NMR metabolic analysis of samples using fuzzy K-means clustering

Cuperlovi´c-Culf, Miroslava; Belacel, Nabil; Culf, Adrian S.; Chute, Ian C.; Ouellette, Rodney J.; Burton, Ian W.; Karakach, Tobias K.; Walter, John A.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1002/mrc.2502
Magnetic Resonance in Chemistry, 47, pp. S96-S104, 2009

Research Article

Magnetic Resonance in Chemistry

Received: 27 October 2008 Revised: 22 February 2009 Accepted: 22 July 2009 Published online in Wiley Interscience: 4 September 2009

(www.interscience.com) DOI 10.1002/mrc.2502

NMR metabolic analysis of samples using fuzzy K-means clustering

Miroslava Čuperlović-Culf,[a]* Nabil Belacel,[a] Adrian S. Culf,[b] Ian C. Chute,[b] Rodney J. Ouellette,[b] Ian W. Burton,[c] Tobias K. Karakach[c] and John A. Walter[c]

The global analysis of metabolites can be used to define the phenotypes of cells, tissues or organisms. Classifying groups of samples based on their metabolic profile is one of the main topics of metabolomics research. Crisp clustering methods assign each feature to one cluster, thereby omitting information about the multiplicity of sample subtypes. Here, we present the application of fuzzy K-means clustering method for the classification of samples based on metabolomics 1D ^{1}H NMR fingerprints. The sample classification was performed on NMR spectra of cancer cell line extracts and of urine samples of type 2 diabetes patients and animal models. The cell line dataset included NMR spectra of lipophilic cell extracts for two normal and three cancer cell lines with cancer cell lines including two invasive and one non-invasive cancers. The second dataset included previously published NMR spectra of urine samples of human type 2 diabetics and healthy controls, mouse wild type and diabetes model and rat obese and lean phenotypes. The fuzzy K-means clustering method allowed more accurate sample classification in both datasets relative to the other tested methods including principal component analysis (PCA), hierarchical clustering (HCL) and K-means clustering. In the cell line samples, fuzzy clustering provided a clear separation of individual cell lines, groups of cancer and normal cell lines as well as non-invasive and invasive tumour cell lines. In the diabetes dataset, clear separation of healthy controls and diabetics in all three models was possible only by using the fuzzy clustering method.

Keywords: fuzzy clustering; sample classification; metabolomics; metabolic profiling; mixture analysis; sample subtypes; ^{1}H NMR; phenotype analysis

Introduction

Functional genomics and systems biology utilise a range of high-throughput molecular methods ('omics') in conjunction with bioinformatics and computational biology in order to provide a new framework for elucidation of disease aetiology. These methods are also attempting to uncover latent connections between seemingly disparate disease states through holistic analysis.[1–4] In this context, an interest in high-throughput analysis of metabolites has grown considerably as the metabolome represents the most direct reflection of the cell state, in contrast to proteomics and transcriptomics in which regulatory effects hamper clear interpretation of the results.[5,6] Measurement of small-molecule metabolites, either endogenous or exogenous, provides a chemical fingerprint of an organism's metabolic state.[1,3,7,8] The results of metabolomic analysis can be used either for sample type determination or for the analysis of metabolite properties. In both of these applications, classification of data is an essential step. In sample analysis, classification is performed in order to (i) determine whether the studied data contain sufficient information to make a distinction between pre-assigned sample types and/or (ii) to determine new sample classes or new relationships between sample classes. In metabolite analysis, classification is performed in order to determine relationships between metabolites – involvement in the same or co-regulated pathways and possible functions of unknown metabolites through the analysis of their clustering partners.[9,10] Only a handful of different clustering and visualisation methods have been used thus far in metabolomics data analysis. Methods for dimensionality reduction are still the most popular in the analysis of samples, although their application has recently been under increasing criticism[11,12] as they only show major trends in data.

Several authors have previously presented the application of some basic clustering tools. Hierarchical clustering (HCL) was shown to be effective for the determination of structurally related metabolites derived from the same biochemical precursors.[13] An application of HCL in sample analysis has also been demonstrated.[7] Work by Hageman *et al*.[6] introduced K-means clustering and bootstrap aggregation to the analysis of metabolite information obtained from a high-throughput analysis. This has shown that K-means clustering with bootstrap aggregation is very robust and highly appropriate for metabolite classification. Several publications presented the application of self-Organised maps

* *Correspondence to: Miroslava Čuperlović-Culf, Institute for Information Technology, National Research Council of Canada, 55 Crowley Farm Road, Suite 1100, Moncton, NB E1A 7R1, Canada.*
E-mail: miroslava.cuperlovic-culf@nrc-cnrc.gc.ca

a *Institute for Information Technology, National Research Council of Canada, Moncton, NB, Canada*

b *Atlantic Cancer Research Institute, Moncton, NB, Canada*

c *Institute for Marine Biosciences, National Research Council of Canada, Halifax, NS, B3H 3Z1, Canada*

(SOM) for sample classification with excellent results, particularly in the analysis of serum samples in clinical applications.[14]

All of these methods are crisp (hard) clustering approaches that are based on the assumption that each data object should be assigned to only one cluster. The restriction of this one-to-one mapping might not be optimal, especially in the analysis of biological data. The adaptability of cells and the diversity in cellular responses to various internal and external stimuli are accomplished through the cooperation and multi-functional properties of a limited number of biological molecules[15–17]. Furthermore, differences between biological samples are often not explicit due to different phenotypical influences.[2] Fuzzy clustering methods allow data objects to be assigned to multiple clusters. The result of fuzzy clustering calculation is the matrix of membership degrees that describe the level of similarity between each feature and each cluster centroid. Several different fuzzy clustering methods utilising different approaches for the calculation of centroids as well as membership values were developed including fuzzy K-means (F-KM),[16,17] fuzzy J-means[15] and fuzzy SOM.[18] These and other fuzzy methods were previously utilised in the analysis of transcriptomics data with excellent results.[15,17] The advantage of fuzzy classification becomes particularly apparent in the analysis of overlapping groups of objects as well as subgroups in conjunction with the separation of major groups.

