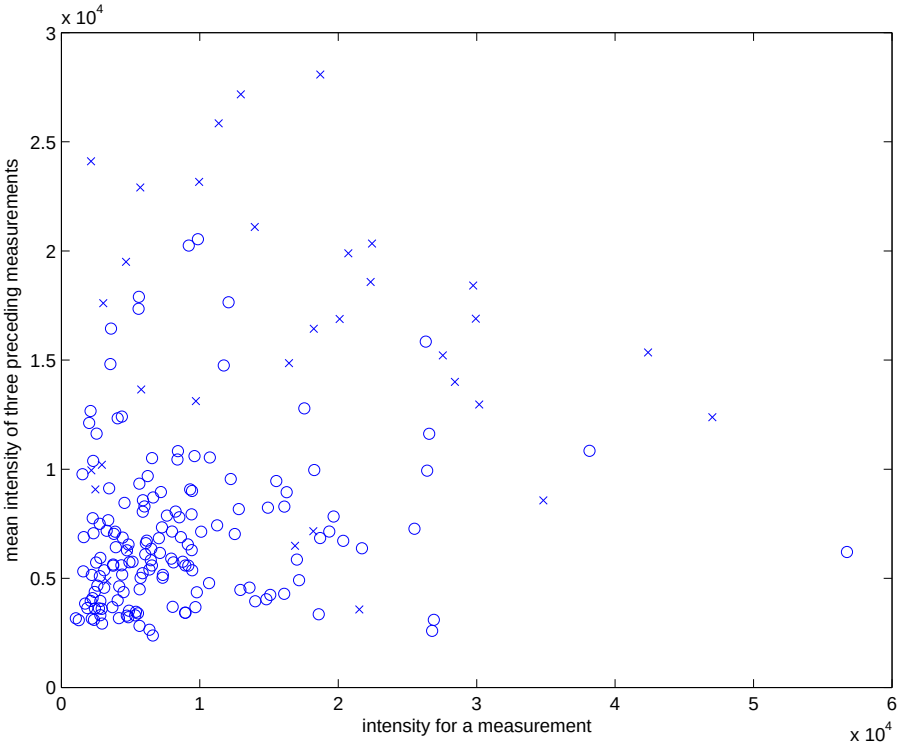

Figure 23: Peak 10: [2054.9 2067.6] - cases for 3 years vs. all controls

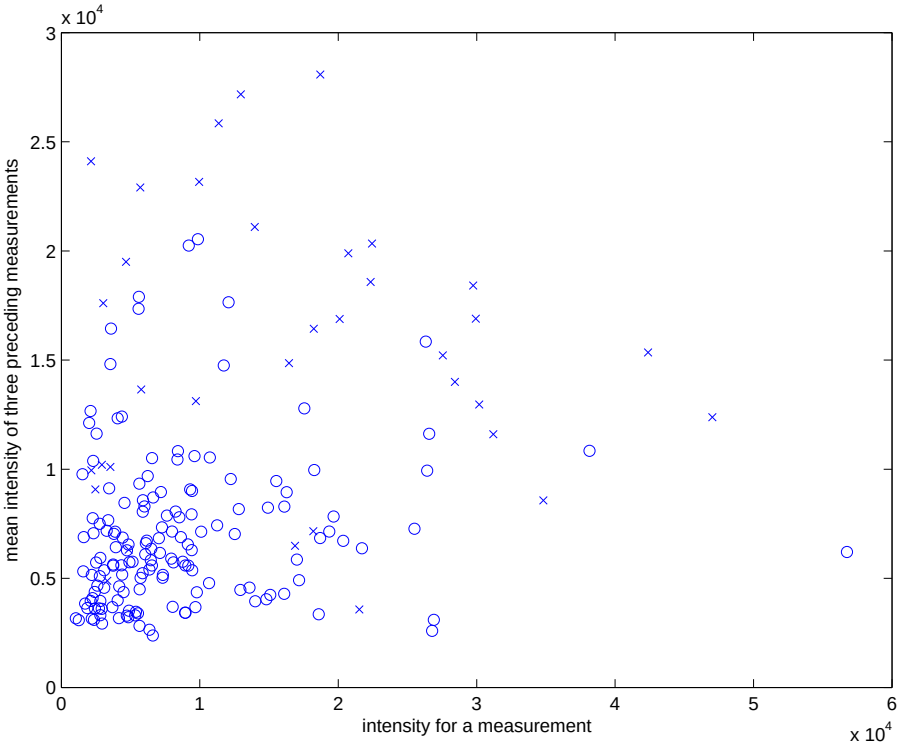

Figure 24: Peak 10: [2054.9 2067.6] - cases for 4 years vs. all controls

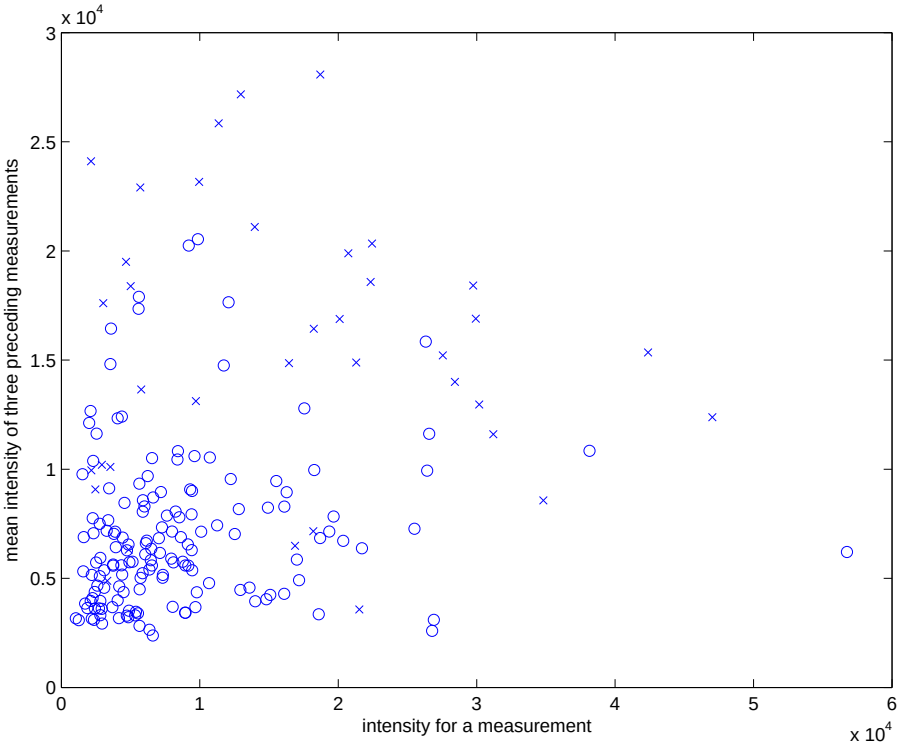

Figure 25: Peak 10: [2054.9 2067.6] - cases for 5 years vs. all controls

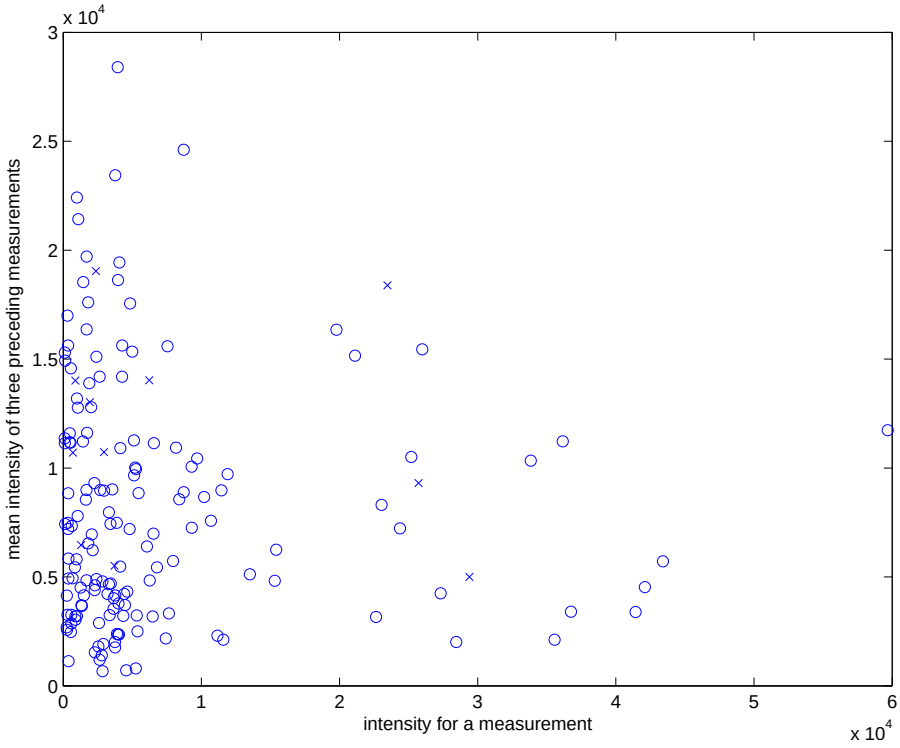

Figure 26: Peak 11: [2447.1 2457.8]- last moment cases vs. all controls

Then we compared the results with those gained on the base of CA125 analysis, see Table 10. The CA125 samples were taken from the same women and in the same moments as serum blood samples. So the comparison could be made in the same conditions and under the same restrictions for serum and CA125 samples using the same procedure of control.

We compared also the best results gained using additional feature reflecting prehistory, with those, which we could achieve without prehistory (in the same conditions). The following table, Table 11, shows the results of comparison:

One can see that using prehistory we always gain better results than in the case, when only current data are used.

