Metabolomics: Open Access

Ullah et al., Metabolomics 2015, 5:1
DOI: 10.4172/2153-0769.1000136

Research Article Open Access

Integrative ^{1}H-NMR-based Metabolomic Profiling to Identify Type-2 Diabetes Biomarkers: An Application to a Population of Qatar

Ullah E[1], Shahzad M[3], Rawi R[1], Dehbi M[2], Suhre K[4], Selim M[5] and Bensmail H[1]*

[1]*Computational Sciences and Engineering, Qatar Computing Research Institute, Education City, Qatar Foundation, Doha, Qatar*
[2]*Qatar Biomedical Research Institute, Education City, Qatar Foundation, Doha, Qatar*
[3]*Information Systems Department, University of Carnegie Mellon, US*
[4]*Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Qatar Foundation, Doha, Qatar*
[5]*Dermatology Department, Hamad Medical Corporation, Doha, Qatar*

Abstract

Diabetes is a leading health problem in the developed world. The recent surge of wealth in Qatar has made it one of the most vulnerable nations to diabetes and related diseases. Recent technological advances in ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy techniques for metabolomics profiling offer a great opportunity for biomarkers discovery. Using this technology, we present in this study, an integrative approach to discover new metabolites and possibly new biomarkers. We performed an integrative analysis of ^{1}H NMR spectras measured in urine, from 348 participants of the Qatar Metabolomics Study on Diabetes (QM- Diab). Our analyses revealed several metabolites that correlate with diabetes and identified specific metabolites affected by anti- diabetes medication, which constraints differentiation be- tween diabetic and control patients.

Keywords: ^{1}H-NMR; Metabolomics; Biomarkers; Diabetes; PCA; ADMM

Introduction

Many chronic diseases like Type II Diabetes (T2D) and its complications may be prevenTable by avoiding factors that trigger the disease process. Accurate prediction and identification using biomarkers will be useful for disease prevention and initiation of proactive therapies to those individuals who are most likely to develop the disease. Recent techno- logical advances in proton ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy techniques for metabolomics profiling offer great opportunity for biomarker discovery [1-18]. Because of experimental issues in the technical equipment, the levels of some metabolites cannot be universally determined. As the number of measured metabolites often exceeds the number of samples, dimensionality reduction methods are required.

In this study, we present a possible analysis workflow for mining ^{1}H-NMR spectrum for a sample of subjects with T2D and controls (see Methods) using robust statistical approaches such as regularized principal component and regularized cluster analysis methods as an integrative approach to discover new metabolites and possibly discover new biomarkers (Figure 1).

QMDiab is a 2012 study from the Dermatology Department of Hamad Medical Corporation in Doha, Qatar. The incentive was the high prevalence of T2D mellitus in Qatar, where the country ranked #21 worldwide in 2013 (International Diabetes Federation, 2014). Metabolite analysis was per- formed on human blood and urine biofluid of 348 subjects with T2D and controls (here we use urine biofluid only) where at least 100 patients were Qatari (173 males and 175 females). The subject characteristics are shown in Table 1.

In the first round of analysis the complete spectra were binned into different bin sizes and normalized using the total peak area normalization method [8]. The qualitative analysis of the major variances in the spectra was performed directly by using a newly developed flexible and robust PCA (we named fPCA) which preprocess noisy and correlated 1H-NMR data (Figure 2). The fPCA was able to cluster all the samples without diabetes but the samples with diabetes had a wide spread. Compounds that are identified by NMR spectrum analysis of original data and loading spectrum of fPC1 were identified (Betaine, Dimethylamine, Glucose, Mannitol, N, N-Dimethylglycine, and b -Alanine). These compounds belong, respectively, to the families of ammonium, amines, sugar and amino acids and can be used as potential biomarkers in human urine for detection of diabetes. Moreover, our study showed that 9 out 178 patients with diabetes had potential Paraquat poisoning based on their abnormal concentration on Citrate, Glutinane and Alanine and 24 out of 178 had Salicylate (Aspirin) detected in their urine. For Asprin abnormality, we conclude that people with diabetes are more encouraged to take Aspirin as it may reduce risk of heart attack due to coronary obstruction, which is a risk many diabetics may develop [19]. Flexible PCA is coded in Java and R and is available upon request from the corresponding author.

Abundances of metabolites are indicative of a variety of conditions, and can provide important insights in a wide variety of biological and clinical investigations. At the same time, interpretation of the spectra gives rise to substantial methodological challenges. The spectra are subject to biological and technical variations, and to uncertainty in identification and quantification of peaks. Nuclear magnetic resonance spectroscopy is a method of choice for identifying and quantifying metabolites in complex biological mixtures, as it is fast, non-destructive and highly reproducible. However interpretation of the spectra is hampered by their complexity, presence of overlapping peaks, and biological variation in the abundance of metabolites. The difficulty is particularly apparent in modern investigations, which

***Corresponding author:** Bensmail H, Computational Sciences and Engineering, Qatar Computing Research, Institute, Education City, Qatar Foundation, Doha, Qatar, Tel: 974-4454-0195; E-mail: hbensmail@qf.org.qa

Received: December 30, 2014; **Accepted:** January 14, 2015; **Published:** January 16, 2015

Citation: Ullah E, Shahzad M, Rawi R, Dehbi M, Suhre K, et al. (2015) Integrative ^{1}H-NMR-based Metabolomic Profiling to Identify Type-2 Diabetes Biomarkers: An Application to a Population of Qatar. Metabolomics 5: 136. doi: 10.4172/2153-0769.1000136

Figure 1: A Qatari population is a palette of ethnicity and nationality. Courtesy of December 2015 issue of bq magazine (bqdoha.com).