In the following, we present two applications of the F-KM method for the classification of breast cell line metabolic profiles and metabolic profiles of urine samples in diabetic patients and appropriate animal models. F-KM is a fuzzy version of standard K-means clustering. In F-KM clustering, each point (in this case sample) has an overall membership, i.e. sum of membership values for all clusters, of 1. This overall membership is apportioned to clusters based on the similarity between the point's (in this case samples) profile (here metabololic fingerprint) and the profile of cluster's centroid. From the membership values, it is then possible to determine different levels of co-clustering between points – based on the top membership, second highest membership etc.

In this work, different clustering methods were tested and compared with F-KM for their ability to separate major classes – in the first dataset – breast cancer and normal cell lines and for second dataset – NMR of urine samples of three species. The methods were further challenged to separate sample subtypes – in cancer cell lines, invasive and non-invasive tumours and in urine samples healthy and diabetic subjects. Fuzzy classification was the only method tested that allowed distinction on both major groups and sample subtypes.

Experimental

Experimental set

All cell lines were obtained from ATCC (Manassas, VA, USA) and cultured as monolayers to 75–85% confluency in T175 cm^2 flasks (Corning) at 37 °C. All media and components were purchased from Invitrogen unless otherwise noted. MCF10a and MCF12a cells were grown in Dulbecco's modified Eagle's medium/Ham's F12 (1 : 1, v/v) supplemented with 2 mM L-glutamine, 1 mM sodium pyruvate, 20 ng/ml epidermal growth factor (Sigma Aldrich), 100 ng/ml cholera toxin (Sigma Aldrich), 0.01 mg/ml bovine insulin (Sigma Aldrich), 500 ng/ml hydrocortisone (Sigma Aldrich), 5% fetal bovine serum (Hyclone) and penicillin/streptomycin (100U/ml and 100µg/ml, respectively). MCF7 cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (Hyclone), 2 mM L-glutamine and penicillin/streptomycin. MDA-MB231 and MDA-MB468 cells were grown in Leibovitz's L-15 medium supplemented with 2 mM L-glutamine and 10% fetal bovine serum (Hyclone).

Cells were washed twice with phosphate-buffered saline (PBS) and then harvested by trypsinisation. Recovered cells were re-suspended in PBS, counted using a haemacytometer, then aliquoted into microcentrifuge tubes as five replicates. The cells were then pelleted in polypropylene microcentrifuge tubes by spinning at 800× g for 2 min and the PBS removed. The number of cells per aliquot/replicate is outlined in Table 1 together with general characteristics of these cell lines.

Lipids were extracted using a modified version of the method described in Ref. [21]. Cell pellets were re-suspended in 500-µl ice-cold 2 : 1 methanol/chloroform mix (v/v), vortexed and then agitated on a shaking platform for 10 min at 4 °C. Two hundred and fifty micro litres of ice-cold chloroform was added and the samples were vortexed. Two hundred and fifty micro litres of ice-cold water was then added and the samples vortexed again. An ultrasonic bath (FS110H, Fisher Scientific) was used to sonicate the samples for 10 min at 25 °C. The samples were then spun in a microcentrifuge at 15 000× g for 5 min at 4 °C. The bottom (chloroform) layers were removed carefully, avoiding pelleted debris using glass Pasteur pipettes and placed in separate polypropylene microcentrifuge tubes. Solvents were evaporated by spinning the samples in a SPD111V SpeedVac concentrator (Thermo). Dried pellets were stored at −80 °C in the dark awaiting subsequent NMR analysis.

^{1}H NMR spectroscopy of lipophilic cell extracts

All ^{1}H NMR experiments were performed on a Bruker DRX-500 spectrometer at 500.13 MHz at 20 °C. Dried samples supplied

Table 1. General characteristics of the selected breast cell lines

Cell line	Type	Characteristics	Morphology	Cell count per aliquot (million); <10% error
MCF12a	Normal[a]	Adherent cells	Epithelial	3
MCF10a	Normal[a]	Adherent cells	Epithelial	10
MB231	Adenocarcinoma	Invasive, metastatic	Epithelial	8
MB468	Adenocarcinoma	Invasive, metastatic	Epithelial	10
MCF7	Adenocarcinoma	Non-invasive	Epithelial	5

The information is obtained from Refs [19,20].
[a] Immortalised normal breast epithelial cells from diploid human breast epithelial cells of two different patients: line MCF12a derived from healthy patients and MCF10a derived from patients with fibrocystic disease.

in vials kept on dry ice were dissolved in 750 µl $CDCl_3$ and placed in Wilmad 535pp 5-mm glass NMR tubes. ^{1}H spectra were obtained with a 5-mm triple-band inverse (TBI) triple-axis gradient probe, tuned and matched for each sample, using a sequence consisting of a 90° pulse (5.7 µs) followed by a 4.365-s acquisition time (AQ) and 2-s relaxation delay (D1), accumulating 128 scans after eight dummy (unrecorded) scans. Spectral width was 15.011 ppm (7507.5 Hz). Receiver gain RG was maintained constant for all spectra. Spectra were processed with MestRe-C software (Mestrelab Research) using an exponential window function with line broadening 0.3 Hz. Phased spectra were referred to the residual $CHCl_3$ peak at 7.26 ppm. The baseline correction was performed by Whittaker smoothing and subtracting (provided in MestRe-C software).

All spectral data were binned using a 0.005 ppm bin size. Only data in the range [7–0] ppm were included in the study. The interval 1.95–1.52 ppm was excluded from the analysis as recommended by Gottschalk *et al*.[22] due to the intensity and chemical shift variation of the signal from residual water. The approximate concentrations of major detectable metabolites were measured from NMR peak integrals determined as direct sums of data points in the spectral range. The integrals were obtained using MestRe-C software (Mestrelab Research).

