Vegas *et al. BMC Bioinformatics* 2016, **17**(Suppl 5):205
DOI 10.1186/s12859-016-1046-1

BMC Bioinformatics

RESEARCH Open Access

Inferring differentially expressed pathways using kernel maximum mean discrepancy-based test

Esteban Vegas[1*], Josep M. Oller[1] and Ferran Reverter[1,2]

From Statistical Methods for Omics Data Integration and Analysis 2014
Heraklion, Crete, Greece. 10-12 November 2014

Abstract

Background: Pathway expression is multivariate in nature. Thus, from a statistical perspective, to detect differentially expressed pathways between two conditions, methods for inferring differences between mean vectors need to be applied. Maximum mean discrepancy (MMD) is a statistical test to determine whether two samples are from the same distribution, its implementation being greatly simplified using the kernel method.

Results: An MMD-based test successfully detected the differential expression between two conditions, specifically the expression of a set of genes involved in certain fatty acid metabolic pathways. Furthermore, we exploited the ability of the kernel method to integrate data and successfully added hepatic fatty acid levels to the test procedure.

Conclusion: MMD is a non-parametric test that acquires several advantages when combined with the kernelization of data: 1) the number of variables can be greater than the sample size; 2) omics data can be integrated; 3) it can be applied not only to vectors, but to strings, sequences and other common structured data types arising in molecular biology.

Keywords: Kernel-based methods, Kernel maximum mean test, Omics data integration

Background

A challenging topic for the bioinformatics community is how to combine data from multiple sources to increase biological knowledge. Integrating data from various different sources is not simply a matter of summing the results of each separate source; rather, it requires the analysis at the same time of all variables from various sources [1–3].

Nowadays, there are many methods to integrating heterogeneous data but kernel-based methods are usually the most powerful [4, 5]. Kernel-based methods have an extensive variety of kernels in which they can be used for each source of data. Thus, a first step to data integration is to choose an appropriate kernel for each type of data and then we combine the kernels for a given statistical task such as classification. The simplest combination of kernels is the positive linear combination of them, but other mathematical operations, such as multiplication and exponentiation, produce valid kernels.

Let us start by recalling the main ideas of kernel-based approaches.

Given a sample space $\mathcal{X}$, we say that k on $\mathcal{X}$ is a real-valued positive definite kernel on $\mathcal{X}$ if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a map such that:

- $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$,
- $\sum_{i,j=1}^{m} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for all $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathcal{X}$ where $i = 1, \ldots, m$.

Thus, 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

*Correspondence: evegas@ub.edu
[1]Department of Statistics, University of Barcelona, Diagonal, 643, 08028, Barcelona, Spain
Full list of author information is available at the end of the article

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

which is an element of a dot product vector space, from now on referred to as a feature space. 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

$$\langle k(\cdot, \boldsymbol{x}), f \rangle = f(\boldsymbol{x})$$

implying $\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle = \langle k(\cdot, \boldsymbol{x}), k(\cdot, \boldsymbol{y}) \rangle = k(\boldsymbol{x}, \boldsymbol{y})$. We can turn our feature space into a Hilbert space $\mathcal{H}_k$ by completion. The space $\mathcal{H}_k$ is the *reproducing kernel Hilbert space* (RKHS), induced by the kernel function k. This remarkable property has important consequences. Indeed, "geometric" intuition can be used to build kernel-based methods, by drawing inspiration from classical statistical methods working in finite dimensional Euclidean spaces. Popular examples of kernel-based methods are kernel principal component analysis (KPCA), kernel ridge regression (KRR), and support vector machines (SVMs) [6, 7].

Mean element

A natural question to raise is how a probability distribution $\mathbb{P}$ is represented in an RKHS $\mathcal{H}_k$. We show that infinite-dimensional counterparts of a fundamental multivariate statistical concept, the mean vector, is particularly appropriate for this purpose. This RKHS-counterpart of the mean vector is known as the mean element.

