Supplementary Material

Fuzzy K Means clustering methodology

Principal component analysis (PCA) is the standard method for visualization of NMR metabolic profiles. This method provides feature separation based on the major variances in the data. A more accurate feature separation based on all the data points can be obtained using clustering methods. Clustering methods can in general be divided into crisp and fuzzy methods. The crisp clustering methods, such as, for example, hierarchical and K-means clustering assign each object (sample) to only one cluster. In fuzzy clustering methods, an indicator variable showing whether an object is a member of a given group/cluster is extended to a weighting factor called membership (*w*). The membership has values between 0 and 1, where membership close to 1 indicates strong association with the cluster and values close to 0 indicate weak or absent association with the cluster. The memberships are calculated for each point and for each cluster. With this approach each point can have significant belonging to multiple clusters, to only one cluster or even to no cluster (defined as membership values equal to 1 divided by number of clusters) thus preventing overfitting. In metabolomics or metabolic profiling the goal of fuzzy clustering of samples is to assign a sample based on its (complete) metabolic signature to a given number of clusters such that any sample can belong to more than one cluster, with a different degree of membership (1).

F-KM is a fuzzy logic extension of the classic, crisp K-Means method (16). For a chosen number of clusters, *c* and dataset matrix *n* x *m*, the F-KM method is used to calculate the *n* x *c* matrix $W = [w_{ik}]$, where w_{ik} is the membership degree of an object (sample, metabolite or spectral bin) i $(i = 1, \dots, n)$ to cluster k $(k = 1, \dots, c)$. The membership values and the centroid positions are calculated from the minimization of the objective function defining the quality of the obtained result. Exact F-KM formalism was described in detail elsewhere (2, 3 and references therein). Briefly, the membership values and the centroid positions are calculated from the objective function $\mathrm{J_m}(W, V)$ defined as:

$$\left(min_{W,V}\right)\mathrm{J_m}(W,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} w_{ik}^{m}\, dist(x_i, v_k)$$

where *W* is the matrix containing membership values; *m* is the fuzziness parameter that regulates the degree of fuzziness in the clustering process; $V=[v_k]$ is a matrix of centroids *i.e.* positions of cluster centres; $X=[x_i]$ is the matrix of point profiles and $dist(x_i, v_k)$ is a measure of distance between data point

and centroid. A range of different distance measures can be applied as part of F-KM. For the datasets analysed in this work, Euclidian distances resulted in the most accurate clustering result and were used. The degree of fuzziness in the clustering process is regulated by the fuzziness parameter, *m*, with *m*=1 giving crisp clustering and with an increasing fuzziness of the result with *m* increasing. Ultimately, at some data dependent value of *m* membership values for all features and for all clusters become $w_{ik}=1/c$ for all *i* and *k*A previously devised empirical rule about the optimal *m* parameter (2) suggests that an optimal *m* value should lead to: a) the median of the top membership values being ≥ 0.5 (prevents the results from being overly fuzzy) and, b) the median of all membership values being ≥ 0 (prevents the results from becoming crisp). The analysis has shown that for these datasets an optimal value of *m* is 2, which is in agreement with the value originally suggested by Bezadek (3) for the general application of F-KM. The F-KM routine from Matlab Central was used for all calculations.

Significant Analysis of Microarrays (SAM)

The SAM method was first developed for the determination of major significantly different gene expression features between different samples based on microarray measurements (4). SAM is, however, a generic statistical tool for major feature selection that can be utilized for different problems including NMR data analysis. The algorithm behind SAM is a robust permutation-based method that relies on variance information present in measurements obtained from all probes of a high throughput measurement. The idea behind SAM was that fluctuations of the high-throughput measurements were feature specific rather than sample specific. In order to account for this, Tusher and co-workers (REF) defined a statistic based on the ratio of change in the feature (*i.e.* gene expression or NMR spectral point) to standard deviation in the data for that feature. To avoid the small variance problem of the T-test, SAM uses a statistic similar to the T-statistic but with an added correction for low value measurements. The "relative difference" $d(i)$ in feature value (gene expression or metabolite concentration or NMR spectral point) is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are defined as the average levels of expression for feature (i) in states *I* and *U*, respectively. The "feature specific scatter" $s(i)$ is the standard deviation of repeated feature measurements:

$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\right\}}$$

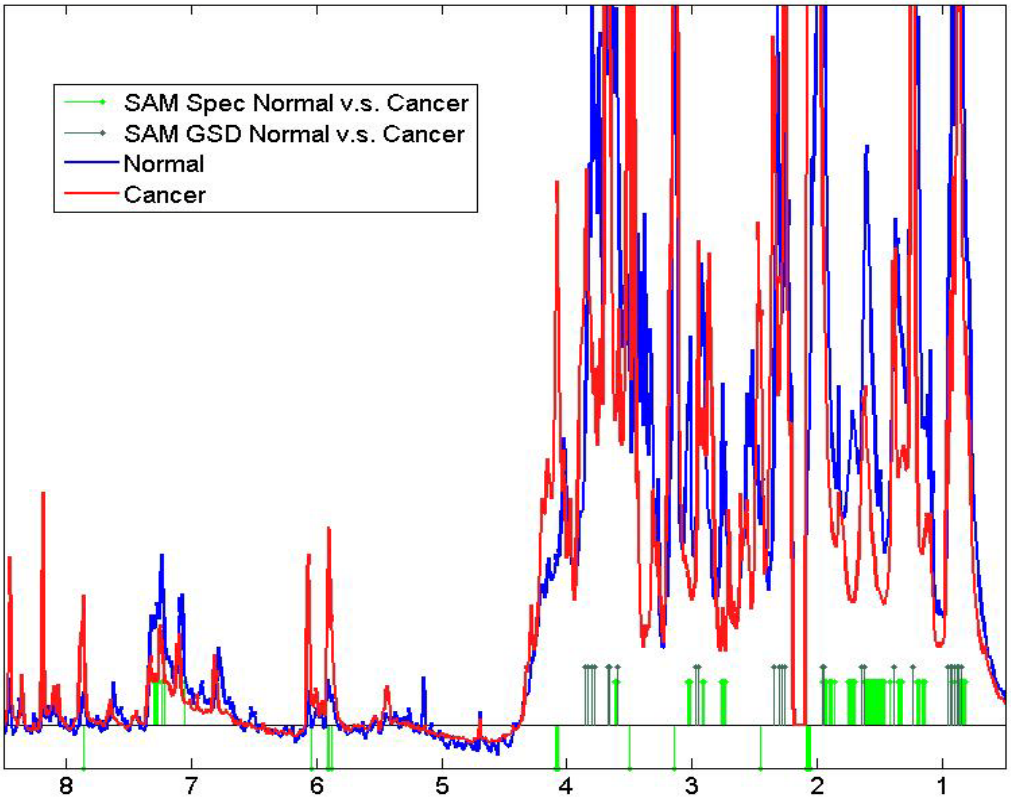
where summations are over the measurements in states *I* (*m*) and *U* (*n*) and $a = (\frac{1}{n_1} + \frac{1}{n_2})/(n_1 + n_2 - 2)$ and n_1 and n_2 are the numbers of measurements in states *I* and *U*.

In the SAM procedure these measures are calculated for each feature in the permutation data. SAM uses a full permutation strategy, sampling across all features and conditions to generate a null distribution. Then, in a permutation test, some elements of the data are permuted (shuffled) to create multiple new pseudo-data sets. The user evaluates, by setting a threshold or visually from a graphical presentation, whether a statistic quantifying departure from the null hypothesis is greater in the observed data than a large proportion of the corresponding statistics calculated on the multiple pseudo-data sets. In this way major differential features between two or more groups of samples can be obtained.