Type II diabetes dataset

Type II diabetes dataset was previously measured and described by Salek *et al*.[23] The data included 270 ^{1}H NMR spectra of urine samples measured using 1D NOESY pulse sequence with water pre-saturation. Measurements were performed at three different instruments with three different magnetic field strengths. Urine samples were collected for wild type and db/db homozygote and heterozygote (with marked diabetic phenotype) male and female mice; obese Zucker (fa/fa) and lean (fa/+) rats and healthy controls and diabetic patient human subjects. The measurements were performed at field strength for ^{1}H of 400 MHz for mouse, 600 MHz for rat and 700 MHz for human samples. Details of the experimental procedure and results are provided in the original publications.[23]

Fuzzy K-means clustering methodology

The crisp clustering methods 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 membership values are calculated for each point with different membership value calculated for each cluster. With this approach, each point can possibly have a significant belonging to multiple clusters, to only one cluster and even to no cluster (if all membership values for the point are equal to one/number of clusters), thus preventing over-fitting. In other words, the goal of fuzzy clustering of samples is to assign a sample based on its 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.

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 $n \times m$, the F-KM method is used to calculate the $n \times c$ matrix $W = [w_{ik}]$, where w_{ik} is the membership degree of an object (sample, metabolite or spectral bin) $i (i = 1, \ldots, n)$ to cluster $k (k = 1, \ldots, c)$. The membership values and the centroid positions are calculated from the minimisation of the objective function defining the quality of the obtained result. The exact F-KM formalism is described in detail elsewhere([15,16] and references therein). Briefly, the membership values and the centroid positions are calculated from the objective function $J_m(W, V)$ defined as

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

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 $\mathrm{dist}(x_i, v_k)$ 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, absolute value distances resulted in the most accurate clustering result and were used. Cluster validity was measured using Rand indexes, R_i (see below). For the NMR spectral profiles of cell lines clustered using F-KM into five sample clusters: $R_i = 0.80$ for absolute value distances; $R_i = 0.74$ for Euclidian distance matrix and $R_i = 0.75$ for cosine dissimilarity matrix. The degree of fuzziness in the clustering process is regulated by the fuzziness parameter, m, with $m = 1$ giving the crisp clustering and with an increasing fuzziness of the result with m increasing ultimately leading to clustering result for all points being $w_{ik} = 1/c$ for all i and k. A previously devised empirical rule about the optimal m parameter[15,24] suggests that an optimal m value should lead to (i) the median of the top membership values being $\geq$0.5 (prevents the results from being overly fuzzy) and (ii) the median of all membership values being $\geq$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[16] for the general application of F-KM. For breast cell line dataset, the median of all membership values was 0.063 and the median of top membership values was 0.894. For type II diabetes dataset, the median of all membership values was once again 0.063 and the median of top membership values was 0.73.

The implementation of F-KM in Partek Genomics Suite (Partek Inc.) was used for the calculations. The F-KM algorithm is freely available as a Matlab routine from Matlab Central.

Quality assessment

Standard external measures, Rand and Jaccard Coefficients, were utilised for the assessment of cluster (U) quality in comparison to the known class labels (P). The coefficients are calculated as

$$\mathrm{Rand} = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

$$\mathrm{Jaccard} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

where n_{11} is the positive agreement term, which represents number of object pairs having the same cluster and the same class, that is, $U_{ij} = 1$, $P_{ij} = 1$; n_{10} is the number of object pairs having the same cluster but a different class, that is, $U_{ij} = 1$, $P_{ij} = 0$; n_{01} is the number of object pairs having a different cluster but

Magnetic Resonance in Chemistry

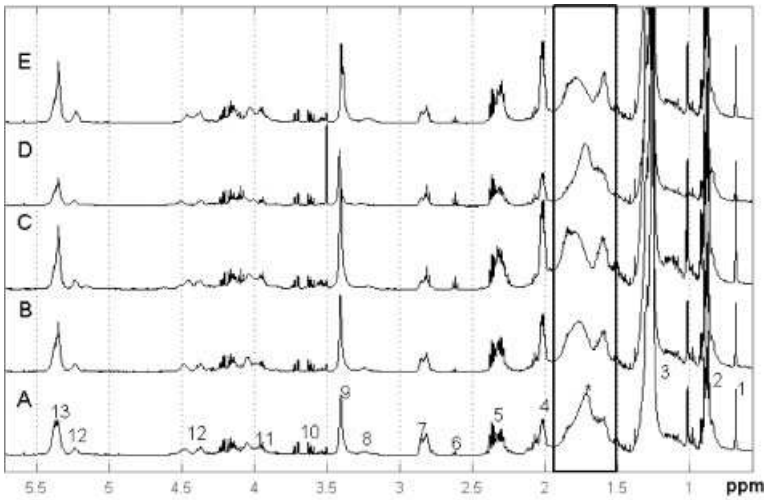

Figure 1. 1D ^{1}H NMR spectra (500 MHz, solvent $CDCl_3$) of the lipophilic fraction for five breast cell lines. The represented spectra are the averages for five replicates of: A-MDA-MB231; B-MDA-MB468; C-MCF10A; D-MCF12A and E-MCF7. The metabolite assignments are given in Table 2.

the same class, that is, $U_{ij} = 0$, $P_{ij} = 1$ and n_{00} is the negative agreement term, i.e. number of object pairs having a different cluster and a different class, that is, $U_{ij} = 0$, $P_{ij} = 0$. The *Rand index* assesses the degree of similarity between U and P via a function of positive and negative agreements in the binary matrices, while Jaccard ignores the negative agreement term.

