PROPORTIONS, PERCENTAGES, PPM: DO THE MOLECULAR BIOSCIENCES TREAT COMPOSITIONAL DATA RIGHT?

David Lovell[1], Warren Müller[1], Jen Taylor[2], Alec Zwart[1] and Chris Helliwell[2]

[1]CSIRO Mathematics, Informatics and Statistics
[2]CSIRO Plant Industry
Corresponding author: David.Lovell@csiro.au

ABSTRACT

The molecular biosciences make many measurements that are (or *are expressed* as) proportions, percentages or some other fraction of a whole. Even though the industrialisation of biology is currently generating this kind of data in overwhelming volumes, few people have broached the issue of compositional data analysis in genomics, transcriptomics, or other "omics" domains. As far as we can tell, the molecular biosciences in 2010 are largely unaware of the compositional data analysis methods that have been motivated by the geosciences and other disciplines.

To raise awareness among molecular bioscientists, this chapter sets out some typical bioscience scenarios giving rise to compositional data. We use a thought-experiment to highlight the difference between compositional data and the phenomena we may naïvely imagine to have observed. Next, we look at the impact of compositional constraints on univariate statistics, multivariate distance metrics, correlation and covariance, paying particular attention to log-transformation of data—a common practice in molecular bioscience. We present software to interactively explore these impacts.

We conclude with two appeals: one, to seek and apply methods that allow the *absolute* abundance of molecular species to be estimated; and two, to be aware of the potential for naïve analysis of compositional data to lead us astray.

1 Introduction

Compositional data analysis has its roots in the geosciences where geologists faced a challenge of how to analyse and interpret measurements of the mineral content of rocks—samples would be described in terms of percentages of different components, or in parts per million (or billion) for trace elements.

It can take a long time for deep methodological knowledge developed in one domain to be applied elsewhere. This chapter is motivated by the concern that there is a lot of compositional data in "the omics" (genomics, transcriptomics, *etc.*) but little awareness of the pitfalls of ignoring compositional constraints, or even that compositional constraints are at play. We are concerned that molecular biology not be led astray by findings that are more to do with artifacts of the measurement and analysis processes than the biological system being measured.

So how widespread is compositional data in the molecular biosciences? Examples include

- **Fixed size/volume samples of different components**, *e.g.*, 1g of tissue (containing different kinds of cells); 1μg of total RNA (containing different species of RNAs); 1μg of metagenomic DNA (containing DNA from different genomes); 1mL of blood (containing different metabolites)
- **Constrained counts**, such as counts of different bases or codons in a fixed-length DNA sequence
- **Proportions**, say, of different k-mers in genomes, gene ontology (GO) terms in samples, or different reads in next-generation sequencing runs.

Compositional data is commonplace in molecular biology; current, evidence of principled approaches to analysing this kind of data is scarce.

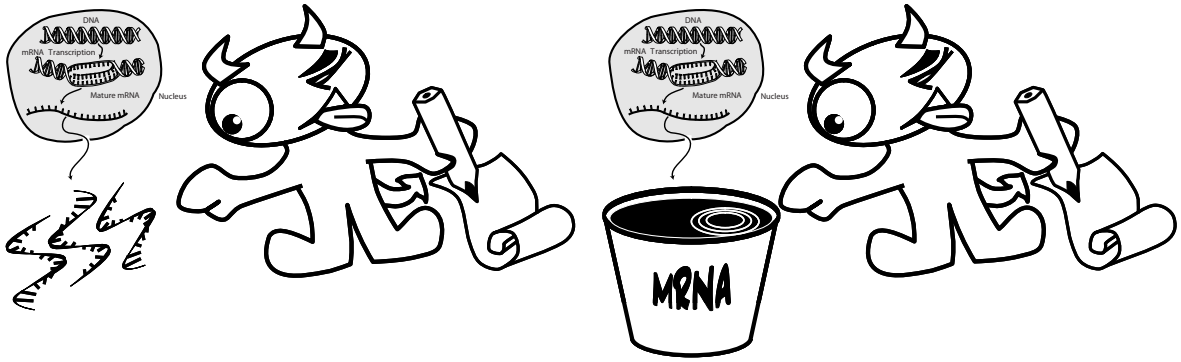

Figure 1: (Left) The Omics Imp tallying the different kinds of messenger RNA molecules emerging from the nucleus of a cell. (Right) The Omics Imp tallying the different kinds of messenger RNA molecules in a full bucket collected from the nucleus of a cell. (Illustration of mRNA courtesy: National Human Genome Research Institute.)

The biosciences can benefit greatly from ground gained in the geosciences, but biology has some important differences to geology as far as compositions are concerned. Molecular biology frequently produces compositions with tens- if not hundreds-of-*thousands* of components, whereas geosciences data usually has much lower dimension (tens to hundreds).

More fundamentally, biologists are generally interested in *living* organisms, whose cells *produce* DNA, RNA, proteins and metabolites. Often, their interest centres on the productivity of these cells. Aitchison (2008) writes that "[when] *we say that a problem is compositional we are recognizing that the sizes of our specimens are irrelevant*"; in the biosciences, the size (sometimes referred to as the *absolute abundance*) of specimens is often highly relevant because it pertains to the productivity of cells. However, many methods of sample preparation or measurement remove information about size, leaving only relative information behind. We begin with a thought experiment to highlight this.

