

Compositions in Life Science Data

Volkmar Liebscher¹

¹Volkmar Liebscher

Ernst-Moritz-Arndt-University Greifswald
Department of Mathematics and Computer Science

Jahnstr. 15a
17487 Greifswald
Germany

`volkmar.liebscher@uni-greifswald.de`

Abstract

The aim of this talk is to convince the reader that there are a lot of interesting statistical problems in presentday life science data analysis which seem ultimately connected with compositional statistics.

Key words: SAGE, cDNA microarrays, (1D-)NMR, virus quasispecies

1 Introduction

It is common biostatistical practice to approach continuous type data by methods from multivariate analysis, especially those designed for Gaussian linear models. In the talk we present several examples from high throughput molecular biology data, as there are

- serial Analysis of Gene Expression (SAGE)
- gene expression microarrays
- nuclear magnetic resonance tomography (NMR)
- pyrosequencing.

In the context of those data, compositions play a crucial rôle as data or estimable parameter. Thus, methods from compositional data analysis might give an edge in further methodological advances in the analysis of that data. Therefore, I am eager to learn more about compositional data analysis in this workshop.

2 Measuring Gene expression — SAGE and Microarrays

An evergreen in today's biology is measuring the strength of expression of some or all genes in a cell type, a tissue or a whole organism.

For that goal, Serial Analysis of Gene Expression (SAGE) is one molecular biology technique based on sequencing shorter, unique parts of available mRNA. It aims at evaluating changes in the population of mRNA. In a first approach, we observe two or a bit more samples from multinomial distributions and ask for differences in the composition of the characters = genes. Typically, one is involved with Fishers exact test and multiple testing. On the other hand, the technique does not work homogeneously on different mRNA types. So biologists just do the gene-by-gene comparisons and do not take the compositional data literally as a representative of the mRNA population.

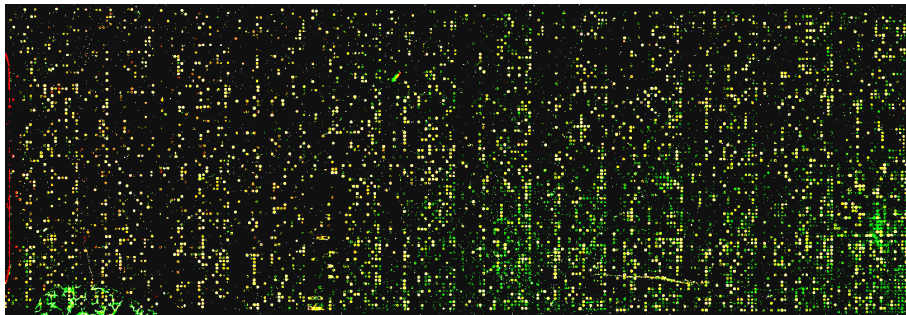


Figure 1: cDNA Microarray, false colour picture

More popular, since cheaper, are cDNA microarrays, as shown in Fig. 1. They target the mRNA too, but using hybridization. In difference to SAGE, the sample number is much higher, and the technique allows for cheaper whole-genome screens. Typically, data are considered to be of continuous type. In fact, at most the proportions of the abundances of the different genes are identifiable, leading again to compositional parameters. By the software of the microarray-scanner a lot of, often sophisticated, methods of normalisation and background correction are applied to provide raw data. Those data are typically not any more of compositional type: May be, some components are negative and normalization might have been nonlinear. But, obviously, the target parameter are the compositions of several mRNA pools, and a compositional approach might be

preferable. In comparison to SAGE, a detailed statistical model seems to be necessary to obtain a reliable statistics. Compositions might help here to deal with zeroes and negative values.

3 Metabonomics — NMR Spectroscopy

Metabonomics aims at quantifying the metabolism of organisms, especially of humans. Certainly, blood or urine contain a sort of fingerprint of the status of our body. A high throughput analysis of chemical compounds in these fluids is therefore a natural approach.