10 Discussion

From the above results we can see that the peak number 10 (m/z = [2414.6 2418.9] delivers the best result, even in comparison with CA125 analysis. Other presented features also seem promising to be used for early diagnostics in combination, if we have more data for training. The other problem is to identify

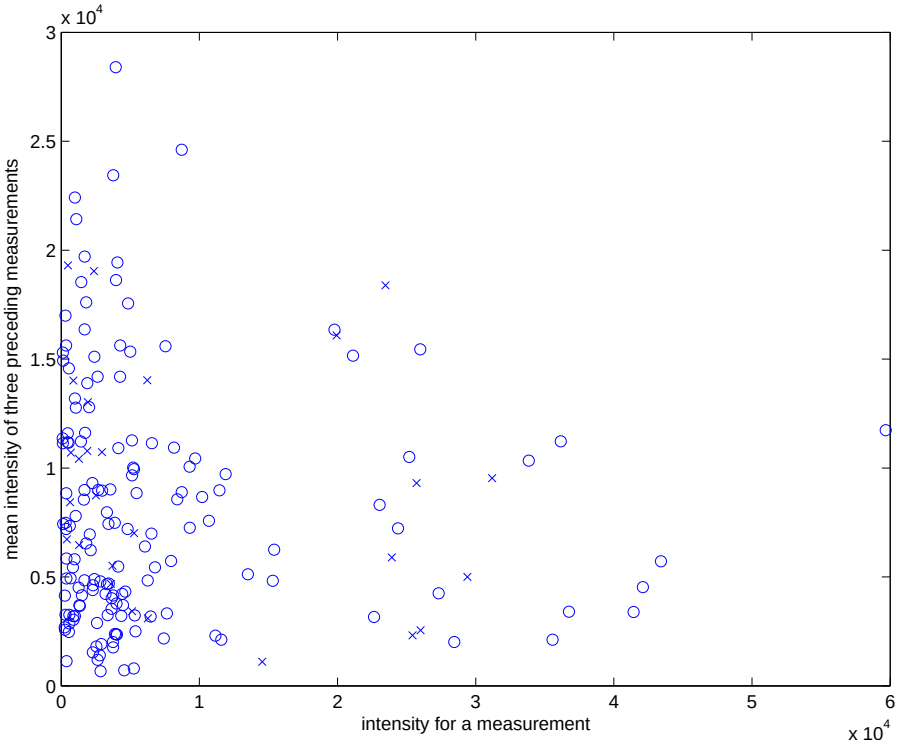

Figure 27: Peak 11: [2447.1 2457.8] - cases for 1 year vs. all controls

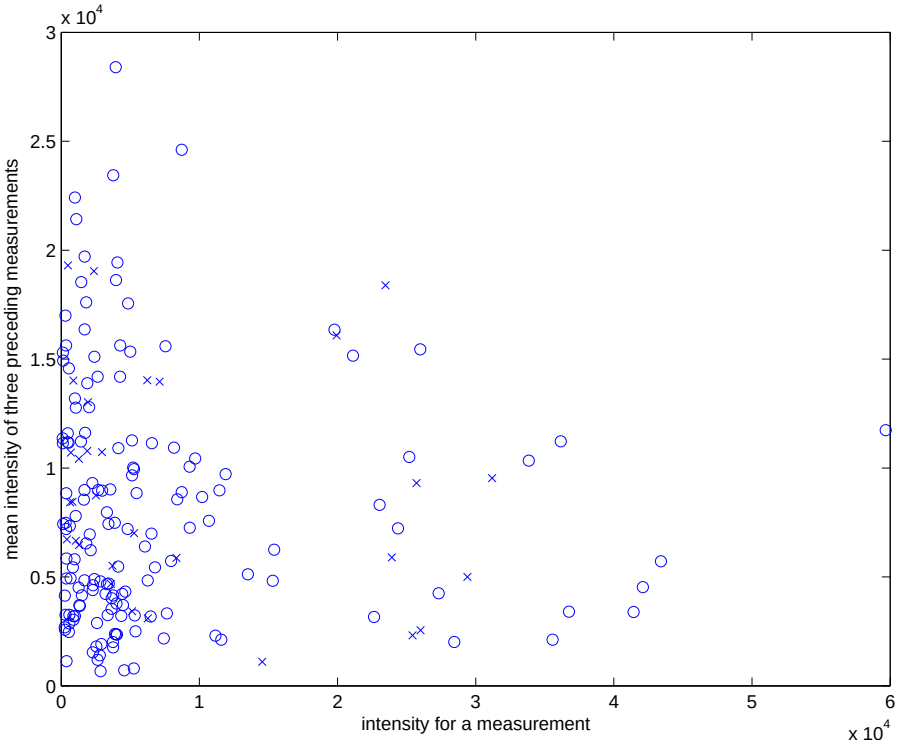

Figure 28: Peak 11: [2447.1 2457.8] - cases for 2 years vs. all controls

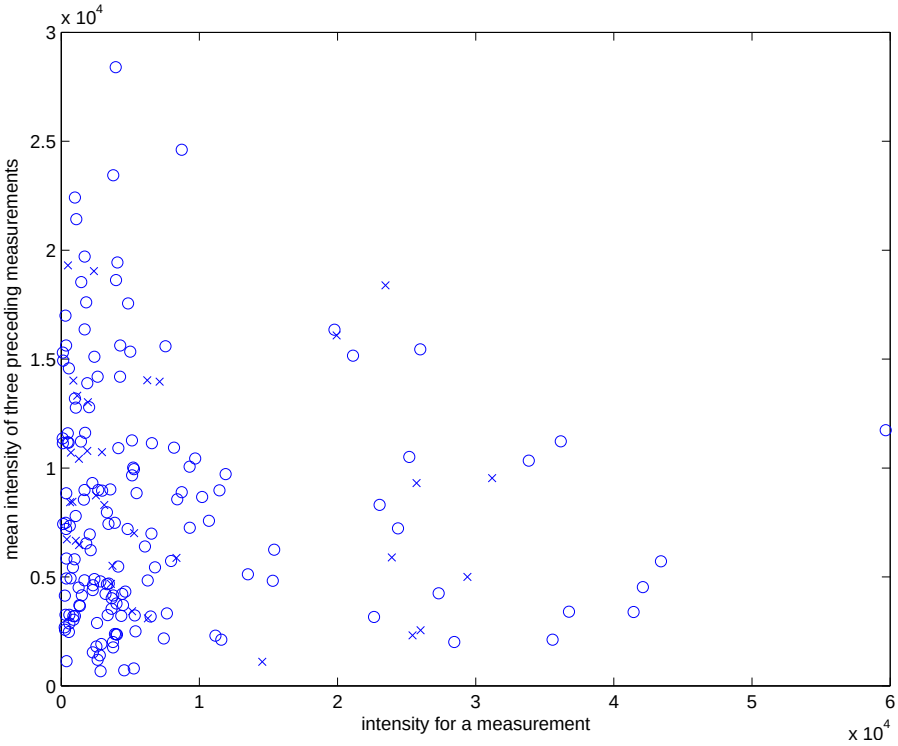

Figure 29: Peak 11: [2447.1 2457.8] - cases for 3 years vs. all controls

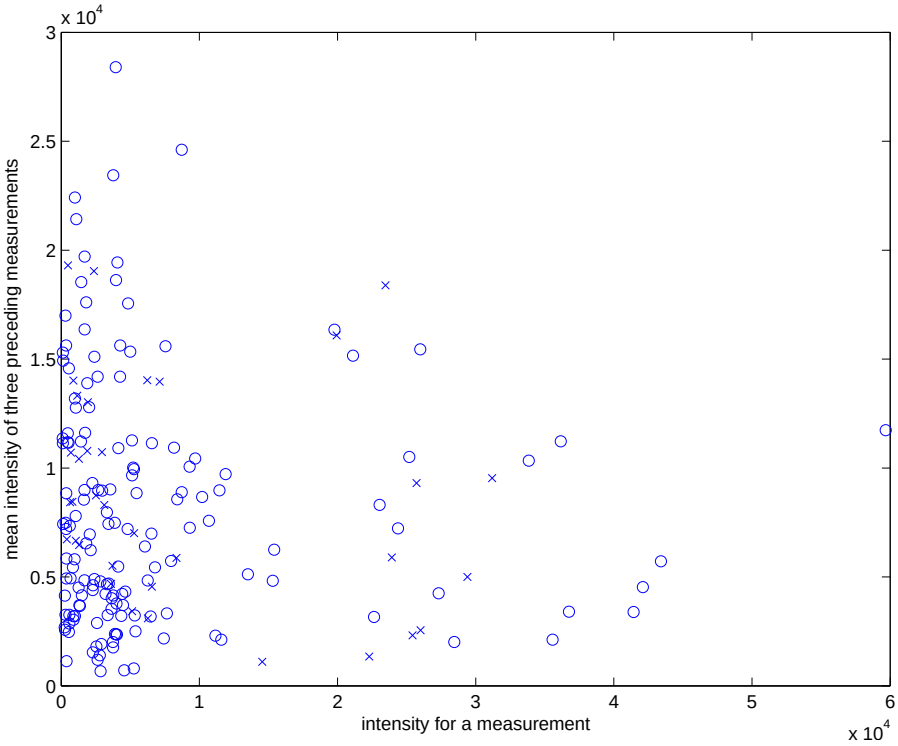

Figure 30: Peak 11: [2447.1 2457.8] - cases for 4 years vs. all controls

