Data Analysis of Human Serum Proteome II
UKCTOCS Data: Pilot Study
CLRC Technical Report 01-09-2006,
November 2005 – September 2006

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

Abstract

The paper describes analysis of human serum proteome to establish several ovarian cancer (OC) biomarkers for early detection of the disease. The data were collected in UKCTOCS study over 7 years - "serial data". The analysis indicates that certain peptides are very important for early diagnostic of ovarian cancer, and they have more predictive power in the early stages than CA125. Further experiments, however, are required to study the dynamics of the control set collected over the same years in order to make reliable classification.

1 Introduction

Recent advances in the analysis of human serum proteome aim to establish novel disease biomarkers that would allow the early detection of diseases. The current techniques include analysis of the serum data using mass spectrometry (m/s). The output of m/s work is a large volume of high-dimensional data and it requires modern methods of data analysis. This report describes several machine learning techniques that have been used in order to establish "proteomic pattern diagnostic". The techniques were applied to a subset of a large bank of serum data collected in the 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, denoising, subtracting the baseline and normalising the data. Subsequently, peak identifications and peak alignment techniques were applied in the latest stages of pre-processing. The remaining sections of the report are devoted to feature selection techniques that can been applied to classify between cases with ovarian cancer (OC) and control (healthy individuals). In addition to using samples from single points in time as features we also experimented with the "prehistory" of the diagnostic process by considering different time slots (0, 1, 2, ... 5 years) before the actual diagnosis is made. The results show that certain peaks and their combinations (and the corresponding peptides) are very important for early diagnostic of OC, and they are more informative in the early stages than CA125. However, additional experiments have shown that different subsets of the data are not comparable, and further research is required to confirm these findings.

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 the Sloan-Kettering Center. In this pilot study the data were divided into two sets. **Set 1** has 266 samples, of which 91 were case samples taken from 19 women with OC, and 175 control samples. The control samples were selected to match case samples (usually 2 controls samples were selected for each case sample). The women in the study were observed for 7 years (1995-2001) and the samples collected in this period were called *serial samples*. Typically, each of the 19 cancer women have 2 to 12 serial samples taken in those years; and most of the healthy patients have just a single sample. For all cancer women, the last sample was taken at the moment when the diagnosis was made. In addition, **Set 2** was also obtained that has 305 samples from 50 healthy women. Each healthy woman has between 5 and 9 serial samples 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 was 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 range of mass-to-charge ratio (m/z) [700, 4000] and high range m/z of [4000, 15000] were obtained separately. In total, 125,206 data points were generated by the mass spectrometer where mass-to-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 a 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.

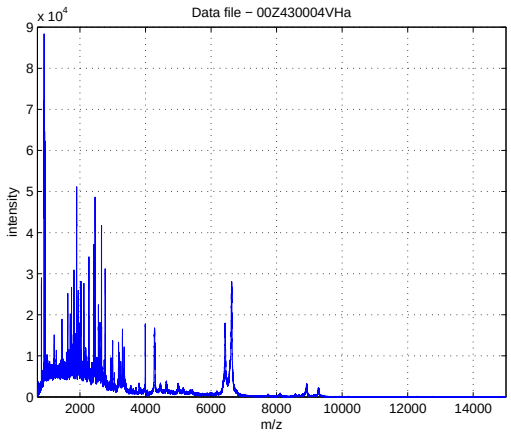

Figure 1: An example of raw MS data

3 Pre-processing

Mass spectrometry instruments are very sensitive and artefacts 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 artefacts and isolate the true protein signal. The goals of pre-processing are to reduce noise, normalise the spectra from different samples and reduce the 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 section we describe our pre-processing of the raw data that includes: calibration, baseline subtraction, smoothing, normalisation and peak 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.

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].

3.2 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

Algorithm 1 Calibration

Require: m/z values of 13 peaks
  fix A_s and B_s in the formula (1) (for example A_s=0.5 and B_s=1.0)
  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

the high frequency noise signal in the spectra by averaging the intensities within a moving window.

3.3 Baseline Subtraction

The goal of baseline subtraction is to remove systematic artefacts, usually due to matrix and chemicals used in the experiments or to detector overload. Ideally the 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 is described in Algorithm 2. Where the algorithm refers to Piecewise Cubic Hermite Polynomial Interpolation of B we use the pchip() function in matlab [2].

3.4 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 in Figure 3.

3.5 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.

Algorithm 2 Baseline Subtraction

Require: Paramiter σ - Maximum mass separation (e.g. $\sigma = 0.0015$ (0.15%))
Require: Number Baseline correction iterations k
Require: Spectrum X
Define the set of intensities in X as X^I and the corresponding intensities (at the same points) for the baseline B as B^I.
Define the set of m/z values in X as X^M and the corresponding m/z values (at the same points) for the baseline B as B^M.
Find all local maxima in X and call set of the m/z locations of these maxima/local peaks P.
Define a new set P_s, which is the peaks from P sorted relitive to peak intensity in decending order.
for Each peak $p_i \in P_s$ **do**
  for Each peak $p_j \in P$ such that $p_j \neq p_i$ **do**
    if $\frac{|p_i - p_j|}{p_i} < \sigma$ **then**
      Remove p_j from P
    end if
  end for
end for
Find the point with minimum intensity between each adjacent peak in P, we call these troughs and denote the set by T
$B = P$
for $i = 1 : k$ **do**
  B = Piecewise Cubic Hermite Polynomial Interpolation of B
  flag = 0
  $j = 1$;
  for Each $x_i \in X^I$ and each coresponding $b_i \in B^I$ **do**
    if $x_i < b_i$ **then**
      if flag == 0 **then**
        $R_t = i$
        flag = 1
      end if
    else
      if flag == 1 **then**
        $R(j) = (R_t, i)$
        $j = j + 1$
      end if
    end if
  end for
  for Each pair $r \in R$ **do**
    Find the point $p = \arg\max_p(B^I_p - X^I_p)$ where p is in the range r
    Create a new point on the baseline by setting $B_p = X_p$.
    Remove any point $k \neq p$ from B if $\frac{|(B^M_p - B^M_k)|}{B^M_p} < \sigma$
  end for
end for
Final baseline B* = Piecewise Cubic Hermite Polynomial Interpolation of B
Calculate baseline subtracted spectra $X'^I = X^I - B^{*I}$