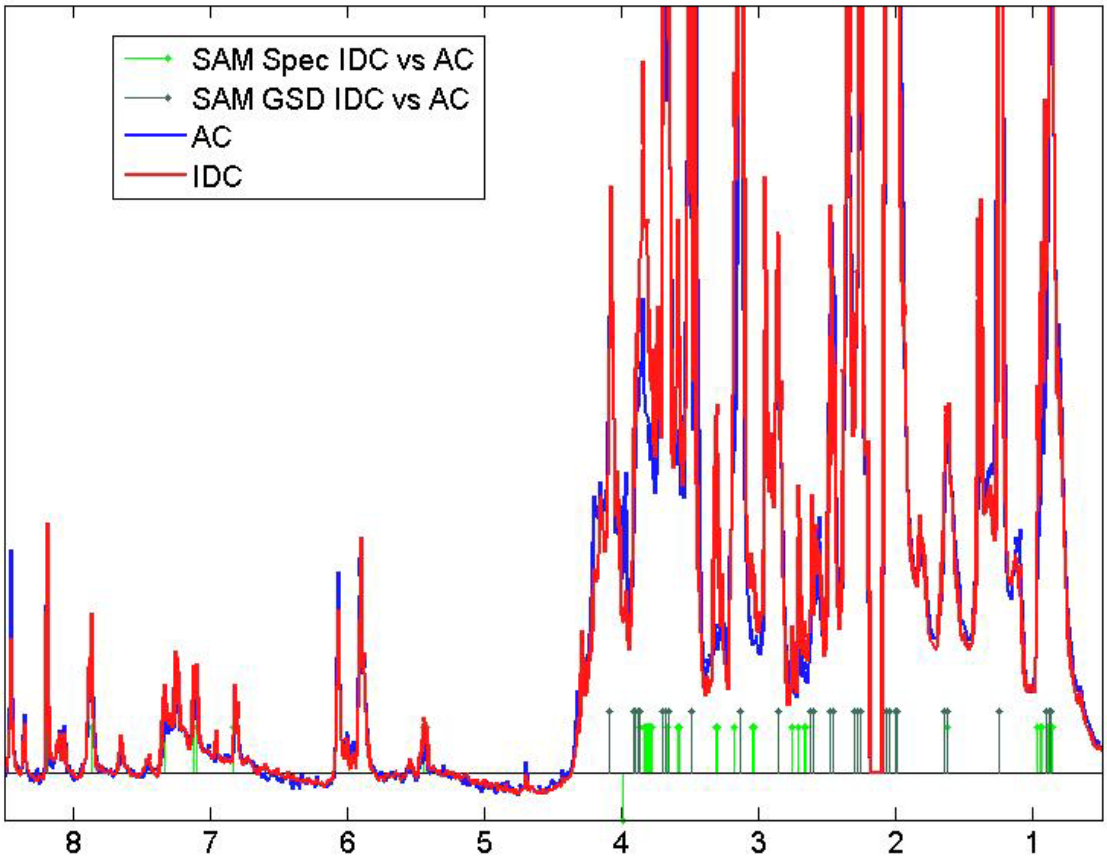
1. M Cuperlovic-Culf, N Belacel, AS Culf, IC Chute, RJ Ouellette, IW Burton, TK Karakach, JA Walter. *Magn Res Chem* 2009, **47**, S96.
2. N. Belacel, , M. Cuperlovic-Culf, , M. Laflamme, , R.J. Ouellette *Bioinformatics*, 2004, **20**, 1-12
3. J.C. Bezdek (1981) Pattern Recognition with Fuzzy Objective Function Algoritms, Plenum Press, New York
4. VG Tusher, R Tibshirani, G.Chu *Proc Amer Acad Sci* 2001, **98**: 5116-5121.