Consider a random variable X taking values in $\mathcal{X}$ and a probability distribution $\mathbb{P}$. The mean element $\mu_{\mathbb{P}}$ associated with X is the unique element of the RKHS $\mathcal{H}_k$, such that, for all $f \in \mathcal{H}_k$

$$\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}_k} = \mathbb{E}_{\mathbb{P}}[f(X)]\,.$$

In statistics, a central concern of the data integration outlined above is often referred to as the two-sample or the homogeneity problem. In this study, we explore a test statistic, known as the maximum mean discrepancy (MMD) [8–11], to test whether two samples are from the same distribution. The MMD test can easily be computed using the "kernel trick". We apply the MMD test to evaluate the differential expression of a set of genes involved in certain metabolic pathways in different conditions. The kernel method allows us integrate metabolomics data with transcriptomic data and so test the homogeneity between conditions, while handling all the available data.

Methods

Maximum mean discrepancy statistic was designed with the aim to determine a function such that its expectation differs when observations come from two different probability distributions. The underlying premise is that if we compute this statistic on samples drawn from different distributions it will measure how these distributions are likely to differ. This consideration leads to the following statistic. Let $\mathcal{X}$ denote our input domain which is assumed to be a nonempty compact set. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{F}$. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions, and let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ and $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be samples composed of independent and identically distributed observations drawn from $\mathbb{P}$ and $\mathbb{Q}$, respectively. The MMD is defined as

$$\mathrm{MMD}[\,\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x})] - \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{y})] \right),$$

and its empirical estimate is defined as

$$\mathrm{MMD}[\,\mathcal{F}, X, Y] = \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{y}_i) \right).$$

By choosing $\mathcal{F}$ to be the unit ball in a universal RKHS [12] we achieve a desirable tradeoff between a class of functions where $\mathrm{MMD}[\,\mathcal{F}, \mathbb{P}, \mathbb{Q}]$ will vanish only if $\mathbb{P} = \mathbb{Q}$ and a class of functions such that the statistic differs significantly from zero for most finite samples X and Y.

When $\mathcal{F}$ is the unit ball in a universal RKHS, Theorem 2.2 in [8] ensures that $\mathrm{MMD}[\,\mathcal{F}, \mathbb{P}, \mathbb{Q}]$ will recognize any discrepancy between $\mathbb{P}$ and $\mathbb{Q}$. Moreover, the finite sample computation of MMD is greatly simplified. Under the assumptions of the aforementioned theorem, the following is an unbiased estimator of $\mathrm{MMD}^2[\,\mathcal{F}, \mathbb{P}, \mathbb{Q}]$:

$$\mathrm{MMD}^2[\,\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(\boldsymbol{y}_i, \boldsymbol{y}_j) - \frac{2}{mn} \sum_{i \neq j}^{m,n} k(\boldsymbol{x}_i, \boldsymbol{y}_j).$$

A two-sample test based on the asymptotic distribution of an unbiased estimate of MMD^2 was introduced in [8]. The estimation of the p-value of the test can be addressed by sampling. From the aggregated data $Z = \{X, Y\}$, we draw randomly without replacement to obtain two new m-samples $\{X^*, Y^*\}$, and compute the test statistic $\mathrm{MMD}^2[\,\mathcal{F}, X^*, Y^*]$ on these new samples. If we repeat this procedure t times, a set of test statistics under the null hypothesis is obtained:

$$\mathrm{MMD}^2[\mathcal{F}, X^{1*}, Y^{1*}],\ \mathrm{MMD}^2[\mathcal{F}, X^{2*}, Y^{2*}], \ldots,\ \mathrm{MMD}^2[\mathcal{F}, X^{t*}, Y^{t*}]$$

Then, we add the original statistic $\mathrm{MMD}^2[\,\mathcal{F}, X, Y]$ to this set, and sort the set in ascending order. Finally, if r denotes the position of $MMD^2[\,\mathcal{F}, X, Y]$ withing this

Table 1 Nutrimouse studies

	coc	fish	lin	ref	sun
wt	4	4	4	4	4
ppar	4	4	4	4	4

The experimental design is balanced. There are 20 wild type (wt) mice and 20 PPAR-deficient (ppar) mice. Eight mice, four wt and four ppar mice, were fed each diet

ordering, the estimation of the p-value is given by $p = \frac{t+1-r}{t+1}$.