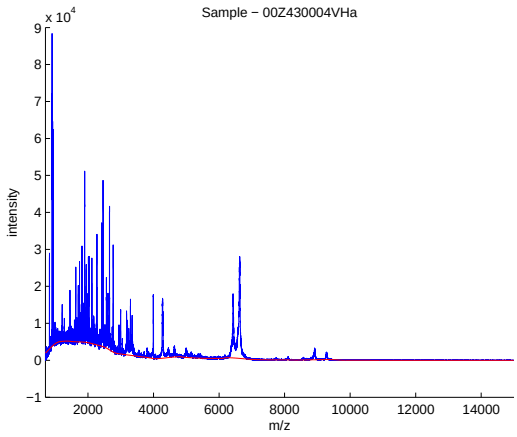

Figure 2: Estimated baseline (shown in red)

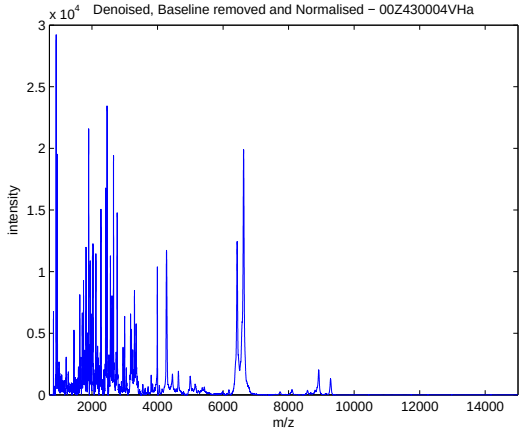

Figure 3: An example of pre-process data after smoothing, baseline subtraction and normalisation

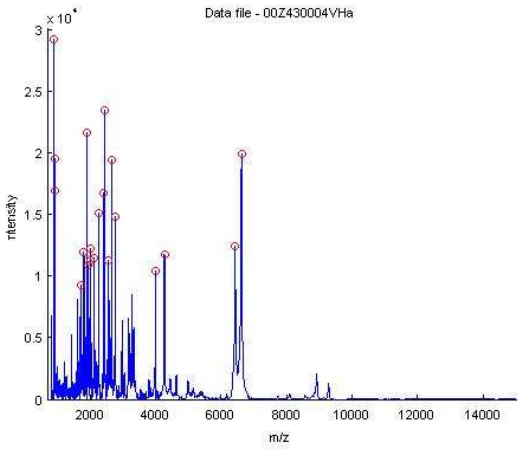

Figure 4: Peak identification

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 a 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 chooses 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 threshold of 9000.

3.6 Peak Alignment

To make an inference about trends across a number of spectra, we need to relate the peaks identified in one spectrum to the peaks found 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 a unique correspondence between them. 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 split into clusters. Cluster definition is done 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 into clusters of order 1. Then we

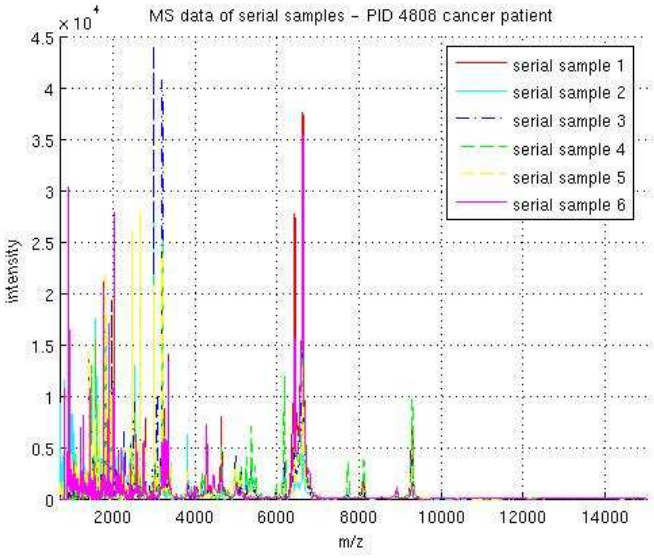

Figure 5: Pre-processed MS data of serial samples - Cancer Patient (ID=4808)

test if each sample has no more than 1 peak in a cluster. If so, the cluster is considered as final. Otherwise, we look at whether there is an interval exceeding $d_2 < d_1$, dividing occurrences of one sample peak within the cluster. If so we divide the cluster of order 1 into smaller ones. Otherwise we consider it as final.

Now for each sample peak we assign the number of its cluster (in cases when there is 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 pre-processing steps described in section 3 were applied to each of 565 raw MS data files. Figure 5 and Figure 6 illustrate an example of pre-processed serial samples from a cancer woman and a normal woman, respectively.

In total, 8372 peaks were identified in all the analysed serum samples. After the pre-processing, 340 peak groups were identified after peak alignment. 48 peak groups were found which are present in more than 20% or so of all samples – see Table 1.

Peak No	m/z value	m/z range
1	6645.9	6634.6 – 6652.8
2	3188.9	3185.3 – 3191
3	2004.2	2001.1 – 2005.5
4	3330.4	3325.2 – 3333.4
5	9307.1	9291.2 – 9318.2
6	2982.3	2978.8 – 2985.2
7	1764.5	1761.8 – 1766.8
8	818.46	817.45 – 818.9
9	4291	4284 – 4296.5
10	3172	3168 – 3176.4
11	2548.5	2544.6 – 2550.7
12	937.35	936.48 – 937.76
13	3280.1	3273.9 – 3282.1
14	2262.3	2259.1 – 2265.2
15	8126.7	8111.5 – 8132
16	1964.1	1961.2 – 1965.6
17	6447.9	6441.3 – 6460.2
18	2562.8	2561 – 2564.2
19	1888.7	1886.1 – 1889.7
20	899.92	898.6 – 900.99
21	1442.4	1440.7 – 1443.3
22	2020.9	2019.7 – 2022.3
23	5010.8	5006.7 – 5014.6
24	8943.1	8938.6 – 8963
25	4652.6	4647 – 4659.3
26	5379.4	5373.1 – 5387.8
27	3788.6	3786.5 – 3791.9
28	2405.1	2404 – 2406.2
29	3229.8	3226.7 – 3234.2
30	1857.6	1855.9 – 1858.4
31	2643.6	2639.7 – 2644.7
32	1388	1385.8 – 1389.3
33	2016.8	2013.6 – 2019.6
34	3206.4	3203.1 – 3208.5
35	2447.5	2446.2 – 2448.7
36	1577.9	1575.9 – 1578.9
37	3297.2	3295.9 – 3300.8
38	1204.9	1203.5 – 1205.7
39	2109.1	2107.4 – 2111
40	2755.2	2750.9 – 2758.2
41	3977.7	3974.7 – 3981
42	8134.4	8132.1 – 8146.8
43	1808.5	1806.3 – 1809.3
44	8935.9	8918.7 – 8938.5
45	2503	2501.6 – 2504.1
46	1937.2	1934.7 – 1939.3
47	2723.82	2720 – 2726.1
48	1477.32	1475.6 – 1478.5

