BMC Systems Biology

RESEARCH Open Access

Kernel-PCA data integration with enhanced interpretability

Ferran Reverter, Esteban Vegas*, Josep M Oller

From High-Throughput Omics and Data Integration Workshop
Barcelona, Spain. 13-15 February 2013

Abstract

Background: Nowadays, combining the different sources of information to improve the biological knowledge available is a challenge in bioinformatics. One of the most powerful methods for integrating heterogeneous data types are kernel-based methods. Kernel-based data integration approaches consist of two basic steps: firstly the right kernel is chosen for each data set; secondly the kernels from the different data sources are combined to give a complete representation of the available data for a given statistical task.

Results: We analyze the integration of data from several sources of information using kernel PCA, from the point of view of reducing dimensionality. Moreover, we improve the interpretability of kernel PCA by adding to the plot the representation of the input variables that belong to any dataset. In particular, for each input variable or linear combination of input variables, we can represent the direction of maximum growth locally, which allows us to identify those samples with higher/lower values of the variables analyzed.

Conclusions: The integration of different datasets and the simultaneous representation of samples and variables together give us a better understanding of biological knowledge.

Background

With the recent rapid advancements in high-throughput technologies, such as next generation sequencing, array comparative hybridization and mass spectrometry, databases are increasing in both the amount and the complexity of the data they contain. One of the main goals of mining this type of data is to visualize the relationships between biological variables that are involved [1]. For instance, visualizing gene expression guides the process of finding genes with similar expression patterns. However, due to the number of genes involved, it is more effective to display the data by means of a low-dimensional plot. Here we focus on the problem of reducing dimensionality and the interpretability of the resulting data representations.

Principal component analysis (PCA) has a very long history and is known to be a very powerful tool in the linear case. PCA is used as a visualization tool for the analysis of microarray data [2] and [3]. However, the sample space that many research problems deal with is considered nonlinear in nature; for example, the sample space of microarray data. One reason for this nonlinearity might be that the interactions of the genes are not completely understood. Many biological pathways are still not fully understood. So, it is quite naive to assume that genes are connected in a linear fashion. Following this line of thought, research into reducing the nonlinear dimensionality for microarray gene expression data has increased. Finding methods that can handle such data is of great importance if we are to glean as much information as possible from them.

Kernel representation offers an alternative to nonlinear functions by projecting the data into a high-dimensional feature space, which increases the computational power of linear learning machines [4] and [5]. Kernel methods enable us to construct different nonlinear versions of any algorithm which can be expressed solely in terms of dot products; this is known as the kernel trick. Kernel machines can be used to implement several learning

* Correspondence: evegas@ub.edu
Department of Statistics, University of Barcelona, Diagonal, 643, 08028 Barcelona, Spain

algorithms but the interpretability of the resultant output representations may be cumbersome, because input variables are only handled implicitly [6].

Nowadays, combining multiple sources of data to improve the biological knowledge available is a challenging task in bioinformatics. Data analysis of different sources of information is not simply a matter of adding the analysis of each separate dataset; instead it consists of the simultaneous analysis of multiple variables in the different datasets [7].

Some of the most powerful methods for integrating heterogeneous data types are kernel-based methods [8] and [9]. We can describe kernel-based data integration approaches as using two basic steps. Firstly, the right kernel is chosen for each data set. Secondly, the kernels from the different data sources are combined to give a complete representation of the available data for a given statistical task. Basic mathematical operations such as multiplication, addition, and exponentiation preserve properties of kernel matrices and hence produce valid kernels. The simplest approach is to use positive linear combinations of the different kernels.

In this work, we analyze the integration of data from several sources of information using kernel PCA, from the point of view of reducing dimensionality and extending previous results [10]. Moreover, we improve kernel PCA interpretability by adding to the plot the representation of the input variables that belong to any dataset. In particular, for each input variable or linear combination of input variables, we can represent the direction of maximum growth locally, which allows us to identify those samples with higher/lower values of the variables analyzed. Therefore the integration of different datasets and the simultaneous representation of samples and variables together give us a better understanding of biological knowledge. This paper starts by briefly reviewing the notion of kernel PCA (Section 2). Section 3 contains our main results: a set of procedures to enhance the interpretability of kernel PCA when multiple datasets are analyzed simultaneously. We then present our results and apply them in parallel to analyze a nutrigenomic study in mouse [11].

Results and discussion

Kernel methods enable us to construct different nonlinear versions of any algorithm which can be expressed solely in terms of dot products, this is the case of kernel PCA. Kernel PCA can be used to reduce dimensionality, thereby improving on linear PCA, but the interpretability of the output representations may be cumbersome because the input variables are only handled implicitly.

In this section, we propose a set of procedures to improve the interpretability of kernel PCA. The procedures are related to the following aspects:

- Representation of input variables.
- Data integration and representation of input variables.
- Representation of linear combinations of input variables.
- Revealing the interpretability of input variables.

To illustrate these procedures we use an example from metabolomics and genomics. The datasets come from a nutrigenomic study in mouse [11]. Forty mice were studied and two sets of variables were acquired: expressions of 120 genes measured in liver cells; and concentrations (in percentages) of 21 hepatic fatty acids (FAs) measured by gas chromatography. Biological units (mice) are cross-classified according to two factors: genotype, which can be wild-type (WT) or PPARα-deficient mice (PPAR); and diet, with 5 classes of diet in accordance with the FA composition.

The oils used for the experimental diet preparation were: corn and rapeseed oils (50:50), as the reference diet (`ref`); hydrogenated coconut oil, as a saturated FA diet (`coc`); sunflower oil, as an $\omega 6$ FA-rich diet (`sun`); linseed oil, as an $\omega 3$ FA-rich diet (`lin`); and corn, rapeseed and fish oils (42.5:42.5:15), as the `fish` diet. In the study, it cannot be assumed that variations in one set of variables cause variations in the other; we do not know a priori if changes in gene expression imply changes in FA concentrations or vice versa. Indeed, the nuclear receptor PPARα, which acts as a ligand-induced transcriptional regulator, is known to be activated by various FAs and to regulate the expression of several genes involved in FA metabolism. It should be noted that the main observations discussed in [11], which were extracted separately from the two datasets by both classical multidimensional tools (hierarchical clustering and PCA) and standard test procedures, are also highlighted by kernel PCA graphical representations.