We compare the performance of the MMD test with those of other methods, such as the Hotelling test [13]. This is a multivariate generalization of the t-test with a multivariate normal distribution and an identical covariance structure. Alternatively, we also run a multivariate generalization of two well-established model-free univariate tests, the Wald-Wolfowitz runs test and Kolmogorov-Smirnov statistic [14], which is based on the idea of minimum spanning tree (MST). A spanning tree of a graph is a spanning subgraph that is a graph so it provides a path between every two nodes of the graph. Moreover, the MST of an edge-weighted graph is a spanning tree whose edges sum to minimum weight. In the multivariate two-sample problem, it can regard an edge-weighted graph that it is based on the N pooled multivariate data in $\mathbb{R}^p$ nodes, where p is the number of variables of the multivariate data, and edges linking all pairs. The edge weight can be estimate by the Euclidean (or any other) distances between the nodes (pairs of multivariate data). Thus, similar nodes have similar distances. The test is based on the construction of the MST of the pooled multivariate data, then it deletes all edges for which the defining nodes originate from different multivariate samples and, finally, it counts the number of disjoint subtrees (R). For large sample sizes, the permutation distribution of the standardized number of subtrees

$$W = \frac{R - E(R)}{\sqrt{var(R)}}$$

approaches the standard normal distribution and the null hypothesis, $\mathbb{P} = \mathbb{Q}$, is rejected for a small number of subtrees [14]. The multivariate Kolmogorov-Smirnov test used the MST to ranking multivariate data. Then, the MST tends to connect nodes (points) that are close. The ranking procedure begins by selecting the root the MST, that is, a node with the largest eccentricity, and then, the nodes are ranked in accordance with the height directed preorder traversal of the tree.

Results and discussion

To illustrate this procedure, we analyze data from a study in the fields of metabolomics and genomics. Specifically, the datasets are drawn from a nutrigenomic study in the mouse [15]. Forty mice were studied and two sets of variables were acquired from liver cells: 1) expressions of 120 genes derived from a nylon macroarray with radioactive labeling; and 2) concentrations of 21 hepatic fatty acids measured by gas chromatography. Biological units (mice) were cross-classified according to two factors: genotype, in either wild-type (wt) or in PPAR-deficient (ppar) mice; and diet, for which five classes (coc, fish, lin, ref, sun) were identified based on fatty acid composition (Table 1). Specifically, the oils used for experimental diet preparation were corn and colza oils (50/50) for a reference diet (ref), hydrogenated coconut oil for a saturated fatty acid diet (coc), sunflower oil for an Omega6 fatty acid-rich diet (sun), linseed oil for an Omega3-rich diet (lin) and corn/colza/enriched fish oils for the fish diet (43/43/14).

For the complete analysis we used a Gaussian kernel and the hyper-parameter was determined by the `sigest` function of the `kernlab` R package [16]. The estimation is based upon the 0.1 and 0.9 quantiles of $||\boldsymbol{x} - \boldsymbol{x}'||^2$. Basically, any value in between these two bounds will produce a good hyper-parameter estimation.

We use the `GSAR` R package [17] to implement the multivariate Kolmogorov-Smirnov test and multivariate Wald-Wolfowitz runs test.