1.1 The Omics Imp and two bioscience experiment paradigms

Small enough to fit into a cell, yet somehow able to wield pencil and paper, The Omics Imp is a molecular accountant *par excellence*. Without disrupting biological processes, the Imp can tally the molecules it observes. Figure 1 (left) shows it counting messenger RNA (mRNA) as it emerges from the nucleus of a cell. The Imp can help experimentalists by counting the different types of mRNA molecules that it sees in a specified time interval. Clearly, these counts are non-negative and constrained only by productivity of the nucleus in the time interval. This vector of counts thus forms a *basis*.

The Imp can also work in other styles of experiment. Figure 1 (right) shows it counting mRNA collected in a bucket *after* emerging from the nucleus of a cell. The Imp can help experimentalists by counting the different types of mRNA molecules it sees in this *full* bucket. Once again, these counts are non-negative but, unlike the scenario in Figure 1 (left), they are constrained by the (arbitrary) size of the bucket, they carry no information about the *absolute* rate of mRNA production, and they are not independent of each other—if the amount of one kind of mRNA in the full bucket increases, the amounts of one or more other kinds of mRNA must decrease. The *sum-constrained* vector of counts produced in this experiment is a *composition*, and this constraint has a profound impact on both the information carried by the counts, and their subsequent interpretation.

Without The Omics Imp, most molecular bioscience measurement processes follow the *bucket-survey* paradigm shown in Figure 1 (right), often using a series of buckets (*e.g.*, extracting a fixed mass of total RNA), with filters (*e.g.*, RNA size fractionation by gel electrophoresis) and multipliers (*e.g.*, polymerase chain reaction (PCR) amplification) between them. However, it is easy to lose sight of the sum-constraints being placed on experimental material, particularly as there is no strong tradition of concern for these issues in molecular biology and bioinformatics.

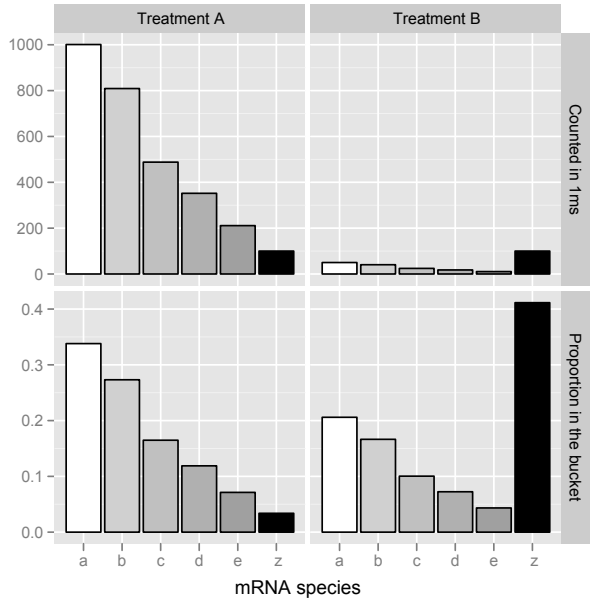

Figure 2: (Fictitious) data obtained by the Omics Imp on different mRNA species (a-z) from a cell nucleus under two different treatments (A and B). The top row shows counts of the mRNAs observed over 1ms. (Note that the Imp counted exactly the same number of mRNA z in both treatments.) The bottom row shows the findings of the corresponding bucket-survey, expressed as the proportion of each mRNA species collected in the Imp's (full) bucket.

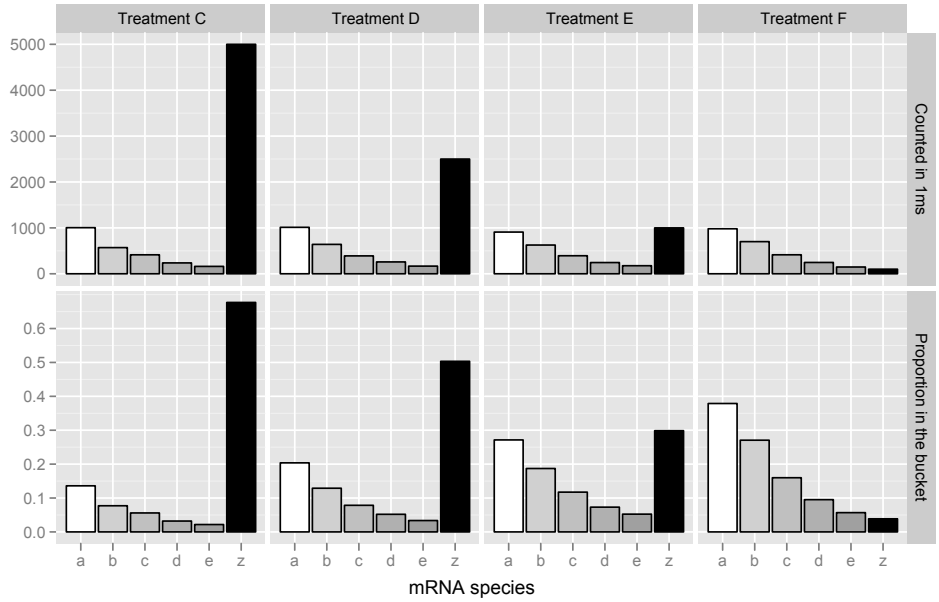

Figure 3: (Fictitious) data obtained by the Omics Imp on different mRNA species (a-z) from a cell nucleus under four different treatments (C-F). Top and bottom rows show the same kind of measurements as Figure 2. Note that the Omics Imp counted approximately the same number of mRNAs a-e in all treatments.