Representation of input variables

In order to achieve interpretability we add supplementary information into kernel PCA representations. We have developed a procedure to represent any given input variable on the subspace spanned by the eigenvectors of $\tilde{C}$ (see Methods).

We can consider that our observations are realizations of the random vector $X = (X_1, \ldots, X_n)$. Then, to represent the prominence of the input variable X_k in kernel PCA, we take a set of points of the form: $\mathbf{y} = \mathbf{a} + s\mathbf{e}_k \in \mathbb{R}^n$, where $\mathbf{e}_k = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^n$, $s \in \mathbb{R}$, and the k-th component is equal to 1 and the others are 0. Then, we can compute the projections of the image of these points, $\tilde{\phi}(\mathbf{y})$, onto the subspace spanned by the eigenvectors of $\tilde{C}$. Taking into account equation (8), the induced curve expressed in matrix form is given by the row vector:

$$\sigma\,(s)^k_{1\times r} = \left(\mathbf{Z}_s^{\mathrm{T}} - \frac{1}{m}\mathbf{1}_m^{\mathrm{T}}K\right)\left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\right)\tilde{\mathbf{V}},$$

where $\mathbf{Z}_s$ is in the form of (7).

In addition, we can represent directions of maximum growth of $\sigma^k(s)$ with respect the variable X_k by projecting the tangent vector at $s = 0$. In matrix form, we have:

$$\left.\frac{d\sigma^k}{ds}\right|_{s=0} = \left.\frac{d\mathbf{Z}_s^{\mathrm{T}}}{ds}\right|_{s=0}\left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\right)\tilde{\mathbf{V}}, \qquad (1)$$

With:

$$\left.\frac{d\mathbf{Z}_s^{\mathrm{T}}}{ds}\right|_{s=0} = \left(\left.\frac{d\mathbf{Z}_s^1}{ds}\right|_{s=0}, \ldots, \left.\frac{d\mathbf{Z}_s^m}{ds}\right|_{s=0}\right)^{\mathrm{T}},$$

and, using the chain rule:

$$\left.\frac{d\mathbf{Z}_s^i}{ds}\right|_{s=0} = \left.\frac{\partial K(\mathbf{y},\mathbf{x}_i)}{\partial yk}\right|_{\mathbf{y}=\mathbf{a}}. \qquad (2)$$

In particular, let us consider the Gaussian radial basis function kernel: $k(\mathbf{x}, \mathbf{z}) = \exp(-c\ ||\mathbf{x} - \mathbf{z}||^2)$, with $c > 0$ a free parameter. Using the notation above, we have:

$$K(\mathbf{y},\mathbf{x}_i) = \exp(-c\|\mathbf{y}-\mathbf{x}_i\|^2) = \exp\left(-c\sum_{t=1}^{n}(y_t - x_{it})^2\right).$$

For the set of points of the form $\mathbf{y} = \mathbf{a} + s\mathbf{e}_k \in \mathbb{R}^n$:

$$\frac{dZ_s^i}{ds}|_{s=0} = \frac{\partial K(y,x_i)}{\partial y_k}|_{y=a} = -2cK(\mathrm{a},\mathrm{x}_i)(a_k - x_{ik}).$$

In addition, if $\mathbf{a} = \mathbf{x}_\beta$ (a training point) then:

$$\left.\frac{d\mathbf{Z}_s^i}{ds}\right|_{s=0} = -2cK(\mathbf{x}_\beta,\mathbf{x}_i)(x_{\beta k} - x_{ik}).$$

To illustrate our procedure we introduce a toy example. We have generated a dataset which has 18 points in 6-dimensional space. Coordinates of the points are selected in order to distinguish 3 groups clearly separated. The group 1 has 6 points such that the sum of X_1 and X_2 coordinates is equal to 15 for each point. Moreover, in this group, there are 3 points such that the sum of X_3, X_4 and X_5 is 0, and is equal to 6 for each the another 3 points. The group 2 has 6 points such that the sum of X_3, X_4 and X_5 coordinates is equal to 0 for each point. Besides, in this group, there are 3 points such that the sum of X_1 and X_2 is 0, and is equal to -4 for each the another 3 points. Finally, the group 3 has 6 points such that the sum of X_1 and X_2 coordinates is equal to 0 for each point. Moreover, in this group, there are 3 points such that the sum of X_3, X_4 and X_5 is 15, and is equal to 24 for each the another 3 points. All coordinates were perturbed with weak gaussian noise in order to introduce a small amount of variability inside each group. At each group the variable X_6 is assigned randomly according to a Gaussian of mean zero and standard deviation 0.5. The configuration of the points is such that we expect that in reduction of dimension only the first dimensions are necessary to reveal the arrangement of the three groups. It can be seen in Figure 1 where the two leading components of kernel PCA are represented. We can see the group 1 (represented by triangles up and circles) on the negative part of the first principal axe, group 2 (represented by plus signs and by cross) in the central part and the group 3 (represented by diamonds and triangles down) on the positive part.

Figure 1 shows samples and the variables from X_1 to X_5 at each sample. Variables are represented by vectors that indicate the direction of maximum growth in each variable. In fact, we can see that the vectors point to those groups characterized by higher values in each variable. For instance, the variables X_1 and X_2 point to the group 1, and the variables X_3, X_4, and X_5 point to the group 3.

Figure 2 shows the variable X_6 at each sample, we can observe that this variable is poorly represented and has no preferred direction towards any group.