Results

NMR spectra of five biological replicates of five types of human breast cell lines – MCF10a, MCF12a, MCF7, MB231 and MB468 (Table 1) were recorded and analysed. Figure 1 shows ^{1}H NMR spectra of the analysed cell lines. Each spectrum represents the average of five biological replicates of the lipophilic fractions. The mean variance of spectra across five replicates was less than 1% for all tested cell lines. The suggested peak assignments based on the published data[18,22,25,26] as well as the analysis of metabolic spectral databases[27–29] are included in Fig. 1 and are defined in Table 2. In addition to the proposed assignments, Table 2 includes the proton chemical shift ranges for the peaks. Further, Table 2 includes mean values across cell replicates of normalised peak integral intensities for each of the five cell lines. Corresponding standard deviations are also provided. The concentration values are presented here only as estimates obtained from peak integrals and are not meant for detailed metabolite analysis. We have excluded the interval 1.52–1.95 ppm from further study due to the intensity and chemical shift variation of residual water signal.[22]

The main resonances observed were from saturated and unsaturated lipids, cholines (choline, and overlapping peaks for phosphocholine and glycerophosphocholine) and cholesterol. Most of the metabolites observed in these measurements are in agreement with previously detected metabolites in other cell line types as well as breast cancer tissues.[22,25]

Cluster analysis of metabolic profiles

Four different classification methods were tested on the breast cancer cell line dataset and on the previously published urine metabolomics data for diabetes analysis.[23] The methods presented in this work included one visualisation method, two crisp clustering methods as well as fuzzy clustering method. Principle component analysis (PCA) is one of the major visualisation methods used in metabolomics analysis. Hierarchical clustering method allows automatic sample separation – without the need for user-defined number of clusters. The other crisp clustering method tested was K-means clustering. This is a standard, highly popular method for separation of objects into user-defined number of clusters. Finally, sample classification was performed using the fuzzy version of K-means clustering – fuzzy K-means (F-KM). The results of PCA and HCL classification for breast cancer cell lines are shown in Fig. 2 and for type 2 diabetes dataset in Fig. 3.

The result of K-means clustering is presented in Fig. 4. K-means method assigns each sample to one cluster with five clusters in breast cell lines dataset (five cell line types) and six clusters in the diabetes dataset (three species with wild type and diabetic model in each).

K-means analysis also allows direct calculation of cluster quality coefficients. In the datasets studied, we had predetermined sample class labels, phenotypes, and thus the cluster quality was calculated using external indices Rand and Jaccard for both studied datasets (Table 3).

F-KM clustering facilitates the identification of subclasses of objects by allowing the objects to belong to more than one group. F-KM clustering was performed using a fuzziness parameter, $m = 2$, and an absolute value distance matrix with five clusters for breast cell line samples and six groups for type 2 diabetes dataset. Direct analysis of membership values (Fig. 5(A)) had shown that normal cell lines (MCF10A and MCF12A) as well as the non-invasive tumour cell line (MCF7) can be easily separated on the basis of the top membership values. The two invasive cell lines (MB231 and MB468) are co-clustered on the basis of the top membership value (for both cell lines, top membership values are for cluster 3); however, the second highest membership values are different for the two invasive cell lines (the second highest membership is to cluster 4 for MB231 and cluster 5 for MB468 cell line). This result shows a similarity between the two invasive cell lines when compared with the normal and non-invasive cancer cell lines. However, further analysis of fuzzy membership values shows that there are observable differences between these two cell lines. Major trends in the top membership values can be

Table 2. Assignment of major peaks in the spectra

	Assignment	Chemical shift (ppm)	MCF10A (n = 5)	MCF12A (n = 5)	MCF7 (n = 5)	MB231 (n = 5)	MB468 (n = 5)
1	C18/C19 cholesterol CH_3	0.65–0.7	0.017	0.015	0.011	0.016	0.016
			(0.001)	(0.007)	(0.004)	(0.004)	(0.002)
2	Triglyceride terminal CH_3	0.79–0.94	0.135	0.150	0.125	0.135	0.136
			(0.007)	(0.009)	(0.009)	(0.007)	(0.006)
3	Lipid, cholesterol $(CH_2)_n$	1.2–1.4	0.6	0.612	0.59	0.580	0.58
			(0.007)	(0.009)	(0.01)	(0.008)	(0.01)
[a]	Lipid CH_2CH_2COO, water	1.5–1.95					
4	Lipid $CH_2CH{=}CH$	1.97–2.1	0.064	0.046	0.068	0.055	0.062
			(0.005)	(0.004)	(0.004)	(0.006)	(0.003)
5	Lipid CH_2COO	2.2–2.4	0.049	0.045	0.054	0.047	0.049
			(0.002)	(0.005)	(0.003)	(0.004)	(0.003)
6	L-methionine[a]	2.58–2.64	0.001	0.002	0.001	0.000	0.000
			(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
7	Lipid $C{=}CHCH_2CH{=}C$	2.75–2.88	0.01	0.017	0.015	0.026	0.019
			(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
8	Lipid CH_2NH_3+	3.2–3.35	0.003	0.004	0.0061	0.0053	0.006
			(0.0003)	(0.001)	(0.0005)	(0.0002)	(0.001)
9	Lipid $N+(CH_3)_3$	3.35–3.45	0.0470	0.0362	0.043	0.0412	0.041
			(0.0008)	(0.0009)	(0.003)	(0.0009)	(0.001)
10	Glycerol or derivatives	3.45–3.75	0.0066	0.01	0.009	0.010	0.008
			(0.0007)	(0.01)	(0.002)	(0.003)	(0.002)
11	Lipid CH_2OPO_2-	3.88–3.98	0.0283	0.017	0.032	0.028	0.028
			(0.0009)	(0.009)	(0.002)	(0.004)	(0.007)
12	Lipid CH_2OCOR	5.18–5.30	0.0053	0.0062	0.008	0.0068	0.008
			(0.0001)	(0.0002)	(0.0003)	(0.0003)	(0.0005)
13	Lipid $CH{=}CH$	5.30–5.45	0.034	0.0337	0.041	0.049	0.045
			(0.001)	(0.0009)	(0.002)	(0.001)	(0.001)

Chemical shift range is included. The peak positions are shown in Fig. 1. Concentration of all metabolites estimated from peak integrals are given for all cell lines. The mean values across replicates normalised to total spectral intensity determined as sum of all integrals are shown, with calculated standard deviations of five replicates shown in parentheses.
[a] Large standard deviation relative to the intensity indicates that L-methionine comes as an impurity in the lipophilic fraction.