Table 1: Most popular peaks among the samples

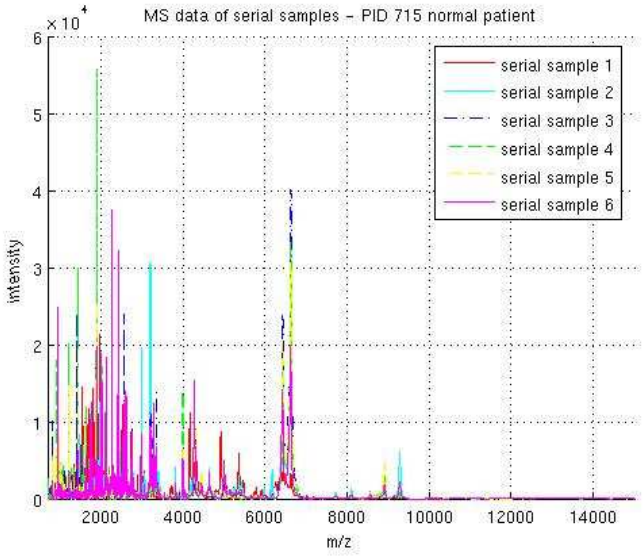

Figure 6: MS data of serial samples - Normal Patient (ID=715)

5 Feature selection techniques

After the pre-processing 340 peak groups were identified out of the original 8372 peaks. We then selected the 48 most frequent peak groups (see section 3.6) and they are presented in Table 1. Our next step is to use these peaks as features in the pattern recognition task. We should be able to derive decision rules in order to separate OC cases from controls.

5.1 The techniques

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 types of error: 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 can be considered as an object for pattern recognition. Patients are grouped in classes according to their diseases or absence of disease. The symptoms are the features and some general information such as age, sex is often added to the set of features. In our case, we shall be using objective measurements (intensities of peaks) as the features space in addition to some clinical and general information.

5.2 Significance level of selected features

The feature space in our data is formed by the intensities of the 48 selected peaks for each MS-sample. That means that for each sample, a 48-dimensional vector of intensity values can be formed to characterize the sample. In this setting a peak is the maximum of intensity at the spectrogram within an interval corresponding to the peak irrespective of whether the peak was large or small.

The first step is to compare the probability distributions of the peak intensities for control and case samples. To estimate the difference for a certain significance level we used the Mann-Whitney The test returns P value of whether two sample sequences could be generated by the same probability distribution, assuming that the samples are independent. Our null hypothesis here is that the samples are coming from the same probability distribution; i.e. there is no difference between the samples.

5.2.1 All examples taken in the last time-point

To begin with 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). In this first set of experiments we considered only the last serial sample (a sample that was taken at the last time-point - the point when the diagnosis was made and the disease was found). The intensities of each of the 48 peaks were calculated using our pre-processing techniques and the significance test was applied. In Table 2 the results are shown for each of the 48 peaks, and for each (last time-point) sample. In column 2 the mean m/z value of the peaks is shown, column 3 shows the m/z interval corresponding to the peak, column 4 shows median intensities of each m/z peak in the control set, column 5 corresponds to peak intensity ratio, which was calculated by dividing the median value of a peak in the cancer group by the median value in the control set, and column 6 gives corresponding P values. For comparison, a line corresponding to CA125 analysis is added.

We see that P values for all peaks are very high in comparison with that of CA125. This means that the 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 probability distribution. The null hypothesis is rejected only for CA125 entry with a very small P value. This picture related to the whole set of samples of control and OC cases, and it would be useful to see if there are some changes in the importance of certain peaks when we consider the samples in dfferent time slots.

5.2.2 Serial samples

Since each serial sample came from the same person but over a period of time using the same UKCTOCS protocol, it was reasonable to use **averages** over serial samples peak intensities and apply the significance test to check whether their probability distributions are different for normal and cancer cases. It meant that we had to exclude people who did not have serial samples but only had a single measurement.

Last-moment samples. In the following experiments we considered only those women that had no less than 4 serial samples. After this restriction there remained only 11 patients with a cancer diagnosis and 49 healthy women.