With kernel MMD, we test whether a fatty acid catabolism pathway is differentially expressed in wt vs ppar mice. We consider a set of 16 genes involved in this catabolic pathway: PECI, MDCI, HPNCL, AOX, BIEN, THIOL, CACP, CPT2, TPα, TPβ, mHMGCoAS, Cyp4a10, Cyp4a14, ACBP,L-FABP, ACOTH and PLTP. Using a permutation procedure based on 2499 repetitions, we obtain a significant p-value (Table 2, Fig. 1). Also the three baseline tests are significant (Table 2). The kernel MMD test shows that fatty acid catabolism genes are differentially distributed in wt vs ppar mice. This result,

Table 2 *P*-values in testing fatty acid catabolism pathway

	Genes				Genes and Fatty acids			
	MMD	Hotelling	mKS	mWW	MMD	Hotelling	mKS	mWW
wt vs ppar	4e-04	1e-13	0.001	0.002	0.0096	3e-12	0.002	0.001
sun vs fish	0.1844	-	0.077	0.211	4e-4	-	0.001	0.001

Columns record the MMD, Hotelling, multivariate Kolmogorov-Smirnov (mKS) and multivariate Wald-Wolfowitz runs test (mWW) *p*-values. The first four columns (left) correspond to the pathway representation based on genes, and the second four (right) correspond to the representation based on the integration of genes and fatty acid levels

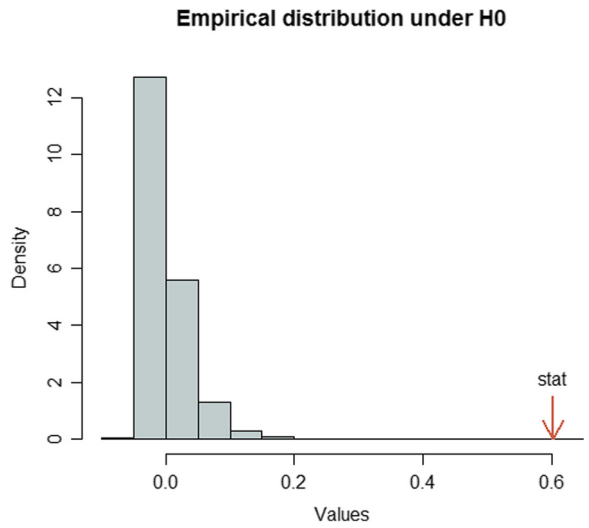

Fig. 1 Empirical distribution of kernel MMD under the null hypothesis. The observed value of the test statistic is indicated by an arrow. The number of repetitions is 2499

moreover, agrees with the data representation obtained by kernel PCA, which is used to explore simultaneously samples and genes. On the one hand, this projection shows a broad separation between wt and ppar mice (Fig. 2, samples only); on the other, each gene involved in the fatty acid catabolism pathway is displayed as an arrow in each sample (Fig. 3, both samples and pathway genes). Locally, arrows indicate the direction of maximum growth of the gene expression [18]. In Fig. 3, all genes present approximately the same direction to the left, with the exception of the ACOTH gene. Notice that wt mice lie to the left of the first axis and ppar lie to the right (Fig. 2), and by taking into account the direction of the vectors (Fig. 3), we can deduce which genes are overexpressed in wt or, in contrast, in ppar mice. Thus, we can see that the ACOTH gene is the only gene to show a higher expression in ppar mice (Fig. 3). Figure 4 (left) shows a heatmap of this set of genes in which we can observe a pattern of expression that agrees with the interpretation based on the representation of genes using kernel PCA.

We exploit the ability of the kernel method to integrate data and so add hepatic fatty acid levels in the pathway to evaluate the test procedure. We consider a set of three fatty acids: C20.5ω.3, C22.5ω.3 and C22.6ω.3 involved in fatty acid catabolism [15]. Thus, we compute the kernel matrix associated with the new feature space (including gene expression and fatty acid levels) by adding the kernel matrix of the gene expression and the kernel matrix of the fatty acid levels. Using a permutation test based on 2499 repetitions, we obtain a significant p-value when testing (Table 2). The heatmap (Fig. 4, right) presents gene expressions in addition to the fatty acid levels, showing that the differences in fatty acids are not so evident between the wt and ppar genotypes. This can be explained by the confounding effect of the diets.