The current raft of nucleotide-counting sequencing technologies (a.k.a. *next-generation*, *short-read*, or *deep* sequencing) also give the impression that a biologist can count, or at least estimate the count of the different types of DNA or RNA sequences produced by a sample of cells. But some thought about the sample preparation and DNA/RNA extraction process should make it clear that there are some different buckets constraining the numbers of molecules under measurement including (1) starting with a fixed weight or volume tissue sample (ignoring cellularity); (2) extracting a fixed weight or volume of DNA/RNA; (3) concluding with a finite (if very large) number of sequence fragment reads.

The terms *under-* and *over-expression* are often used in gene expression analysis to refer to mRNAs that are less/more expressed in comparison to some reference situation. These mRNAs are also described as being *down-/up-regulated* by processes that control their level of expression. Figure 2 emphasises the perils of conflating these terms with under- and over-*production*. The bucket-survey suggests that, in comparison to Treatment A, mRNA z is over-expressed in Treatment B, even though it is being produced at exactly the same rate in both situations. To make comprehensive statements about gene expression, we have to know the total amount of mRNA being produced (or, as Aitchison terms it, the *size* of the specimen) as well as the relative abundances of different mRNA species.

Relative abundance data can also make statistically independent components appear correlated. In the fictitious data of Figure 2 the absolute amount of mRNA z remained constant in both Treatment A and B while mRNAs a-e changed dramatically. Figure 3 shows the opposite scenario: the absolute abundances of mRNAs a-e remain constant across Treatments C-F while mRNA z changes dramatically. A naïve interpretation of the mRNA proportions in the bucket would describe mRNAs a-e as positively correlated with each other and negatively correlated with mRNA z across the four treatments—the proportions of mRNAs a-e increase together while the proportion of mRNA z decreases. All this despite the fact that the absolute number of copies of mRNAs a-z are statistically independent in the four treatments. This is another manifestation of the sum constraint imposed by looking at the contents of a full bucket. We repeat: if the amount of one kind of mRNA in the full bucket increases, the amounts of one or more other kinds of mRNA must decrease.

2 The impact of compositional constraints in the omics

Biology is complex, and biological systems have many sources of variability. Probability and statistics provide a principled framework for dealing with variation and uncertainty in the biosciences but, to paraphrase Aitchison (1986, back cover), standard statistical procedures usually lead to misinterpretation and doubtful inferences when applied to compositions.

In this section, we look at how various statistics used in the molecular biosciences are affected by *closure*, the constraint that all components sum to a constant. We will assume the reader has a grasp of compositional data analysis concepts and terminology set out earlier in this book and, for simplicity, deal only with compositions of positive components that sum to 1.

Where possible, we will relate the statistics and statistical methods to prevailing practices of analysing bioscience data, particularly the use of log-transformation which, as we shall see, is no panacea for the analysis of purely relative information.

2.1 Univariate impact of closure

Even though many molecular bioscience measurements are highly multivariate, *univariate* statistical methods—such as the t-test—are workhorses in the omics, albeit with additional machinery to incorporate prior knowledge or mitigate the effects of multiple testing. We can gain insight into the impact of sum-constraints on univariate statistics of these very high-dimensional mixtures by working with a two-part composition $\mathbf{x} = (x_1, x_2)$ derived from a basis $\mathbf{w} = (w_1, w_2)$ through its closure: $\mathbf{x} = \mathcal{C}(\mathbf{w})$. We reduce the dimension of the mixture by making x_1 be the aggregation of all components *except* x_2.

As Witten and Tibshirani (2007) point out, despite the wealth of available methods, biologists show a fondness for fold-change and the t-statistic for univariate analysis—presumably because of

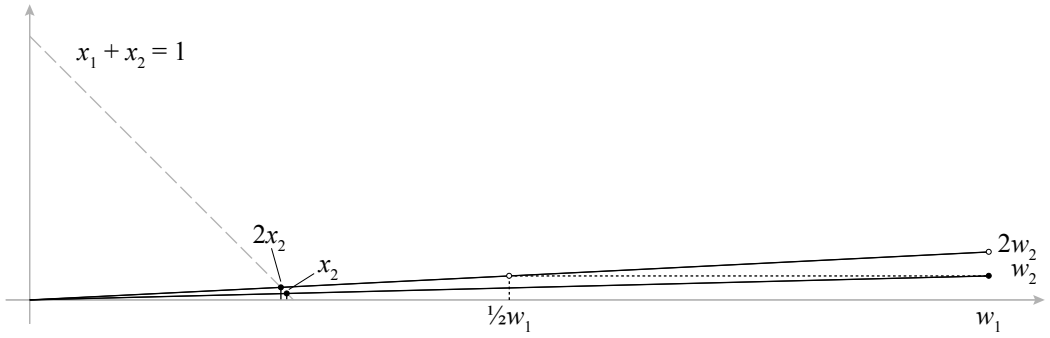

Figure 4: Did a two-fold change in x_2 occur because w_2 doubled? Or because w_1 halved? Two compositions (filled points) lie on the simplex (dashed line) corresponding to $x_1 + x_2 = 1$: the left has twice the amount of x_2 than the right. The two hollow points show two basis vectors that could give rise to a doubling of x_2: the right uses twice the amount of w_2 to give a two-fold change in x_2; the left uses *half* the amount of w_1 to the same end. We cannot tell from the compositional data alone what caused the two-fold change in x_2.

their simplicity and interpretability. However, as we can see in Figure 4 and (in more detail) in Lovell et al. (2010), two-fold change in a component of a composition (*i.e.*, x_2) does not, on its own, tell us what has happened in the underlying basis. We can say that "$x_2 \mapsto kx_2$ implies $w_2 \mapsto kw_2$" provided (1) x_2 is a relatively small component (2) the rest of the basis stays the same (3) $\log k$ is "small." This means that we are reasonably safe to do univariate statistics on components of very large compositions that are not changing dramatically.

