{ Some parts of the Paper read at HSSS in Pavia, 1999 }

# Pattern recognition in case of two-group data with uncomplete assignment information: A study on coding and non coding DNA sequences

Anna Bartkowiak
Institute of Computer Science, University of Wrocław,
Przesmyckiego 20, Wrocław 51–151 Poland
e-mail aba@ii.uni.wroc.pl http://www.ii.uni.wroc.pl/~aba/

Stanisław Cebrat and Paweł Mackiewicz
Institute of Microbiology, University of Wrocław, PL
http://www.microb.uni.wroc.pl/genetics/private/cebrat/cebrat.html

## 1 Introduction

### 1.1 DNA sequences – the code for life functions of the organism

The mystery contained in the sequences (ORF's) is that they may code a life function of the organism. For some ORFs it is exactly known, which life functions are governed by the code written in the ORF. In such case the sequence contained in the ORF is called *gene*.

The DNA code is a great challenge to contemporaneous research, especially for biochemistry and genetics, also for the information theory and data analysis in general.

The 'life' information hidden in chromosomes is coded in so called DNA (deoxyribonucleic acid) sequences with four bases (nucleotides) denoted by A, G, C, T. This is the alphabet of the code. The code is organized in triplets called codons.

It is known that the coding information is contained in some pieces of the sequences called ORFs (Open Reading Frames). Each ORF starts with a specific codon (the `start codon ATG`) and ends with one of three specific `stop codons: TAG, TAA, TGA`. The coding information is contained within the ORF. Each of the 61 codons (including the start codon) codes for one aminoacid.

The DNA sequences, in particular the order and succession of the codons have been extensively investigated using various methods of mathematical statistics and stochastic processes.

A review of the problems – from the point of mathematical statistics – may be found in Braun & Müller (1998). Some specific topics, to mention a few out of many published in mathematical journals, were considered by Avery & Henderson (1999), Muri (1998), Prum & *al.* (1995), Kamb & *al.* (1995).

### 1.2 The yeast genome

In our investigation we are concerned with the yeast genome. It contains 16 chromosomes, containing sequences totalling about 13 millions bases, which yields really a great amount of data.

The DNA sequences of the yeast chromosomes are exactly known and are put into databanks accessible through the internet (genome-ftp.stanford.edu or http://mips.biochem.mpg.de).

As for June 1999 there were known together $n = 7472$ ORFs. From these $n_1 = 2733$ were known as coding a life function. For the remaining $n_2 = 4739$ ORFs the meaning of the code was not exactly recognized, or recognized only partially.

The biochemists have found – by some alignment methods, that some of the ORFs with not recognized life function are very similar to some genes. Based on this similarity six classes of homologies with genes were established. The classes and their frequencies encountered in our data are:

| Homology Class | Frequency |
|---|---|
| H1 | 290 |
| H2 | 857 |
| H3 | 705 |
| H4 | 860 |
| H5 | 421 |
| H6 | 1606 |
| Total | 4739 |

ORFs contained in class H1 are highly similar to some genes, while ORFs from classes H5 and H6 are most dissimilar.

The problem we are concerned is the following one: Is it possible – on the basis of their statistical properties – to recognize genes, i.e. sequences coding some life functions of yeasts, in the second group of data?

### 1.3 The data

We have statistical data which were gathered in the Institute of Microbiology, University of Wrocław, by prof. S. Cebrat and his team.

Firstly, each ORF was represented in a *spider -plot.* Two exemplary spider plots – for a coding and non coding ORFs are shown in Figure 1.

