Identifying and Implementing the Underlying Operators for Nuclear Magnetic Resonance based Metabolomics Data Analysis

Ashwin Manjunatha
Kno.e.sis Center
Wright State University
Dayton, OH 45435

Paul Anderson
Air Force Research Laboratory
Wright-Patterson AFB
Dayton, OH 45433

Ajith Ranabahu
Kno.e.sis Center
Wright State University
Dayton, OH 45435

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, OH 45433

Abstract

The science of metabolomics is a relatively young field that requires intensive signal processing and multivariate data analysis for interpretation of experimental results. The lack of integration and standardization for metabolomics compounded by the complexity of the experimental data has lead to a fragmented research community. While efforts have been undertaken to approach these problems, the efforts to develop a set of standards for reporting processing and analysis procedures has stalled.

In this paper, we propose a set of fundamental operators for nuclear magnetic resonance(NMR) based metabolomics. These operators are implementation independent, and can be used to easily and precisely describe the processing and analysis steps that led to research conclusions. This formalization can facilitate inter-lab communication, and due to its simplicity, it is easily adapted by the metabolomics community. A Domain Specific Language (DSL) is also included to demonstrate an implementation of these operators. The DSL is simple, convenient for a domain scientist, and can be easily transformed into multiple target platforms.

1 Introduction

Metabolomics, the measurement of metabolite concentrations and fluxes in various biological systems, is one of the most comprehensive of all bionomics. Unlike proteomics and genomics that assess intermediate products, metabolomics assesses the end product of cellular function, metabolites. Changes occurring at the level of genes and proteins (assessed by genomics and proteomics) may or may not influence a variety of cellular functions. But metabolomics, by contrast, assesses the end products of cellular metabolic function. For instance, a disease or foreign compound may interfere at the genomic or proteomic level, while it will always manifest itself at the metabolomic level. In contrast to various other proteomic, genomic, and metabolomic analyses, NMR spectroscopy is non-invasive, non-destructive, and requires little sample preparation [14].

The first use of multivariate statistical techniques to high resolution NMR spectra was the classification of spectra from rat urine according to type of organ toxin which had been administered [8]. Further, NMR spectroscopy of biofluids has been shown to be an effective method in metabolomics to identify variations in biological states [9]. These applications rely on algorithmic spectral processing techniques to be successful.

A typical ^{1}H NMR spectrum of pure proteins, biofluids, or tissue may contain thousands of overlapping resonances, thus, the quantification and analysis is inherently complex. To overcome these complexities, the field of NMR-based metabolomics employs a variety of computationally intensive algorithms that range from signal processing to pattern recognition techniques. The analysis of an NMR spectroscopic dataset is often divided into five steps: (1) standard post-instrumental processing; (2) normalization (3) quantification of spectral features; (4) scaling; and (5) multivariate statistical modeling.

After standard post-instrumental processing, normalization is performed on a per-spectrum basis to make the samples directly comparable to each other [4]. This is designed to remove artificial differences, such as variable dilution of the samples, which is a common problem in NMR spectroscopic studies of

urine where many toxins can cause large increases or decreases in urinary volume; however, normalization is less important when samples are highly regulated, such as plasma, which is highly regulated by homeostasis (i.e., maintenance of physiological conditions required to maintain life).

Quantification of spectral features, step (3), is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). A common method of quantification employed by the NMR community is known as binning or bucketing, which divides a NMR spectrum into several hundred regions. This technique is performed to (1) minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition [16]; and (2) reduce the dimensionality for multivariate statistical analyses.

There are several alternatives to spectral binning that still provide data dimension reduction. Examples of these include PARS [7], curve-fitting method for direct quantification [5], peak alignment tools in HiRes [20], and targeted profiling [18]. These techniques identify peaks or specific peak patterns in the spectra that are conserved across spectra. After the patterns have been identified, they are quantified by determining the peak area or amplitude. The accuracy of these algorithms is dependent on the spectral resolution, the quality of the peak alignment, and the breadth of spectroscopic pattern databases.