A natural extension of the above procedure is the representation of linear combinations of input variables. Details can be found in section 3.2. With the aim to show this property we displayed in Figure 3 the samples and the linear combinations $X_1 + X_2$ and $X_3 + X_4 + X_5$ at each sample. Linear combinations are represented by

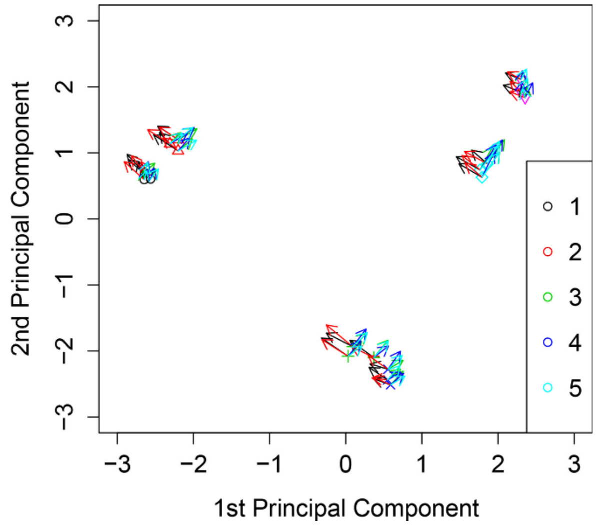

Figure 1 Kernel PCA analyzing the toy example. Variables are represented by vectors that indicate the direction of maximum growth in each variable.

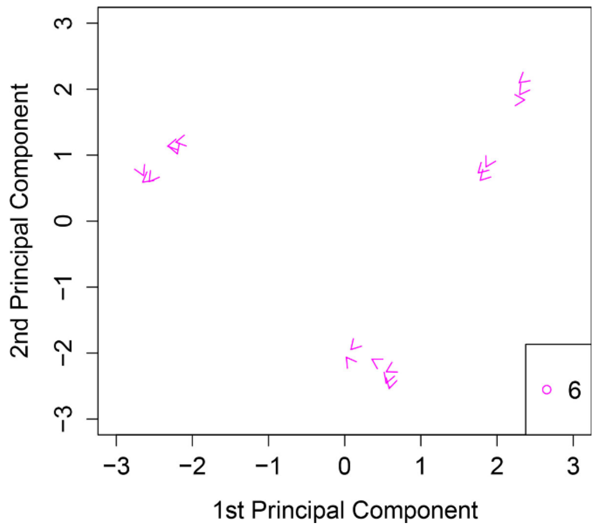

Figure 2 Kernel PCA analyzing the toy example. Variable X_6 is poorly represented and the direction of maximum growth of this variable shows no trend to any group.

vectors that point to the direction of maximum growth in each of the linear combinations. We can observe that at each sample vectors point to those groups with higher values in each of linear combinations. For example, vectors representing $X_1 + X_2$ point to group 1, and vectors representing $X_3 + X_4 + X_5$ point to group 3.

Analyzing the nutrigenomic dataset

We illustrate the representation of variables by analyzing the dataset in [11]. We apply kernel PCA and representation of variables to the genomic data and FA data.

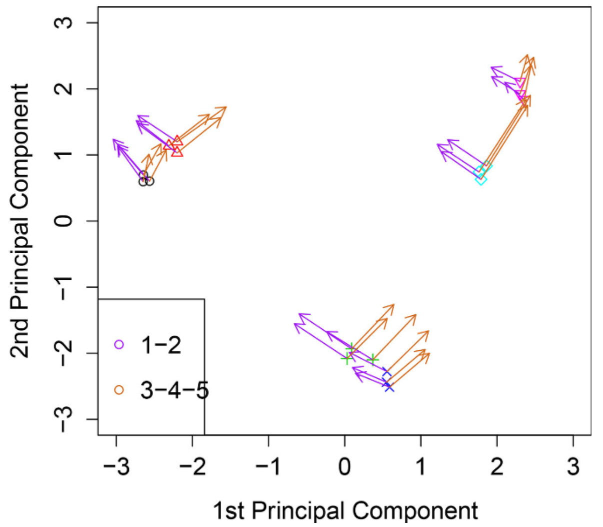

Figure 3 Kernel PCA analyzing the toy example. Linear combinations of variables are represented by vectors that indicate the direction of maximum growth in each of the linear combinations.

Firstly, we compute kernel PCA by analyzing only gene expression level data. Figure 4 shows the two leading axes of kernel PCA. We can observe that the genotypes are clearly separated (WT samples are represented in black and PPAR samples in red). Diet representation is: `ref` diet is represented by the letter x; `coc` diet by circles; sun diet by diamonds; `lin` diet by plus signs; and `fish` diet by triangles). Figure 4 shows the *AOX* (blue vector) and *CAR1* (green vector) genes. Vectors indicate the direction of maximum growth of the gene expression at each sample point. Thus, we can observe that *AOX* increases towards WT and *CAR1* towards PPAR. These results are in agreement with those found in [11] and [12]. Figure 5 and Figure 6 show the profiles of the medians of the expression of *AOX* and *CAR1* grouped by genotype. We can observe that these profiles agree with the kernel PCA representation.

Secondly, to compare results, we compute kernel PCA analyzing only FA levels. In Figure 7 we can observe that the sample points are separated by genotype, but we can also observe that the samples with `coc` diet (a diet with hydrogenated coconut oil as a saturated FA diet) form a cluster. Figure 7 shows C20.2ω.6 (green vector) and C16.0 (blue vector) FAs. It reveals higher levels of C20.2ω.6 towards PPARα-deficient clustered samples (red) and that levels of C16.0 are higher towards the WT cluster of samples (black).

These results are also in agreement with those found in [11] and [12]. Figure 8 and Figure 9 show the profiles of the medians of the concentrations of C16.0 and C20.2ω FAs, grouped by genotype. We can observe that these profiles agree with the kernel PCA representation.