We also studied the effect of the diet on the catabolism pathway. In particular, we compare sun vs fish diets. In this case, the number of samples is less than the number of variables (genes+fatty acids). Kernel methods allow us to avoid this issue in contrast to the classical Hotelling test that does not.

In addition, the heatmap of the genes (Fig. 5, left) shows an effect of the type of mouse (wt/ppar) but not of the diet. However, when the fatty acid levels are included in the analysis (Fig. 5, right), we observe a different pattern of expression between the diets; that is, the fish diet promotes the levels of this set of fatty acids. Using a permutation test based on 2499 repetitions, we

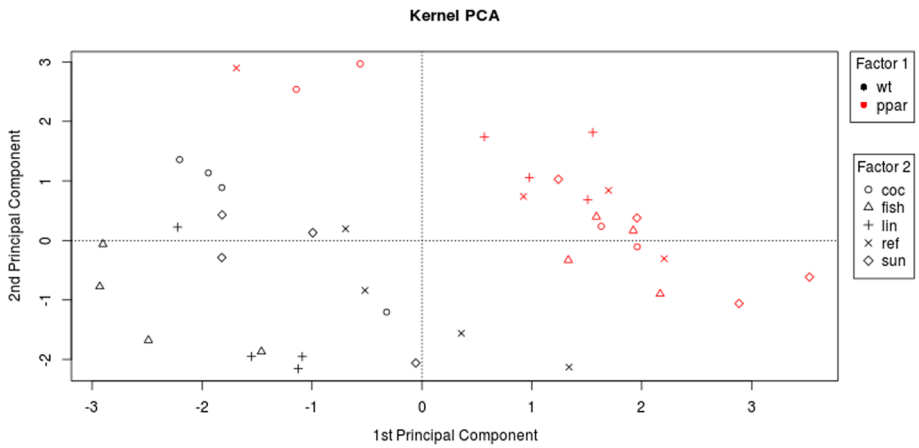

Fig. 2 Kernel PCA of gene expression. The wt samples are represented in *black* and the ppar samples in *red*. Diets represented as follows: (ref) diet by *letter x*; (coc) diet by *circles*; (sun) diet by *diamonds*; (lin) diet by *plus signs*; and (fish) diet by *triangles*

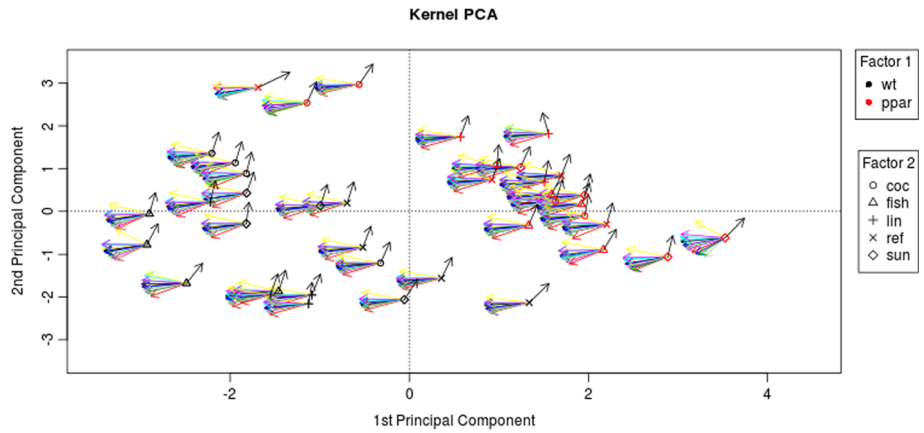

Fig. 3 Kernel PCA of gene expression. Kernel PCA of gene expression which shows 16 genes correspond to the fatty acid catabolism pathway. All genes have approximately the same direction (*black vector*) to *left* except ACOTH gene

obtain a non-significant p-value when the pathway is represented only by the gene expressions. In contrast, when the fatty acid levels are added, the p-value is significant. The multivariate Kolmogorov-Smirnov statistic and Wald-Wolfowitz runs test have similar p-values (Table 2).