In practice, this is exactly the situation with "spike-in" experiments, where a known concentration of a readily identifiable molecule (or cocktail of molecules) is added to samples of a mixture containing unknown concentrations of molecules, usually to assess the sensitivity of different microarray technologies (McCall and Irizarry, 2008) or infer sample mRNA concentrations (Bissels et al., 2009).

We believe that samples in many gene expression experiments will be much more variable than in carefully controlled spike-in evaluations. Are the amounts of mRNA produced by the cells under study similar enough in size and composition not to confound univariate statistical analysis? We do not know. But we suspect that the mRNA products of cells from different tissues (*e.g.*, brain/liver, cancer/non-cancer) or tissues at different stages of life (*e.g.*, dormant/germinating) are quite different in both size and composition, calling into question the whole paradigm of testing for "significant differential expression" using only measures of relative abundance.

2.2 Impact of closure on multivariate distance metrics

Molecular bioscience is replete with multivariate data, including microarray data in which each sample is represented by a point in a space of as many dimensions as there are spots on the array. Multivariate distance metrics underpin clustering methods (*e.g.*, hierarchical clustering) by telling us how "close" multivariate points are to each other. This section explores how different multivariate distance metrics are affected by compositional constraints by using two basis vectors, $\mathbf{w}$ and $\mathbf{W}$, where

$$\begin{aligned} W_i &= K_i \cdot w_i, \qquad K_i > 0 \\ &= \mathrm{g_m}(\mathbf{K}) \cdot k_i \cdot w_i \end{aligned}$$

so that $\mathbf{k}$ is a *logcontrast*. $\mathbf{x}$ and $\mathbf{X}$ are the corresponding closures of $\mathbf{w}$ and $\mathbf{W}$, so that $\mathbf{X}$ is equal to the perturbation $\mathbf{x} \oplus \mathbf{k}$ and

$$x_i = x_i / \mathbf{x} \cdot \mathbf{j}, \qquad X_i = k_i x_i / \mathbf{x} \cdot \mathbf{k}.$$

2.2.1 Aitchison's distance between compositions

For clarity, we will work with squared distances. Lovell et al. (2010, p. 16) show that Aitchison's distance can be written in terms of the logcontrast perturbation vector $\mathbf{k}$:

$$\mathrm{d}_a^2(\mathbf{x}, \mathbf{X}) \quad = \quad \sum_i \log^2 k_i.$$

Aitchison's distance tells us only about *relative* differences (*i.e.*, *ratios*) of corresponding components.

2.2.2 Euclidean distance between compositions

The (squared) Euclidean distance between compositions $\mathbf{x}$ and $\mathbf{X}$ is

$$\begin{aligned} \mathrm{d}_e^2(\mathbf{x}, \mathbf{X}) &= \sum_i (x_i - X_i)^2 \\ &= \sum_i \left(\frac{w_i}{\mathbf{w} \cdot \mathbf{j}} - \frac{k_i \cdot w_i}{\mathbf{w} \cdot \mathbf{k}} \right)^2 \end{aligned}$$

and does not lend itself to simplification without some assumptions about $\mathbf{k}$. If we assume that $\mathbf{k}$ is such that $\mathbf{w}$ and $\mathbf{W}$ each sum to the same amount we can write

$$\mathrm{d}_e^2(\mathbf{x}, \mathbf{X}) \quad = \quad \sum_i (1 - k_i)^2 x_i^2.$$

d_e is bounded by $\sqrt{2}$, and depends on both the perturbation and the composition that was perturbed.

Imagine very high dimensional compositions, *e.g.*, RNA-seq counts of thousands of different mRNA species. By necessity, each component will be a very small proportion of the total number of mRNA sequence reads. Consequently, the Euclidean distance between any two such compositions will be very small also, even though they may have components that are many-fold different in relative abundance. Aitchison's distance, with its focus on the ratio of corresponding components, will emphasise these differences in relative abundance much more effectively.

2.2.3 Euclidean distance between logged compositions

A common approach to analysing count data is to adopt a log-linear Poisson model (McCullagh and Nelder, 1989). It is also common in the biosciences to analyse and present strictly positive data using a logarithmic transformation, without necessarily referring to an underlying probabilistic model. "Logged data" is the commonplace in microarray analysis and now, RNA-seq data analysis (Robinson and Smyth, 2007; Marioni et al., 2008).