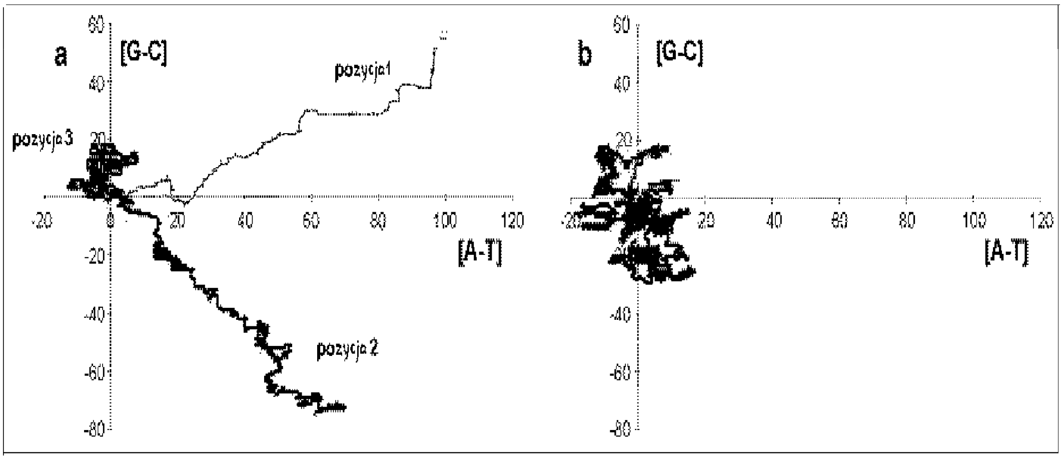


Figure 1: *Spider plots. Left: for a coding Orf; Right: for a non-coding ORF*

Next each ORF was characterized by 13 variables $v_1 - v_13$ which were gathered from the spider plot:
$v_1 - v_3$ - angles of the legs,
$v_1 \leftarrow \alpha_1 = \arctan \frac{G_1-C_1}{A_1-T_1}$,
$v_2 \leftarrow \alpha_2 = \arctan \frac{G_2-C_2}{A_2-T_2}$,
$v_3 \leftarrow \alpha_3 = \arctan \frac{G_3-C_3}{A_3-T_3}$,
$v_4 - v_9$ - normalized values of the ultime points,
$v_4 \leftarrow \frac{A_1-T_1}{\sqrt{L}} \quad v_5 \leftarrow \frac{G_1-C_1}{\sqrt{L}} \quad \}u_1$,

$v_6 \leftarrow \frac{A_2-T_2}{\sqrt{L}} \quad v_7 \leftarrow \frac{G_2-C_2}{\sqrt{L}} \quad \}u_2,$
$v_8 \leftarrow \frac{A_3-T_3}{\sqrt{L}} \quad v_9 \leftarrow \frac{G_3-C_3}{\sqrt{L}} \quad \}u_3,$
$v_{10}$- length of the entire ORF – in codons,
$v_{10} \leftarrow L,$
$v_{11} - v_{13}$ - normalized length for each of the legs, namely
$v_{11} \leftarrow u_1 = \sqrt{v_4^2 + v_5^2} = \sqrt{(\frac{A_1-T_1}{\sqrt{L}})^2 + (\frac{G_1-C_1}{\sqrt{L}})^2},$
$v_{12} \leftarrow u_2 = \sqrt{v_6^2 + v_7^2} = \sqrt{(\frac{A_2-T_2}{\sqrt{L}})^2 + (\frac{G_2-C_2}{\sqrt{L}})^2},$
$v_{13} \leftarrow u_3 = \sqrt{v_8^2 + v_9^2} = \sqrt{(\frac{A_3-T_3}{\sqrt{L}})^2 + (\frac{G_3-C_3}{\sqrt{L}})^2}.$

In Figure 2 we show how the length and the angle of a leg was taken from the 'spider' plot:

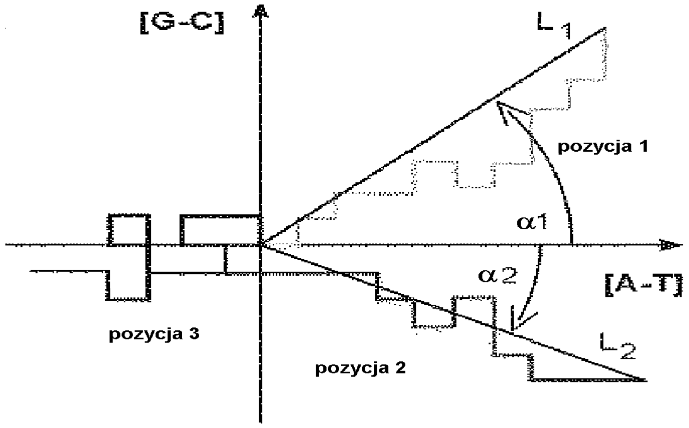


Figure 2: *Spider plots. Measuring length and angle of a leg*

In such way we got a rectangular table $\mathbf{X}_{n \times p}$. Each row of this table characterizes one ORF, ready for statistical analysis.