Peak No	m/z	Range	Median intensity (normalised)	Median ratio	p-value
1	6645.9	6634.6 – 6652.8	128.76	1.0377	0.91387
2	3188.9	3185.3 – 3191	34.128	1.8127	0.012547
3	2004.2	2001.1 – 2005.5	49.143	1.0238	0.42938
4	3330.4	3325.2 – 3333.4	23.911	1.4172	0.024758
5	9307.1	9291.2 – 9318.2	8.0021	1.1403	0.55193
6	2982.3	2978.8 – 2985.2	24.246	1.2549	0.55193
7	1764.5	1761.8 – 1766.8	33.79	1.1368	0.30418
8	818.46	817.45 – 818.9	15.114	1.3889	0.93188
9	4291	4284 – 4296.5	14.438	0.7757	0.77347
10	3172	3168 – 3176.4	12.663	1.1826	0.271
11	2548.5	2544.6 – 2550.7	15.017	0.90985	0.92633
12	937.35	936.48 – 937.76	8.433	1.4712	0.76281
13	3280.1	3273.9 – 3282.1	8.6912	1.9269	0.22269
14	2262.3	2259.1 – 2265.2	20.548	1.1628	0.89039
15	8126.7	8111.5 – 8132	4.1111	0.93207	0.40534
16	1964.1	1961.2 – 1965.6	15.402	1.398	0.24038
17	6447.9	6441.3 – 6460.2	59.694	0.93121	0.86837
18	2562.8	2561 – 2564.2	11.957	1.2926	0.23074
19	1888.7	1886.1 – 1889.7	10.356	0.77269	0.71784
20	899.92	898.6 – 900.99	9.9228	0.85271	0.98191
21	1442.4	1440.7 – 1443.3	8.3327	1.1578	0.70502
22	2020.9	2019.7 – 2022.3	41.263	1.2567	0.67928
23	5010.8	5006.7 – 5014.6	11.029	0.81472	0.10889
24	8943.1	8938.6 – 8963	5.2508	1.1853	0.79223
25	4652.6	4647 – 4659.3	11.702	0.74588	0.70243
26	5379.4	5373.1 – 5387.8	6.411	0.84285	0.39363
27	3788.6	3786.5 – 3791.9	4.5351	1.1878	0.11763
28	2405.1	2404 – 2406.2	10.96	1.1516	0.51299
29	3229.8	3226.7 – 3234.2	10.381	1.1831	0.11763
30	1857.6	1855.9 – 1858.4	10.173	0.87123	0.62399
31	2643.6	2639.7 – 2644.7	9.0744	1.0591	0.7469
32	1388	1385.8 – 1389.3	7.8566	0.95715	0.47123
33	2016.8	2013.6 – 2019.6	50.994	1.0689	0.66655
34	3206.4	3203.1 – 3208.5	7.4657	1.5213	0.034938
35	2447.5	2446.2 – 2448.7	6.8544	1.1567	0.84373
36	1577.9	1575.9 – 1578.9	5.4893	0.96582	0.66655
37	3297.2	3295.9 – 3300.8	7.6509	1.1204	0.30914
38	1204.9	1203.5 – 1205.7	6.0051	0.90137	0.9402
39	2109.1	2107.4 – 2111	12.717	0.52881	0.36711
40	2755.2	2750.9 – 2758.2	7.8938	1.1092	0.64387
41	3977.7	3974.7 – 3981	1.2874	0.78671	0.77614
42	8134.4	8132.1 – 8146.8	4.1111	0.87697	0.3209
43	1808.5	1806.3 – 1809.3	4.1082	1.3571	0.34172
44	8935.9	8918.7 – 8938.5	5.1757	1.2048	0.73373
45	2503	2501.6 – 2504.1	8.1174	0.87972	0.95131
46	1937.2	1934.7 – 1939.3	9.0519	1.2904	0.60438
47	2723.8	2720 – 2726.1	8.5138	1.6115	0.57074
48	1477.3	1475.6 – 1478.5	5.8744	1.0542	0.51976
		CA125	2.5518	1.7988	1.34912e-006

Table 2: All samples from 19 women with cancer and 218 healthy women

In the first set of experiments we took into account only the last sample for these 11 cases and 49 controls and applied the significance test to them. The results are shown in Table 3. The results show that peak 2 has become comparable with CA125; that is, we can reject the null hypothesis for this peak (feature), and assume that the controls and cases are not coming from the same probability distribution.

Average of 3 samples excluding the last one.

In the next set of experiments we tried to establish if the diagnosis can be made long before the results of the last sample became available. For this purpose, we excluded the last sample (the sample taken at the last time-point) and compared the control and case samples only over the 3 samples preceding the last one by taking the averages of intensities for controls and cases. The test results are shown in Table 4.

One can see that P values for 8 of the peaks are smaller in comparison with CA125 indicating that the controls and cases are unlikely to come from the same distribution. The null hypothesis is rejected for those peaks. Peaks are 9,12,14,19,21,28,38 and 41.

However, the difference in distributions does not mean that a feature can be used for reliable diagnosis. For example, it might be due to the fact that distributions are indeed different, but overlapping, and the overlapping distribution inevitably implies errors. Different distributions for the cases and controls may have some important biological meaning. However, from the clinical point of view, we need to obtain features with "good" classification abilities. Further investigation is required to analyse these peaks in order to rely on them for early diagnosis.

5.3 Prehistory parameter

According to the classical scheme of pattern recognition the classes are fixed, and the task is to develop the decision rules that would allow the recognition of future objects as belonging to one class or another (in a two-class problem). In our case we have two classes: the control and case individuals. However, what exactly constitutes the case (OC) in our study? To begin with we can consider as cases (OC) 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 the rest of the samples were healthy. But, since we shall have different number of samples changing over the period of time, we include an input parameter to characterise *prehistory* among other input parameters.

We consider as an **object for recognition** not an individual, but **a moment of sampling**. We include only those cases (moments), which had at least k preceding measurements from the same person (usually k is a fixed constant equal to 2 or 3).

In this two-class problem, the control group formed the first class (healthy); it included all moments of healthy women sampling, with at least k predecessors in time. Then we fixed some time interval T and defined the second class (OC) in the following way. We included in this class those sampling moments (from the group of OC cases), after which the disease was found not later than in time T and not less than k samples preceded it. Thus, if $T = 0$, we refer to class two only the sampling moments when the disease was detected by other means.