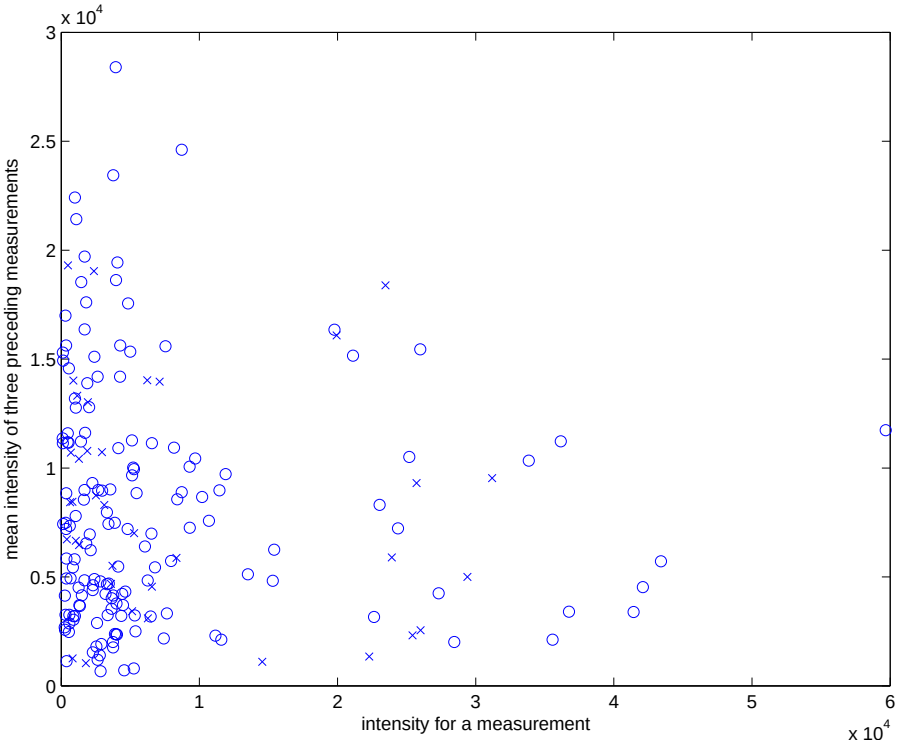

Figure 31: Peak 11: [2447.1 2457.8] - cases for 5 years vs. all controls

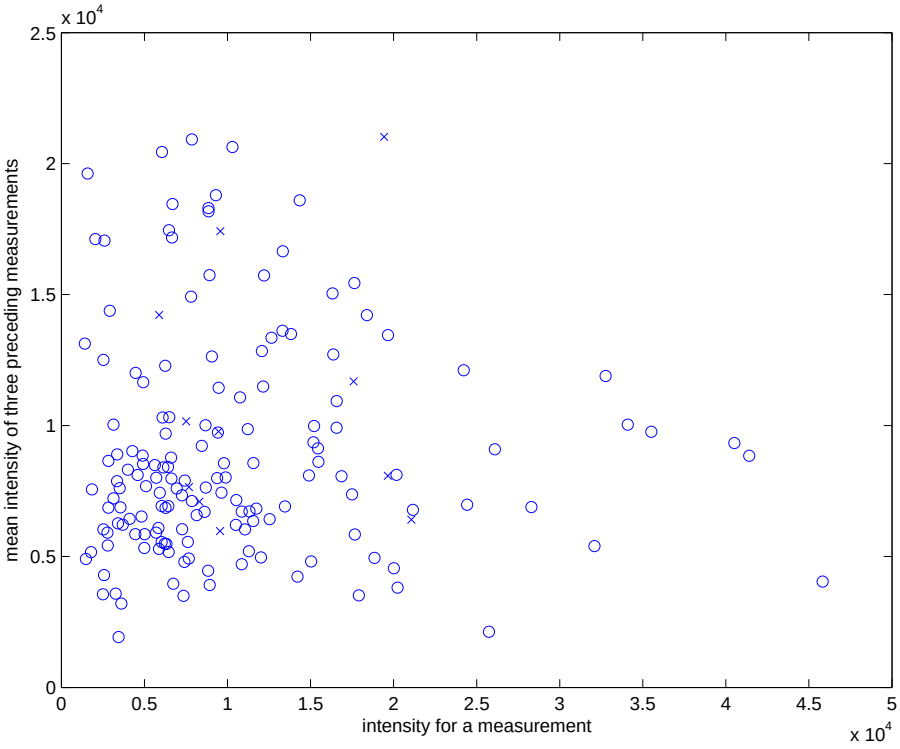

Figure 32: Peak 7: [2302.9 2308.4] - last moment cases vs. all controls

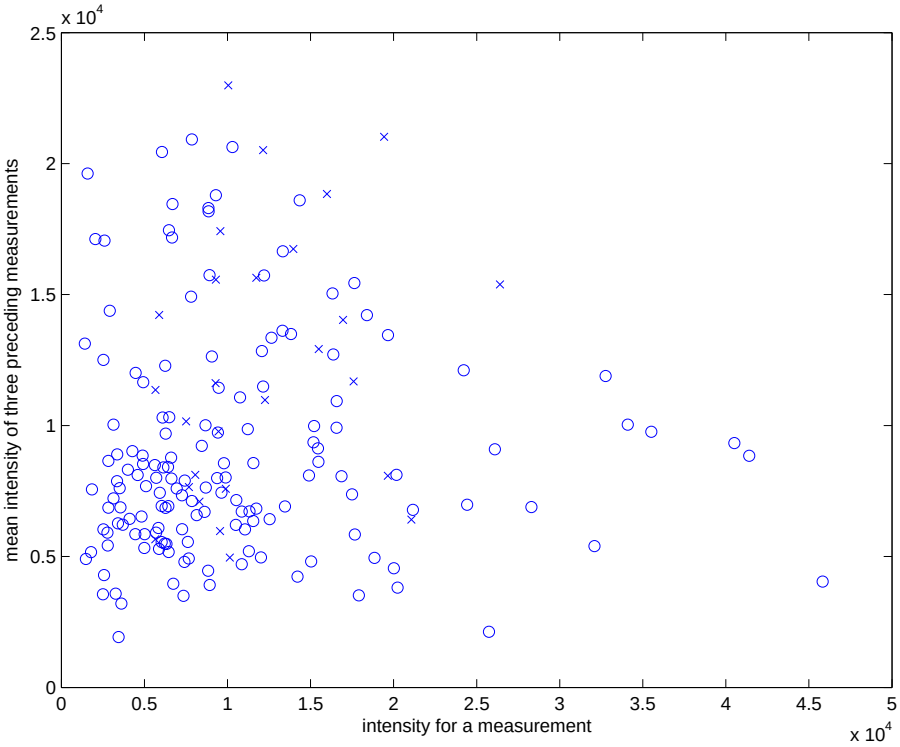

Figure 33: Peak 7: [2302.9 2308.4] - cases for 1 year vs. all controls

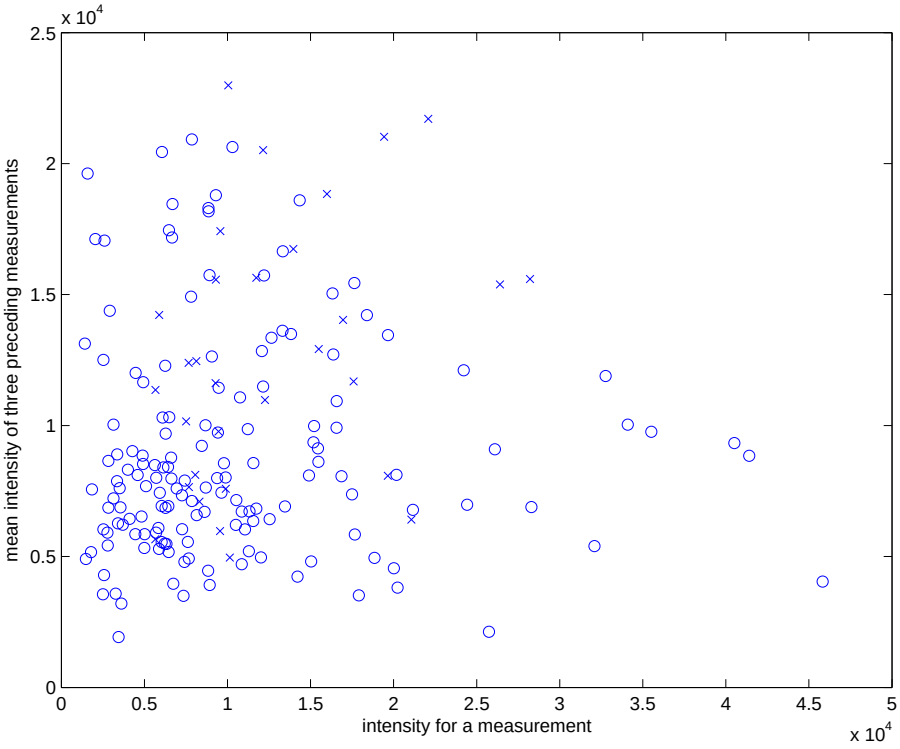

Figure 34: Peak 7: [2302.9 2308.4] - cases for 2 years vs. all controls

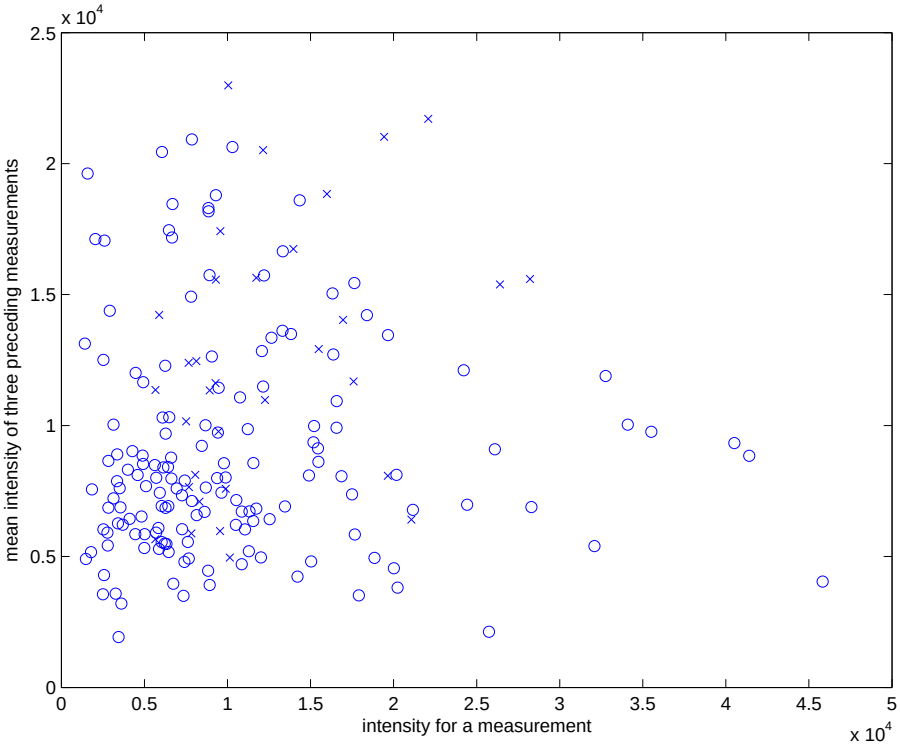

Figure 35: Peak 7: [2302.9 2308.4] - cases for 3 years vs. all controls

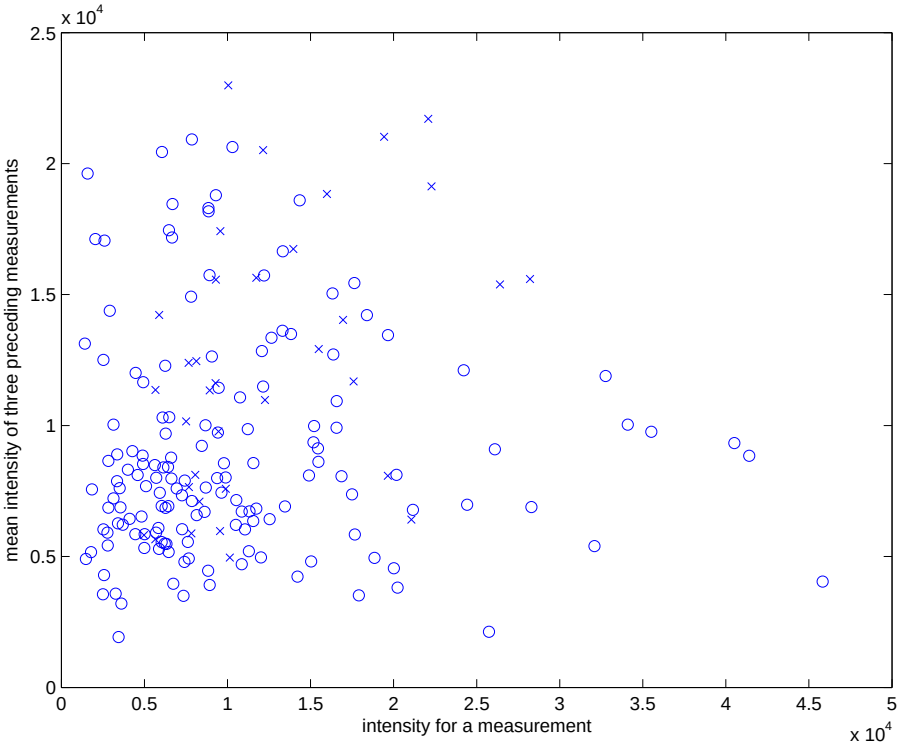

Figure 36: Peak 7: [2302.9 2308.4] - cases for 4 years vs. all controls

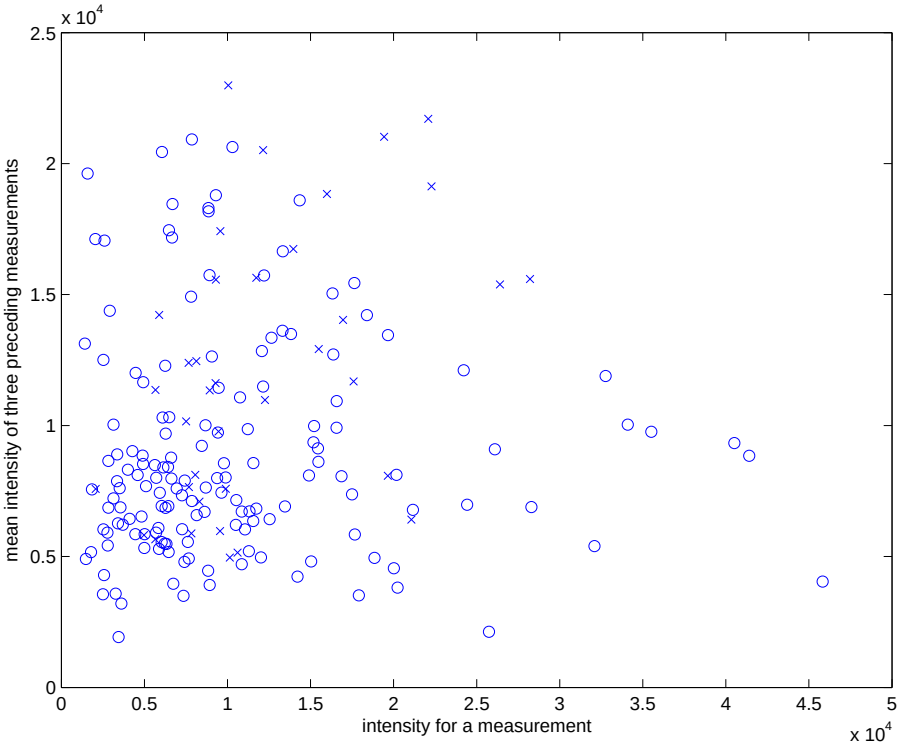

Figure 37: Peak 7: [2302.9 2308.4] - cases for 5 years vs. all controls