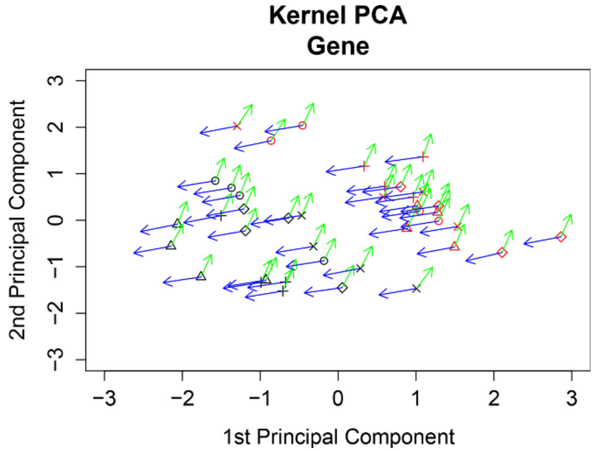

Figure 4 Kernel PCA of gene expression. The genes *AOX* (blue vector) and *CAR1* (green vector) are represented at each sample point. WT samples are represented in black and PPAR samples in red. Diet representation is: (ref) diet by the letter x; (coc) diet by circles; (sun) diet by diamonds; (lin) diet by plus signs; and (fish) diet by triangles.

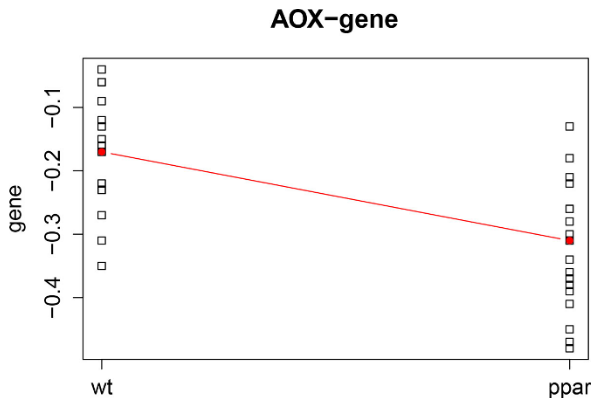

Figure 5 AOX gene profile. Profile of the median gene expression of the *AOX* gene.

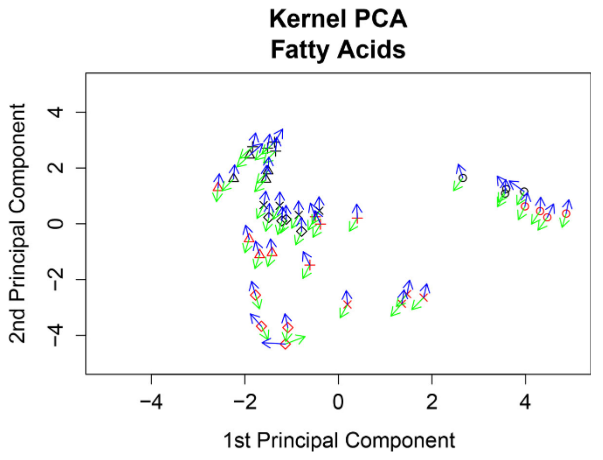

Figure 7 Kernel PCA of fatty acid concentrations. The fatty acids C16.0 (blue vector) and C20.2ω.6 (green vector) are represented at each sample point. WT samples are represented in black and PPAR samples in red. Diet representation is: (ref) diet by the letter x; (coc) diet by circles; (sun) diet by diamonds; (lin) diet by plus signs; and (fish) diet by triangles.

Data integration and representation of input variables

The kernel formalism allows us to combine heterogeneous datasets for data fusion. Basic algebraic operations such as addition, multiplication and exponentiation preserve the key properties of symmetry and positive semi-definiteness, and thus allow a simple but powerful algebra of kernels. If k_1 and k_2 are kernels defined respectively on $\mathcal{X}_1 \times \mathcal{X}_1$ and $\mathcal{X}_2 \times \mathcal{X}_2$, then their direct sum:

$$(k_1 \oplus k_2)(x_1, x_2, x'_1, x'_2) = k_1(x_1, x'_1) = k_2(x_2, x'_2)$$

is a kernel on $(\mathcal{X}_1 \times \mathcal{X}_2) \times (\mathcal{X}_1 \times \mathcal{X}_2)$. Here, $\mathbf{x}_1, \mathbf{x}'_1 \in \mathcal{X}_1$ and $\mathbf{x}_2, \mathbf{x}'_2 \in \mathcal{X}_2$.

This construction can be useful if the different parts of the input have different meanings and should therefore be dealt with differently. In that case, we can split the inputs into two parts, $\mathbf{X}_1$ and $\mathbf{X}_2$, and use two different kernels for these parts. This is the case when we are integrating two separate datasets. In consequence, our procedure can easily be extended to data fusion. Firstly, we reduce the dimension of the entire data $(\mathbf{x}_{1i}, \mathbf{x}_{2i})$, $i = 1, ..., m$, by applying kernel PCA with the kernel K given by $k_1 \oplus k_2$. Secondly, to find the coordinates of a test point:

$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2),$$

we proceed by analogy with (8), so that (7) becomes:

$$Z = \left(K\left(y_1, x_{1i}, y_2, x_{2i}\right)\right)_{m \times 1} = \left(k_1\left(y_1, x_{1i}\right) + k_2\left(y_2, x_{2i}\right)\right)_{m \times 1}.$$

When we integrate two datasets, we can represent any given input variable that belongs to one of the datasets. Let us suppose that we wish to represent the variable X^l_k that belongs to the dataset $l = 1, 2$. Then (2) becomes:

$$\left.\frac{d\mathbf{Z}^i_s}{ds}\right|_{s=0} = \left.\frac{\partial K_l(\mathbf{y}_l, \mathbf{x}_{li})}{\partial y_{lk}}\right|_{\mathbf{y}_l=\mathbf{a}_l}.$$

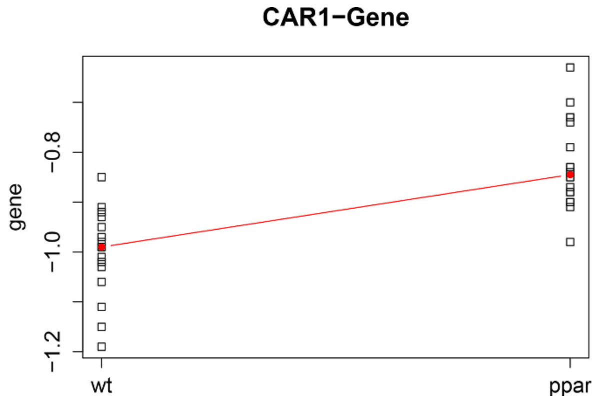

Figure 6 CAR1 gene profile. Profile of the median gene expression of the *CAR1* gene.