With this in mind, let us look at the Euclidean distance between $\log \mathbf{x}$ and $\log \mathbf{X}$.

$$\begin{aligned} \mathrm{d}_e^2(\log \mathbf{x}, \log \mathbf{X}) &= \sum_i (\log x_i - \log X_i)^2 \\ &= \sum_i (\log \mathbf{k} \cdot \mathbf{x} - \log k_i)^2 \\ &= D \log^2 \mathbf{k} \cdot \mathbf{x} + \sum_i \log^2 k_i \\ &= D \log^2 \mathbf{k} \cdot \mathbf{x} + \mathrm{d}_a^2(\mathbf{x}, \mathbf{X}) \qquad \geq \mathrm{d}_a^2(\mathbf{x}, \mathbf{X}) \end{aligned} \tag{1}$$

So the Euclidean distance between logged compositions is closely related to Aitchison's distance but, like d_e, still depends on both the perturbation by $\mathbf{k}$ and the composition $\mathbf{x}$ that was perturbed. Furthermore, the Euclidean distance between logged compositions also depends explicitly on D.

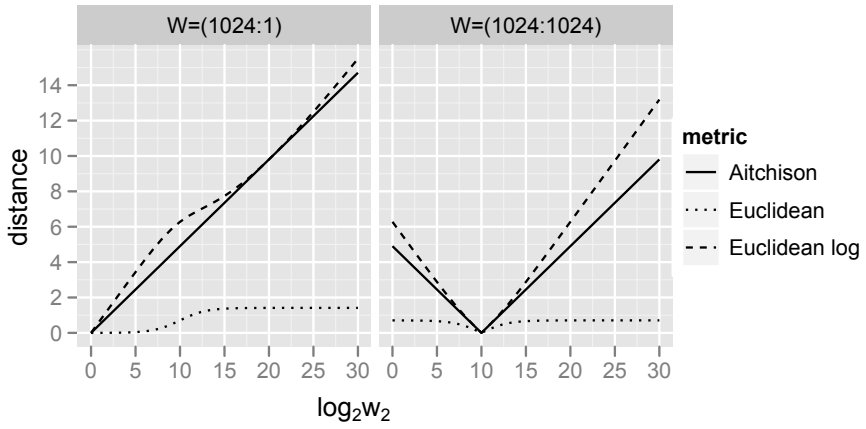

Figure 5: Plots of three different distance metrics between composition $\mathcal{C}(1024, w_2)$ and $\mathcal{C}(\mathbf{W})$ for $\mathbf{W} = (1024, 1)$ in the left panel, and $\mathbf{W} = (1024, 1024)$ in the right. Note that the Euclidean distance to $\mathcal{C}(1024, 1)$ approaches $\sqrt{2}$, the length of the edge of a D-dimensional simplex. The Euclidean distance to $\mathcal{C}(1024, 1024)$, the barycentre of the simplex, approaches $\sqrt{(D-1)/D} = 1/\sqrt{2}$.

Microarray data are conventionally dealt with on a $\log_2$ scale. So even though compositional constraints may be at play in these data (because they are derived from fixed weights of total RNA), distance-based analyses (*e.g.*, hierarchically-clustered heatmaps) are likely to be quite similar to what we would get with a purely compositional approach—more of which in Section 2.2.5

2.2.4 Three distance metrics on some two-part compositions

To provide additional insight into the three distance metrics we have considered, we apply them to the simplest possible compositions: those with only two parts. We use two basis vectors $\mathbf{w} = (w_1, w_2)$ and $\mathbf{W} = (W_1, W_2)$, and compare distances between their corresponding compositions. We fix $\mathbf{W}$ and w_1 while sweeping through a range of w_2 values and observe the distances between $\mathbf{x} = \mathcal{C}(\mathbf{w})$ and $\mathbf{X} = \mathcal{C}(\mathbf{W})$ as a function of w_2. Figure 5 shows $\mathrm{d}_a(\mathbf{x}, \mathbf{X})$, $\mathrm{d}_e(\mathbf{x}, \mathbf{X})$, and $\mathrm{d}_e(\log \mathbf{x}, \log \mathbf{X})$.

As we are working with only two components, we can go a little further in understanding when $\mathrm{d}_e(\log \mathbf{x}, \log \mathbf{X}) = \mathrm{d}_a(\mathbf{x}, \mathbf{X})$, *i.e.*, when $\mathbf{k} \cdot \mathbf{x} = 1$ (Equation 1). Since $\mathbf{x}$ is a two-part composition, $x_2 = 1 - x_1$. Also, because $\mathbf{k}$ is a logcontrast, we know that $k_2 = 1/k_1$. These facts mean that $\mathbf{k} \cdot \mathbf{x} = 1$ when $k_1 = 1$ or when $k_1 = (1 - x_1)/x_1$. In the left panel of Figure 5 we see these situations at $\mathbf{x} = \mathcal{C}(1024, 1)$ and $\mathbf{x} = \mathcal{C}(1024, 1024^2)$.

The three main points to observe from Figure 5 are that Euclidean distance does not reflect relative changes in components very well, Aitchison's distance clearly depends only on the relative abundance of corresponding components, and the Euclidean distance between log-components is bounded below by Aitchison's distance.

2.2.5 Three distance metrics on some high-dimensional compositions

While we cannot think of a way to systematically visualise the behaviour of different distance measures on higher dimensional compositions, we can look at a particular data set that has already been used to exemplify differences between distance metrics.

Vêncio et al. (2007) proposed Aitchison's simplicial distance as an alternative to Euclidean distance in clustering digital gene expression data, and demonstrated that Aitchison's distance clustered simulated RNA-seq data in essentially the same way as a Euclidean clustering of data from the corresponding microarray experiment. They also showed that Aitchison's distance clustered simulated

RNA-seq data more interpretably than clustering of that data based on Euclidean distance.