Our goal is to investigate statistically the formal difference between the genes and the rest of the ORFs on the basis of some features characterizing the frequency and consecution of appearing the bases and codons. Results obtained by S. Cebrat and his team (see, e.g. Cebrat Cebrat & *al.* 1997, 1998, and Mackiewicz & *al.*) are encouraging.

### 1.4 Methods of our analysis

The specificity of our analysis is that we have two groups of data, for which it can be said:

One group contains the coding ORFs and the situation with them is more or less clear.

The other group contains ORFs which may be coding or non–coding sequences. Thus the ORFs contained in that group have no clear group assignment. We suppose that at least a part of them is not coding (there may be some mutants or some relicts from past mutations). Nonetheless we will try to separate both groups of data.

We will apply here canonical discriminant analysis with orthogonal jitter. This is described in Section 2.

Next, in Section 3, we will estimate the proportion of coding and not–coding ORFs in the yeast genome – by using a mixture decomposition of the first canonical variate.

# 2 Canonical Discriminant Variate with Orthogonal Jitter

## 2.1 Considering entire set of 7472 ORFs

We consider generally two groups of data defined as:

**Group 1.** *Genes*, $n = 2733$,

**Group 2.** Other Orf's, called in the following *non–genes*, $n = 4739$.

For these data the canonical discriminant analysis was performed. We have used for that purpose a special method 'canonical analysis with orthogonal jitter' described in the paper [3].

Let $\mathbf{x} = (x_1, \ldots, x_p)$ denote generally the analyzed data vector. We seek for a new variable

$$z = \mathbf{xu}$$

being a linear combination of the observed variables $x_1, \ldots, x_p$ and such that the derived $z$ discriminates mostly among the considered two groups of data. The discrimination power of the derived variable is defined as the ratio of the between group to the within group variance of $z$ (Fisher's discriminant criterion).. It is known, that in the case of two-group data only **one** such variate can be obtained. The vector $\mathbf{u}$ establishing the transformation $z = \mathbf{xu}$ indicates at the same time a direction in the $p - variate$ data space, which yields a greatest separation between data points belonging to different groups.

We built a second variate establishing a second direction in the data space; this second direction does not have any discriminative power, none the less it is very useful in seeing more clearly the analyzed points. We call this variate 'orthogonal jitter'.

Using the described method we have calculated the first canonical discriminant variate and the orthogonal jitter for the considered two groups: Genes and Other ORFs, taking into account the entire data set with all 7472 ORFs.

In Figure 3, top, we show scatterplot based on these two variables. The genes are depicted using red color; the remaining ORFs are represented by black dots. The genes are overlying the other ORFs.

One can see how helpful is the second variate. Albeit it has no discriminative power, it is very helpful in detecting outliers.

One may notice also that the non-recognized ORFs are scattered much more as the recognized genes.

In Figure 3, bottom, we show a scatterplot containing all recognized genes ($n_1 = 2733$).

Again one may notice the great spread, mainly in the northern direction. On the other hand, the distribution seems to be somehow cut in the south-west direction of the figure. It seems as not all of the points belonging to the genes were displayed.

Generally, looking at the scatterplot exhibited in the top of Figure 3 one may notice that there exist a shift in the location of the two groups of data: points denoting genes are located more to the right.

One may notice in the same exhibit that the non-genes are much more scattered. There are at least two big outliers belonging to the second group of data.

Thus: the orthogonal jitter permitted in our case to identify enormously atypical observations, which – considering only the first discriminant variate would remain undiscovered.