Population Characteristics	T2D n=178	No-T2D n=170
Age (years)	54.0 (34.8-70.7)	38.5(23.3-62.5)
Gender (% female)	75 (44.1%)	98 (55.1%)
Ethnicity		
Arab (%)	85 (50.0%)	115 (64.6%)
South Asian (%)	65 (38.2%)	34 (19.1%)
Filipino(%)	13 (7.6%)	22 (12.4%)
Other or mix (%)	7 (4.1%)	7 (3.9%)

Table 1: Subject characteristics. Arab: Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Morocco, Oman, Palestine, Qatar, Saudi Arabia, So- malia, Sudan, Syria, Tunisia, United Arab Emirates and Yemen South Asian: India, Bangladesh, Nepal, Pakistan, Sri Lanka. Values represent median (90% range) or number of subjects (%).

require an accurate and fast analysis of spectra from hundreds and even thousands of biological samples. Statistical inference is the only approach that can yield objective and reproducible conclusions from such data. At present the statistical tools available for this task are of limited performance.

Diabetes is usually a lifelong chronic disease characterized by an above-average concentration of sugar in the blood and urine. It is characterized as a disease of affluence and hence affects a considerable portion of the population of the developed world. Diabetes is caused by a reduction in the insulin production by the pancreas or a decreased response of body cells to insulin. The prevalence of diabetes in Qatar is higher in females than in males [1]. Risk factors also increase with age, obesity, hypertension, heart diseases and smoking habits [1]. Family history also effects a person's predisposition to diabetes, which shows that there is a significant genetic component. In Qatar, there are a lot of marriages between close cousins and this is a cause for concern. Qatar Diabetes Association (QDA), which was set up by Qatar Foundation (www.qf.org.qa), is leading the fight against diabetes by educating the general population about the risk factors.

Approach

Technologies to measure high-throughput biomedical data in proteomics, chemometrics, and genomics have led to a proliferation of high-dimensional data that pose many statistical challenges. As metabolites, are biologically interconnected, the variables, in these data sets are not only far larger than the sample size but are often highly correlated and noisy. More generally, methods such as PLS, PCA and SPCA can be used as dimension reduction techniques that finds projections of the data that maximize the covariance between the data and the response [15]. During the last decade, several work have been proposed to encourage sparsity in these projections, or loadings vectors, to select relevant features in high-dimensional data [13,14]. There are several motivations for regularizing the PCA loadings vectors. Several authors have shown that the PCA projection vectors are asymptotically inconsistent in high-dimensional settings and encouraging sparsity in the loadings has been shown to yield consistent projections [11-15]. However, the computational cost is expensive when requiring a large number of loading so it is desirable to find an approach, which regularize loading scores, reduce features and boost the computation of PCA. The PCA loading vectors can be used as a data compression technique when making future predictions; sparsity further compresses the data. As many variables in high-dimensional data are noisy and irrelevant, sparsity presents a method for automatic feature selection. This leads to results that are easier to interpret and visualize. While sparsity in PCA is important for high-dimensional data, there is also a need for more general and flexible regularized methods. Consider our NMR spectroscopy as a motivating example. This high-throughput data measures the spectrum of chemical resonances of all the latent metabolites, or small molecules, present in a biological sample. Typical experimental data consists of discretized, functional, and non-negative spectra with variables measuring in the thousands for only a small number of samples. Additionally, variables in the spectra have complex dependencies arising from correlation at adjacent chemical shifts, metabolites resonating at more than one chemical shift, and overlapping resonances of latent metabolites. Because of these complex dependencies, there is a long history of using PCA to reduce the NMR spectrum for supervised data [16]. Classical PCA or Sparse PCA, however, are not optimal for this type of data as they do not account for the non-negativity or functional nature of the spectra and do not encourage sparsity or group sparsity.

In this paper, we seek a more flexible framework for regularizing the PCA loadings that are computationally efficient and fast for analyzing high-dimensional ^{1}H NMR data that encourage sparsity, group sparsity, or smoothness, and also leads to a more computationally efficient and fast numerical algorithm.

Characteristics	Arab n = 200		South Asian n = 99		Filipino n = 35	
	Type II Diab n = 85	Non Type II Diab n = 115	Type II Diab n = 65	Non Type II Diab n = 34	Type II Diab n = 13	Non Type II Diab n = 22
Age (years)	53.9 (34.271.2)	39.1 (22.664.4)	52.6 (35.269.1)	39.0(25.057.6)	49.3(37.863.0)	37.2(23.257.8)
Gender (% female)	51 (60.0%)	70 (60.9%)	11 (16.9%)	13 (38.2%)	11 (84.6%)	13 (59.1%)
Smoking (%)	8 (9.4%)	10 (8.7%)	6 (9.2%)	2 (5.9%)	1 (7.7%)	2 (9.1%)

Table 2: Subject characteristics stratified by ethnicity.

Characteristics	Female n = 173 Type II Diabetes n = 75	Female n = 173 Non Type II Diabetes n = 98	Male n = 175 Type II Diabetes n = 98	Male n = 175 Non Type II Diabetes n =95
Age (years)	52.6 (33.770.6)	36.5 (19.561.2)	54.4 (34.971.1)	41.7 (25.964.3)
Ethnicity				
Arab (%)	51 (68.0%)	70 (71.4%)	34 (35.8%)	45 (56.3%)
South Asian (%)	11 (14.7%)	13 (13.3%)	54 (56.8%)	21 (26.3%)
Filipino (%)	11 (14.7%)	13 (13.3%)	2 (2.1%)	9 (11.3%)
Other or mix (%)	2 (2.7%)	2 (2.0%)	5 (5.3%)	5 (6.3%)

Table 3: Subject characteristics stratified by gender.