Lovell et al. (2010) replicate the results obtained by Vêncio *et al.*, and show that the Euclidean distance between the *logarithm* of the RNA counts gives essentially identical results to those using Aitchison's distance. (This was the observation that led us to look more closely at the relationship between d_a and d_e on log-transformed data (Equation 1).)

As we mentioned in Section 2.2.3, we would expect a distance-based analysis of microarray fluorescence data on a $\log_2$ scale to be quite similar to a purely compositional approach. Vêncio et al. (2007) advocate Aitchison's distance as a metric for RNA-seq and other forms of enumeration-based gene expression data on the grounds that they are compositional data. As we shall see in the next section, we think issues to do with correlation and covariance provide even stronger empirical reasons for using compositional methods.

3 Impact of closure on correlation and covariance

Compositional constraints are notorious for their impacts on the covariance and correlation structures of data (Aitchison, 1986, Section 3.3). A major motivation for the investigations described here is to understand better how compositional constraints in omics data might affect our estimates of covariance when working with log-transformed data, as is common practice in the omics.

3.1 Closure, covariance, correlation and log-transformed data

To help us understand the impact of constraints on correlation and covariance where there are large numbers of components D (as is the case in most omics settings), we will work in terms of three parts: components 1 and 2 are measurements whose pairwise correlation is of interest, and component 3 represents "the rest", *i.e.*, the other $D-2$ measurements aggregated together.

We consider first how $\mathrm{Cov}(\log x_1, \log x_2)$— the covariance between the two components of interest in the composition—relates to $\mathrm{Cov}(\log w_1, \log w_2)$— the covariance between the two components of interest in the composition's *basis*. We use the fact that

$$\mathrm{Cov}(A+C, B+C) \quad = \quad \mathrm{Cov}(A,B) + \mathrm{Cov}(A,C) + \mathrm{Cov}(B,C) + \mathrm{Var}(C)$$

to write:

$$\begin{aligned} \mathrm{Cov}(\log x_1, \log x_2) &= \mathrm{Cov}\left(\log(w_1/t), \log(w_2/t)\right) \\ &= \mathrm{Cov}(\log w_1, \log w_2) - \mathrm{Cov}(\log w_1, \log t) - \mathrm{Cov}(\log w_2, \log t) + \mathrm{Var}(\log t) \end{aligned} \quad (2)$$

This highlights the effect of variation in the *size* t of the D-part basis (Aitchison, 1986, Section 9.2). Equation 2 shows explicitly that $\mathrm{Cov}(\log x_1, \log x_2) = \mathrm{Cov}(\log w_1, \log w_2)$ when t is constant, in other words, *when* $\mathbf{w}$ *is already a composition* (but one constrained to sum to t rather than 1).

Unfortunately, the *correlation* between the (log) components of interest cannot be expressed as neatly Equation 2. However, we note that $\mathrm{Corr}(\log x_1, \log x_2) = \mathrm{Corr}(\log w_1, \log w_2)$ when t is constant.

Symbolic understanding of the relationship between $\log \mathbf{x}$ and $\log \mathbf{w}$ is useful, but we also felt a need to visualise possible relationships between bases and compositions to explore the extent to which $\mathrm{Cov}(\log x_1, \log x_2)$ could mislead us about what's going on in the underlying basis.

3.2 A simulation approach to understanding the impact of closure

Simulation has been used previously to explore compositional data (*e.g.*, Skala (1977); Brehm et al. (1998)) but, as far as we know, not to investigate the properties of *log-transformed* compositions. Simulation gives us complete control over the statistical properties of the data at the expense of losing connection to real experimental data. Unfortunately, at this time, we know of no one who has actual experimental data from both bases and their corresponding compositions in a molecular biology setting.

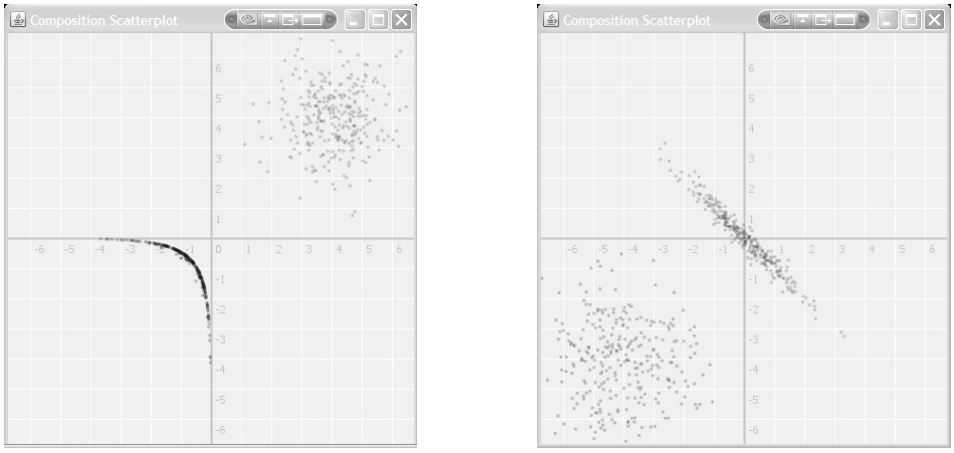

Figure 6: Two screenshots of interactive simulation software (see Lovell et al. (2010)) showing 300 samples from a trivariate log-normal basis (upper right point clouds), and the corresponding composition (lower left point clouds). Only components 1 and 2 are shown, and axes are drawn on a log scale. The left screenshot shows samples where $\log \mu_{1,2,3} = (4, 4, 0)$ and $\log \sigma_{1,2,3} = (0, 0, 0)$. While the sample correlation between $\log w_1$ and $\log w_2$ is zero, the correlation observed between $\log x_1$ and $\log x_2$ is around -0.75. The right screenshot shows samples where $\log \mu_{1,2,3} = (0, 0, 4)$ and $\log \sigma_{1,2,3} = (0, 0, 0)$. While the sample correlation between $\log w_1$ and $\log w_2$ is -0.98, the correlation observed between $\log x_1$ and $\log x_2$ is around 0.