The R source code and example can be consulted at [19] so the experiment can be reproduced.

Conclusion

MMD is a non-parametric test that acquires several advantages when apply the kernelization of the test: 1)

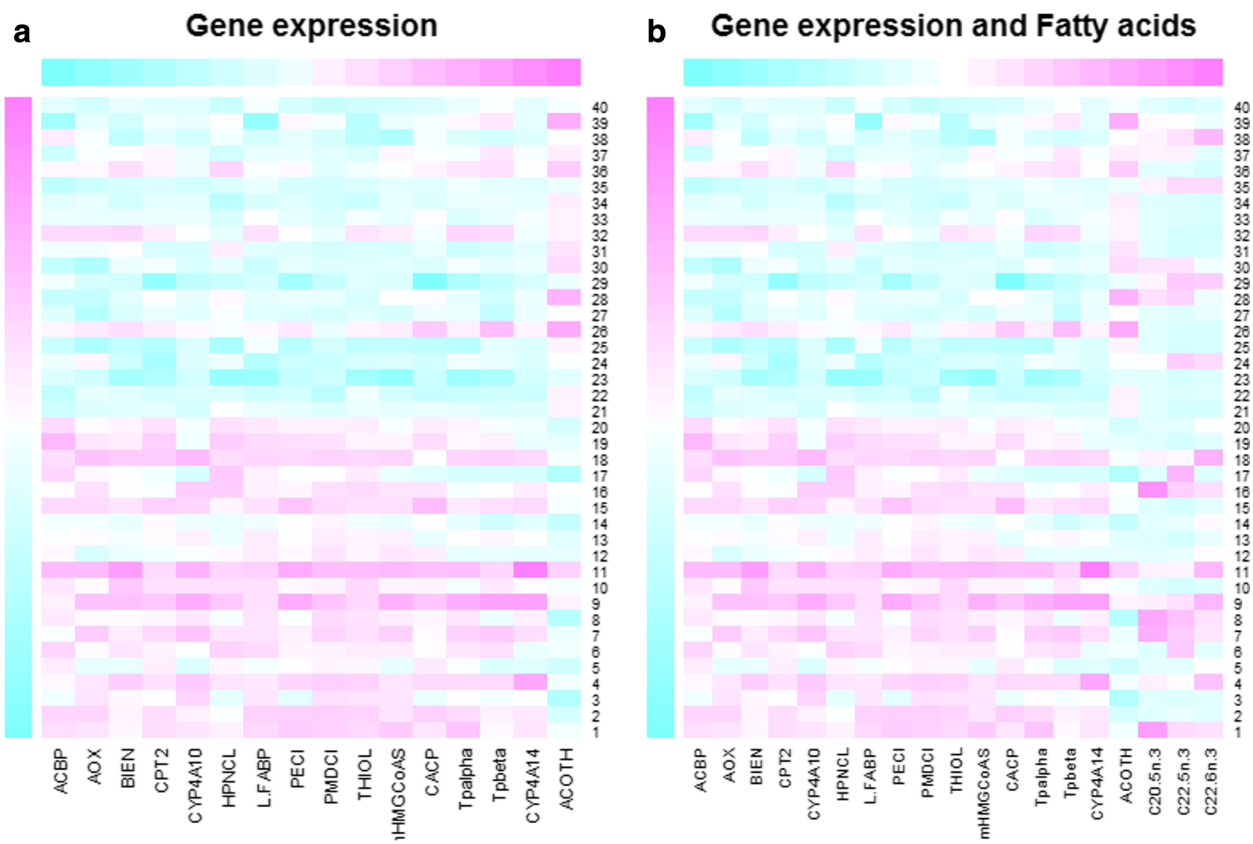

Fig. 4 Heatmaps. Figure shows the expression of the fatty acid catabolism pathway between wt and ppar genotype from gene expression (*left*) and gene expression and fatty acids (*right*). Mice from 1 to 20 are wt and from 20 to 40 are ppar