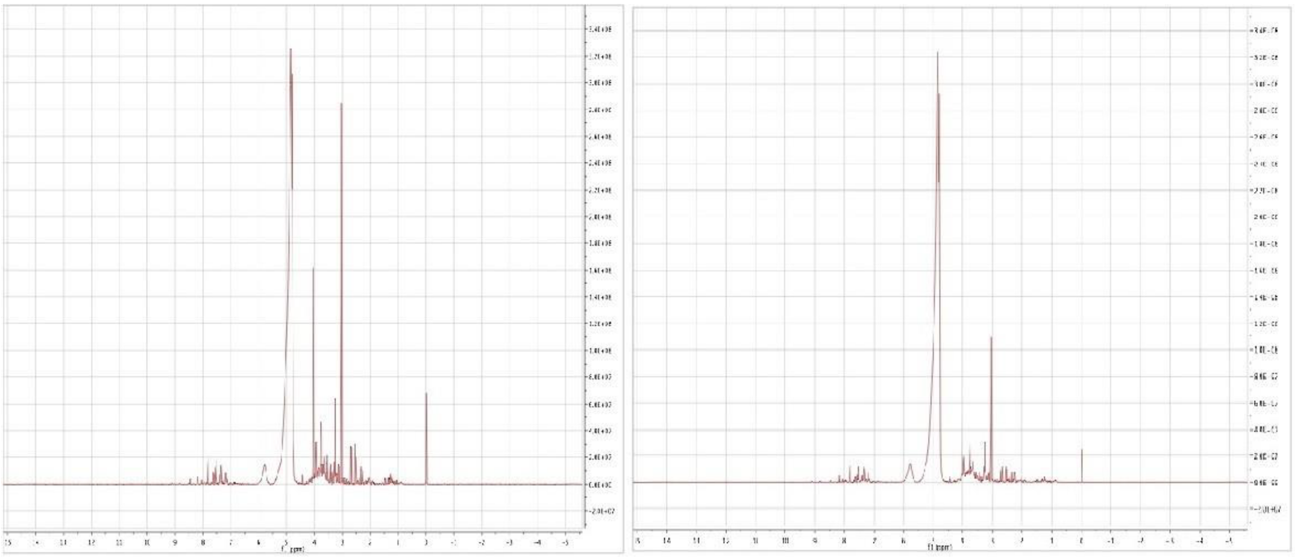

Figure 2: An example of binned spectra.

Methods

QMDiab Study

This study was embedded in the Qatar Metabolomics Study on Diabetes (QMDiab), a cross-sectional case-control study with 348 subjects (Tables 1-3). The work was a joint collaboration between Hamad Medical Corporation and Weill Cornell Medical College Qatar. Patients were asked to enroll between February and June 2012. The study has been approved by the Institutional Review Board (IRB) of Hamad Medical Corporation and Weill Cornell Medical College Qatarand is accordance with the Helsinki Declaration of 1975. Written informed consent was obtained from all participants. The study measured metabolites in 348 individuals within the age of 17 to 81. The metabolites were measured in the three body fluids non-fasting blood plasma, urine, and saliva. In the time from February to June 2012, 1107 samples were taken from the participants, comprising 1563 metabolites including amino acids, peptides, carbohydrates and lipids, as well as age, gender, ethnicity, weight, height, Body Mass Index (BMI) and personal history of T2D [17].

The samples were analyzed by the three companies Metabolon Inc., Chenomx Inc., and Biocrates Life Sciences AG. The respective companies utilized liquid/gas chromatography with mass spectrometry injections, targeted profiling using NMR, and Multiple Reaction Monitoring (MRM). The study found that all variables of ethnicity, gender and smoking had a strong effect on a diabetes risk factor, advanced glycation end products. Women, Arabs, Filipinos, and smokers were more strongly affected than men, south Asians, and non or irregular smokers [17].

Statistical Analysis

NMR binned data: When dealing with high resolution NMR spectra it is in general impracticable to work with the entire data points of the spectra which are usually in the order of 32Kb and bigger. The most common strategy used to reduce the number of variables consists in dividing each spectrum in a defined number of regions, the so called bins. Several binning strategies are available today, from regular binning, where bins have fixed width, to more sophisticated strategies such as gaussian or dynamic adaptive binning [8]. Here we used regular binning to preprocess the high resolution data ~ 65536 data points in a single spectrum and remove any anomalies. This was motivated by the fact that when dealing with an array of NMR spectra, whilst regular binning of a number of bins over stacked spectra containing spectra will generate a matrix, it is not possible to generate a similar matrix using directly deconvolved peaks (peak list) since the number and position of peaks varies from spectrum to spectrum. In our case, we have used a binning approach which automates the binning of assembled NMR spectrum using imposed alignment of each spectra. In fact, we had 354 files that contained NMR coordinates. Each file had approximately 65,000 data points. This means that our algorithm had to iterate through 374 x 65,000 = 22,620,000 (22 and 1/2 million) data points. Each bin gives rise to a new value which is representative for the bin. We used a bin interval of 0.007 ppm. Using JAVA, we iterated through all x values in this interval and calculated the mean and standard deviation. After this we considered values inside m ± 3σ and calculated their mean. The obtained matrix after processing the data was of size 348 x 2960.