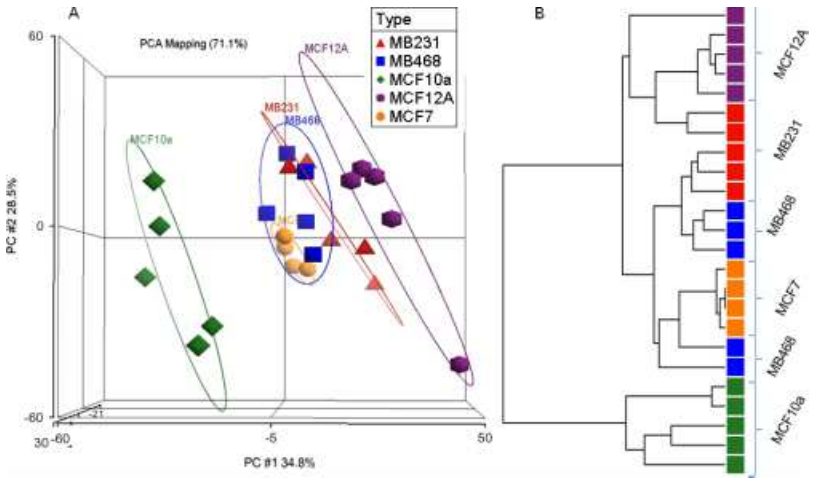

Figure 2. A. PCA of ^{1}H NMR spectral profiles of metabolites for the replicates of five cell lines. Each data point represents a sample with replicates for each cell type and total of 24 samples (outlier replicate of MCF7 cell line was removed from this analysis). The total variation explained by PCs 1, 2 and 3 is 71.1%. The dispersion matrix is calculated using covariance. Ellipses represent two standard deviation from the centroid for each cell line group. B. Hierarchical clustering result for the 24 cell lines samples. The clustering was performed using Euclidian distance calculation. Colour version of the figure is available as Supplementary Material on MRC Web site. Coding is given in the legend.

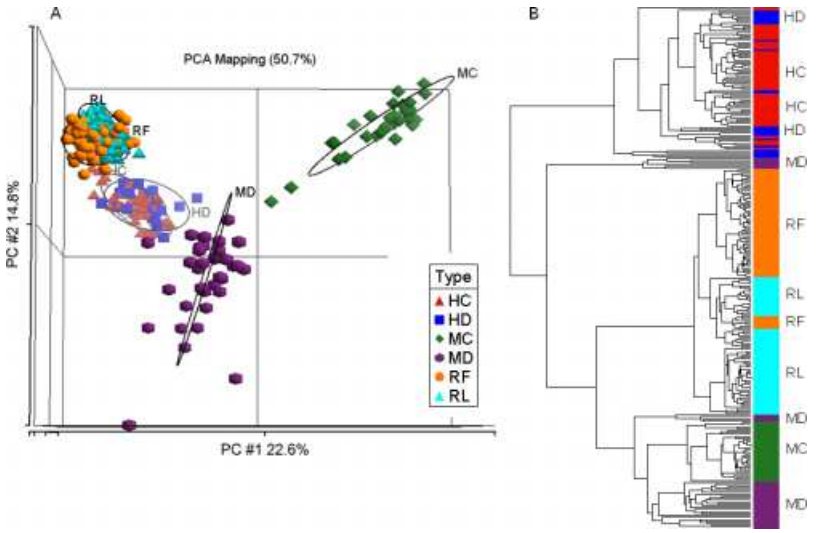

Figure 3. A. PCA of ^{1}H NMR spectral profiles of metabolites for the urine samples for rat, mouse and human healthy and diabetic models. Each data point represents a sample. The total variation explained by PCs 1, 2 and 3 is 50.7%. The dispersion matrix is calculated using covariance. Ellipses represent two standard deviation from the centroid for each cell line group. B. Hierarchical clustering result for the 270 urine samples. The clustering was performed using Euclidian distance calculation. Colour version of the figure is available as Supplementary Material on MRC Web site. Coding is given in the legend.

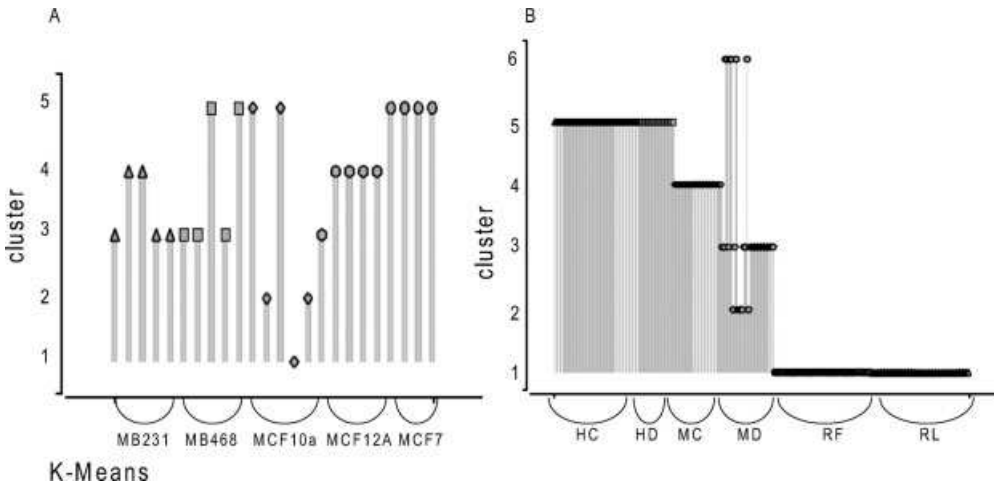

Figure 4. K-means clustering of (A) ^{1}H NMR spectral profiles of metabolites for the replicates of five cell lines clustered into five clusters and (B) ^{1}H NMR spectral profiles of metabolites for the urine samples for rat, mouse and human healthy control and diabetes models clustered into six clusters. The clusters were calculated using absolute value distance matrix.

observed by PCA of membership values (Fig. 5(B)), i.e. by PCA of the output of F-KM analysis. PCA scores plot of membership values provides clear and easy visualisation of the clustering results. This visual presentation shows clear separation between different cell lines while indicating similarity between the two invasive cancer cell lines (shown in red and blue). PCA in this context shows global membership trend for each sample. Samples that are close together in Fig. 5(B) have similar membership values and are thus similarly clustered according to F-KM; samples that are far apart in the PCA plot are dissimilar across all groups.