Peak No	m/z	Range	Median intensity (normalised)	Median ratio	p-value
1	6645.9	6634.6 – 6652.8	109.91	1.3044	0.058581
2	3188.9	3185.3 – 3191	32.328	2.1588	4.0009e-005
3	2004.2	2001.1 – 2005.5	50.051	0.74715	0.63293
4	3330.4	3325.2 – 3333.4	23.781	1.5151	0.013008
5	9307.1	9291.2 – 9318.2	9.0218	0.84706	0.5927
6	2982.3	2978.8 – 2985.2	37.431	0.92362	0.98476
7	1764.5	1761.8 – 1766.8	32.599	0.99532	0.66037
8	818.46	817.45 – 818.9	3.6504	1.4574	0.45623
9	4291	4284 – 4296.5	11.739	1.896	0.56656
10	3172	3168 – 3176.4	21.53	0.61438	0.3205
11	2548.5	2544.6 – 2550.7	13.76	0.94219	0.83355
12	937.35	936.48 – 937.76	3.72	1.4551	0.86349
13	3280.1	3273.9 – 3282.1	4.7254	3.6682	0.089078
14	2262.3	2259.1 – 2265.2	52.141	0.34515	0.00026334
15	8126.7	8111.5 – 8132	6.6725	0.24549	0.00082802
16	1964.1	1961.2 – 1965.6	16.371	1.3771	0.072526
17	6447.9	6441.3 – 6460.2	65.526	0.79236	0.24387
18	2562.8	2561 – 2564.2	10.453	1.5603	0.030867
19	1888.7	1886.1 – 1889.7	3.3067	1.7441	0.30168
20	899.92	898.6 – 900.99	4.9943	2.0381	0.28469
21	1442.4	1440.7 – 1443.3	2.2961	3.4512	0.14641
22	2020.9	2019.7 – 2022.3	28.773	2.0808	0.051338
23	5010.8	5006.7 – 5014.6	18.361	0.4756	0.00032864
24	8943.1	8938.6 – 8963	2.8324	2.3457	0.23623
25	4652.6	4647 – 4659.3	13.125	0.50243	0.075617
26	5379.4	5373.1 – 5387.8	11.455	0.37381	0.019768
27	3788.6	3786.5 – 3791.9	4.6076	1.1248	0.20735
28	2405.1	2404 – 2406.2	12.48	1.1483	0.66037
29	3229.8	3226.7 – 3234.2	11.572	1.0838	0.38471
30	1857.6	1855.9 – 1858.4	6.6334	1.1379	1
31	2643.6	2639.7 – 2644.7	9.889	0.90584	0.75986
32	1388	1385.8 – 1389.3	8.548	1.3192	0.38465
33	2016.8	2013.6 – 2019.6	29.697	2.0643	0.016078
34	3206.4	3203.1 – 3208.5	7.1968	1.788	0.00094961
35	2447.5	2446.2 – 2448.7	5.8873	1.3501	0.37928
36	1577.9	1575.9 – 1578.9	4.7759	1.0354	0.51599
37	3297.2	3295.9 – 3300.8	6.7969	2.3786	0.12643
38	1204.9	1203.5 – 1205.7	2.8539	0.9402	0.46785
39	2109.1	2107.4 – 2111	21.028	0.3624	0.0079191
40	2755.2	2750.9 – 2758.2	9.1377	1.0668	0.87853
41	3977.7	3974.7 – 3981	1.3066	0.77425	0.30224
42	8134.4	8132.1 – 8146.8	6.6502	0.24549	0.00067246
43	1808.5	1806.3 – 1809.3	2.5218	2.7493	0.020507
44	8935.9	8918.7 – 8938.5	2.793	2.3457	0.20735
45	2503	2501.6 – 2504.1	12.073	0.71463	0.33947
46	1937.2	1934.7 – 1939.3	11.269	1.0444	0.98476
47	2723.8	2720 – 2726.1	6.3559	2.4107	0.17497
48	1477.3	1475.6 – 1478.5	5.2225	1.2024	0.24387
		log(CA125)	2.5376	1.8341	3.38873e-005

Table 3: The last samples of 11 cases and 49 controls

Peak No	m/z	Range	Median intensity (normalised)	Median ratio	p-value
1	6645.9	6634.6 – 6652.8	117.92	1.1567	0.056078
2	3188.9	3185.3 – 3191	38.679	1.7735	0.033957
3	2004.2	2001.1 – 2005.5	31.59	1.2821	0.0079191
4	3330.4	3325.2 – 3333.4	19.758	1.0977	0.96952
5	9307.1	9291.2 – 9318.2	10.719	0.91455	0.86349
6	2982.3	2978.8 – 2985.2	27.34	1.5884	0.0056035
7	1764.5	1761.8 – 1766.8	24.557	1.2184	0.037309
8	818.46	817.45 – 818.9	18.629	0.93978	0.75986
9	4291	4284 – 4296.5	24.726	0.53162	0.00018079
10	3172	3168 – 3176.4	18.069	1.7874	0.29338
11	2548.5	2544.6 – 2550.7	20.638	0.63469	0.0020994
12	937.35	936.48 – 937.76	28.75	0.24828	7.0766e-005
13	3280.1	3273.9 – 3282.1	11.558	0.50569	0.040938
14	2262.3	2259.1 – 2265.2	27.389	0.49778	7.0766e-005
15	8126.7	8111.5 – 8132	4.8222	1.2498	0.34922
16	1964.1	1961.2 – 1965.6	18.527	1.3124	0.17497
17	6447.9	6441.3 – 6460.2	51.881	0.91305	0.35914
18	2562.8	2561 – 2564.2	18.36	0.68069	0.27618
19	1888.7	1886.1 – 1889.7	36.665	0.37339	8.3028e-005
20	899.92	898.6 – 900.99	23.841	1.2071	1
21	1442.4	1440.7 – 1443.3	40.669	0.35208	0.00019505
22	2020.9	2019.7 – 2022.3	21.802	2.2768	0.00054444
23	5010.8	5006.7 – 5014.6	13.052	0.97731	0.52841
24	8943.1	8938.6 – 8963	6.6211	0.5884	0.044861
25	4652.6	4647 – 4659.3	12.753	0.9767	1
26	5379.4	5373.1 – 5387.8	8.5377	1.8729	0.0028799
27	3788.6	3786.5 – 3791.9	5.6136	0.90823	0.68828
28	2405.1	2404 – 2406.2	34.105	0.26639	3.7779e-006
29	3229.8	3226.7 – 3234.2	9.5406	0.84373	0.46786
30	1857.6	1855.9 – 1858.4	30.881	0.36835	0.018782
31	2643.6	2639.7 – 2644.7	18.126	2.3734	0.012326
32	1388	1385.8 – 1389.3	8.6749	1.2453	0.75986
33	2016.8	2013.6 – 2019.6	25.087	1.9826	0.00082802
34	3206.4	3203.1 – 3208.5	11.795	0.99402	0.46786
35	2447.5	2446.2 – 2448.7	15.435	2.7871	0.0046922
36	1577.9	1575.9 – 1578.9	8.1234	1.2911	0.037309
37	3297.2	3295.9 – 3300.8	6.1088	0.74868	0.46786
38	1204.9	1203.5 – 1205.7	23.582	0.26449	8.9887e-005
39	2109.1	2107.4 – 2111	13.567	0.69986	0.024177
40	2755.2	2750.9 – 2758.2	11.126	0.50283	0.00062697
41	3977.7	3974.7 – 3981	5.4204	0.081267	2.2237e-005
42	8134.4	8132.1 – 8146.8	4.8214	1.2498	0.3205
43	1808.5	1806.3 – 1809.3	7.5761	3.1465	0.0030647
44	8935.9	8918.7 – 8938.5	6.6858	0.57481	0.040938
45	2503	2501.6 – 2504.1	7.8815	1.812	0.17497
46	1937.2	1934.7 – 1939.3	12.944	1.205	0.16897
47	2723.8	2720 – 2726.1	7.1069	1.0678	0.73094
48	1477.3	1475.6 – 1478.5	7.672	1.0153	0.38996
		log(CA125)	2.4508	1.7286	0.00029419