Sparse PCA with elastic net: Consider the linear regression model with n observations and p predictors. Let $Y = (y_1 ,....,\quad)^T$ be the response vector and $X = [X_1, ..., X_p]$, $j = 1, ..., p$ the predictors, where $Xj = (x_{1j}, ...,x_{n\,j})^T$. After a location transformation we can assume all the Xj and Y are centered. The lasso is a penalized least squares method, imposing a constraint on the ℓ_1 norm of the regression coefficients. Thus, the lasso estimates β_{lasso} are obtained by minimizing the lasso criterion

$$\beta^{\wedge}_{lasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{1}^{p} |\beta_j| \tag{1}$$

where λ is non-negative. The lasso continuously shrinks the coefficients toward zero, and achieves its prediction accuracy via the bias variance trade-off. Due to the nature of the ℓ_1 penalty, some coefficients will be shrunk to exact zero if λ is large enough. The elastic net [12] generalizes the lasso to overcome these drawbacks, while enjoying its other favorable properties. For any non-negative λ_1 and λ_2, the elastic net estimates $\hat{\beta}$ are given as follows:

$$\beta^{\wedge}_{elastic} = (1+\lambda_2)\left\{\arg\min_{\beta} \|Y - \sum_{j=1}^{p} X_j\, \beta_j\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{1}^{p} |\beta_j|\right\}^2 \tag{2}$$

The connection between robust regression method and PCA have been discussed by [11] and the problem becomes equivalent to the following optimization problem

$$\hat{\beta} = \arg\min_{\beta} \|Z_i - X\beta\|^2 + \lambda\|\beta\|^2 + \lambda_1 \|\beta\|_1 \tag{3}$$

where $\|\beta\|_1 = \sum_1^p \beta_j$ is the ℓ_1-norm of $\beta, Z_i = U_i D_{ii}$ the i^{th} principal component. Approximated principal component are given by $X\hat{V}_i$ where $\hat{V}_i = \frac{\hat{\beta}}{\|\beta\|}$ and large enough λ_1 gives a sparse $\hat{\beta}$, hence a sparse $\hat{V}_i$.

Algorithm1 summarizes the steps of SPCA. From an algorithmic point of view, to find the solutions in (3), each of the corresponding optimization problems can be seen as a Lasso problem by introducing new observations and then use Least Angle Regression algorithm (LARS) or coordinate-descent (Gauss-Seidel) algorithm. It is interesting to note that (i) for p $\succeq$ n the augmented data set has p + n observations and p variables, which can slow the computation considerably; (ii) if the original design matrix is normalized, there is no guarantees the augmented design matrix will behave similarly, which can cause a loss of a part of the interpretation of the data; and (iii) the coordinate-descent algorithm proceeds by one at a time philosophy, e.g. it minimizes the loss function of β_j while maintaining components $\beta_{k,}$ k ≠ j fixed at their actual values, in this case we cannot develop Gauss-Seidel for a grouped variable selection problem. To overcome these limitations, we derive a unified alternating direction method of multipliers based algorithm to handle sparse principal component selection which aims at selecting important components and penalizing the others through β [20]. We propose a doubly regularized model with a general penalty term of the form

$$\frac{\mu}{2}\beta^t Q\beta + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|$$

so the flexible elastic equation to minimize, given a fixed A=[α_1,…, α_k], from Algorithm 1 becomes:

$$(\alpha_{j-}\beta)^T X^T X(\alpha_{j-}\beta) + \frac{\mu}{2}\beta^t Q\beta + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \tag{4}$$

where λ , μ ≥ 0 are two tuning parameters, $\omega = (\hat{\omega}_2, ..., \hat{\omega}_{2\,p})^t$ and $Q = (q_{i\,j})_{1=i,\, j=p}$ are weights associated with the ℓ_1 and ℓ_2 norms respectively, which are fixed in advance.

The advantages of our algorithm are: (1) Provide a general frame to deal with the limitations of unweighed versions of lasso-type estimates. A weighted version possesses the oracle properties of selecting the subset of interesting variables with a proper choice of the weights and increasing the number of hits and decreasing the number of false positives. (2) Combine the strengths of Lasso and a quadratic penalty designed to capture additional structure on the features in high dimensional setting which is frequent in high-throughput generated from ^{1}H-NMR spectroscopy. (3) Develop an easy and fast algorithm using the Alternating Direction Method of Multipliers (ADMM) approach to find optimal estimator without augmenting or normalizing data (see next section).

Alternating Direction Method of Multipliers (ADMM): Recently, the alternating direction method of multipliers has been revisited and successfully applied to solving large scale problems arising from different applications. In this section we give an overview of ADMM. Consider the following optimization problem:

minimize f (β) + g(ξ)

subject to β - ξ = 0, (5)

where f and g are two convex functions and β, $\xi \in R^{p.}$ In this optimization problem, we have two sets of variables, with separable objective. The augmented Lagrangian for this problem is:

$$L_\tau(\beta,\xi,\delta) = f(\beta) + g(\xi) + \delta^t(\beta-\xi) + (\tau/2)\|\beta-\xi\|_2^2,$$

where δ is the dual variable for the constraint β-ξ=0 and τ>0 is a penalty parameter. The augmented Lagrangian methods were developed in part to bring robustness to the dual ascent method, and in particular, to yield convergence without strong assumptions like strict convexity or finiteness of *f* and *g*.

At iteration *k*, the ADMM algorithm consists of the three steps:

$$\beta^{k+1} := \arg\min_{\beta} L_t(\beta, \xi^k, \delta^k), \text{ // } \beta\text{- minimization} \tag{6}$$

$$\xi^{k+1} := \arg\min_{\xi} L_\tau(\beta^{k+1}, \xi, \delta^k), \text{ // } \xi \text{ – minimization} \tag{7}$$

$$\delta^{k+1} := \delta^k + \tau(\beta^{k+1} - \xi^{k+1}) \text{ . //dual-update} \tag{8}$$