When it comes to simulating data from a three-part basis, we are confronted by having to assume a distribution for $\mathbf{w}$. Clearly, w_1, w_2, w_3 must all be positive. For simplicity, we decided to ensure w_1, w_2 could have some straightforward statistical dependence while remaining statistically independent of w_3. We see this as "one step along" from the simplest scenario of completely independent parts.

We think that a trivariate log-normal distribution for $\mathbf{w}$ is the simplest but most general way to create a three-part basis that can be used to explore the impact of closure on the pair-wise relationship between w_1, w_2. In this scenario $\log \mathbf{w} \sim \mathcal{N}_3(\mu, \mathbf{\Sigma})$, where $\mu^{\mathrm{T}} = (\mu_1\ \mu_2\ \mu_3)$ and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

To understand the impact that parameter changes had in different parts of this 7-dimensional space ($\mu_{1,2,3}$, $\sigma_{1,2,3}$, ρ_{12}) we developed interactive plotting software in Java shown in Figure 6 (Lovell et al., 2010). We have used this software to find two extreme (but plausible) situations that characterise how the analysis of log-transformed compositional data could lead to incorrect inferences about the relationship between the components of interest, x_1 and x_2:

$w_1, w_2 \gg w_3$: When the basis is dominated by the components of interest, $\log x_1$ and $\log x_2$ tend to move towards their upper limits, *i.e.*, the boundary defined by $\log x_2 = \log(1 - x_1)$ for $x_1 \in (0, 1)$. This imposes a negative bias on the $\mathrm{Corr}(\log x_1, \log x_2)$ in comparison to $\mathrm{Corr}(\log w_1, \log w_2)$. This can be explored by sweeping ρ_{12} through its range, with $\log \mu_{1,2,3} = (4, 4, 0)$ and $\log \sigma_{1,2,3} = (0, 0, 0)$ (see Figure 6 left). This situation is easy to detect in compositional data because $x_1 + x_2$ will be close to 1. However, nothing can be done to infer the relationship between the basis variables w_1 and w_2 using the compositional data alone.

$w_3 \gg w_1, w_2$: When the basis is not dominated by the components of interest, the degree of correspondence between $\mathrm{Corr}(\log x_1, \log x_2)$ and $\mathrm{Corr}(\log w_1, \log w_2)$ depends on the variance of $\log w_3$. As σ_3 increases, $\mathrm{Corr}(\log x_1, \log x_2)$ tends to be positively biased in comparison to $\mathrm{Corr}(\log w_1, \log w_2)$. This can be explored by sweeping $\log \sigma_3$ through its range, with $\log \mu = (0, 0, 4)$, $\log \sigma_{1,2} = (0, 0)$ and $\rho_{12} = -0.98$ (see Figure 6 right). While it is again easy to detect

this situation in compositional data (this time $x_1 + x_2$ will be close to 0), there is nothing in that data to tell us about the variance of $\log w_3$.

In summary, the correlation between log-transformed components of interest in the composition will be approximately the same as that of their counterparts in the basis only when $w_3 \gg w_1, w_2$ and $\mathrm{Var}(\log w_3)$ is small—in other words, when w_1 and w_2 are small parts of a relatively constant total. *We are unable to tell when that is the case using the compositional data alone.*

4 Implications

We have explored the potential for sum-constrained data to lead analyses of omics data astray. We have seen that *provided the components of interest are relatively small parts of mixture samples that remain relatively constant in size and composition*, univariate statistics, distances on log-transformed components, and correlations between log-transformed components will not lead us to draw radically different conclusions to analyses on unconstrained data. The main problem is: *we can't tell when that proviso holds using compositional data alone.*

We believe this has two main implications. First, that, wherever possible, experimentalists should gather additional information that allows the absolute abundance of the components under study to be inferred. Second, that when only relative abundance information exists, data should be analysed using appropriate compositional methods.

4.1 Gathering information to infer absolute abundance

In our introduction, we stressed that absolute abundance of specimens (*e.g.*, mRNAs, organisms, *etc.*) is often very important in the biosciences. We introduced The Omics Imp as a means to show how different experimental paradigms can determine whether absolute abundance can be inferred, and how relative abundance alone does not tell us about how many copies of an mRNA are being produced.

Miura et al. (2008) and Kanno et al. (2006) describe methods to measure mRNA absolute abundance, so that a cell's transcriptome could be described in terms of the counts of each different kind of transcript present in that cell. To the best of our knowledge, application of these methods is not yet commonplace, but we hope that this chapter will serve as an argument for these, and other absolute abundance techniques to be employed more often in the pursuit of understanding biological systems.

