What is principal component analysis?

Markus Ringnér

Principal component analysis is often incorporated into genome-wide expression studies, but what is it and how can it be used to explore high-dimensional data?

Several measurement techniques used in the life sciences gather data for many more variables per sample than the typical number of samples assayed. For instance, DNA microarrays and mass spectrometers can measure levels of thousands of mRNAs or proteins in hundreds of samples. Such high-dimensionality makes visualization of samples difficult and limits simple exploration of the data.

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set[1]. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal. By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped.

Saal *et al.*[2] used microarrays to measure the expression of 27,648 genes in 105 breast tumor samples. I will use this gene expression data set, which is available through the Gene Expression Omnibus database (accession no. GSE5325), to illustrate how PCA can be used to represent samples with a smaller number of variables, visualize samples and genes, and detect dominant patterns of gene expression. My aim with this example is to leave you with an idea of how PCA can be used to explore data sets in which thousands of variables have been measured.

Principal components

Although understanding the details underlying PCA requires knowledge of linear algebra[1], the basics can be explained with simple geometrical interpretations of the data. To allow for such interpretations, imagine that the microarrays in our example measured the expression levels of only two genes, *GATA3* and *XBP1*. This simplifies plotting the breast cancer samples according to their expression profiles, which in this case consist of two numbers (**Fig. 1a**). Breast cancer samples are classified as being either positive or negative for the estrogen receptor, and I have selected two genes whose expression is known to correlate with estrogen receptor status[3].

PCA identifies new variables, the principal components, which are linear combinations of the original variables. The two principal components for our two-dimensional gene expression profiles are shown in **Figure 1b**. It is easy to see that the first principal component is the direction along which the samples show the largest variation. The second principal component is the direction uncorrelated to the first component along which the samples show the largest variation. If data are standardized such that each gene is centered to zero average expression level, the principal components are normalized eigenvectors of the covariance matrix of the genes and ordered according to how much of the variation present in the data they contain. Each component can then be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the samples when projected onto the component. Here, genes were centered in all examples before PCA was applied to the data. The first component in **Figure 1b** can be expressed in terms of the original variables as $PC1 = 0.83 \times GATA3 + 0.56 \times XBP1$. The components have a sample-like pattern with a weight for each gene and are sometimes referred to as eigenarrays. Methods related to PCA include independent component analysis, which is designed to identify components that are statistically independent from each other, rather than being uncorrelated[4].

Dimensional reduction and visualization

We can reduce the dimensionality of our two-dimensional expression profiles to a single dimension by projecting each sample onto the first principal component (**Fig. 1c**). This one-dimensional representation of the data retains the separation of the samples according to estrogen receptor status. The projection of the data onto a principal component can be viewed as a gene-like pattern of expression across samples, and the normalized pattern is sometimes called an eigengene. So for each sample-like component, PCA reveals a corresponding gene-like pattern containing the same variation in the data as the component. Moreover, provided that data are standardized so that samples have zero average expression, the eigengenes are eigenvectors to the covariance matrix of the samples.

So far we have used data for only two genes to illustrate how PCA works, but what happens when thousands of genes are included in the analysis? Let's apply PCA to the 8,534 probes on the microarrays with expression measurements for all 105 samples. To get a view of the dimensionality of the data, we begin by looking at the proportion of the variance present in all genes contained within each principal component (**Fig. 1d**). Note that although the first few components have more variance than later components, the first two components retain only 22% of the original variance and 63 components are needed to retain 90% of the original variance. On the other hand, 104 components are enough to retain all the original variance—a much smaller number than the original 8,534 variables. When the number of variables is larger than the number of samples, PCA can reduce the dimensionality of the samples to, at most, the number of samples, without loss of information.

To see whether the variation retained in the first two components contains relevant information about the breast cancer samples, each

Markus Ringnér is in the Division of Oncology, Department of Clinical Sciences, Barngatan 2, Lund University, 221 85, Lund, Sweden.
e-mail: markus.ringner@med.lu.se

