A Proposed Statistical Protocol for the Analysis of Metabolic Toxicological Data Derived from NMR Spectroscopy

Benjamin J. Kelly
*Wright State University - Main Campus*

Paul E. Anderson
*Wright State University - Main Campus*

Nicholas V. Reo
*Wright State University - Main Campus, nicholas.reo@wright.edu*

Nicholas J. DelRaso

Travis E. Doom
*Wright State University - Main Campus, travis.doom@wright.edu*

*See next page for additional authors*

Follow this and additional works at: [http://corescholar.libraries.wright.edu/knoesis](http://corescholar.libraries.wright.edu/knoesis)

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation

This Conference Proceeding is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Authors
Benjamin J. Kelly, Paul E. Anderson, Nicholas V. Reo, Nicholas J. DelRaso, Travis E. Doom, and Michael L. Raymer

This conference proceeding is available at CORE Scholar: http://corescholar.libraries.wright.edu/knoesis/533
A Proposed Statistical Protocol for the Analysis of Metabolic Toxicological Data Derived from NMR Spectroscopy

Benjamin J. Kelly 1, Paul E. Anderson 1, Nicholas V. Reo 2, Nicholas J. DelRaso 3, Travis E. Doom 4, and Michael L. Raymer 1*

1Wright State University
Dept. of Computer Science and Engineering
Dayton, OH 45435

2Wright State University
Dept. of Biochemistry and Molecular Biology
Boonshoft School of Medicine
Kettering, OH 45429

3Air Force Research Laboratory
Human Effectiveness Directorate
Wright-Patterson AFB, OH 45433

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a non-invasive method of acquiring a metabolic profile from biofluids. This metabolic information may provide keys to the early detection of exposure to a toxin. A typical NMR toxicology data set has low sample size and high dimensionality. Thus, traditional pattern recognition techniques are not always feasible. In this paper, we evaluate several common alternatives for isolating these biomarkers. The fold test, unpaired t-test, and paired t-test were performed on an NMR-derived toxicological data set and results were compared. The paired t-test method was preferred, due to its ability to attribute statistical significance, to take into consideration consistency of a single subject over a time course, and to mitigate the low sample, high dimensionality problem. We then grouped the resulting statistically salient potential biomarkers based on their significance patterns and compared results to several known metabolites affected by the tested toxin. Based on these results, we present a statistical protocol of sequential t-tests and clustering techniques for identifying putative biomarkers. We then present the results of this protocol applied to a specific real world toxicological data set.

1 Introduction

Metabolomics, the evaluation of metabolites for the purpose of understanding cellular function, is one of the more comprehensive of all biomomics. Unlike proteomics and genomics that assess intermediate products that may or may not contribute to variation in cellular function, metabolomics assesses the end product of cellular function.

Proton nuclear magnetic resonance (NMR) spectroscopy of biofluids has been shown as an effective method in metabolomics to identify biological variation [9]. Unlike various other proteomic, genomic, and metabolomic analyses, NMR spectroscopy is non-invasive, non-destructive, and requires little preparation [10].

The resulting NMR spectrum is generally separated into several hundred bins to minimize effects from slight variations in peak positions. A typical metabolomics experiment will have only a small number of samples from a statistical perspective(< 10). This combination of high dimensionality and low sample size make traditional pattern recognition techniques infeasible; therefore, techniques that are sensitive to low sample sizes are necessary.

Principal component analysis (PCA) is an unsupervised method commonly used to visualize NMR-derived metabolomic data [7]. PCA transforms the data into a new coordinate system where each new dimension is computed as a linear combination of the original values. These new dimensions are called principal components, and are orthogonal vectors that explain the variance in the data. Various literature reports describe the use of PCA to illustrate the clustering of metabolites using the top principal components, with the idea that metabolites can separate groups based on their ability to differentiate as a possible biomarker [1, 5, 6].

Another commonly used alternative is the Student’s t-test that has been used both in conjunction with and independent of PCA. The t-test attributes statistical significance to PCA results [5, 6]. Independent of PCA, the t-test has been used to assess differences between treatment and control groups with the intent to identify biomarkers [13].