Error	*Case T = 0.*			
Param.	1/1	2/4	2/2	4/4
Type 1	3 (27%)	7 (63%)	8 (71%)	9 (82%)
Type 2	8 (5.2%)	10 (6.5%)	7 (4.6%)	0 (0%)
	Case T = 1.			
Param.	1/4	1/3	2/5	4/4
Type 1	7 (26%)	8 (30%)	10 (37%)	15 (55%)
Type 2	30 (19%)	25 (16%)	13 (8.4%)	7 (4.5%)
	Case T = 2.			
Param.	1/5	1/4	1/3	2/2
Type 1	8 (26%)	9 (29%)	10 (32%)	15 (48%)
Type 2	40 (26%)	35 (23%)	29 (19%)	8 (5.2%)
	Case T = 3.			
Param.	1/5	1/4	1/1	3/4
Type 1	9 (27%)	10 (30%)	12 (36%)	17 (52%)
Type 2	40 (26%)	35 (23%)	18 (12%)	8 (5.2%)
	Case T = 4.			
Param.	1/5	1/4	1/3	1/1
Type 1	11 (31%)	12 (34%)	13 (37%)	14 (40%)
Type 2	48 (31%)	43 (28%)	33 (21%)	18 (12%)
	Case T = 5.			
Param.	1/5	1/3	1/1	3/5
Type 1	13 (35%)	15 (41%)	16 (43%)	19 (51%)
Type 2	48 (31%)	33 (21%)	19 (12%)	14 (9.1%)

Table 12: Classification results using CA125