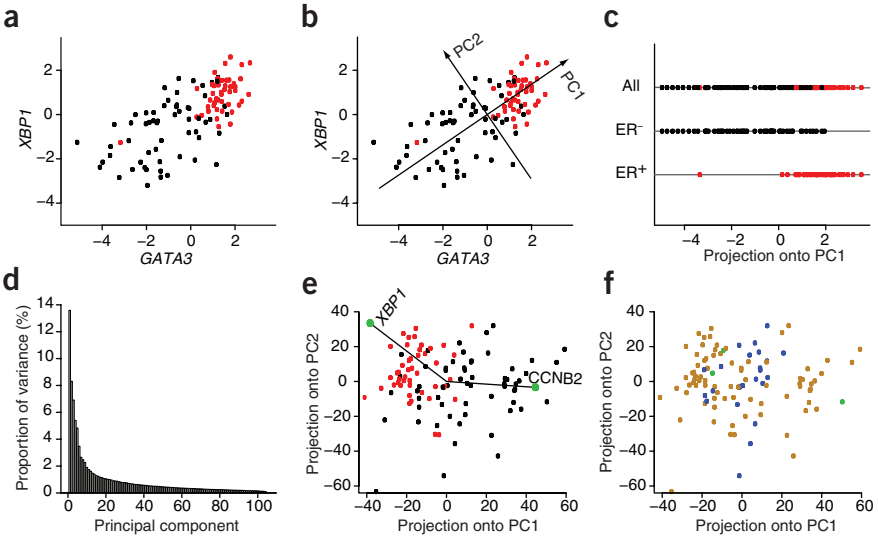

Figure 1 Principal component analysis (PCA) of a gene expression data set. (**a**) Each dot represents a breast cancer sample plotted against its expression levels for two genes. (In **a–c**, **e,** samples are colored according to estrogen receptor (ER) status: ER$^+$, red; ER$^-$, black). (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread. (**c**) Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER$^+$, ER$^-$ and all samples separately. (**d**) The variance of the principal components when PCA is applied to all 8,534 genes with expression levels for all samples. (**e**) PCA biplot with samples plotted in two dimensions using their projections onto the first two principal components, and two genes plotted using their weights for the components (green points). The scale shown is for the samples; for the genes, the scale should be divided by 950. (**f**) Samples plotted as in **e** but colored according to *ERBB2* status (blue, *ERBB2*$^+$; brown, *ERBB2*$^-$; green, unknown).

sample is projected onto these components in **Figure 1e**. The result is that the dimensionality can be reduced from the number of genes down to two dimensions, while still retaining information that separates estrogen receptor–positive from estrogen receptor–negative samples. Estrogen receptor status is known to have a large influence on the gene expression profiles of breast cancer cells[3]. However, note that PCA did not generate two separate clusters (**Fig. 1e**), indicating that discovering unknown groups using PCA is difficult. Moreover, gene expression profiles can also be used to classify breast cancer tumors according to whether they have gained DNA copies of *ERBB2* or not[3] and this information is lost when reducing this data set to the first two principal components (**Fig. 1f**). This reminds us that PCA is designed to identify directions with the largest variation and not directions relevant for separating classes of samples. Also, it is important to bear in mind that much of the variation in data from high-throughput technologies may be due to systematic experimental artifacts[5–7], resulting in dominant principal components that correlate with artifacts.

As the principal components have a sample-like pattern with a weight for each gene, we can use the weights to visualize each gene in the PCA plot[8]. Most genes will be close to the origin in such a biplot of genes and samples, whereas the genes having the largest weights for the displayed components will extend out in their respective directions[9]. Biplots provide one way to use the correspondence between the gene-like and sample-like patterns revealed by PCA to identify groups of genes having expression levels characteristic for a group of samples. As an example, two genes with large weights are displayed in **Figure 1e**.

Applications in computational biology

An obvious application of PCA is to explore high-dimensional data sets, as outlined above. Most often, three-dimensional visualizations are used for such explorations, and samples are either projected onto the components, as in the examples here, or plotted according to their correlation with the components[10]. As much information will typically be lost in two- or three-dimensional visualizations, it is important to systematically try different combinations of components when visualizing a data set. As the principal components are uncorrelated, they may represent different aspects of the samples. This suggests that PCA can serve as a useful first step before clustering or classification of samples. However, deciding how many and which components to use in the subsequent analysis is a major challenge that can be addressed in several ways[1]. For example, one can use components that correlate with a phenotype of interest[9] or use enough components to include most of the variation in the data[11]. PCA results depend critically on preprocessing of the data and on selection of variables. Thus, inspecting PCA plots can potentially provide insights into different choices of preprocessing and variable selection.

PCA is often implemented using the singular value decomposition (SVD) of the data matrix[1]. The sample-like eigenarray and the gene-like eigengene patterns are both uncovered simultaneously by SVD[10,12]. Many applications beyond dimensional reduction, classification and clustering have taken advantage of global representations of expression profiles generated by this decomposition. Applications include identifying patterns that correlate with experimental artifacts and filtering them out[6], estimating missing data, associating genes and expression patterns with activities of regulators and helping to uncover the dynamic architecture of cellular phenotypes[7,10,12]. The rapid growth in technologies that generate high-dimensional molecular biology data will likely provide many new applications for PCA in the years to come.

ACKNOWLEDGMENTS

I wish to thank the Swedish Foundation for Strategic Research for support through the Lund Strategic Centre for Clinical Cancer Research (CREATE Health).

1. Jolliffe, I.T. *Principal Component Analysis* (Springer, New York, 2002).
2. Saal, L.H. *et al. Proc. Natl. Acad. Sci. USA* **104**, 7564–7569 (2007).
3. Perou, C.M. *et al. Nature* **406**, 747–752 (2000).
4. Comon, P. *Signal Process.* **36**, 287–314 (1994).
5. Coombes, K.R. *et al. Nat. Biotechnol.* **23**, 291–292 (2005).
6. Nielsen, T.O. *et al. Lancet* **359**, 1301–1307 (2002).
7. Li, C.M. & Klevecz, R.R. *Proc. Natl. Acad. Sci. USA* **103**, 16254–16259 (2006).
8. Gabriel, K.R. *Biometrika* **58**, 453–467 (1971).
9. Landgrebe, J. Wurst, W. & Welzl, G. *Genome Biol.* **3**, RESEARCH0019 (2002).
10. Alter, O., Brown, P.O. & Botstein, D. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
11. Khan, J. *et al. Nat. Med.* **7**, 673–679 (2001).
12. Holter, N.S. *et al. Proc. Natl. Acad. Sci. USA* **97**, 8409–8414 (2000).