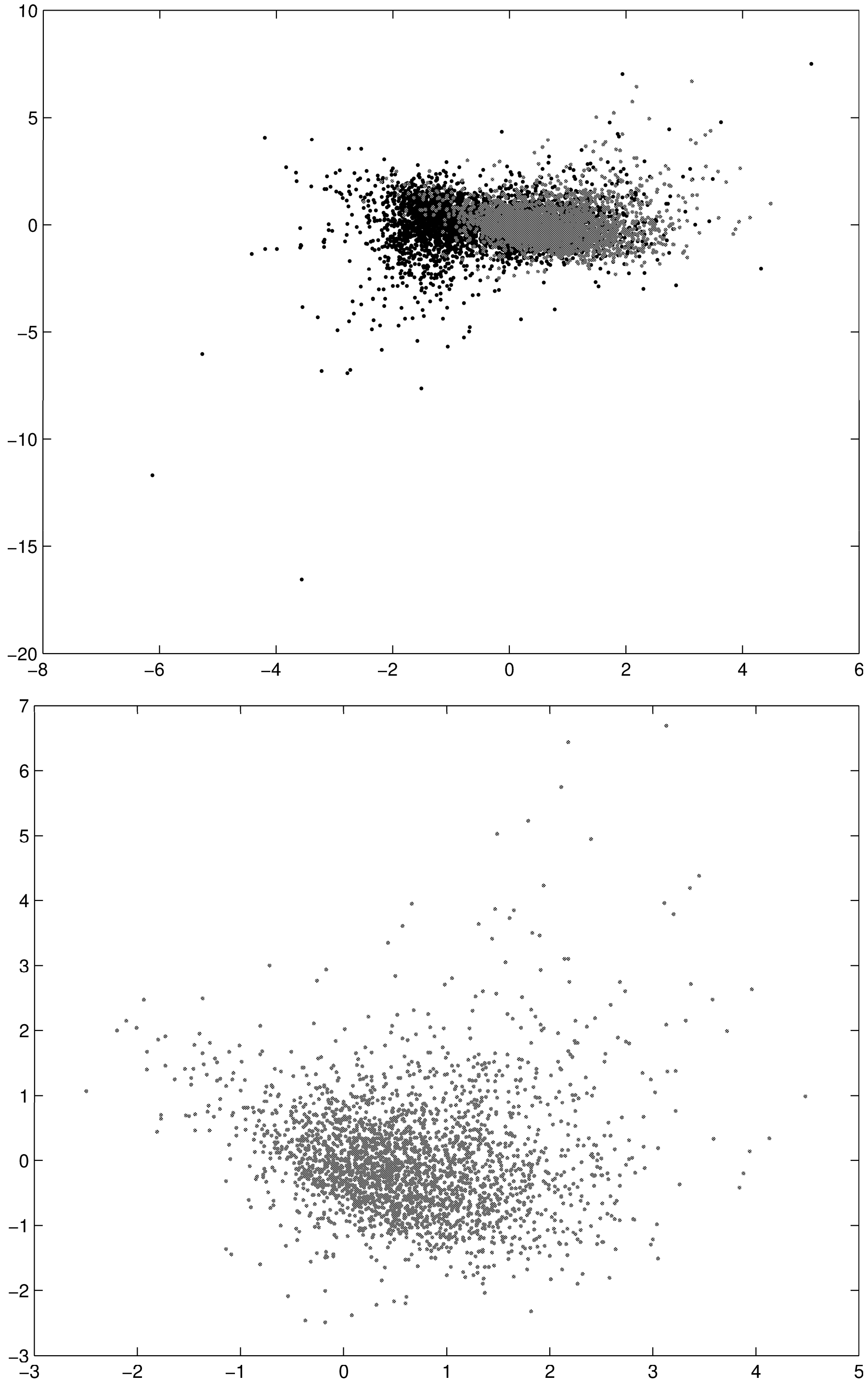


*Figure 3. Top: Genes as red squares and remaining ORFs as black dots; a total of $n = 7472$ points are exhibited; Bottom: Only points corresponding to recognized genes are displayed*

## 2.2 Differentiation between Genes and classes of Homologies

In previous subsection we have carried out the analysis using the entire data set.

Now we turn to the groups of homologies described in Subsection 1.2 of the paper.

We consider in turn each class of homologies and match it with a sample of n=1000 genes (for a sampling scheme, see [2]).

The methodology of investigation is the same as in previous Subsection. However, let us say it clearly, the evaluations of the first discriminant variate and of the variate denoting orthogonal jitter were carried out separately for each pair of groups (each pair composed from the sample group of genes and one homology class).

It is interesting to see, how – with increasing homology class number – the separation between both groups of data is increasing.

Below we show scatterplots obtained when taking into account very near homologies (H1∪H5), then the most distant H5 and H6 classes.