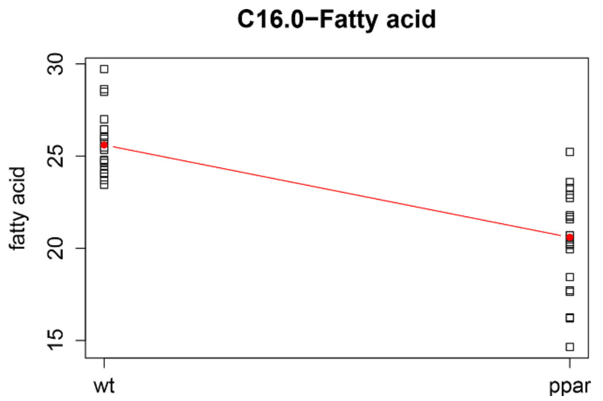

Figure 8 C16.0 fatty acid profile. Profile of the median concentrations of the C16.0 fatty acid.

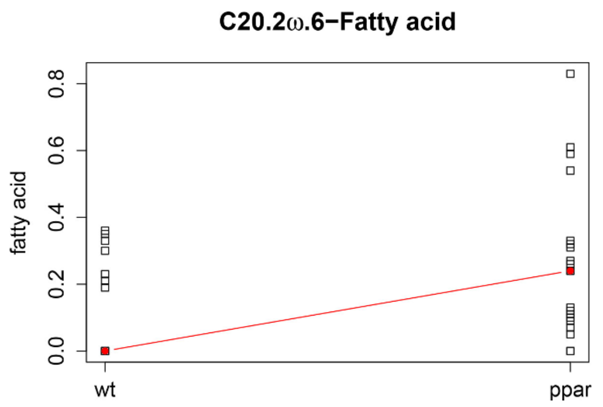

Figure 9 C20.2ω.6 fatty acid profile. Profile of the median concentrations of the C20.2ω.6 fatty acid.

Then, formula (1) allows us to display variables that belong to any of the datasets over the kernel PCA representation of samples, simultaneously.

Analyzing the nutrigenomic dataset

Continuing with the same nutrigenomic study, we compute kernel PCA by analyzing both datasets simultaneously; that is, gene expressions and FA concentrations. We observe that the genotypes are clearly separated (WT is represented in black and PPAR in red) and also mice with the `coc` diet form a cluster of both genotypes; see Figure 10. Also, Figure 10 shows AOX (black vector) and CAR1 (green vector) genes, and C20.2ω.6 (blue vector) and C16.0 (red vector) FAs. It reveals higher expression of CAR1 and higher concentrations of C20.2ω.6 towards the PPAR cluster. In contrast, AOX gene expression and concentrations of C16.0 are higher towards the WT cluster. These results are in agreement with those found in the individual kernel PCAs above.

Representation of linear combinations of input variables

A natural extension of the above procedure is the representation of linear combinations of input variables. This may be useful for representing gene modules or gene networks. Let us suppose that we wish to represent the linear combination: $X_{k_1} + X_{k_2} + \cdots + X_{k_l}$, where $k_1, k_2,...,k_l \in \{1, 2, ..., n\}$, with $ki \neq kj$, $i,j = 1, ..., l$. Then, when K is the Gaussian radial basis function kernel, (2) becomes:

$$\frac{dZ_s^i}{ds}|_{s=0} = \sum_{t=1}^{l} \frac{\partial K(\mathrm{y}, \mathrm{x}_i)}{\partial y_{k_t}}|_{\mathrm{y}=\mathrm{a}}.$$

Then, formula (1) allows us to represent any linear combination of input variables.

Analyzing the nutrigenomic dataset

To illustrate this procedure we have analyzed the genes GSTpi2, CYP3A11 and CYP2c29. These genes are involved in the functioning of detoxification [12]. We perform kernel PCA analyzing both dataset simultaneous and represent the sum of the expressions of the genes GSTpi2, CYP3A11 and CYP2c29. Figure 11 shows sample points and the vector corresponding to the sum of the three gene expressions is attached to each point. The vector indicates the direction of maximum growth of the sum of the expressions. We observe that the sum of the expressions increases towards the `fish` diet. This is in agreement with the findings in [12].

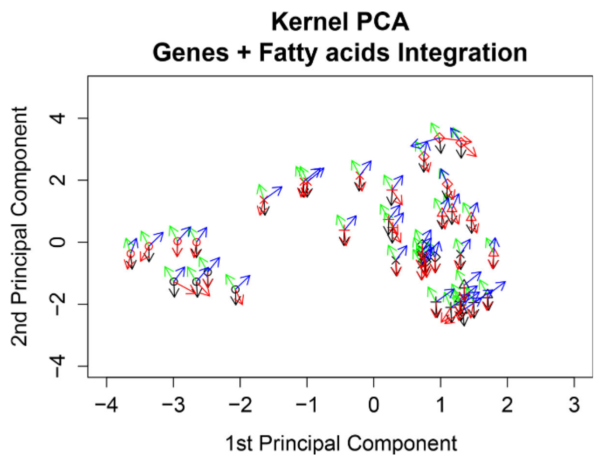

Figure 10 Kernel PCA analyzing gene expression and fatty acid concentrations simultaneously. The genes AOX (black vector) and CAR1 (green vector) and fatty acids C20.2ω.6 (blue vector) and C16.0 (red vector) are represented at each sample point. The WT samples are represented in black and the PPAR samples in red. Diet representation is: (ref) diet by the letter x; (coc) diet by circles; (sun) diet by diamonds; (lin) diet by plus signs; and (fish) diet by triangles.

Revealing the interpretability of input variables

Our procedure for representing input variables on the two-dimensional subspace expanded by the two main eigenvectors of $\tilde{C}$, displays the variables as vectors whose direction is the direction of maximum growth of the variable at a given point; in particular, at the sample points.

