On the frequency distribution of amino-acids composing a gene

Anna Bartkowiak[1 2]

[1] Institute of Computer Science, University of Wrocław, PL
[2] Wroclaw High School of Applied Informatics, PL

INTRODUCTORY. A gene may be viewed as a DNA sequence coding genetic information on vital functions of the organism. For the eukaryotes (both baker's yeast and humans belong to this group) the coding of the genetic information is similar, with quantitative difference only: a yeast genome contains $\approx\ 6\ 000$ and a human genome $\approx\ 60\ 000$ genes appropriately. The genetic code uses a 4-letter alphabet A,T, G, T organized in triplets ('codons') constituting 23 amino-acids. In the following we consider the frequency distribution of the amino-acids coding the essential genetic information contained in so called ORFs (Open Reading Frames). The data were gathered from the 7th chromosome, where 548 unquestionable ORFs were found. We consider only 20 amino-acids (the STOP codons were omitted from our analysis). Both for the yeast and the human cells the genome sequencing is fully known. Quoting after H.M. Johnson (2004): "Yeast is a superb model for understanding the basic functions of human cells, which have to do nearly everything yeast cells do." ... "As the human genome is sequenced, we will be able to compare human genes with those of yeast. When a similar gene is located, its function in humans can be deduced through experiments with yeast, which is much more amenable to genetic manipulation."

UNIVARIATE DISTRIBUTIONS. The boxplots show nearly symmetrical IQR boxes with numerous outliers. The variance-to-mean ratio fluctuates around 1.0, which indicates a Poisson-like or binomial-like distribution. Skewness is moderate, however excess kurtosis is large: for 15 (out of 20) amino-acids the excess kurtosis is > 2.0, the largest value being equal to 18.0; remind: for the normal distribution the excess kurtosis equals zero). For consecutive amino-acids, the estimated probabilities of their appearing in the ORF's sequence range from 0.013 to 0.097 .

MULTIVARIATE DISTRIBUTION. Converting the frequencies into percentages we obtain a real-valued data table X of size 548×20. Subsequent rows constitute individual data vectors for subsequent ORFs; at the same time the data vectors may be considered as individual data points located in R^d. The elements of X satisfy two linear restrictions, thus the covariance matrix is degenerate, which may be clearly shown by applying PCA or SVD. Thus the obtained data table can not be immediately subjected to parametric multivariate analysis using kinds of normal or multivariate t approximations. However, we are at liberty to use graphical and distance-based methods.

In Matlab we find splendid interactive tools for visualization of a data matrix (like the mesh-, waterfall-, and surf-plots) permitting to view the entire data in one glance. By rotating the plot, we may see the data in various perspectives. Applying permutations of rows and columns, we may get a more pronounced display of grouping among some instances. In particular, we might be interested to get some support for the hypothesis, that gene composition may be dependent on the gene location in the chromosome.

From the distance-based methods we found Kohonen's self-organizing maps especially appealing. A map (som) yields a planar visualization of multivariate data, with topology preservation of most its points. Operating with color (Ultsch's method) we may 'see' the 'shape' of the data cloud. Moreover, applying smoothed data histograms (SDH, Pampalk et al.) we may draw on top of the map isolines reflecting the probability density distribution of the data points in the multivariate (in our case: R^{20}) data space. We have here indeed many possibilities to find clusters of similar data vectors (genes); moreover, the attractive for biologists possibility, to visualize the found clusters in a planar map.