SOME STATISTICAL PROBLEMS ARISING WHEN INVESTIGATING DNA SEQUENCES

Anna Bartkowiak* *University of Wroclaw, PL*

aba@ii.uni.wroc.pl

http://www.ii.uni.wroc.pl/~aba

Stanisław Cebrat and Paweł Mackiewicz *University of Wrocław, PL*

cebrat@angband.microb.uni.wroc.pl

http://www.microb.uni.wroc.pl/genetics/private/cebrat

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.
In our investigation we are concerned with the yeast genome. It contains 16 chromosomes, containing sequences totalling about 13 millions bases, which makes a terrific amount of data for analysis. The DNA sequences of the yeast chromosomes are exactly known and are put into databanks accessible through the internet.
Presenting the essence of the topic we can say – in a very big abbreviation – that:
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`) and ends with one of three specific `stop codons`. The coding information is contained within the ORF. It is written using the remaining (out of 64 possible) codons. Each of the 61 codons (including start codon) codes for one aminoacid. The same amino acid can be coded by different triplets–codons.
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).
The problem we are concerned with is the following one: It is known, that in the yeast genome the coding information is contained in the ORFs. So are also known the ORF sequences. The generic problem is to find out which information (which life function) is coded in the given ORF.
For some ORFs their life functions are known exactly (even have their names) – in such case the ORF is called a `gene`. For some other ORFs their functions (i.e. what exactly are they coding) is not known or even there is a supposition that many of them may not code for proteins.
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 & *al*. 1997, Cebrat & *al*. 1998, and Mackiewicz & *al*. 1997) are encouraging.
The subject of our present investigation is a data table containing 13 traits for 7472 ORFs identified in the yeast genome. Out of these there are 2733 recognized genes and 4739 not exactly recognized.

The traits (variables) were established by Cebrat & *al*. from the proposed by them `spider graphs` obtained by specific `DNA walks`.
Because the data are quite large we have to be careful with the statistical analysis.
What we have done till now? We have drawn a *representative* sample about 1000 genes from the entire set of genes (using the method proposed by Bartkowiak, 1996) and next we have compared the sample with some parts of the remaining ORFs. This was done using the method of discriminant analysis for 2 groups of data. For graphical display we have used the canonical discriminant variate and another variate jittered in an orthogonal direction – this permitted us to obtain a two-dimensional display, also to notify some outliers.
The preliminary analysis has shown that the remaining ORFs may be viewed as a mixture of subgroups with differentiated coding probability (this has been suspected by the microbiologists). Some of the subgroups overlap with the *gene* group, some show quite a difference. In Fig. 1 and Fig. 2 we show the display of the combined subgroup 1 and 2 of the mixture (Figure 1), and the fifth subgroup of the mixture (Figure 2) – opposed to the genes. The figures exhibit the most differentiated view of the compared subgroup and the gene group – when considering 13 traits characterizing their ORFs.

`Acknowledgement` The research was sponsored partially by the (Polish) Scientific Research Committee (KBN) in the form of grants *no.* 8 T11 C031 16 and *no.* 6 PO4A 030 14.

References

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

Bartkowiak, A. (1996) Sampling a multi–trait representative sample. *Biometrical Letters*, **33**, No 2, 59–69.

Bartkowiak A. and Szustalewicz, A. (1997) Detecting outliers by a grand tour. *Machine Graphics & Vision*, **6**, 487–505.

J. V. Braun, J. V. and H-G. Müller, H-G. (1998) Statistical methods for DNA sequence segmentation. *Statistical Science*, **13**, No. 2, 142–162.

S. Cebrat, S., Mackiewicz, P. and Dudek, M.R. (1998) The role of the genetic code in generating new coding sequences inside existing genes. *BioSystems*, **45**, 165–176.

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

Mackiewicz, P., Kowalczuk, M., Fita, M., Cebrat, S. and Dudek, M. R. (1997) Asymmetry of coding versus noncoding strand in coding sequences of different genomes. *Microbial & Comparative Genomics*, **2**, No. 4, 259–268.

Muri, F. (1998) Modelling bacterial genomes using hidden Markov models. *COMPSTAT 1998, Invited and Contributed Papers*, 89–100.

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

Prum, B., Rodolphe, F. and de Turckheim, E. (1995) Finding words with unexpected frequencies in deoxyribonucleic Acid sequences. *J.R. Statist. Soc. B*, **57**, No. 1, 205–220.

Anna Bartkowiak, Institute of Computer Science, University of Wroclaw, Przesmyckiego 20, Wroclaw 51-151 Poland.

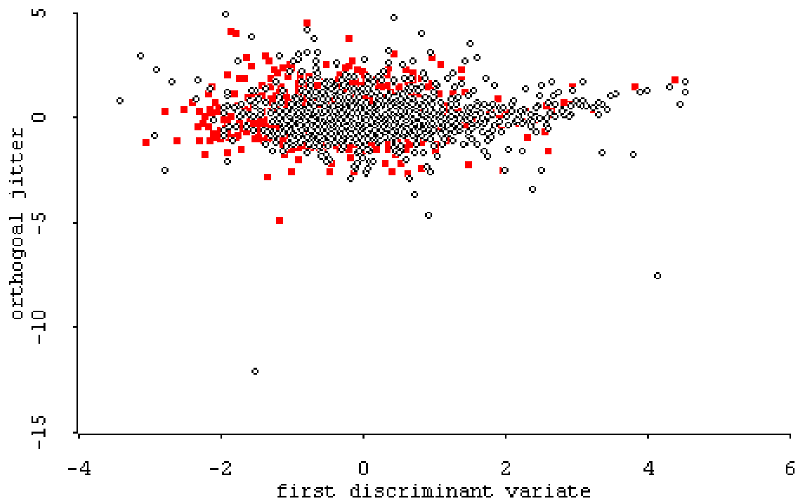

Figure 1. Two groups of DNA data. Filled squares denote ORFs identified as genes. Open circles denote ORFs of the first and second subgroups without known functions. The graph exhibits $n_0 = 1000$ genes and $n_1 = 1147$ other ORFs. The open circles are overlaying the filled squares.

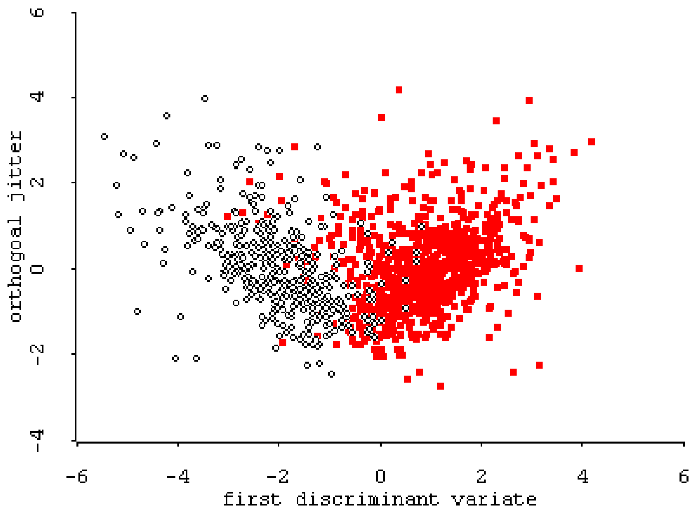

Figure 2. As Figure 1; the filled squares denote the same $n_0 = 1000$ recognized genes; but the open circles ($n_5 = 421$) denote here ORFs of the fifth subgroup.