So, if we set a direction in this plane, given by a vector w, we can search for input variables whose representation on the kernel PCA plane are correlated with this direction. Let us suppose that we observe clusters of samples in the kernel PCA representation; then an interesting direction can be given by the vector defined by any two cluster centroids.

Once we have selected a vector w, we denote w_i *as* the parallel vector of w attached to the image given by kernel PCA of the sample point $\mathbf{x}_i$, $i = 1, ..., m$. For any variable X_k, we now compute its vector representation in kernel PCA using formula (1); we denote this vector as $\left.\frac{d\sigma^k}{ds}\right|_{s=0}$. Therefore, for each sample point,

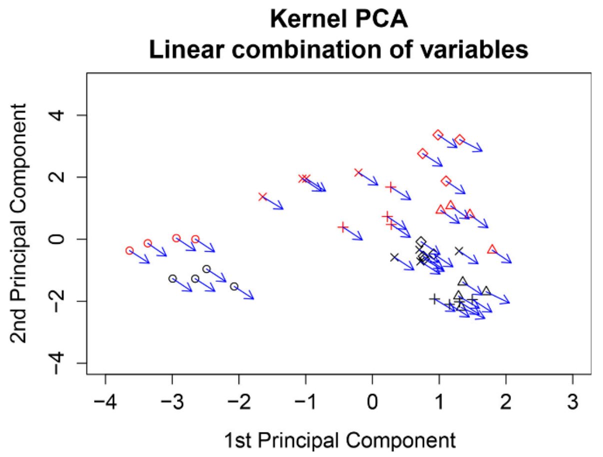

Figure 11 Representation of linear combinations of input variables. The sum of the expression of the genes: GSTpi2, CYP3A11 and CYP2c29 is represented. These genes are associated with detoxification. Wild type samples are represented in black and PPAR samples in red. Diet representation is: (ref) diet by the letter x; (coc) diet by circles; (sun) diet by diamonds; (lin) diet by plus signs; and (fish) diet by triangles.

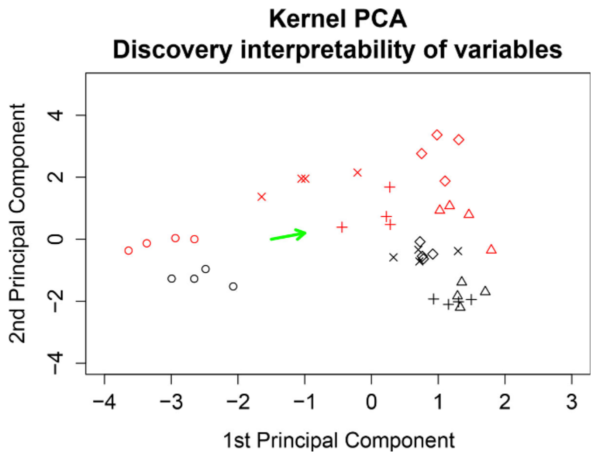

Figure 12 Kernel PCA analyzing gene expression and fatty acid concentrations simultaneously. The green vector represents variables that are expressed less in samples with the coc diet. It is defined by two cluster centroids: the left-hand cluster is the coc diet; and the right-hand cluster is comprised of the other diets.

$\mathbf{x}_i$, $i = 1, ..., m$, we have two vectors, one corresponding to the direction w_i, and other corresponding to the X_k representation, $\left(\frac{d\sigma^k}{ds}\Big|_{s=0}\right)_{\mathbf{x}_i}$. After this, to measure the strength of the correlation between X_k and w, we average the cosine of the angles between each pair of vectors, that is:

$$R_k := \frac{1}{m}\sum_{i=1}^{m}\cos\left(w_i, (\frac{d\sigma^k}{ds}|_{s=0})_{xi}\right).$$

Finally, we order all the variables according to R_k and we can select those with higher values and also those with lower values. Thus, in this way, for each sample cluster, we can find the correlated variables with higher and lower values. Knowledge of such variables can improve the biological interpretability of the results.

A natural extension of this procedure is to take as w the vector corresponding to one of the input variables. Then, if we know that a certain input variable is useful for interpreting the kernel PCA representation, we can search for other input variables whose representation on the kernel PCA plane are correlated with this feature. If we are integrating multiple datasets, we can search for correlated variables in each dataset.

Analyzing the nutrigenomic dataset

To illustrate this procedure. We have selected a preferred direction in the kernel PCA plane. Figure 12 shows this direction (green vector). This direction represents variables that are less expressed in samples with the coc diet than in those with other diets. Tables 1 and 2 summarize the genes and FAs that are most correlated with the selected direction.

In Table 1, we can observe that FAs with negative correlation, such as C16.1ω.7, C20.3ω.9 and C18.1ω.7, represent FAs with higher concentrations in samples with the coc diet. In contrast, FAs that are positively correlated, such as C22.4ω.6, C18.2ω.6, C18.3ω.3 and C22.5ω.6, represent FAs with higher concentrations in samples with other types of diet. Furthermore, in Table 2, we can observe that genes with negative correlation at the top of the table, such as S14, ACC2 and LPL, are more highly expressed in samples with the coc diet, whereas genes at the bottom of the table, that are positively correlated, are less expressed in the coc diet samples. These results are in agreement with those found in [12].

Conclusions

With the rapidly increasing amount of genomic, proteomic, and other high-throughput data that is available, the importance of data integration has increased significantly recently. Biologists, medical scientists, and clinicians are also interested in integrating the high-throughput data that has recently become available with previously existing clinical, laboratory and biological information.

Kernel methods, in particular kernel PCA, constitute a powerfully methodology because they allow us to reduce dimensionality and integrate multiple datasets, simultaneously. Moreover, in this paper we have introduced a set of procedures to improve the interpretability of kernel PCA representations. The procedures are related to the following aspects: 1) representation of variables; 2) linear combination of representations of variables; 3)

Table 1 Fatty acids: correlation with the preferred direction.