Scaling is designed to control the weighting of features before a multivariate statistical or pattern recognition technique is applied [4, 17]. Scaling techniques are applied to the entire data set on an individual feature basis. A number of scaling techniques are commonly used, including mean-centering [4], auto-scaling [4], Pareto scaling [6, 11], and logarithmic scaling [1]. Both normalization and scaling are highly context dependent, and therefore, no single approach is optimal for all types of experiments [4].

The 5th step, multivariate statistical analysis is often divided into unsupervised and supervised analyses. Unsupervised exploratory data analysis is commonly accomplished via principal component analysis (PCA). A number of approaches are available for identification of the presence or absence of a toxic response and for characterization of biomarkers (sets of variables) associated with that response. These supervised techniques include linear discriminant analysis (LDA), logistic regression, t-tests, and partial least squares discriminant analysis (PLS-DA) [19, 13].

The work presented herein identifies the common fundamental operators on metabolomics data and formalizes these operators such that a mathematical expression can be constructed describing the analysis protocol. Hence, the contributions of this paper are;

1. A set of fundamental operators to describe NMR-based metabolomics.
2. A DSL and a related set of tools that partially implements these operators.

Since the operators are fundamental in nature, we show that the DSL representation can easily be converted to a Cloud based environment as well as a desktop environment.

2 Motivation

Computing and the Internet have had such an impact on the modern life sciences that it has become cliche to call biology a data driven science. In select research areas, such as molecular evolution and genomics, bench scientists, mathematicians, computer scientists and others have come together, resulting in transformative changes to the manner in which experimental data is collected and analyzed. Where the research community has converged around a handful of key web resources (NCBI, EMBL, etc.), the result has been standardization of tools, data formats, analysis techniques, and even experimental methods. The fact remains, however, that not all areas of study have achieved this level of integration.

In contrast to the fields of proteomics and genomics, metabolomics is a relatively new field of study that has not benefited from the level of integration and standardization of more developed fields; however, this is not due to the lack of impact or importance. It can be argued that the endpoint of all biological processes is the facilitation and regulation of metabolism, thus observation of metabolite levels can provide keen insight into the condition of an organism. Metabolomics provides a comprehensive snapshot of all metabolite levels in a fluid or tissue and is recognized for its broad domain of applicability in clinical, pre-clinical, environmental, and diagnostic research areas.

The lack of integration and standardization for metabolomics is compounded by the complexity of the experimental data generated and the diversity of experimental instruments and analysis techniques in use. The selection of appropriate and accessible tools for the preprocessing, exploration, visualization, and statistical analysis of this highly dimensional data can have a profound impact on the research conclusions.

Scientists, working groups, and professional societies throughout the field have called for standards governing data storage and experiment reporting. While efforts have been undertaken [3, 12, 15], the development of these standards has stalled. Faced with a paucity of standards and resources metabolomics researchers employ a variety of proprietary and in house tools. Few, if any, of these proprietary software pack-

ages are universally adopted. The result is fragmentation in the research community.

While scientists and computer scientists will often disagree upon the best implementation of various techniques, such as scaling and normalization, these remain fundamental operations on NMR spectroscopic data. This abstracted level of commonality can be exploited to standardize the field of metabolomics, which will reduce the fragmentation by improving inter-lab communication. Herein, we propose a set of domain specific operators for the field of NMR-based metabolomics. These operators are implemented using a Domain Specific Language (DSL) to illustrate the ability to define the processing independent of the target platform.

DSLs are indeed in use in the many scientific domains. Scientists and biologists are familiar with DSL driven scientific software tools that provide friendly environments for their particular needs. Matlab [10] is one such commercial software that provides specific data structures and modules that biologists need in their routine workflows. Scientists typically run Matlab in a desktop environment and hence they are constrained with respect to computational power. Moving to a distributed environment may require mastering a set of new technologies and many scientists are hesitant to move away from the convenience of domain specific tools such as Matlab. We make two observations in this context.