In this paper, several pattern recognition alternatives will be evaluated on an NMR-derived toxicological data set. The fold test, common among microarray data analysis, along with two variations on the Student’s t-test, commonly used on both NMR and microarray data analysis, will be per-
formed on the same data set and results will be compared. These tests are performed on each bin individually, thus mitigating the problems of high dimensionality. The proposed methodology will begin by performing a paired t-test to identify statistically salient bins. These bins will then be grouped based on their significance patterns and several metabolites that are known to be affected by the specific toxin will be compared to these grouping results. Agreement between the results and known metabolites would confirm this sequence of tests as an option when attempting to isolate biomarkers from NMR-derived toxicological data.

2 Methods

2.1 Data Acquisition and Preparation

Treated rats were administered a single, sub-lethal, 50 mg/kg dose of alpha-naphthylisothiocyanate (ANIT) in corn oil vehicle via gavage, while control rats were given an identical volume of corn oil vehicle only. Urine samples were collected pre-dose and at 24, 48, 72, and 96 hours post-dose. Proton NMR spectra were acquired on a Varian INOVA operating at 600 MHz.

NMR spectral data were processed using Varian software, employing exponential multiplication, Fourier transformation, and baseline flattening. Spectra were subdivided into 280 regions of 0.04 ppm width for integration using Varian Binning software. The water and urea signals in urine spectra were excluded from the analyses. The integrated bin areas were then normalized to the TSP signal intensity. A second normalization was performed by summing 256 bins over the entire range, which excludes the region of TSP signal.

The resulting data were comprised of a set of 256 features (bins) for each of 5 time points (pre-dose, 24hr, 48hr, 72hr, and 96 hr post-dose) for 5 treated and 13 control rats.

2.2 Fold Test

The fold test is often used for identifying differentially expressed genes in microarray analysis and is applicable for NMR data [3, 8]. It is a direct comparative approach in which the log ratio of the two conditions are evaluated and the resulting ratios is compared to an arbitrary cut-off. A two-fold difference is a common cutoff to determine an interesting difference, though no statistical significance can be attributed to this cutoff.

For this data set, the mean of log ratios of pre-dose measurements to each of the 24, 48, 72, and 96 hour measurements of the same bin were compared to a cut-off ratio of 2. All bin/time points from treatment rats that exhibited a two-fold ratio were retained.

2.3 Student's t-test

The t-test can be more informative than the fold test and is used prominently in microarray and NMR data analysis [1, 2, 3, 5, 8]. The t-test considers the variance of measurements and assigns statistical significance to the result.

A simple two-sample t-test between the pre-dose measurements and each subsequent time point is not applicable for this data set, as the assumption of independent sets of data cannot be met. Instead, a two-sample t-test versus a set of controls using the differences between the pre-dose measurements and each subsequent time point must be performed. These differences are assumed to be nearly normally distributed [2]. With a small sample size (n < 15), it may not be safe to assume that treatment measurements and control measurements have equal variances. Thus, a two-sample t-test assuming unequal variances was performed.

The differences between pre-dose measurements and the 24, 48, 72, and 96 hour measurements for each bin were calculated for both the treatments and the controls. These differences were then compared using a two-sample, unpaired t-test assuming unequal variances with an alpha value of 0.01.

Another frequently used alternative is the paired t-test [1, 5]. A paired t-test between the pre-dose measurements and each following time point takes advantage of the fact that the pre-dose measurements and the subsequent time point measurements are taken from the same subject. Any inherent variability between subjects is mitigated by each subject serving as its own control [8].

For this data set the pre-dose measurements and the 24, 48, 72, and 96 hour measurements for each bin from treatment were compared using a paired t-test with alpha = 0.01.

The nature of measuring the same sample at several time intervals and testing multiple hypotheses on these measurements requires an adjusted p-value. A simple and common approach is to perform a Bonferroni correction. The Bonferroni correction is conservative and assumes independence of the test statistics [4].