In Figure 4 we show scatterplot exhibiting the differentiation between the genes (only a sample of 1000 genes was taken) and the homologue groups of ORFs H1 and H2 (we have taken these classes together, because class H1 is very small).

As we know already, the classes H1 and H1 contain ORFs which are much similar to genes. The same may be noticed when looking at the Figure 4 where it can be seen, that both groups are much overlapping.

One may also notice in Figure 4 two big outliers belonging to the H1∪H2 group.

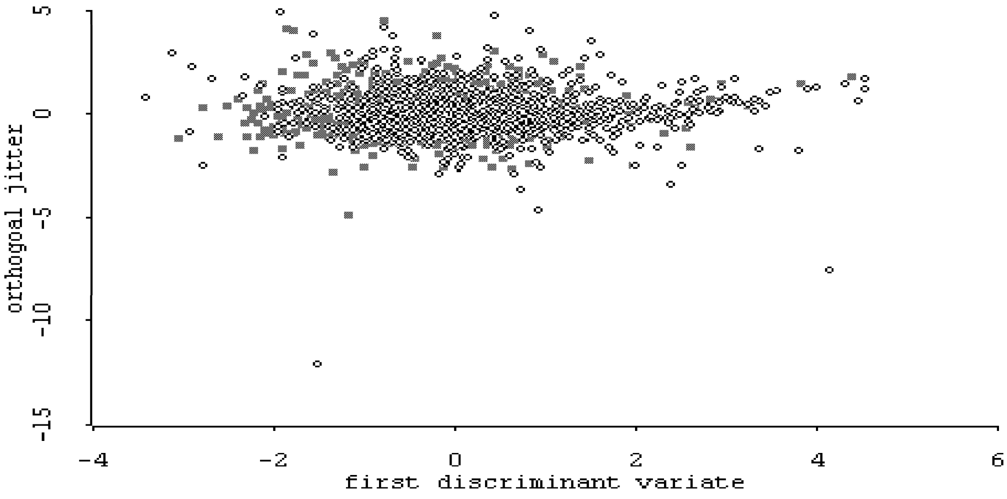


Figure 4: *Displaying two groups of DNA data: Filled red squares denote a sample of 1000 ORFs identified as genes. Open circles denote* $n_1 = 1147$ *ORFs belonging to H1∪H2. The open circles are overlying the filled squares*