Feature	Error	with prehistory	without prehistory
Peak 10	Type 1	0 (0%)	4 (36%)
1/3	Type 2	7 (4.5%)	34 (22%)
Peak 7	Type 1	4 (36%)	5 (45%)
2/5	Type 2	11 (7.1%)	25 (16%)
Peak 11	Type 1	2 (18%)	7 (64%)
2/5	Type 2	16 (17%)	36 (24%)
CA125	Type 1	3 (27%)	4 (36%)
1/1	Type 2	8 (5.2%)	12 (7.9%)

Table 13: Classification results using prehistory

Our Peak	Tempst's Peak	Identified Peptide
11 [2447.1, 2457.8]	2451.11	Transthyretin (K) ALGISPFHEHAEVVFTANDSGPR
7 [2302.9, 2308.4]	2305.20	C4a (R) GLEEELQFSLGSKINVKVGGNS
8 [1926.8, 1933.3]	1927.94	apoA-IV SLAELGGHLDQQVEEFR
9 [2043.0, 2055.3]	2052.89	apoA-I ATEHLSTLSEKAKPALEDL (R)
2 [3236.2, 3244.3]	3239.22	Fibrinogen α SYKMADEAGSEADHEGTHSTKRGHAKSRPV (F
5 [1796.4, 1819.2]	1807.78	apoA-I ELQEGARQKLHELQE
12 [1894.2, 1901.6]	1895.99	C4a RNGFKSHALQLNNRQI (R)