In this data set, like microarray data sets, that assumption cannot be met. The Bonferroni correction may still be used in situations of dependence, but the correction becomes highly conservative and type-II error is likely to increase [11]. This was considered acceptable and dividing the desired alpha by the number of hypotheses tested was used. The adjusted alpha of 0.0025 was used for both the unpaired and paired t-tests.

2.4 Removal of Vehicle Effect

The corn oil vehicle may also have an effect, though possibly brief, on their metabolomic profile. These effects have the potential to confound results, causing a bin/time point to be considered significantly different from the pre-dose mea-
surement when that significant difference has nothing to do with the injected toxin, but rather the vehicle within which the toxin was administered. To account for this effect, the vehicle control subjects were administered the same vehicle injections without any toxin.

For this data set, the paired t-test as described in Section 2.3 was performed on each of the subsequent time points on the treatment subject group and the control subject group. Any bin/time point that was considered significantly different from the pre-dose measurement in either group was considered to be exhibiting a possible vehicle effect. In the event that a bin/time point exhibits both a toxin response and a vehicle effect simultaneously, an unpaired t-test between the treatment and control groups at that bin/time point was performed. If there was a significant difference (a=0.01) between the treatment and control groups at that bin/time point, the bin/time point was not removed because there appeared to be a toxin response beyond the vehicle effect. If there was not a significant response, the bin/time point was removed from further analysis.

2.5 Grouping of Similar Bins

Once differentially responsive bins have been identified, it is useful to determine which are reacting in a similar fashion. The output from the aforementioned statistical tests can be used to group together bins that behave similarly. This is useful in identifying multiple bins that represent the same compound, and can also function as a first step in identifying biomarkers. Once redundant bins have been identified, one of them can be selected as a representative.

The p-values associated with each bin/time point combination can be abstracted to define a pattern for each bin. For each bin/time point combination, an 'N' can be assigned for a non-significant change from pre-dose. If a significant change was detected, then a designation of 'S' can be used along with a sign indicating the direction of the change (i.e. a positive significant change is designated '+S'). Applying these rules to each bin, provides a response pattern for each bin. For example, a pattern of N +S +S N, indicates no significant difference from pre-dose at 24 hours, a positive significant difference at 48 and 72 hours, and no significant difference at 96 hours. This describes a response to the toxin at 48 and 72 hours, followed by a recovery.

3 Results

The difference between the three methods was analyzed by comparing the responsive bin/time points produced by each method. Table 1(a) shows that the paired and unpaired method results in approximately the same number of responsive bin/time points. The two-fold method found fewer responsive bin/time points. Table 1(b) shows that

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U</td>
<td>F</td>
</tr>
<tr>
<td>38</td>
<td>45</td>
<td>10</td>
</tr>
</tbody>
</table>
(a) Cardinality

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P ∩ U</td>
<td>P ∩ F</td>
<td>U ∩ F</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
(b) Intersection

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>U</td>
<td>F</td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
(c) Set difference (row - column)

Table 1. Comparing the differentially responsive bin results from the paired t-test (P), unpaired t-test (U), and two-fold test (F).

![Figure 1. Example of time point removed due to vehicle effect.](image)

The p-values for the paired t-test of the 24 hr time points in both the treatment group (Figure 1(a)) and the control group (Figure 1(b)) were less than the adjusted alpha of 0.0025. The p-value for the unpaired t-test between the treatment and control groups was 0.534.

there was considerable overlap between the paired and unpaired methods, and that there were only 5 bin/time points that all three methods selected. Table 1(c) shows that both
the unpaired and paired methods select bin/time points that
the other method does not consider significant. These
differences are examined in the discussion section.

Of the entire set of 1024 bin/time points, 3.4% were
identified as vehicle effects where both the treatment group
and control group’s time point measurement were signifi-
cantly different from the pre-dose measurement. Of these
35 bin/time points, 62.9% were removed from analysis due
to vehicle effect, (example in Figure 1). The remaining
37.1% were not removed because the treatment group was
significantly different from the control group, suggesting a
response above any vehicle effect.