Supplementary Figures

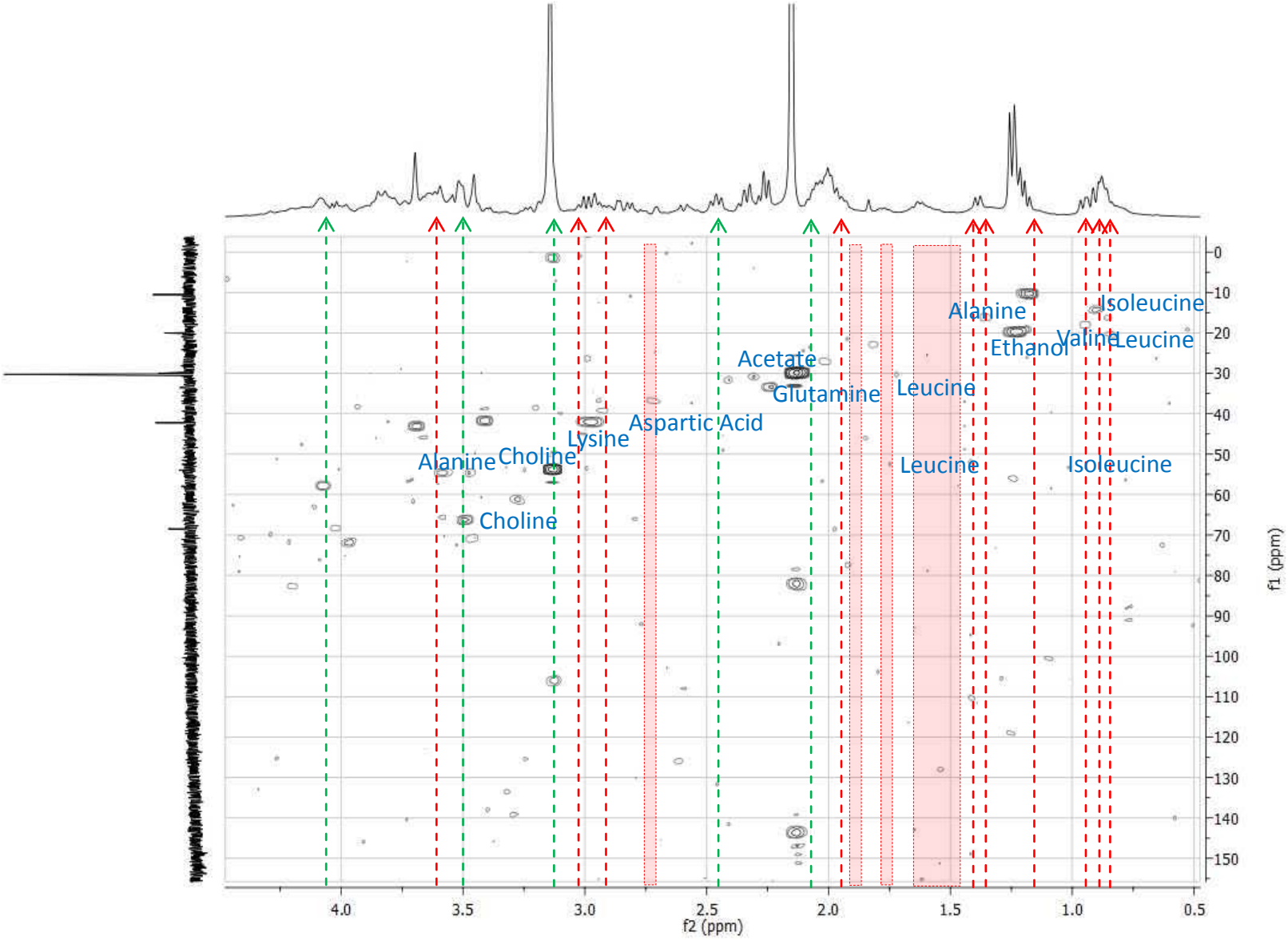
Supp. Figure 1. Major differential features between cancer and normal cell types determined with SAM. Indicated are SAM features obtained for spectral data (light green) and peaks determined using global spectra deconvolution method (GSD) (olive green).



Supp. Figure 2. SAM determined major differential features between ER+ invasive ductal carcinoma (IDC) and ER- adenocarcinoma (AC) cell types. Indicated are SAM features obtained for spectral data (light green) and GSD data (olive green).



Supp. Figure 3. Major differential features between cancer and normal cell types determined with SAM. Indicated are SAM-obtained significantly higher intensity features in cancer cell lines (green) and in normal cell lines (red). The major features are shown with a HMQC experiment background.



Supp. Figure 4. Major differential features between IDC and AC cell types determined with SAM. Indicated are significantly higher intensity features in IDC cell lines (blue) and in AC cell lines (orange). The major features are shown with a HMQC experiment background.