1. In the first step of the ADMM algorithm, we fix ξ and δ and minimize the augmented Lagrangian over β.

2. In the second step, we fix β and δ and minimize the augmented Lagrangian over ξ.

3. Finally, we update the dual variable δ .

If we consider the scaled dual variable η = (1/τ) and the residual r = η - ξ, the ADMM algorithm can be expressed on its scaled dual form as (we will use the scaled form in the paper):

$$\beta^{k+1} := \arg\min_{\beta} \left\{ f(\beta) + (\tau / 2 \left\| \beta - \xi^k \right\|_2^2 \right\}; \tag{9}$$

$$\xi^{k+1} := \arg\min_{\xi} \left\{ g(\xi) + (\tau / 2 \left\| \beta^{k+1} - \xi + \eta^k \right\|_2^2 \right\}; \tag{10}$$

$$\eta^{k+1} := \eta^k + \beta^{k+1} - \xi^{k+1}. \tag{11}$$

Stopping criteria: The primal and dual residuals at iteration k have the forms:

$$e_{pri}^k = (\beta^k - \xi^k), e^k = -\tau(\eta^k - \eta^{k-1}).$$

The ADMM algorithm terminates when the primal and dual residuals satisfy stopping criterion. A typical stopping criterion is given in [5] where the authors propose to terminate when $\left\| e_{pri}^k \right\| \le \varepsilon^{pri}, \left\| e_{dual}^k \right\| \le \varepsilon^{dual}$. The tolerances ε^{pri} >0 and ε^{dual} >0 can be chosen using an absolute and relative criterion, such as $\varepsilon^{pri} = \sqrt{\ } \varepsilon^{abs} + \varepsilon^{rel} \max\left\{ \left\| \beta^k \right\|_2, \left\| \eta^k \right\|_2 \right\}$ and $\varepsilon^{dual} = \sqrt{p}\varepsilon^{abs} + \varepsilon^{rel} \tau \left\| \eta^k \right\|_2$, where ε^{abs} >0 and ε^{rel} >0 are absolute and relative tolerances. A reasonable value for the relative stopping criterion is $\frac{\mu}{2}\beta^t Q\beta + \lambda \sum_{j=1}^{p} \omega_i \left| \beta_j \right|$ =10^{-3} or 10^{-4}, depends on the scale of the typical variable (see [5] for details).

SPCA with ADMM: In this section we derive an efficient Alternating Direction Method of Multipliers algorithm for an elastic net approach of sparse PCA estimators with a more general penalty term of the form $\frac{\mu}{2}\beta^t Q\beta + \lambda \sum_{j=1}^{p} \omega_i \left| \beta_j \right|$.

To check the importance of a variable, we estimate its coefficient $\hat{\beta}$ solution of the generic problem:

$$\hat{\beta}_{fPCA}(\lambda, \mu) = \arg\min_{\beta} (\alpha_j - \beta)^T X^T X (\alpha_j - \beta) + \frac{\mu}{2} \beta^t Q \beta + \lambda \sum_{j=1}^{p} \omega_j \left| \beta_j \right| \tag{12}$$

or equivalently

$$\hat{\beta}_{fPCA}(\lambda, \mu) = \arg\min_{\beta} \frac{1}{2} \left\| y^{**} - X^{**} \beta \right\|_2^2 + \frac{\mu}{2} \beta^t Q \beta + \lambda \sum_{j=1}^{p} \omega_j \left| \beta_j \right| \tag{13}$$

where λ ,μ are two non negative tuning parameters, Q is a positive semi-definite matrix, $y^{**}=\Sigma^{1/2} \alpha_j$, $X^{**}=\Sigma^{1/2}$ and Σ is the covariance matrix of X.

Equation (13) combines the strengths of regularized techniques of type Lasso and a quadratic penalty designed to capture additional structure on the features. When $\hat{\omega}_j$ = 1, it is straightforward to show that all type of lasso models (Lasso, Enet, Slasso, L1Cp and Wfusion) are particular case of (13) using an augmented data reparameterization of the form

$$X^*_{(n+p)\times p} = \begin{pmatrix} X \\ \sqrt{\mu L^t} \end{pmatrix}; \text{Q=LL}^t; \; y^*_{(n+p)} = \begin{pmatrix} y^{**} \\ 0 \end{pmatrix},$$

Therefore any efficient algorithm developed to find the whole solution path of the Lasso like least angle regression or coordinate descent algorithm can be applied. Unfortunately, the good properties of the two optimization techniques are overshadowed by the difficulties (i), (ii) and (iii). To deal with those problems, we propose to solve (13) using the ADMM algorithm. The idea is simple and straightforward. First, we propose to re-write (13) on the following ADMM form:

$$\frac{1}{2} \left\| y - X\beta \right\|_2^2 + (\mu / 2) \beta^t Q \beta + \lambda \sum_{j=1}^{p} \hat{\omega}_j \left| \xi_j \right|$$

subject to β - ξ = 0. (14)

If we write $f(\beta) = (1/2) \left\| y - X\beta \right\|_2^2 + (\mu/2)\beta^t Q\beta$, g (ξ) = $\lambda \sum_{j=1}^{p} \hat{\omega} \left| \xi_j \right|$ and $\hat{\omega}_j = \left(\left| \hat{\beta}_j \right| + 1/n \right)^{-1}$ then (13) becomes (5). In this case f and g are two convex functions. Apply- ing the ADMM algorithm to (14), we have to perform the following two steps at each iteration:

The β- minimization step.

This step updates β^k by:

$$\beta^{k+1} := \arg\min_{\beta} \left\{ f(\beta) + (\tau/2) \left\| \beta - \xi^k + \eta^k \right\|_2^2 \right\}$$

$$:: +(\tau/2) \left\| \beta - \xi^k + \eta^k \right\|_2^2 \right\}$$

$$: +(\tau/2) \left\| \beta - \xi^k + \eta^k \right\|_2^2 \right\}$$

The ξ - minimization step.