In the type 2 diabetes dataset, the top membership values clearly separate samples from three species and also the subtypes of rat and mouse samples. Human samples are co-clustered on the basis of the top membership only (Fig. 6(A)). Closer inspection of second membership values shows similarities within rat and mouse samples of two subtypes and also shows differences in some samples between two subtypes of human subjects. The PCA representation of membership data (Fig. 6(B)), i.e. PCA of the output of F-KM analysis of type II diabetes data clearly shows this result. The between species separation as well as separation of rat and mouse subtypes is apparent and there is some separation between two human subject subgroups.

Discussion

Several different unsupervised data analysis methods were compared. Methods included visualisation algorithm PCA, crisp clustering HCL and K-means as well as fuzzy clustering method F-KM. All these different algorithms were tested on two metabolomics data sets. First dataset was measured as part of this work and included 1D ^{1}H NMR measurements of lipophilic extracts of replicates of five different breast cell lines. The second previously published dataset included NMR spectra of urine samples of three different species – mouse, rat and human – with healthy and diabetic-like

Table 3. Jaccard and Rand coefficients for clustering for breast cancer cell lines (absolute value distance measure; five clusters)

Breast cell lines clustering result		
	K-means	Fuzzy K-means ($m = 2$) – top memberships
Rand	0.757246	0.807971
Jaccard	0.247191	0.311688
Type 2 diabetes dataset		
	K-means	Fuzzy K-means ($m = 2$) – top memberships
Rand	0.846785	0.91962
Jaccard	0.53749	0.645322

The perfect classification result has both indices equal to 1.

phenotypes for each species with a total of 270 samples. The appraisal of clustering methods was performed by using the direct comparison of results and also by using two cluster quality coefficients.

Classical methods PCA and HCL of NMR of cell lines provided good separation of normal and cancer cell lines (Fig. 2) as well as separation of three species studied in the diabetes dataset (Fig. 3). However, in the cell line dataset, PCA did not give any indication of subtypes, i.e. tumour invasive and non-invasive cell lines (Fig. 2(A)). In the type II diabetes dataset, PCA analysis leads to excellent separation of three species, good separation of different mouse phenotypes and very poor separation of rat and human subjects (Fig. 3(A)). PCA and related visualisation methods focus on the major trends in the data and do not take into consideration overall changes in profiles. The PCA shows that observing only major trends cannot provide an accurate determination of subtypes and, although the PCA method provides a highly pictorial representation of the data, it does not lead to optimal sample classification. In other words, major changes in metabolic profiles are indicative of major phenotypes such as cancer or normal cell lines or different animal species. However, analysis methods must consider more subtle changes in metabolic profiles in order to determine sub-phenotypes of samples.

The HCL method provides clusters represented as dendograms without the need for user-defined number of clusters. However, the HCL method resulted in relatively poor sample separation in the cases studied here. In the cell lines datasets, HCL provided a good separation of normal cell lines for most replicates but this method co-clustered MCF7 and MB468 cell lines as well as MB231 and MCF12A at the same dendrogram level (Fig. 2(B)). In the analysis of type II diabetes dataset, HCL provided a result that was similar to that provided by PCA (Fig. 3(B)), i.e. HCL separated samples from the three species well, leading to good separation of samples from two groups of mice, fairly good separation of rat phenotypes and poor separation of human subjects. K-means algorithm is a standard, crisp clustering method with user-defined number of clusters. In the cases studied here, the number of clusters was known and therefore K-means method appeared to be a good approach. However, K-means gave a very poor result relative to known classes for cell line samples (Fig. 4(A)). For type II diabetes dataset, K-means clustering led to good separation into three species, but any separation of within-species phenotypes was only possible for the mouse model, and, even in this case, there were some very pronounced miss-classifications (Fig. 4(B)).

Finally, for the breast cell lines dataset, the F-KM clearly shows distinct metabolic profiles for all five cell lines (Fig. (5)). The membership values of each sample to all clusters (Fig. 5(A)) show that even the analysis of only the top memberships allows clear separation of MCF10A, MCF12A and MCF7 cell lines. MB231 and MB468 are co-clustered on the basis of the top membership values but are clearly distinguishable from the second membership values. This result shows that MB231 and MB468 cell lines are more similar to one another than to the other cell lines but using F-KM it is still clearly possible to distinguish between these two types. The PCA representation of membership values (Fig. 5(B)) shows this result pictorially. It should be noted that cell lines were grown under optimal conditions with one set of conditions (i.e. media) being used for MCF12A and MCF10A; second media for MB231 and MB468 and third media for MCF7 cell lines. Although these are standard conditions regularly utilised for the comparison of different cell lines, they can influence final metabolic profiles. However, it is clear that other differences in the profiles dominate the classification as the cell lines that are grown under the same conditions such as MCF12A and MCF10A are separated into two