Table 4: Peak intensities averaged over 3 samples preceding the last one

Peak No.	Mean m/z	p-value (Set1 vs Set 2 Controls)
9	4291	0.31881
12	937.35	0.17195
14	2262.3	1.8228e-011
19	1888.7	0.79389
21	1442.4	0.12274
28	2405.1	3.9514e-022
38	1204.9	0.30664
41	3977.7	0.1657
	CA125	0.0115

Table 5: P-value for instability of each peak, instability is identified by comparing control samples from Set 1 and Set 2, which we would expect to have the same distribution (this is our null hypothesis). Small p-values indicate unstable peaks.

There were 11 moments at time $T = 0$. If, however, the time $T = 3$ years, we refer to the second class only the sampling moments, that were obtained 3 years or less before the disease was detected. Or, in the case $T = \infty$ all sampling moments of a woman who developed OC are included in the second class. There were 33 sampling moments at $T = 3$. We shall call the parameter k (number of preceding samples) as the **prehistory parameter**.

In summary, the feature space (input parameters) is formed by the intensities of the selected 48 peaks on each MS-spectrogram; the objects of recognition were the moments of sampling; and to take into consideration the prehistory, the peak intensities were added by an average value over k preceding samples ("prehistory"). 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 parameter, responsible for prehistory. Of course it would be more accurate to normalise this difference by mean squared deviation in the prehistory: $\frac{X_i(t_j)-X_i^*}{\delta}$ where X_i^* is an average intensity of k preceding moments. But the very low number of preceding samples for an individual makes unreliable any estimation of mean squared deviation.

5.4 Unstable peaks

Since we have two sets of healthy indeviduals (serial controls and matched single controls), it would be interesting to compare their p-values. The results in Table 5 show that for 2 out of the 8 peaks we can reject the null-hypothesis, and only for for the remaining 6 can we not reliably state that the controls in set 1 and set 2 come from the same distribution. Since these samples were handled in different time, in different parts of the country, by different people and very likely diffferent protocols, we shall reject the 8 peaks with small p-values, and will be dealing only with the remaining 6 peaks(9,12,19,21,38,41): $m/z = [3977.7, 937.35, 1888.7, 1204.9, 4291, 1442.4]$.

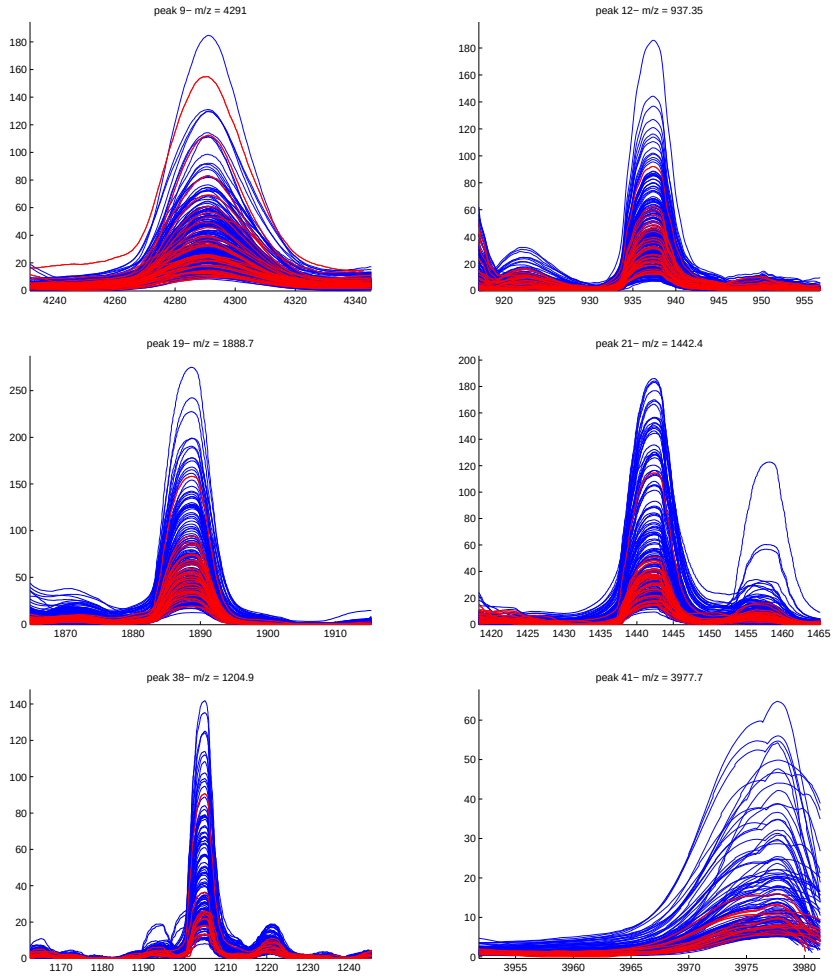

Figure 7: The 6 'stable' peaks (peaks 9,12,19,21,38 and 41)

These 6 peaks are shown in Figure 7, where the red lines represent m/z values ranges in Table 1.