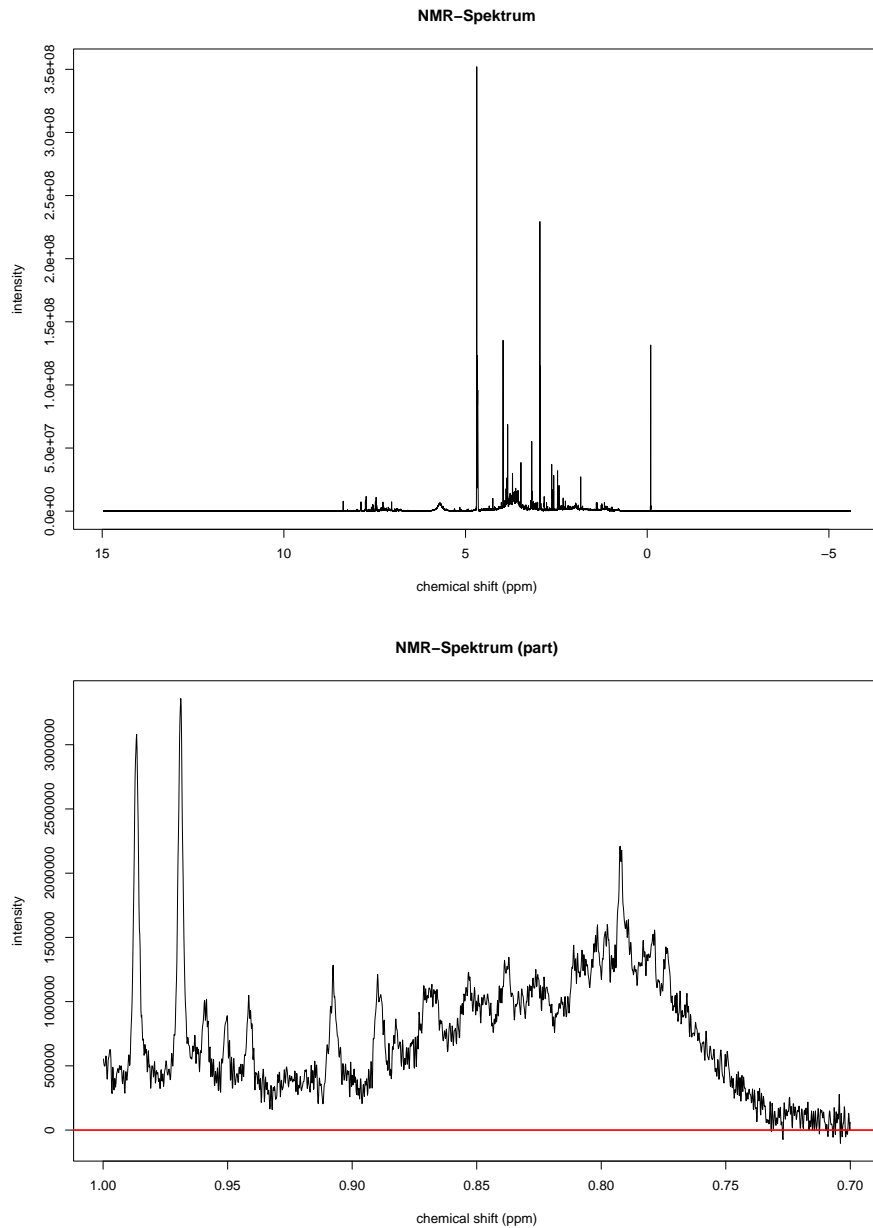


Figure 2: NMR Spektrum, full (above) and part (below). For normalization, the peak around zero should be exactly zero.

Nuclear magnetic resonance tomography (NMR) is a quantum physical technique to identify chemi-

cal compounds in a liquid and sometimes their structure, too. A typical NMR-spectrum is displayed in Fig. 2. The underlying target is the composition of chemical compounds in the fluid. In contrast to this, the real data display several drawbacks:

- First, data are obtained by Fourier transform of relaxation signals. Thus, data are negative sometimes (see Fig. 2, below).
- Second, there are noninformative parts of the spectrum (e.g. related to water) which shall be removed.
- Third, the spectrum may be shifted by some amount from measurement to measurement.
- Fourth, every substance has its own, characteristic pattern which is distributed over quite different parts of the spectrum. Thus one need further modelling or experiments, to describe these characteristic features.

Ultimately, one shall search for a *sparse*, interpretable description of the spectrum by the composition of the important compounds.

4 Virology — Fast Sequencing

These data stem from a group working in animal virology. Especially the avian flu has attracted a lot of research capacity. Due to similarities between the animal and the human virus and the special structure of the influenza viruses (the consist of 8 segments) there is a constant fear that the viruses might cross to produce a highly pathogenetic human virus. Thus one is interested in the sort and evolution of mutations in the virus. There is a recent hypothesis that inside an infected body there might be several mutated variants of the virus (quasispecies) present. This could enhance the adaptation rate of the virus population to new hosts and new drugs, since there would be no need for getting new mutations since they are already present.

454 Sequencing is a recent pyrosequencing technique which provides a huge number ($\approx 1.6 \cdot 10^6$) of sequences of short to medium size DNA fragments (≈ 100 base pairs long). The quality of the data is quite high, with some problems in detecting long repeats (regions of a genome where one nucleic acid occurs repeatedly). In this respect, this technique is similar to SAGE.

Besides its use for sequencing large genomes there is also the possibility to use this technique for detecting mutational changes in a virus population. Of course, all we could infer is the composition of the possible mutations in the virus population. Since there are sequencing errors, one cannot take the empirical composition from the data but has to provide a sparser, but reliable representation of the composition of mutations in the population.

5 Summary

There are an abundance of data from modern life science high throughput techniques, where compositions play a crucial rôle. Typical problems these data present are:

- high complexity
- negative data
- informative zeros in compositions, the goal is a *sparse* composition representation compatible with the data.