4.2 Analysing compositional omics data appropriately

There are circumstances where omics data are truly relative (*e.g.*, metabolite concentrations within the bloodstream), or when interest genuinely centres on comparing relative amounts (*e.g.*, the nucleotide or codon composition of samples of genomic DNA). There are also many circumstances where measurements have been made in a ways that ensure that data carry only relative information (*e.g.*, RNA-seq or microarray data obtained from fixed volumes of total RNA). In their seminal paper on RNA-seq, Mortazavi et al. (2008) explicitly render their data compositional by working in terms of reads per kilobase of exon per million mapped sequence reads (RPKM). (By working with fixed weight aliquots of mRNA and using a sequencing platform that has limits (albeit very large ones) on the number of sequences that can be read, the data were already constrained to be compositional.)

In these situations, we think much more needs to be done to apply compositional data analysis methods instead of analysis techniques that assume data are unconstrained. We have shown that simply log-transforming the compositional data is not a panacea—we need to be sure that the components of interest are relatively small parts of mixture samples that remain relatively constant in size and composition, and this cannot be determined using compositional data alone.

Aitchison (1986) has pioneered methods for compositional data analysis, founded upon *logratios* of components. We conjecture that bringing these into play with omics data would mean, for example

- Working with (log) ratios of fluorescence intensities *between* spots within a microarray. This would be an explicitly multivariate treatment of the data rather than, say, the conventional approach of multiple univariate analyses that seek to test for significant differential expression. (One of the beliefs that has to be abandoned in working with compositional data is the idea that a single component means anything in isolation—it is only meaningful *relative* to other components.) We wonder also whether adopting this approach would obviate or simplify the process of microarray normalisation that seeks to render arrays comparable within and across experiments.
- Working with (log) ratios of mRNA counts within RNA-seq runs.
- Adopting Aitchison's distance as a metric for compositional comparison. Given the relationship between Aitchison's distance and Euclidean distance with log-transformed data (Equation 1), and the fact that omics data is often log-transformed before hierarchical clustering or other distance-based methods are applied, this may not lead to dramatically different results across the board. However, in areas that use Euclidean distance on (untransformed) compositional data, we expect the application of Aitchison's distance to provide more meaningful insights.

We can see that omics data poses challenges to compositional data analysis methods. Datasets often contain zero measurements—either because a component was not present, or because it was present but not sampled, or because some measurement error occurred. The problem of zeros becomes more pernicious the less that samples have in common, *e.g.*, metagenomic samples drawn from very different environments. Of course, this is not so much a defect of compositional data analysis methods as a sharp reminder that comparing samples with different attributes is an ill-posed problem.

A second challenge posed by omics data to compositional (indeed *any*) analysis methods is the paucity of independent samples (n) in comparison to the abundance of measurements (p). Modern bioscience data is as notoriously high-dimensional as modern bioscience data collection is underfunded, and $p \gg n$ datasets are commonplace. The industrialization of biology has us, at present, in a situation where it is feasible to make millions of measurements on a few individuals, but not *vice versa*.

We acknowledge these challenges, both methodological and financial. However, our primary aim is to ensure that bioscientists are not lead astray by artifacts of the measurement process, and we hope through this, and subsequent publications that awareness will be raised about the need to handle compositional data from the molecular biosciences appropriately.

Acknowledgments.

We gratefully acknowledge colleagues who provided us with feedback on our ideas, in particular, Rob Knight (University of Colorado, Boulder) and Ian Saunders (CSIRO).

5 REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (2008, May 27-30). The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. See Daunis-i Estadella and Martin-Fernandez (2008). http://dugi-doc.udg.edu/handle/10256/706.

Bissels, U., S. Wild, S. Tomiuk, A. Holste, M. Hafner, T. Tuschl, and A. Bosio (2009, December). Absolute quantification of microRNAs by using a universal reference. *RNA 15*(12), 2375–2384.

Brehm, J., S. Gates, and B. Gomez (1998, July). A Monte Carlo comparison of methods for compositional data analysis. In *1998 Annual Meeting of the Society for Political Methodology*. http://polmeth.wustl.edu/retrieve.php?id=295.

Daunis-i Estadella, J. and J. E. Martin-Fernandez (Eds.) (2008, May 27-30). *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Girona (Spain).

Kanno, J., K. Aisaki, K. Igarashi, N. Nakatsu, A. Ono, Y. Kodama, and T. Nagao (2006). "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays. *BMC Genomics 7*, 64.

Lovell, D., W. Müller, J. Taylor, A. Zwart, and C. Helliwell (2010, March). Caution! Compositions! Can constraints on omics data lead analyses astray? Technical Report EP10994, CSIRO. http://www.csiro.au/David.Lovell.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008, September). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research 18*, 1509–1517.

McCall, M. N. and R. A. Irizarry (2008, October). Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Research 36*(17), e108.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed ed.). London: Chapman and Hall.

Miura, F., N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki, and T. Ito (2008). Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics 9*, 574.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008, July). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods 5*(7), 621–628.

Robinson, M. D. and G. K. Smyth (2007, November). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics 23*(21), 2881–2887.

Skala, W. (1977). A mathematical model to investigate distortions of correlation coefficients in closed arrays. *Mathematical Geology 9*(5), 519–528.

Vêncio, R., L. Varuzza, C. de B Pereira, H. Brentani, and I. Shmulevich (2007). Simcluster: clustering enumeration gene expression data on the simplex space. *BMC Bioinformatics 8*(1), 246.

Witten, D. M. and R. Tibshirani (2007, November). A comparison of fold-change and the t-statistic for microarray data analysis. Technical report, Stanford University. http://www-stat.stanford.edu/ tibs/ftp/FCTComparison.pdf.