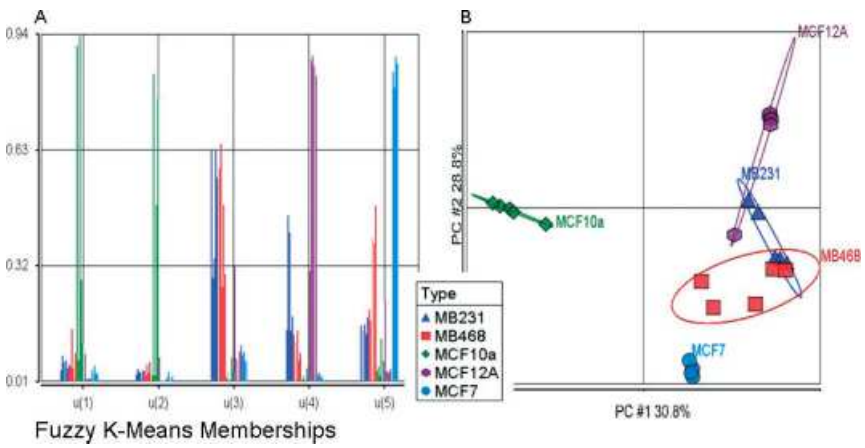

Figure 5. Membership values for each sample determined by fuzzy K-means clustering of samples from breast cell line NMR spectra. Higher membership values represent stronger allegiance to a given cluster. The cell lines studied are normal (MCF10A and MCF12A); invasive cancers (MB231 and MB468) and non-invasive cancer (MCF7). (A) Exact membership values for each sample across all 5 clusters; (B) PCA analysis of membership values showing major trends in the clusters.

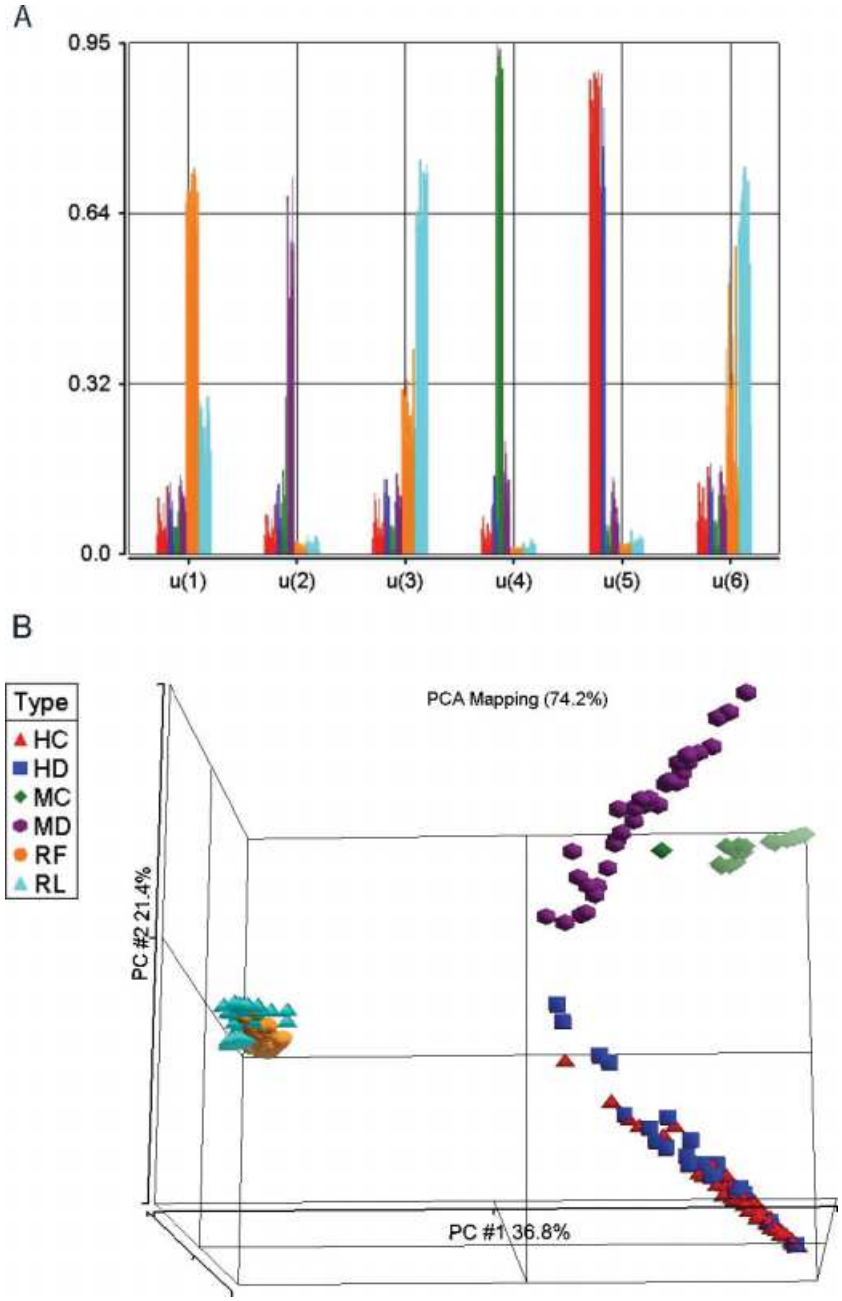

Figure 6. Membership values for each sample determined by fuzzy K-means clustering of samples from type 2 diabetes dataset. Higher membership values represent stronger allegiance to a given cluster. The dataset included human control (HC), human diabetes patients (HD, mouse wildtype (MC) and db/db mice (MD), obese Zucker rats (RF) and lean Zucker rats (RL). (A) Exact membership values for each sample across all six clusters; (B) PCA analysis of membership values showing major trends in the cluster results.

distinguishable clusters using all of the tested methods. These five cell lines were used in this work for the comparison of different analysis methods and thus it is not our intention to make any conclusions regarding the biology of these cell lines based on this experiment. Regardless, this work does provide a proof of concept and validation of the clustering method, which can now be applied to cell lines or tissues grown under more appropriate conditions for direct biological comparison.

Similar results in terms of the ability of different analysis methods to separate different sample phenotypes was observed in the second, type 2 diabetes dataset. The problem in this dataset is that the NMR measurements were performed under different conditions for the three species studied. Therefore, the phenotypical differences within species are not equally pronounced in the three groups. F-KM analysis resulted in an improved clustering based on both the Jaccard and Rand index values (Table 2) as well as the analysis of membership values. From the top membership values (Fig. 6(A)), it was possible to clearly separate phenotypes in mouse and rat and this was also observed from the PCA plots of membership values (Fig. 6(B)). Differences between human subjects are less clear and the two groups are co-clustered on the basis of the top membership classification. However, closer inspection of second and third membership values shows separation between two human subject groups

that can be observed in the PCA representation of membership values.