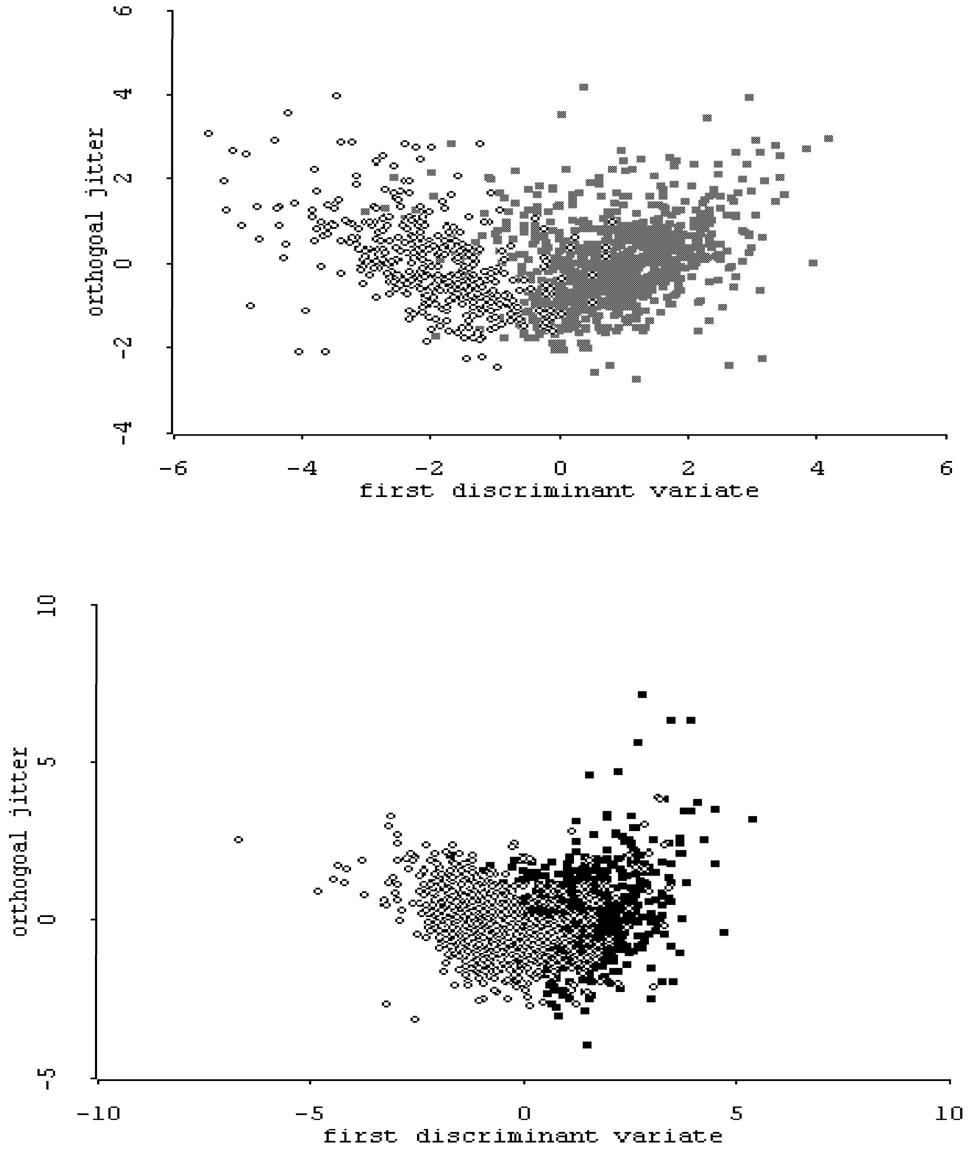


*Figure 5. Displaying homologies classes H5 (top) and H6 (bottom) together with a sample of n=1000 genes. Notice the differentiation visible in each plot. Top:* $n_5 = 421$ *ORFs belonging to H5. Bottom:* $n_6 = 1606$ *ORFs belonging to H6. The open circles (homologies) are overlying the filled red squares denoting genes*

## 3 Decomposition of the density of the canonical discriminant variate into a mixture of two gaussians

We have stated at the begin of our paper that we have to do essentially with two groups of data. One of these groups contains recognized genes. The other group contains the rest of the ORFs; among these may be genes which code some life function of the yeast organism, and possibly there are some other DNA sequences which do not code any life function.

One might expect the the fact of coding and not coding life functions may be somehow reflected in the statistical characteristics of the ORFs. In particular one might imagine that the genes occupy mostly one region of the data space, and the non-genes are somehow shifted in that space to other location. The scatterplots exhibited in Figure 3, top, Figure 4 and Figure 5 confirms that supposition.

On the other hand, the canonical variate is a linear function of several (in our case of 13) originally observed variates, and as such should be distributed normally (or very near to the gaussian distribution).

Taking into account these presumptions we have investigated the distribution of the first discriminant variate Z. We have taken into account all the 7472 ORFs. We have approximated the density of Z by a non–parametric method by performing a kernel density smoothing. We have used for this purpose the KDE package by Frederic Udina [13, 12].

Let $x_1, \dots, x_n$ denote the observed values of a variable $X$. The kernel density esitmate of $X$ is then defined as [12]

$$\hat{f}_{h(x) = \frac{1}{nh}} \Sigma_{i=1,\dots,n} K(\frac{x - x_i}{h}). \tag{1}$$

To apply this definition we have to chose and define the function $K$, called the *kernel function*, and the bandwidth $h$.

In our case we have substituted for the observed values $x_1, \dots, x_n$ the values $z_1, \dots, z_n$ calculated as the values of the first canonical discriminant variate.

We have stated that the choice of the kernel had little impact on the shape of the obtained density. We have chosen *biweight* kernel.

The smoothed density is shown in all 3 exhibits of Figure 6. One can see clearly, that the density looks like composed from two components.

We have assumed that these components are gaussians $\varphi(\mu, \sigma)$ with unknown mean $\mu$ and standard deviation $\sigma$, different for each component.