(1) There is an increase in the available computing power and distributed computing tools. These tools however have sharp learning curves often discourage scientists from adopting them.

(2) User friendly and domain specific tools are deemed important by scientists. The convenience of such tools is often preferred over their apparent lack of performance. The performance issues can often be alleviated by adding more computing resources rather than code optimization as outlined by the so called *Carbon vs Silicon* argument. In the subsequent section, We describe the set of fundamental operators we have identified for NMR-based metabolomics.

3 Formalizing Fundamental Operators

The definition of the fundamental operators for NMR-based metabolomics will provide a common language that will facilitate inter-lab communication by precisely described the processing and analysis. Some of these operators include:

- **Normalization (N):** This family of operators are performed on a per-spectrum basis to make the samples directly comparable to each other.

Two common sub-operators of this family include Sum normalization (N_{sum}) and normalization by weight (N_{weight}).

- **Correction (C):** This family of operators remove errors introduced by measuring equipments such as baseline shift. Sub-operators of this family include baseline correction ($C_{baseline}$) and phase correction (C_{phase}).

- **Quantification (Q):** This family of operators reduce the dimensionality of the data and attempt to extract or approximate metabolite concentrations. Sub-operators of this family include binning ($Q_{binnnig}$) and targeted profiling ($Q_{targetedprofiling}$).

- **Scaling (S):** This family of operators control the weighting of features before a multivariate statistical or pattern recognition technique is applied. Sub-operators of this family include auto-scaling ($S_{autoscaling}$), Pareto-scaling ($S_{paretoscaling}$), and mean-centering ($S_{meancentering}$)

- **Mining(M):** This family of operators selects the significantly responding metabolites/features for a given experiment. Sub-operators of this family include t-test (M_{ttest}), and partial least squares with variable selection (M_{pls}).

- **Visualize(V):** This family of operators output a visualized representation of the data and/or results. Sub-operators of this family include principal component analysis (V_{PCA}) and partial least squares scores plot (V_{PLS}).

- **Transformation(T):** This family of operators perform data transformations, such as Fourier transforms ($T_{fourier}$).

These operators operate on Matrices (S) or Vectors (s). For example $N_{sum} : s \rightarrow s$ where $s \epsilon S$.

The primary objective of these operators is to provide an uniform mathematical language to describe a NMR data processing task. As an example Equation 1 is a pure function oriented representation of doing a base line correction on Fourier transformed, phase corrected and auto-scaled data set S where S' is the processed data set. This representation (and other equivalent symbolic representations) are suitable for scientific exchanges since they formerly indicate the operations and their order.

$$S' = C_{baseline}(Q_{autoscaling}(C_{phase}(T_{fourier}(S))))) \quad (1)$$

Since Equation 1 may not be intuitive as to the order of the operations, one may use an alternative representation that resembles a workflow. Equation 2 uses $\rightarrow$ to denote an input to a operator.

$$S' = S \rightarrow T_{fourier} \rightarrow C_{phase} \rightarrow Q_{autoscaling} \rightarrow C_{baseline}$$
$$(2)$$

Another convenient representation is the pseudocode style, as illustrated in Program 1, which is readily converted to the DSL described in Section 4.

Program 1 A Pseudocode representation of a processing task

$S_1 = T_{fourier}(S)$
$S_2 = C_{phase}(S_1)$
$S_3 = Q_{autoscaling}(S_2)$
$S' = C_{baseline}(S_3)$
$F = M_{pls}(S')$
$V_{pca}(S', F)$

4   Using a DSL to represent the Fundamental Operators

We now present the details of our first attempt in implementing a subset of these operators as a DSL. While this DSL is created by restricting the Ruby base language, one may implement these operators by many other means, e.g. Matlab functions or C macros. We selected a DSL for its readability and the gentle learning curve although one may implement them in more specialized formats. The primary goal of this work is to provide a set of fundamental operators that can be used to represent a NMR process independent of an implementation.