This step updates ξ^k by:

$$\xi^{k+1} := \arg\min_{\xi} \left\{ g(\xi) + (\tau / 2 \left\| \beta^{k+1} - \xi + \eta^k \right\|_2^2 \right\}$$

$$:= \arg\min_{\xi} \left\{ \lambda \sum_{j=1}^{p} \hat{\omega}_j \left| \xi_j \right| + \frac{\tau}{2} \left\| \beta^{k+1} - \xi + \eta^k \right\|_2^2 \right\},$$

We show in the appendix that the solution consists of updating each component ξ_j^k for j = 1,..., p by:

$$\xi_j^{k+1} := sign(\beta_j^{k+1} + \eta_j^k) \max\left(\left| \beta_j^{k+1} + \eta_j^k \right| - \frac{\lambda \hat{\omega}_j}{\tau}, 0 \right)$$

$$:= S_{\frac{\lambda \hat{\omega}_j}{\tau}} \left(\beta_j^{k+1} + \eta_j^k \right),$$

$$\text{where, } S_\kappa(a) = (1 - \kappa / |a|) + a = \begin{cases} a - \kappa \text{ if } a > \kappa \\ 0 \text{ if } |a| \le \kappa \\ a + \kappa \text{ if } a < -\kappa \end{cases}$$

is the soft thresholding function introduced and analyzed by [6]. The dual-update step is straightforward and consists of updating η^k by η^{k+1}:= η^k +β^{k+1} - ξ^{k+1}. It is worth to notice that since τ > 0, μ ≥ 0, X^tX and Q are positive semi-definite matrices, (X^tX+μQ+ τI_p) is always invertible. If p > n, let M = μQ + τI_p, to alleviate the cost of calculations, we can exploit the Woodbury formula for $(X^tX + M)^{-1}$. Algorithm 2 shows the complete details of the flexible elastic-net with ADMM and and Algorithm 3 summarizes the flexible PCA.

Tuning parameters selection: In practice, it is important to select appropriate tuning parameters in order to obtain a good prediction precision and to control the amount of sparsity in the model. Choosing the tuning parameters can be done via minimizing an estimate of the out-of-sample prediction error. If a validation set is available, this can be estimated directly. Lacking a validation set one can use ten-fold cross validation. In our experiments l takes 100 logarithmically equally spaced values, μ ϵ {0,0.1,1,10,100} and γ ϵ {0.5,1,2.5,5,25}.

Results

fPCA, was applied to examine similarities and/or differences in the

^{1}H-NMR spectra. A flexible principal component is a weighted linear combination of each of the original NMR variables so that the original data matrix is compressed into a smaller number of variables; the NMR data may be compressed into three to four fPCs in cases where the changes between groups or due to specific treatments are quite large. Figure 3 shows projection of processed urine samples with uniform 0.007 ppm bin widths on the first and second fPC axes and Figure 4 summarizes the loading scores. From this projection, fPCA analysis shows two perpendicular clustered groups with an overlap in diabetic and non-diabetic samples.

Interestingly, after further analysis, we identified that the overlap summarizes patients with controlled diabetes. Any supervised learning algorithm may lead to invalid results due to huge overlap of diabetic and non-diabetic samples and provide a one cluster summary. Here, flexible principal component 1 (fPC1) provide the maximum variance across diabetic and non-diabetic samples while principal component 2 (fPC2) summarizes maximum variance across samples within diabetic or non-diabetic samples. Sixty metabolites were identified by ^{1}H-NMR spectrum analysis of original data and loading spectrum of fPC1 and fPC2. Twenty four metabolites from major energy sources such as carbohydrates, lipids, and proteins, are identified by NMR spectrum analysis of original data and loading spectrum of fPC1 can be used as potential biomarkers in human lipids for detection of diabetes. In total, 24 metabolites were detected at statistically different concentrations (Table 4).

Many compounds detected at higher levels for T2D were the end products of gluconeogenesis, including glucose and its polymer. Glycine-betaine (betaine) and glutamate, three of the major osmoprotectants used by S. Typhimurium, were found at higher concentrations. Other compounds more abundant

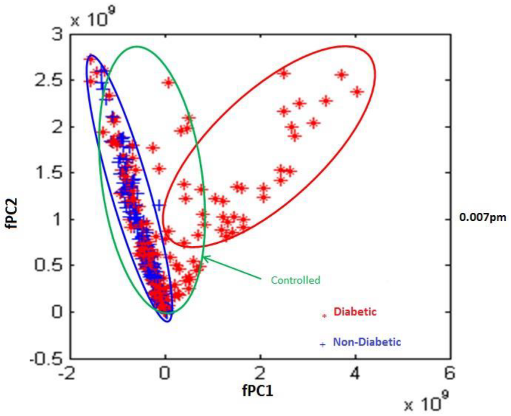

Figure 3: The X axis represents projection on the first flexible principal component and the Y axis represents the orthogonal component. Clustering of the blue stars to the left of the zero line indicates the urine metabolomic diagnostic test is highly sensitive in determining the presence of diabetes disease. (Blue: Non-diabetic, Red: Diabetic)

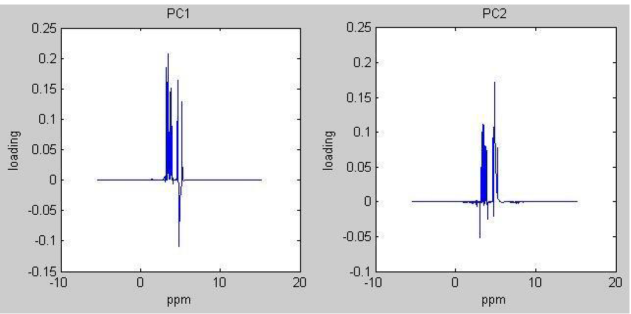

Figure 4: Loading scores for first and second flexible components.