Thus we have assumed the model:

$$z \sim f_1 \varphi(\mu_1, \sigma_1) + (1 - f_1) \varphi(\mu_2, \sigma_2). \tag{2}$$

The parameters $f_1$ and $1 - f_1$ are called mixing coefficients.

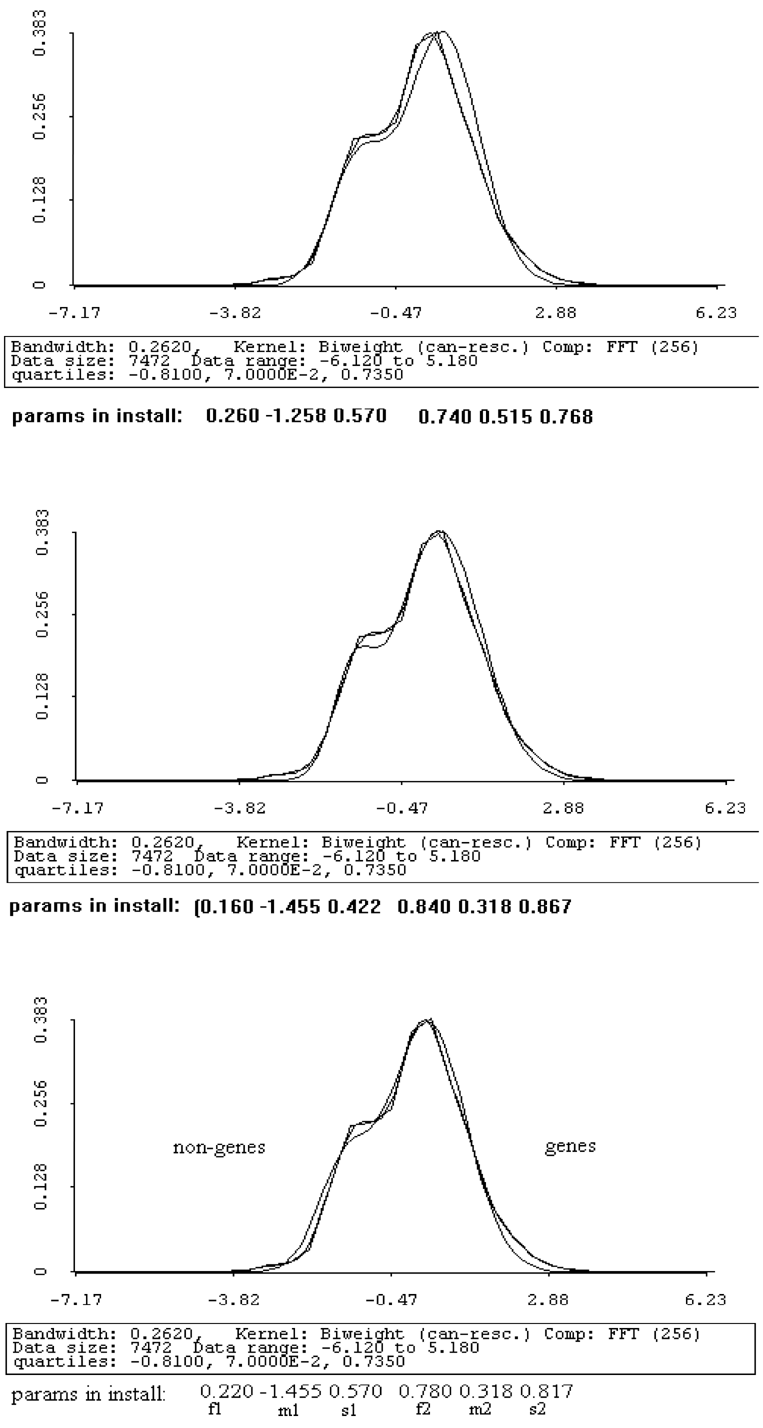
We have tried to estimate the parameters $f_1, \mu_1, \sigma_1, \mu_2, \sigma_2$ appearing in the above equation. This was done in an interactive mode by trial and error. Depending on some other parameters needed by the smoothing technique obtained slightly different density curves. Three of them are shown in Figure 6.

Looking at the exhibits in Figure 6 one may see quite clearly that the density visible in the plots looks really like a mixture composed from two normal distributions.