6 Recognition results

As a result of these restrictions we have considered six time-points $T = 0, 1, 2, ...5$ and in each case there should be no less than 3 preceding measurements, i.e. $k = 3$. The number of the sampling moments for Class 1 (controls) were kept the same (154), and the sampling moments for Class 2 (OC) for each time slot is given in Table 6.

The number of objects in class 2 (OC) was very low, and any division into training and test sets would result in unreliable predictions. To have a fair estimation of recognition results we had to use the method called leave-one-out.

T	0	1	2	3	4	5
Class 2 (OC)	11	27	31	33	35	37

Table 6: Number of sampling moments in each time slot

The idea is to leave one object out, execute the training procedure using all the other objects and then test the object that was left out. The procedure is repeated leaving out all objects one by one and the number of errors and correct answers is calculated. This serves as a fair estimation of recognition quality. In our case, if a sample of a person was left out, then all samples of the same person were 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.

Another problem with this low number of samples, but with a rather large number of features is known as "overfitting": 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. The solution is to decrease considerably the number of features.

A series of experiments were conducted in order to find good separation using just **one feature** (one of the 6 peak intensities). However, it was impossible to find a good discriminative peak to make reliable diagnostics.

Recognition with the prehistory parameter. Then we turned to the idea of using the second parameter - prehistory - to discriminate classes in two dimensional space with coordinates ${X_i}^*(t_j)$ and $X_i(t_j)$, where ${X_i}^*$ is an average of preceding sampling moments (prehistory), and X_i is the $i-th$ peak intensity. Then a reasonable discrimination was obtained in this 2-dimensional space.

As we said before, the considered time-points were $T = 0, 1, 2, 3, 4, 5$ years. For $T = 0$ we considered as the OC cases only those sampling moments when the disease was found by other means at the last moment. In this case we used for recognition 11 sampling moments in the class **OC** (because for every woman only one sampling moment was taken) and 154 sampling moments in the class **healthy** (because several samples from one woman were included).

For $T = 3$ it means that the disease was found no later than 3 years after sampling. In this case we had for recognition 33 sampling moments in the class OC (because now one patient could be sampled several times within time interval T) and again 154 sampling moments in the class healthy. And so on for all different values of T. In addition, of course, no less than 3 sample measurements preceded the time-point of diagnosis.

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

One can see in Figure 8, the results of discrimination with CA125, where the first coordinate (x) represents intensities of CA125 at the last moment when the actual diagnosis was made, and the second coordinate (y) represents the average intensities of the 3 preceding sampling moments. There were 11 OC cases and 154 controls. Figure 9 shows a simmilar diagram for peak 41.

We compared also the best results using an additional feature reflecting prehistory with those which we could achieve without prehistory (under the same conditions). The best results were achieved with $T = 1$ and and a weighting