Metabolite Diabetes detected by fPC1	Metabolite Diabetes detected by fPC2
2-Hydroxyisobutyrate	2-Hydroxyisobutyrate
3-Hydroxyisovalerate	-
Acetate	Acetate
Acetone	-
Betaine	Betaine
Creatine	-
Creatinine	-
Dimethylamine	Dimethylamine
Glucose	Glucose
Glycine	Glycine
Glycolate	Glycolate
Hypoxanthine	-
Isopropanol	-
Lactate	-
Maleate	Maleate
Mannitol	Mannitol
Methanol	Methanol
Methylamine	Methylamine
N,N-Dimethylglycine	N,N-Dimethylglycine
Succinate	-
Tartrate	Tartrate
Taurine	Taurine
β-Alanine	-
π -Methylhistidine	-

Table 4: Compounds detected in actual samples. Left column summarizes the ones detected by fPC1 and right column summarizes the ones detected by fPC2. Compound detected with fPC1 are potential biomarkers in human urine for diabetes. Compound detected with fPC2 indicating most variations in the normal and diabetic samples. Compounds in red have been reported abnormal in human urine in HMDB.

Blood metabolite	actual sample	fPC1 (mM)	fPC2 (mM)
1,3-Dihydroxyacetone	X	0.4195	0.1324
1,3-Dimethylurate	X	1.3649	1.1966
1,6-Anhydro-b -D-glucose	X		
1,7-Dimethylxanthine	X	0.1613	0.117
1-Methylnicotinamide	X		
2-Hydroxyisobutyrate	X	0.0761	0.038
2-Oxobutyrate	X	0.0716	0
2-Oxoglutarate	X	0.0631	0
3-Aminoisobutyrate	X		
3-Hydroxyisovalerate	X	0.0221	0
3-Hydroxymandelate	X		
3-Indoxylsulfate	X		
4-Hydroxyphenylacetate	X		
4-Pyridoxate	X	0.2208	0
5,6-Dihydrouracil	X	0.6077	0
Acetate	X	0.0601	0.0072
Acetoin	X	0.0215	0

Acetone	X	0.0518	0
Alanine	X		
Adenine	X	0.1256	0
Arabinitol	X	3.6563	0.1111
Asparagine	X		
Benzoate	X		
Betaine	X	10.4309	9.8544
Butanone	X	0.0261	0
Caffeine	X	0.1369	0.0722
Carnitine	X		
Choline	X		
Citrate	X		
Creatine	X	1.0648	0
Creatine phosphate	X	1.3224	0
Creatinine	X	0.633	0
Cytosine	X	0.0336	0
Dimethylamine	X	0.2206	0.4833
Dimethyl sulfone	X	0.2033	0.1388
Ethanol	X		
Ethanolamine	X		
Ethylene glycol	X		2.2361
Formate	X		
Fumarate	X		0.1383
Galactose	X		
Glucose	X	156.7578	146.6056
Glucuronate	X	6.3225	0
Glycine	X	15.3768	15.733
Glycolate	X	58.6403	52.3104
π -Methylhistidine	X	0.1438	0

Table 5: Compounds detected in actual samples and loadings of fPC1 and fPC2.

in gestational diabetes mellitus were 3-hydroxyisovalerate and 2-hydroxyisobutyrate, probably due to altered biotin status and amino acid and/or gut metabolisms (the latter possibly related to higher BMI values). The major compounds detected at higher levels were the upper TCA cycle intermediates succinate, and the transhydrogenase Lactate/malate, which has dual metabolic functions, named: Delta(1)-piperideine-2-carboxylate/Delta(1)-pyrroline-2-carboxylate reductase, the first member of a novel subclass in a large family of NAD(P)-dependent oxidoreductases [7]. The compounds identified by analysis of loading spectrum of fPC2 are the compounds that indicate most variations in the normal and diabetic blood samples (Table 6). The fPCA was able to cluster all the samples without diabetes. The samples with diabetes had a wide spread. An important observation in this case is the overlap of diabetic and non-diabetic samples. The overlap may be due a result of controlled diabetes of diabetic samples, which resulted into normal concentrations of metabolites compared to the metabolite concentrations for non-diabetic samples. The distribution of samples in fPC1, fPC2 space show that diabetic and non-diabetic samples are distributed along fPC1 and variation among the samples within the diabetic and non-diabetic groups is along fPC2. The spectra for fPC1 and fPC2 are shown in Figure 4. The spectra were processed to compute the concentration of metabolites.

Since the reference compound was present in all the metabolites, it was not included in fPC1 and fPC2. Therefore, the concentration of compounds in fPC1 and fPC2 could not be computed. But, the maximum concentration of a compound was calculated and is shown in Table 5. Moreover, 9 people out of 178 had potential Paraquat poisoning based on their abnormal concentration on citrate, glutinane and alanine and 24 out of 178 had Salicylate (Aspirin) detected in their blood and urine. For Asprin abnormality, we conclude that people with diabetes are more encouraged to take Aspirin as it may reduce risk of heart attack due to coronary obstruction, which is a risk many diabetics concurrunce (Figure 5).