For both datasets, F-KM classification based on the top membership values resulted in better quality of clustering results based on both Jaccard and Rand indices when compared with the K-means method (Table 2). F-KM membership values also provided additional data regarding sample sub-phenotypes. From the analysis of fuzzy membership values, it was possible to separate major phenotype differences as well as minor phenotype subgroups in both datasets tested. For breast cell lines, it was established that differences can be observed between cancer and normal as well as invasive and non-invasive cancer cell lines. In the type II diabetes dataset, fuzzy membership values allow clear separation of three species, as well as two phenotypes in mouse and rat models. For human type II diabetes and healthy phenotypes, membership values lead to better sample separation than other methods; however, even with the F-KM method, it is rather difficult to make clear separation of these phenotypes. This result indicates that other environmental or clinical influences (such as age, body mass index, etc.) can create problems for human subject classification regardless of the method used. Further work will focus on the exploration of different pre-processing methods in conjunction with fuzzy classification.

Conclusions

In conclusion, we report the application of the fuzzy clustering method in the analysis of metabolomic data. From the analysis of fuzzy membership values, it was possible to separate samples based on major phenotypical differences as well as minor phenotype subgroups. The F-KM method resulted in the best, most accurate classification based on sample class labels of all the methods tested and provided additional information that led to sub-phenotype classification.

From these results, it is clear that the analysis method of choice for metabolomics data must include complete spectra in the sample classification calculations and should be able to provide information about both major as well as minor cluster memberships for all samples. The F-KM method follows these requirements and allows more accurate sample classification when compared with standard methods.

Acknowledgements

Authors would like to thank Drs R. Salek and J. Griffin for providing the type 2 diabetes dataset. This work was supported by the National Research Council Atlantic Initiative and Atlantic Innovation Fund.

References

[1] E. Holmes, I. D. Wilson, J. K. Nicholson, *Cell* **2008**, *134*, 714.
[2] J. Loscalzo, I. Kohane, A. L. Barabasi, *Mol. Syst. Biol.* **2007**, *3*, 124.
[3] M. Ala-Korpela, *Expert Opin. Mol. Diagn.* **2007**, *7*, 761.
[4] C. Yang, A. D. Richardson, J. W. Smith, A. Osterman, *Pac. Symp. Biocomput.* **2007**, *12*, 181.
[5] J. L. Griffin, J. P. Shockcor, *Nat. Rev.* **2004**, *4*, 551.
[6] J. A. Hageman, R. A. van den Berg, J. A. Westerhuis, H. C. J. Hoefsloot, A. K. Smilde, *Critic Rev. Anal. Chem.* **2006**, *36*, 211.
[7] O. Beckonert, M. E. Bollard, T. M. D. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon, J. K. Nicholson, *Anal. Chim. Acta* **2003**, *490*, 3.
[8] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, C. M. Slupsky, *Anal. Chem.* **2006**, *78*, 4430.
[9] J. Quackenbush, *Science* **2003**, *302*, 240.
[10] A. M. Bowcock, *Nature* **2007**, *447*, 645.
[11] S. Halouska, R. Powers, *J. Magn. Res.* **2006**, *178*, 88.
[12] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, F. A. van Dorsten, *Metabolomics* **2008**, *4*, 81.
[13] Y. Tikunov, A. Lommen, C. H. Ric de Vos, H. A. Verhoeven, R. J. Bino, R. D. Hall, *Plant Physiol.* **2005**, *139*, 1125. Bovy A.G.
[14] V. P. Mäkinen, P. Soininen, C. Forsblom, M. Parkkonen, P. Ingman, P. Kaski, P. H. Groop, M. Ala-Korpela, *Mol. Syst. Biol.* **2008**, *4*, 167.
[15] N. Belacel, M. Cuperlovic-Culf, M. Laflamme, R. J. Ouellette, *Bioinformatics* **2004**, *20*, 1.
[16] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press: New York, **1981**.
[17] A. P. Gasch, M. B. Eisen, *Genome Biol.* **2002**, *2*, 0059-1–0059-22.
[18] R. D. Pascual-Marqui, A. D. Pascual-Montano, K. Kochi, J. M. Carazo, *Pattern Recognit.* **2001**, *34*, 2395.
[19] M. Lacroix, G. Leclercq, *Breast Cancer Res. Treat.* **2004**, *83*, 249.
[20] R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J. P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J.Wang, W. L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, J. W. Gray, *Cancer Cell* **2006**, *10*, 515.
[21] E. G. Bligh, W. J. Dyer, *Can. J. Biochem. Physiol.* **1959**, *37*, 911.
[22] M. Gottschalk, G. Ivanova, D. M. Collins, A. Euastece, R. O'Connor, D. F. Brougham, *NMR Biomed.* **2008**, DOI: 10.1002/nbm.1258.
[23] R. M. Salek, M. L. Maguire, E. Bentley, D. V. Rubtsov, T. Hough, M. Cheeseman, D. Nunez, B. C. Sweatman, J. N. Haselden, R. D. Cox, S. C. Connor, J. L. Griffin, *Physiol. Genomics* **2007**, *29*, 99.
[24] D. Dembele, P. Kastner, *Bioinformatics* **2003**, *19*, 973.
[25] N. J. Waters, E. Holmes, C. J. Waterfield, R. D. Farrant, J. K. Nicholson, *Biochem. Pharmacol.* **2002**, *64*, 67.
[26] T. L. Whitehead, T. Kieber-Emmons, *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *47*, 165.
[27] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, J. L. Markley, *Nucleic Acids Res.* **2007**, *36*, D402.
[28] S. L. Robinette, F. Zhang, L. Bruschweiler-Li, R. Brüschweiler, *Anal. Chem.* **2008**, *80*, 306.
[29] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, L. Querengesser, *Nucleic Acids Res.* **2007**, *35*, (Database issue), D521.