Table 14: Comparison with Tempst's group findings

the peaks with certain proteins, peptides or proteases. We have made an attempt to establish very narrow intervals of m/z values in order to find out the corresponding protein biomarkers. The further experiments would require to confirm or disconfirm these results.

We also compared the 15 most popular peaks identified in Table 1 with those published by Paul Tempst's group [4]. Seven of our peaks match with their results, see Table 14 and two of the seven peaks (i.e. peaks 11 and 7) are useful for classification.

11 Acknowledgements

This work has been supported by MRC grant S505/65: "Proteomic analysis of the Human Serum Proteome for Population Screening, Diagnosis and Biomarker Discovery", and the Royal Society grant.

12 References

1. U. Menon, Pilot Study Document, 2005.

2. Rice Wavelet Toolbox, http://www-dsp.rice.edu/software.

3. Steven J. Skates, Usha Menon, Nicola MacDonald, Adam N. Rosenthal, David H. Oram, Robert C. Knapp and Ian J. Jacobs, Calculation of the Risk of Ovarian Cancer from Serial CA-125 Values for Preclinical Detection in Postmenopausal Women, Journal of Clinical Oncology, Vol 21, Issue 90100 (May), 2003.

4. Josep Villanueva, David R. Shaffer, John Philip, Carlos A. Chaparro, Hediye Erdjument-Bromage, Adam B. Olshen, Martin Fleisher, Hans Lilja, Edi Brogi, Jeff Boyd, Marta Sanchez-Carbayo, Eric C. Holland, Carlos Cordon-Cardo, Howard I. Scher and Paul Tempst, Differential exoprotease activities confer tumor-specific serum peptidome patterns, J. Clin. Invest. 116:271-284, 2006.