The current implementation provides abstractions on top of Apache Pig, a platform for analyzing large data sets over the map-reduce framework, Hadoop [2]. Due to its underlying map-reduce architecture and its fault-tolerant file system, Hadoop is ideal for analyzing large spectroscopic data sets. The layered architecture of the implementation is illustrated in Figure 1. Note that since the language is based on fundamental operators, the workflow represented by the DSL can be converted to other forms (e.g. a Matlab based script running on a desktop or .net based program running on the Windows Azure Cloud) in a lossless manner. The Metabolink toolkit contains the compiler/generators to convert the DSL script into concrete implementations that run on target platforms.

The drag and drop style graphical user interface, depicted in Figure 1, would be a layer of convenience over a textual language. This is important in the context of scientific workflows due to the high complexity of the workflows and the difficulty of visualizing them. The success of tools like Taverna is evidence to the effectiveness of drag and drop style workflow composers in the scientific computing domains.
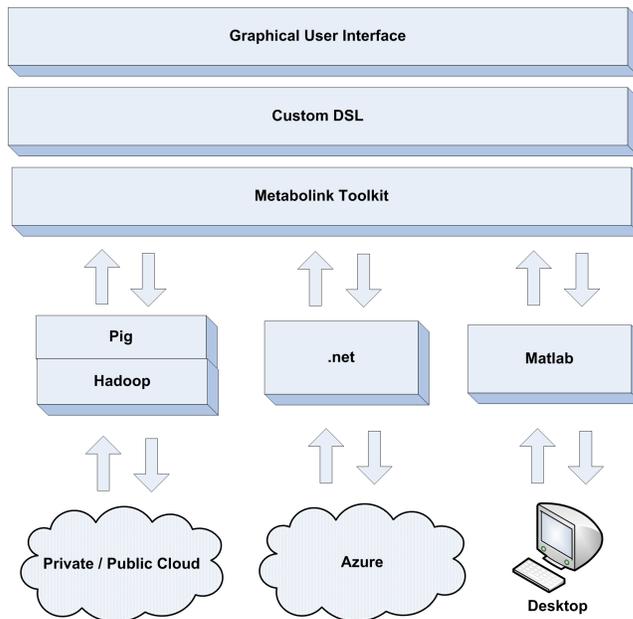
Figure 1: Layered Architecture of the Implementation

Listing 1 outlines a simple *mini workflow* where a data file is loaded, filtered, sum normalized and written back to a new file. The variables *raw_data_file* and *normalized_data_file* represent the input and the output files respectively. Other function references are self explanatory. Equation 3 shows the mathematical represenation of the script in Listing 1.

$$S' = S \rightarrow Q_{filter} \rightarrow N_{sum} \qquad (3)$$

Listing 1: Filtering and Sum normalization implemented using the DSL

```
# load data
original_data =
load_data_from_csv(raw_data_file)

# filter out a range
filtered =
range_filter({:min=> 20, :max => 50},
             original_data)

# sum normalize
normalized = sum_normalize(filtered)

# write the file
store(normalized_data_file, normalized)
```

In order to contrast the effort in implementing this in PIGLatin, Listing 2 shows one of the simplest hand written PIGLatin scripts. This script implements sum normalization.

Listing 2: Sum normalization implemented using the PIG

```
A = LOAD '$filename' USING PigStorage(',')
 AS (colnum:int, value:double);

B = GROUP A BY colnum;

C = FOREACH B GENERATE group,
        SUM(A.value);

D = COGROUP A by colnum inner,
     C by $0 inner;

F = FOREACH D GENERATE group,
    FLATTEN (A),FLATTEN (C);

G = FOREACH F GENERATE $0,($2/$4)*100;

STORE G INTO '$filename_processed'
  USING PigStorage (',');
```

There are two observations from these code comparisons.
(1) The PIGLatin script is not intuitive, i.e. its not obvious from the script as to its function.
(2) Creating the PIGLatin script requires a different pattern of thinking and reasoning that needs to be obtained through practice.
It is clearly intuitive for the biologist to follow the first script rather than the second.