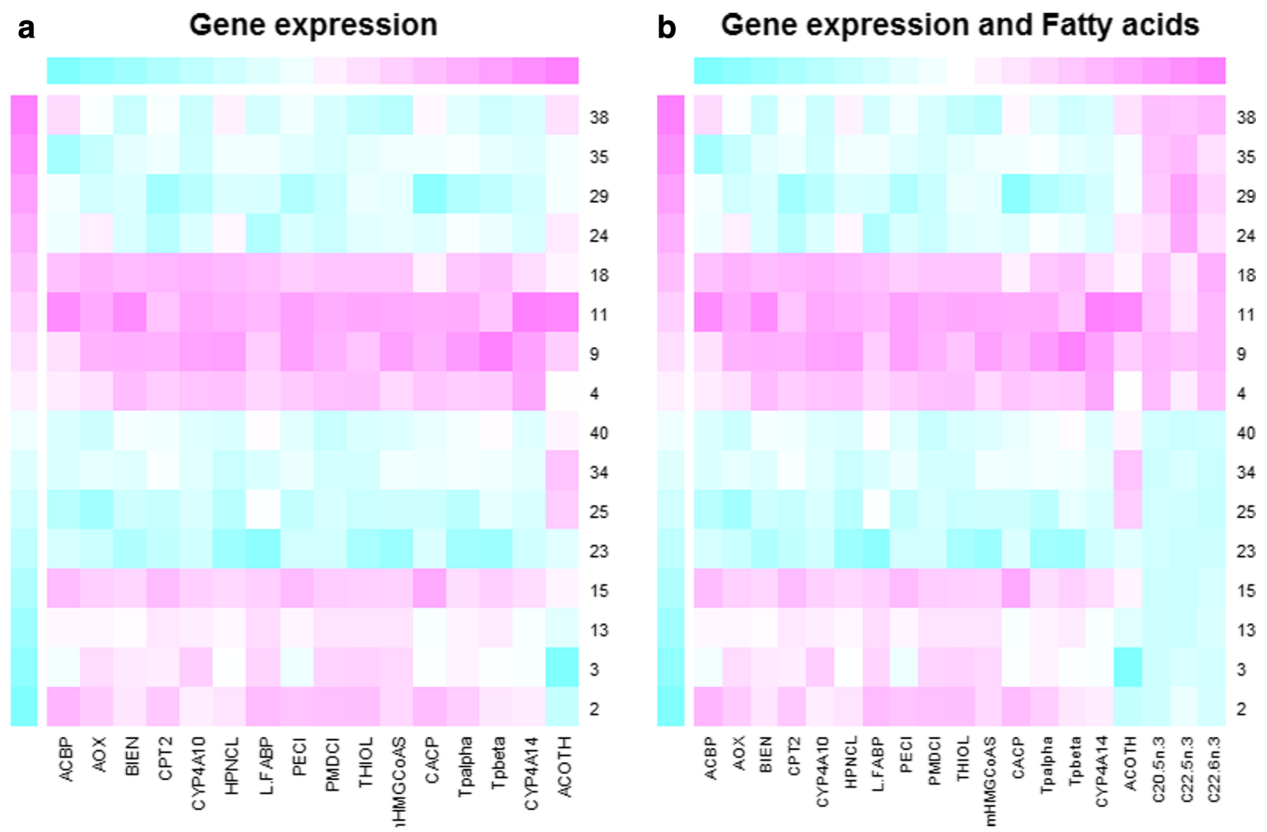

Fig. 5 Heatmaps. Figure shows the expression of fatty acid catabolism pathway between sun and fish diet from gene expression (*left*) and gene expression and fatty acids (*right*). Mice 2, 3, 13, 15, 23, 25, 34, 40 were fed the sun diet and the others the fish diet

the number of variables can be greater than the sample size; 2) omics data can usefully be integrated; 3) it can be applied not only to vectors, but to strings, sequences and other common structured data types arising in molecular biology. Our results indicate that the kernel MMD can be used to identify differentially expressed pathways; however, further studies with several sets of pathways are needed in order to assess its overall performance. This study suggests that kernel MMD is a useful approach to the analysis of pathway differential expression, since it takes into account all the genes involved in the pathway and, moreover, offers the possibility of integrating several data types.