FA	mean	sd
C16.1ω.7	-0.927	0.100
C20.3ω.9	-0.917	0.336
C18.1ω.7	-0.907	0.270
C14.0	-0.898	0.131
C18.3ω.6	-0.862	0.372
C18.1ω.9	-0.695	0.132
C16.1ω.9	-0.480	0.224
C16.0	-0.295	0.265
C20.1ω.9	0.176	0.401
C22.5ω.3	0.198	0.346
C20.3ω.3	0.235	0.383
C20.5ω.3	0.300	0.219
C20.3ω.6	0.386	0.227
C18.0	0.392	0.171
C22.6ω.3	0.453	0.151
C20.2ω.6	0.601	0.306
C20.4ω.6	0.664	0.360
C22.4ω.6	0.684	0.367
C18.2ω.6	0.718	0.290
C18.3ω.3	0.727	0.482
C22.5ω.6	0.731	0.499

Fatty acids. Mean and standard deviation of the *Rk* measure of the strength of correlation with the preferred direction.

Table 2 Genes: correlation with the preferred direction.

gene	mean	sd
S14	-0.998	0.002
ACC2	-0.997	0.004
LPL	-0.997	0.005
ap2	-0.996	0.006
NGFiB	-0.996	0.005
i.FABP	-0.995	0.007
COX1	-0.993	0.012
CIDEA	-0.993	0.012
MDR1	-0.991	0.016
Lpin	-0.991	0.007
MTHFR	-0.991	0.012
Lpin1	-0.989	0.009
i.BAT	-0.988	0.014
PPARg	-0.986	0.025
ACAT2	-0.984	0.013
CYP2b10	-0.978	0.022
hABC1	-0.976	0.021
ACC1	-0.975	0.012
SPI1.1	0.353	0.042
GSTpi2	0.587	0.038

Gene codes. Mean and standard deviation of the R_k measure of the strength of correlation with the preferred direction.

data integration and representation of variables; and 4) revealing the interpretability of input variables. Our procedure is a kernel-based exploratory tool for data mining that enables us to extract nonlinear features while representing variables.

Methods

Given a sample space $\mathcal{X}$, a real valued positive definite kernel k on $\mathcal{X}$ is a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, y) = k(y, x)$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0$ for all $m \in \mathbb{N}, \alpha_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathcal{X}\ i = 1, \ldots, m$, and kernel is zero is attained if all the coefficients α_j are zero. A kernel can be interpreted as a similarity measure of the samples and allow us to identify each $\boldsymbol{x} \in \mathcal{X}$with a real function given by

$$\phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}} = \left\{ f : \mathcal{X} \to \mathbb{R} \right\}$$
$$\boldsymbol{x} \mapsto \phi(\boldsymbol{x})(\cdot) = k(\cdot, \boldsymbol{x})$$

which is an element of a dot product vector space that will be called feature space [5]. It consists of all functions

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot,\ \boldsymbol{x}_i)$$

for any $m \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}, \alpha_1, \ldots, \alpha_m \in \mathbb{R}$. It has the reproducing property

$$< k(\cdot,\ \boldsymbol{x}), f >= f(\boldsymbol{x})$$

Implying $\left\langle \varphi(x), \varphi(y) \right\rangle = \left\langle k(\cdot, x), k(\cdot, y) \right\rangle = k(x, y)$. After completion we can turn our feature space into a Hilbert space $\mathscr{H}_k$ [5]. The space $\mathscr{H}_k$ is the *reproducing kernel Hilbert space* (RKHS) induced by the kernel function k.

Given any φ and any set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, the Gram or kernel matrix of k with respect $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ is the $m \times m$ matrix K with elements $K_{ij} = \left\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \right\rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Let us define

$$\bar{\phi} := \frac{1}{m} \sum_{i=1}^{m} \phi(\boldsymbol{x}_i)$$

then, the points

$$\tilde{\varphi}(x_i) = \varphi(x_i) - \bar{\varphi} \tag{3}$$

will be centered. Let $\tilde{K}$ be denote the kernel matrix of centered points, $\tilde{K}_{ij} = \left\langle \tilde{\phi}(\boldsymbol{x}_i), \tilde{\phi}(\boldsymbol{x}_j) \right\rangle$, Because we do not have the centered data (3), we cannot compute $\tilde{K}$ explicitly, however we can express it in terms of its noncentered counterpart K [5]. Using the vector $\mathbf{1}_m = (1, \ldots, 1)^{\mathrm{T}}$, we get the more compact expression

$$\tilde{K} = K - \frac{1}{m}K\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{T}K + \frac{1}{m^2}(\mathbf{1}_m^{\mathrm{T}}K\mathbf{1}_m)\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}.$$

In $\mathscr{H}_k$ the covariance matrix takes the form

$$\tilde{C} = \frac{1}{m}\sum_{j=1}^{m}\tilde{\phi}(\boldsymbol{x}_j)\,\tilde{\phi}(\boldsymbol{x}_j)^{\mathrm{T}}.$$

We have to find eigenvalues $\tilde{\lambda} \geq 0$ and nonzero eigenvectors $\tilde{\mathbf{V}} \in \mathcal{H}_k \backslash \{0\}$ satisfying

$$\tilde{\mathbf{C}}\tilde{\mathbf{V}} = \tilde{\lambda}\tilde{\mathbf{V}} \tag{4}$$

To find the solutions of (4) we solve the dual eigenvalue problem

$$\tilde{K}\tilde{\alpha} = m\tilde{\lambda}\tilde{\alpha}, \tag{5}$$

with $\tilde{\boldsymbol{\alpha}}$ being the expansion coefficients of an eigenvector (in $\mathscr{H}_k$) in terms of the centered points (3)

$$\tilde{\mathbf{V}} = \sum_{i=1}^{m}\tilde{\alpha}_i\tilde{\phi}(\boldsymbol{x}_i). \tag{6}$$

The solution $\tilde{\boldsymbol{\alpha}}^k, k = 1, ..., r,$ are normalized by normalizing the corresponding vector $\tilde{\mathbf{V}}^k$ in $\mathscr{H}_k$, which translates into $\tilde{\lambda}_k \left\langle \tilde{\boldsymbol{\alpha}}^k, \tilde{\boldsymbol{\alpha}}^k \right\rangle = 1$.