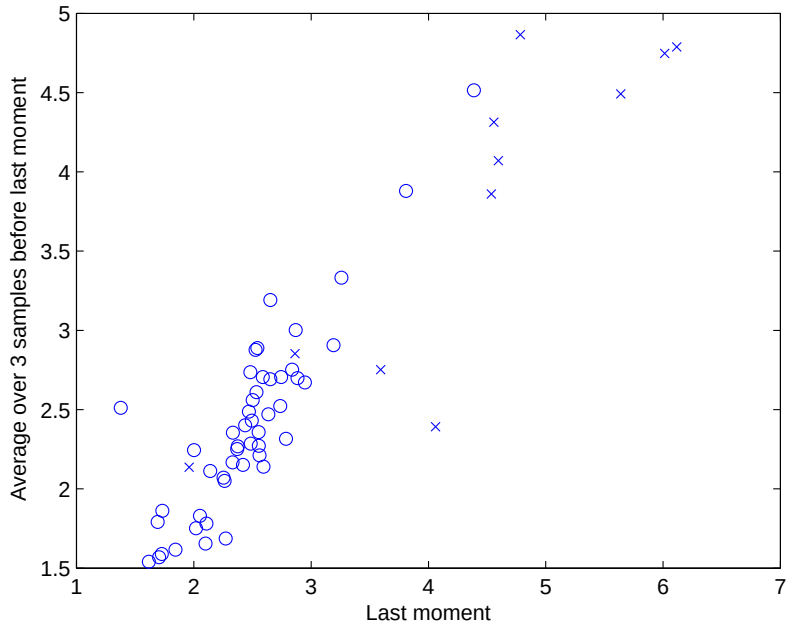

Figure 8: CA125 - last–moment cases vs. all controls

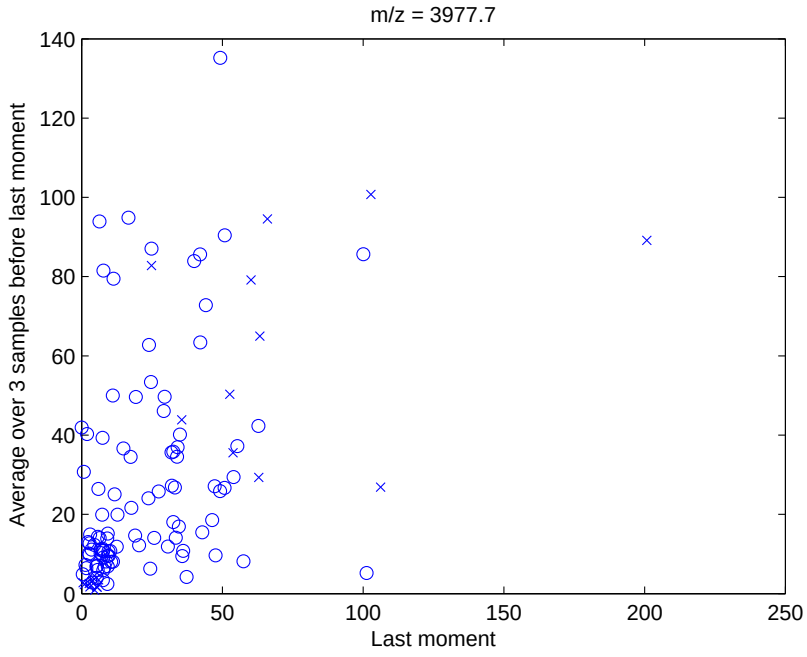

Figure 9: Peak 41 - $m/z = 3977.7$ - last–moment cases vs. all controls

Peak	T	W	Error	With Prehistory	without prehistory
1	1	1/5	Type 1	40.7%(11/27)	48.1%(13/27)
			Type 2	45.5%(70/154)	58.4%(90/154)
2	1	1/5	Type 1	18.5%(5/27)	44.4%(12/27)
			Type 2	31.2%(48/154)	51.9%(80/154)
3	1	1/5	Type 1	55.6%(15/27)	66.7%(18/27)
			Type 2	53.2%(82/154)	52.6%(81/154)
4	1	1/5	Type 1	18.5%(5/27)	44.4%(12/27)
			Type 2	37.0%(57/154)	55.2%(85/154)
5	1	1/5	Type 1	51.9%(14/27)	66.7%(18/27)
			Type 2	46.8% (72/154)	69.5%(107/154)
6	1	1/5	Type 1	33.3%(9/27)	44.4%(12/27)
			Type 2	38.3%(59/154)	49.4%(76/154)
CA125	1	1/5	Type 1	18.5%(5/27)	11.1%(3/27)
			Type 2	18.8%(29/154)	22.7%(35/154)

Table 7: Classification results for single peaks where $T = 1$ and $W = 1/5$

factor equal to 1/5. This weighting factor regulates the error proportion of **type 1** (an OC sample classified as healthy), and **type 2** (a healthy sample classified as OC). This parameter consists of two parts: number of neighbours and the voting threshold. For example, the parameter value 2/5 means that 5 neighbours neighbours were considered, and the classification was the 'case' if 2 or more of 5 neighbours were 'cases'. Table 7, shows the results of comparison. It is obvious that using prehistory we always gain better results than when only "current" sampling moments are used.

It is clear from Table 7 that using single peaks for descrimination between OC and controls does not produce better Type 1 errors than CA125. In an attempt to combat this we introduced classification for multiple peaks. We use the pattern recognitions algorithm as outlined above. The best results produced by this method are shown in Table 8. In several cases we have found combinations of peaks with better Type 1 and Type 2 errors than CA125. Once again it is clear that using prehistory produces better results.

Peaks	T	W	Error	With Prehistory	without prehistory
12,38,41	1	1/5	Type 1	7.4%(2/27)	29.6%(8/27)
			Type 2	29.2%(45/154)	45.5%(70/154)
12,21,38	1	1/5	Type 1	18.5%(5/27)	18.5%(5/27)
			Type 2	26.6%(41/154)	38.3%(59/154)
9,12,19,38,41	2	1/5	Type 1	3.2%(1/31)	35.5%(11/31)
			Type 2	40.9%(63/154)	55.9%(89/154)
12,38,41	2	2/5	Type 1	19.4%(6/31)	77.4%(24/31)
			Type 2	14.9%(23/154)	18.8%(29/154)
9,12,19,38,41	3	1/5	Type 1	3%(1/33)	30.3%(10/33)
			Type 2	41.6%(64/154)	57.8%(89/154)
12,38,41	3	2/5	Type 1	18.1%(6/33)	24.2%(8/33)
			Type 2	14.9%(23/154)	50%(77/154)
12,41	4	1/5	Type 1	2.9%(1/35)	28.6%(10/35)
			Type 2	30.5%(47/154)	53.2%(82/154)
12,38	4	2/5	Type 1	14.3%(5/35)	28.6%(10/35)
			Type 2	17.5%(27/154)	53.9%(83/154)
12,38,41	5	1/5	Type 1	0%(0/37)	21.6%(8/37)
			Type 2	31.2%(48/154)	53.2%(82/154)
12,38	5	2/5	Type 1	10.8%(4/37)	21.6%(8/37)
			Type 2	18.8%(29/154)	60.3%(93/154)
CA125	1	1/5	Type 1	18.5%(5/27)	11.1%(3/27)
			Type 2	18.8%(29/154)	22.7%(35/154)

Table 8: Comparison results with and without prehistory parameter for $T = 1, 2, \ldots, 5$ and weighting factor W

peak No.	Mean m/z	Tempst peak	Peptide
9	3977.7	3970.97(6.7)	(R) QAGAAGSRMNFRPGVLSSRQLGLPGPPDVPDHAAYHPF.
12	937.35	942.43 (5.1)	HWESASLL.
19	1888.7	1895.99(7.3)	RNGFKSHALQLNNRQI (R)
21	1204.9	1206.57(1.7)	EGDFLAEGGGVR
38	4291	(No HMR)	
41	1442.4	1449.76(7.4)	THRIHWESASLL.

Table 9: Serum peptide signatures for OC

7 Discussion

From the above results we can see that the peaks 9,12,19,21,38 and 41 are the best candidates for discriminating the distributions between the classes. These peaks could be candidates for potential biomarkers. Introduction of the prehistory parameter significantly improved the recognition abilities of the peaks. While CA125 gives good discrimination at the last moment, other presented peaks give good recognition results several years before OC is diagnosed. In other words, they seem to be very promising for early diagnosis of OC.

The experiments have shown that using of combinations of several peaks would bring the error rate down, but to confirm this we would need to experiment with a larger set of samples.

Our next problem is to identify the peaks with certain proteins, peptides or proteases in order to find out the corresponding protein biomarkers. At this stage we compared the 48 most popular peaks identified in Table 1 with the peaks identified by J. Villanueva et al 2006 [4]. If we look at m/z values of five of our peaks we can see that they are close to the peptides that were found in [4]. They are presented in Table 9.

However, the results we are reporting here depend on the quality of the data we used, and our experiments also showed that the different data sets (Set 1 - control and Set 2 - also control) are not comparable, and we cannot rely on the Set 2 serial control data.

Further experiments are being conducted and will be reported shortly.

8 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 Royal Society grant 15955 "Efficient Randomness Testing".

9 References

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

2. Matlab Bioinformatics Toolbox, Matlab Version 7.2.0.232(R2006a)

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.