5   Conclusion

Introduction of a set of fundamental operators for NMR-based metabolomics is indeed a valuable generalization that provides a means of formal definition of the processing task. Although these operators may not be exhaustive, they can act as a basis to build domain specific languages and tooling that immensely benefits the scientists. These operators can be easily implemented to take advantage of Clouds and other scalable computing environments without exposing complex details of such environments.

References

[1] ML Anthony, BC Sweatman, CR Beddell, JC Lindon, and JK Nicholson. Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Molecular pharmacology*, 46(1):199, 1994.

[2] A. Bialecki, M. Cafarella, D. Cutting, and O. OMalley. Hadoop: a framework for running applications on large clusters built of commodity hardware. 2005. Avialable online at http://hadoop.apache.org.

[3] A.L. Castle, O. Fiehn, R. Kaddurah-Daouk, and J.C. Lindon. Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Briefings in Bioinformatics*, 7(2):159, 2006.

[4] A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson, and J.C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem*, 78(7):2262–2267, 2006.

[5] DJ Crockford, HC Keun, LM Smith, E. Holmes, and JK Nicholson. Curve-fitting method for direct quantitation of compounds in complex biological mixtures using 1H NMR: application in metabonomic toxicology studies. *Anal. Chem*, 77(14):4556–4562, 2005.

[6] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold. *Multi-and megavariate data analysis*. Umetrics Umeå, 2001.

[7] J. Forshed, R.J.O. Torgrip, K.M. Åberg, B. Karlberg, J. Lindberg, and S.P. Jacobsson. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of pharmaceutical and biomedical analysis*, 38(5):824–832, 2005.

[8] KP Gartland, CR Beddell, JC Lindon, and JK Nicholson. Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine. *Molecular pharmacology*, 39(5):629, 1991.

[9] J.L. Griffin. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Current opinion in chemical biology*, 7(5):648–654, 2003.

[10] D. Hanselman and B.C. Littlefield. *Mastering MATLAB 5: A comprehensive tutorial and reference*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1997.

[11] E. Holmes, AW Nicholls, JC Lindon, S. Ramos, M. Spraul, P. Neidig, SC Connor, J. Connelly, SJP Damment, J. Haselden, et al. Development of a model for classification of toxin-induced lesions using 1H NMR spectroscopy of urine combined with pattern recognition. *NMR in Biomedicine*, 11(4-5):235–244, 1998.

[12] J.C. Lindon, J.K. Nicholson, E. Holmes, H.C. Keun, A. Craig, JT Pearce, S.J. Bruce, N. Hardy, S.A. Sansone, H. Antti, et al. Summary recommendations for standardization and reporting of metabolic analyses. *Nature biotechnology*, 23(7):833, 2005.

[13] H. Martens and T. Naes. *Multivariate calibration*. John Wiley & Sons Inc, 1992.

[14] N.V. Reo. NMR-based metabolomics. *Drug and chemical toxicology*, 25(4):375–382, 2002.

[15] D.V. Rubtsov, H. Jenkins, C. Ludwig, J. Easton, M.R. Viant, U. G
"unther, J.L. Griffin, and N. Hardy. Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics*, 3(3):223–229, 2007.

[16] M. Spraul, P. Neidig, U. Klauck, P. Kessler, E. Holmes, JK Nicholson, BC Sweatman, SR Salman, RD Farrant, E. Rahr, et al. Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of pharmaceutical and biomedical analysis*, 12(10):1215–1225, 1994.

[17] R.A. Van Den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, and M.J. Van Der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):142, 2006.

[18] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, and C.M. Slupsky. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem*, 78(13):4430–4442, 2006.

[19] H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 1:391–420, 1966.

[20] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T.R. Brown. HiResa tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics*, 22(20):2562, 2006.