Blood metabolite	Actual sample	fPC1 (mM)	fPC2 (mM)
Guanidoacetate		19.2443	16.7526
Hippurate	X		
Histidine	X		
Histamine		0.1448	0
Hypoxanthine	X	0.1254	0
Isopropanol	X	0.017	0
Lactate	X	0.2972	0
Lysine	X		
Maleate	X	0.2852	0.1811
Mannitol	X	18.0916	0.9322
Methanol	X	1.1467	0.9029
Methylamine	X	0.0838	0.0072
Methylguanidine		0.1761	0.1834
N,N-Dimethylglycine	X	0.028	0.0368
N-Methylhydantoin		0.2124	0
N-Nitrosodimethylamine		0.2762	0
O-Acetylcarnitine	X		
O-Acetylcholine		0.1567	0
O-Phosphocholine	X		
Propionate		0.0089	0
Propylene glycol	X		
Pyroglutamate	X		
Salicylate	X		
Sarcosine		0.2763	0.3998
Serine		0.0698	0
Succinate	X	0.0456	0
Sucrose	X		
Tartrate	X	0.308	0.3615
Taurine	X	41.1061	48.1989
Threonine	X		
Thymine		0.0636	0
Trigonelline	X		
Trimethylamine N-oxide	X		
Tyrosine	X		
Trans-Aconitate		0.1153	0.0789
Trimethylamine		0.0404	0.0392
Trimethylamine N-oxide		6.421	6.0661
Uracil	X		
Urea	X		
Uridine	X		
Valine	X		
Xylose	X		
Cs-Aconitate	X		
Trans-Aconitate	X		
β -Alanine	X	0.4513	0
τ -Methylhistidine			

Table 6: Compounds detected in actual samples and loadings of fPC1and fPC2.

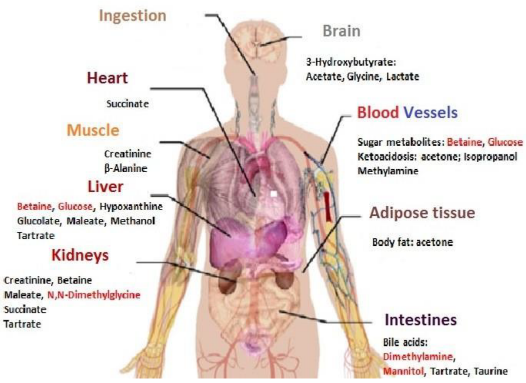

Figure 5: Metabolite biomarkers for Diabetes incident and metabolic classes. Red metabolites are the perturbed metabolites that were found associated with incident diabetes and their link to organ-specific processes and pathways, courtesy of [10]

Conclusion

In this study, we presented an integrative analysis that revealed metabolites correlated with diabetes for a subset of Qatari population and we, furthermore, identified specific metabolites affected by medication, which constraints differentiation between diabetic and control patients. Despite significant advances, no single profiling method currently allows simultaneous analysis of all of the metabolites in the metabolome. Ultimate achievement of our study is to present an integrative statistical method for mining raw ^{1}H NMR data. Challenges appear in handling big data where number of peaks is larger than the number of samples which limited the use of traditional statistical methods. Our next work is the continuation of the development of computational methods for the analysis of complex ^{1}H NMR datasets and their integration with equally complex genomic, transcriptomic, and proteomic profiles as well as metabolome integrated network analysis.

References

1. Bener A, Zirie M, Janahi IM, Al-Hamaq AO, Musallam M, et al. (2009) Prevalence of diagnosed and undiagnosed diabetes mellitus and its risk factors in a population-based study of Qatar. Diabetes Res Clin Pract 84: 99-106.
2. Bruice PY (2011) Organic Chemistry, Prentice Hall,US.
3. Bener A, Alsaied A, Al-Ali M, Hassan AS, Basha B, et al. (2008) Impact of lifestyle and dietary habits on hypovitaminosis D in type 1 diabetes mellitus and healthy children from Qatar, a sun-rich country.Ann Nutr Metab 53: 215-222.
4. Bener A, Zirie M, Al-Rikabi A (2005) Genetics, obesity, and environmental risk factors associated with type 2 diabetes.Croat Med J 46: 302-307.
5. Chu AB, Boyd P, Parikh SN, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3:1-122.
6. Donoho D, Johnstone I (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425455.
7. Muramatsu H, Mihara H, Kakutani R, Yasuda M, Ueda M, et al. (2005) The putative malate/lactate dehydrogenase from Pseudomonas putida is an NADPH-dependent delta1-piperideine-2-carboxylate/delta1-pyrroline-2-carboxylate reductase involved in the catabolism of D-lysine and D-proline. J Biol Chem 280:5329-5335.
8. Hackstadt AJ, Hess AM (2009) Filtering for increased power for microarray data analysis.BMC Bioinformatics 10: 11.
9. Nicholson JK, Lindon JC (2008) Systems biology: Metabonomics.Nature 455: 1054-1056.
10. Suhre K, Meisinger C, Doring A, Altmaier E, Belcredi, et al. (2010) Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS ONE 5:e13953.
11. Zou H, Hastie T, Tibshirani R (2006) Sparse Principal Component Analysis. J Comput Graph Stat15: 262-286.
12. Zou H, Hastie T (2005) Regularization and Variable Selection via the Elastic Net. J Roy Statist Soc Ser 67:301-320.
13. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P (2008) A sparse PLS for variable selection when integrating omics data.Stat Appl Genet Mol Biol 7: Article 35.
14. Chun H, Kele S (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol 72: 3-25.
15. Allen GI, Peterson C, Vannucci M, Maleti Ä, Savati M (2013) Regularized Partial Least Squares with an Application to NMR Spectroscopy.Stat Anal Data Min 6: 302-314.
16. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data.Trends Biotechnol 22: 245-252.
17. Do KT (2013) Metabolomic analysis of multiple body fluids in the qatar metabolomics study of diabetes. Master's thesis, Mnchen, Deutschland: Institute of Computational Biology, Helmholtz Zentrum.
18. Suhre K (2014) Metabolic profiling in diabetes.J Endocrinol 221: R75-85.
19. De Berardis G, Sacco M, Strippoli GF, Pellegrini F, Graziano G, et al. (2009) Aspirin for primary prevention of cardiovascular events in people with diabetes: meta-analysis of randomised controlled trials.BMJ 339: b4531.
20. Anbari ME, Alam S, Bensmail H (2014) COFADMM: A Computational Features Selection with Alternating Direction Method of Multipliers. Procedia Comput Sci 29: 821-830.