Consider a test point $\boldsymbol{y}$. To find its coordinates we compute projections of centered φ-images of $\boldsymbol{y}$ onto the eigenvectors of the covariance matrix of the centered points,

$$\begin{aligned}\left\langle \tilde{\varphi}(y), \tilde{\mathrm{V}}^k \right\rangle &= \left\langle \varphi(y) - \tilde{\varphi}, \tilde{\mathrm{V}}^k \right\rangle \\ &= \sum_{i=1}^{m}\tilde{\alpha}_i^k \left\langle \varphi(y) - \bar{\varphi}, \varphi(x_i) - \bar{\varphi} \right\rangle \\ &= \sum_{i=1}^{m}\tilde{\alpha}_i^k \left(\left\langle \varphi(y), \varphi(x_i) \right\rangle - \left\langle \bar{\varphi}, \varphi(x_i) \right\rangle - \left\langle \varphi(y), \tilde{\varphi} \right\rangle + \left\langle \bar{\varphi}, \bar{\varphi} \right\rangle \right) \\ &= \sum_{i=1}^{m}\tilde{\alpha}_i^k \left\{ K(y, x_i) - \frac{1}{m}\sum_{s=1}^{m}K(x_s, x_i) \right. \\ &\quad \left. - \frac{1}{m}\sum_{s=1}^{m}K(y, x_s) + \frac{1}{m^2}\sum_{s,t=1}^{m}K(x_s, x_t) \right\}.\end{aligned}$$

Introducing the vector

$$\mathrm{Z} = \left(K\left(y, x_i\right)\right)_{m\times 1}. \tag{7}$$

Then,

$$\begin{aligned}\left(\left\langle \tilde{\varphi}(y), \tilde{\mathrm{V}}^k \right\rangle\right)_{1\times r} &= \mathrm{Z}^{\top}\tilde{\mathrm{V}} - \frac{1}{m}1_m^{\top}K\tilde{\mathrm{V}} - \frac{1}{m}(\mathrm{Z}^{\top}1_m)1_m^{\top}\tilde{\mathrm{V}} + \frac{1}{m^2}(1_m^{\top}K1_m)1_m^{\top}\tilde{\mathrm{V}} \\ &= \mathbf{Z}^{\top}\left(\mathrm{I}_m - \frac{1}{m}1_m1_m^{\top}\right)\tilde{\mathrm{V}} - \frac{1}{m}1_m^{\top}K\left(\mathrm{I}_m - \frac{1}{m}1_m1_m^{\top}\right)\tilde{\mathrm{V}} \\ &= \left(\mathrm{Z}^{\top} - \frac{1}{m}1_m^{\top}K\right)\left(\mathrm{I}_m - \frac{1}{m}1_m1_m^{\top}\right)\tilde{\mathrm{V}},\end{aligned} \tag{8}$$

where $\tilde{\mathbf{V}}$ is a $m \times r$ matrix whose columns are the eigenvectors $\tilde{\mathbf{V}}^1, \ldots, \tilde{\mathbf{V}}^r$.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FR suggest the combination of Kernel PCA and variable representation and designed the study. JMO was responsible for the development of the procedure. EV implemented the approach, coded the procedures and prepared analysis and results. FR and EV wrote the manuscript.

Acknowledgements

This article has been published as part of *BMC Systems Biology* Volume 8 Supplement 2, 2014: Selected articles from the High-Throughput Omics and Data Integration Workshop. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/8/S2. This work was partially supported by COST-BMBS, Action BM1006 "Next Generation Sequencing Data Analysis Network", SeqAhead.

Declarations

The publication costs for this article were partially funded by MEC-MTM2008-00642.

Published: 13 March 2014

References

1. Gorban AN, Kegl B, Wunsch DC, Zinovyev A: *Principal Manifolds for Data Visualization and Dimension Reduction* Springer Publishing Company; 2007.
2. Pittelkow YE, Wilson SR: **Visualisation of Gene Expression Data -the GE-biplot, the Chip-plot and the Gene-plot.** *Statistical Applications in Genetics and Molecular Biology* 2003.
3. Park M, Lee JW, Lee JB, Song SH: **Several biplot methods applied to gene expression data.** *Journal of Statistical Planning and Inference* 2008, **138**:500-515.
4. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis* Cambridge University Press; 2004.
5. Scholkopf B, Smola AJ: *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond* Cambridge MIT Press; 2002.
6. Li X, Shu L: **Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis.** *Expert Systems with Applications* 2009, **36**:7644-7650.
7. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J: **Data Integration in Genetics and Genomics: Methods and Challenges.** *Human Genomics Proteomics* 2009.
8. Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble S: **A statistical framework for genomic data fusion.** *Bioinformatics* 2004, **20(16)**:2626-2635.
9. Daemen A, Gevaert O, De Moor B: **Integration of clinical and microarray data with kernel methods.** *Proceedings of the 29th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC '07)* Lyon, France; 2007, 5411-5415.
10. Reverter F, Vegas E, Oller JM: **Kernel Methods for Dimensionality Reduction Applied to the "Omics" Data.** In *Principal Component Analysis -Multidisciplinary Applications* Sanguansat P, InTech 2012.
11. Martin PG, Guillou H, Lasserre F, D'ejean S, Lan A, Pascussi JM, Sancristobal M, Legrand P, Besse P, Pineau T: ***Novel aspects of PPARα*-mediated regulation of lipid and xenobiotic metabolism revealed through a multrigenomic study.** *Hepatology* 2007, **54**:767-777.
12. Gonzalez I, Dejean S, Martin PGP, Goncalves O, Besse P, Baccini A: **Highlighting relationships through Regularized Canonical Correlation Analysis: application to high throughput biology data.** *Journal of Biological Systemsn* 2009, **17(2)**:173-199.

doi:10.1186/1752-0509-8-S2-S6
Cite this article as: Reverter *et al.*: **Kernel-PCA data integration with enhanced interpretability.** *BMC Systems Biology* 2014 **8**(Suppl 2):S6.