Abbreviations

KPCA: kernel principal component analysis; KRR: kernel ridge regression; MMD: maximum mean discrepancy; MST: minimum spanning tree; PCA: principal component analysis; RKHS: reproducing kernel Hilbert space; SVM: support vector machine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EV implemented the approach and coded the procedures. FR suggested the application of Kernel MMD for differential pathway expressions. JMO prepared the analysis and discuss the results. FR and EV wrote the manuscript. All authors read and approved the final manuscript.

Declarations

This article has been published as part of BMC Bioinformatics Volume 17 Supplement 5, 2016: Selected articles from Statistical Methods for Omics Data Integration and Analysis 2014. The full contents of the supplement are available online at http://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-5. Funding for publication of this article was partially supported by grant 2014 SGR 464 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain).

Author details

[1]Department of Statistics, University of Barcelona, Diagonal, 643, 08028, Barcelona, Spain. [2]Center of Genomic Regulation, Parc de Recerca Biomedica de Barcelona, Dr. Aiguader, 88, 08003, Barcelona, Spain.

Published: 6 June 2016

References

1. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics : HGP. 2009. doi:10.4061/2009/869093.
2. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J. Data integration in the era of omics: current and future challenges. BMC Syst Biol. 2014;8(Suppl 2):1. doi:10.1186/1752-0509-8-S2-I1.
3. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype – phenotype interactions. Nat Rev Genet. 2015;16(2):85–97. doi:nrg386810.1038/nrg3868.
4. Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble S. A statistical framework for genomic data fusion. Bioinformatics. 2004;20(16): 2626–635. doi:10.1093/bioinformatics/bth294.

5. Daemen A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. In: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE; 2007. p. 5411–415. doi:10.1109/IEMBS.2007.4353566.
6. Scholkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA, USA: MIT Press; 2001.
7. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. New York, NY, USA: Cambridge University Press; 2004.
8. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ. Integrating structured biological data by Kernel Maximum Mean Discrepancy. Bioinformatics. 2006;22(14):49–57. doi:10.1093/bioinformatics/btl242.
9. Drewe P, Stegle O, Hartmann L, Kahles A, Bohnert R, Wachter A, Borgwardt K, Rätsch G. Accurate detection of differential RNA processing. Nucleic Acids Res. 2013;41(10):5189–98. doi:10.1093/nar/gkt211.
10. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC Genomics. 2013;14:826. doi:10.1186/1471-2164-14-826.
11. Gretton A. A Kernel Two-Sample Test. J Mach Learn Res. 2012;13:723–73.
12. Steinwart I. On the influence of the kernel on the consistency of support vector machines. J Mach Learn Res. 2001;2:67–93. doi:10.1162/153244302760185252.
13. Hotelling H. A generalized t test and measure of multivariate dispersion. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, Calif: University of California Press; 1951. p. 23–41. http://projecteuclid.org/euclid.bsmsp/1200500217.
14. Friedman J, Rafsky L. Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Stat. 1979;7:697–717.
15. Martin PGP, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi JM, Sancristobal M, Legrand P, Besse P, Pineau T. Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology (Baltimore, Md). 2007;45(3):767–77. doi:10.1002/hep.21510.
16. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 Package for Kernel Methods in R. J Stat Softw. 2004;11(9):1–20.
17. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics. 2014;30:360–8.
18. Reverter F, Vegas E, Oller JM. Kernel-PCA data integration with enhanced interpretability. BMC Syst Biol. 2014;8(Suppl 2):6. doi:10.1186/1752-0509-8-S2-S6.
19. The Kernel Source R Code. https://eib.stat.ub.edu/tiki-index.php?page_ref_id=73.