Below each exhibit we have put the parameters of the two gaussians (i.e. their means and standard deviations) and the mixing coefficients $f_1$ and $f_2 = 1 - f_1$.

How many genes are in our data? Taking $n \times f_2$ we obtain we obtain 5529, 6276 and 5828 as a guess. Remind, only 2733 were identified so far.

Thus much unknown waits still for discovery.



```
Bandwidth: 0.2620,   Kernel: Biweight (can-resc.) Comp: FFT (256)
Data size: 7472  Data range: -6.120 to 5.180
quartiles: -0.8100, 7.0000E-2, 0.7350
```

**params in install: 0.260 -1.258 0.570 0.740 0.515 0.768**

```
Bandwidth: 0.2620,   Kernel: Biweight (can-resc.) Comp: FFT (256)
Data size: 7472  Data range: -6.120 to 5.180
quartiles: -0.8100, 7.0000E-2, 0.7350
```

**params in install: (0.160 -1.455 0.422 0.840 0.318 0.867**

```
Bandwidth: 0.2620,   Kernel: Biweight (can-resc.) Comp: FFT (256)
Data size: 7472  Data range: -6.120 to 5.180
quartiles: -0.8100, 7.0000E-2, 0.7350
```

| params in install: | 0.220 | -1.455 | 0.570 | 0.780 | 0.318 | 0.817 |
|---|---|---|---|---|---|---|
| | f1 | m1 | s1 | f2 | m2 | s2 |

*Figure 6. Distribution of the first canonical variate viewed as composed from two gaussians. 3 trials are shown. The derived gaussians contribute to their sum in proportions f1 and f2. Symbols m1:s1 and m2:s2 denote means and standard deviations of the two contributing gaussians in each trial.*

**Acknowledgements**

The work of the first author was partially supported by the State Research Committee KBN grant 8T11 C 031 16.

## References

[1] P.J. Avery and D.A. Henderson, Fitting Markov chain models to discrete state series such as DNA sequences. Appl. Statistics, V. 48, Part 1, 1999, 53–61.

[2] A. Bartkowiak: Sampling a multi–trait representative sample. *Biometrical Letters*, V. 33, No 2, 1996, 59–69.

[3] Bartkowiak A., Szustalewicz A., Two non-conventional methods for visualization of multivariate two-group data. Manuscript, pp. 1–14, December 2000, Submitted.

[4] Bartkowiak A., Szustalewicz A., Detecting outliers by a grand tour, *Machine Graphics & Vision*, V. 6, 1997, 487–505.

[5] J. V. Braun, H-G. Müller, Statistical methods for DNA sequence segmentation. Statistical Science, Vol.13, No. 2., 1998 142–162.

[6] S. Cebrat, P. Mackiewicz, M.R. Dudek, The role of the genetic code in generating new coding sequences inside existing genes. Biosystems 1998, 165–176.

[7] S. Cebrat, M. Dudek, A. Rogowska, Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences. J. Appl. Genet. V. 38, No. 1, 1997, 1–9.

[8] P. Mackiewicz, M. Kowalczuk, M. Fita, S. Cebrat, M.R. Dudek, Asymmetry of coding versus noncoding strand in coding sequences of different genomes. Microbial. & Comparative Genomics, V2, No. 4, 1997, 259–268.

[9] F. Muri, Modelling bacterial genomes using hidden Markov models. COMPSTAT 1998, Invited and Contributed Papers, 89–100.

[10] A. Kamb, Ch. Wang, et al., Software Trapping: A strategy for finding genes in large genomic regions. Computers and Biomedical Research 28, 1995, 140–153.

[11] B. Prum, F. Rodolphe, E. de Turckheim, Finding words with unexpected frequencies in deoxyribonucleic Acid sequences. J.R. Statist. Soc. B, V. 57, No. 1., 1995, 205–220.

[12] F. Udina, Interactive graphics for nonparametric curve estimation. Manuscript. Department d'Economia, Universitat Pompeu Fabra, Barcelona, Spain, March 31, 1996, pp. 1–39.

[13] F. Udina, Interactive graphics for nonparametric curve estimation. *The XVIIIth International Biometric Conference IBC'96, Amsterdam July 1–5, 1996, Invited Papers*, 51–55.