Grouping each bin into distinct groups based on their
pattern produced 9 patterns (shown in Table 2). Approxi-
mately 80% of the bins do not contain one time point that
is deemed significantly differentiating (pattern: N N N N).
The remaining 20% were grouped into 8 patterns.

4 Discussion

4.1 Comparison of Methods

The fold test performed as expected. Only the most ex-
treme differences between the pre-dose and later time point
measurements were selected. The fold test had no sensitiv-
ity for consistency and had no ability to attribute statistical
significance to any selected bin/time points.

The unpaired t-test was able to reward consistency of
the group and attribute statistical significance to the results.
This not only dramatically increased the number of selected
bin/time points as compared to the fold test but it also in-
creased the quality bin/time points as measured by statisti-
cal significance.

The paired t-test was able to provide the same improve-
ments over the fold test while providing a more effect-
ive method of vehicle removal and rewarding consistency
among subjects, not just groups. The situation described in
Section 2.4, where there appears to be a toxin response be-
yond a vehicle effect, is identified by the paired t-test more
often and several instances of the unpaired failing to re-
move a suspected vehicle effect also occur. Another situa-
tion where the paired t-test appears to have an advantage
is when the treatment group’s differences from the pre-dose
measurements vary between positive and negative values.
Here, the control group’s set of differences sum to nearly
0 while the treatment group’s set of differences group do
not. The unpaired t-test may view this as a significant dif-
ference while the paired t-test will look at the highly varied
treatment values and conclude there is nothing statistically
significant. A third example arises when the average dif-
ference between the timepoint of interest and the pre-dose
measurement are small in the treatment and control. The
control group measurements are highly varied, resulting in
an average near 0 while the treatment group measurements
are small, upward, and uniform. The unpaired t-test does
not recognize this consistency across the time points.

4.2 Grouping of Similar Bins

The grouping of similar bins is useful for identifying re-
dundant bins from the same compound. In the simplest
case, if a compound is responding to the treatment at a given
time point, all of the redundant bins should show up as re-
sponsive. This is oversimplified because different metabo-
lites can affect the same bin. Thus making the problem of
identifying redundant bins difficult. However, if redundant
bins can be identified, one representative bin can be used
for further analysis. To test, five compounds were located in
the spectrum: citrate, Alpha-ketoglutarate (α KG), benzoyl-
glycine (hippurate), phenylacetylglucose (PAG), and crea-
tine.

Citrate should be found at bin 198 and 194. Bin 194
was grouped as a member of the pattern N - S N N (i.e. it
indicates a negative response at 48 hours). Bin 198 was
identified as N N N N, indicating no response; however,
it's p-value at 48 hours was 0.0036, just above the adjusted
alpha value of 0.0025. This was due to our conservative
Bonferroni correction. Thus, a less conservative multiple
test correction is needed. Several less conservative meth-
ods of adjusting the p-value for multiple testing have been
discussed as alternatives to Bonferroni and may be appro-
priate methods in attempting to decrease the number of false
negatives (e.g. the 48 hr time point of bin 198) [4, 11].
The compounds α KG, PAG, and creatine also fall victim to
the correction. The compound hippurate has three bins that
should be responding in unison: 47, 55, and 157. All of
these bins were identified as having a pattern of N - S N N.

4.3 PCA of Resulting Bin/Time Points

To assess the results, the significant bin/time points
were visually and quantitatively compared to the significant
bin/time points using principal component analysis and J2
criterion. The J2 criterion calculates the degree of which
samples are well clustered around their means [12].

Figure 2 shows a visual comparison of the results. The
5 Conclusion

While all 3 methods isolated statistically significant bins, or in the case of the fold test informative bins, there were some clear differences. The paired t-test with vehicle effect removal appeared to be the more complete of the three tests, having distinct advantages over the other 2 methods for which examples have been shown. With a less conservative correction for multiple hypothesis testing, the paired t-test in combination with our vehicle removal technique, is a viable option for isolating possible biomarkers with some statistical significance. These possible biomarkers can then be effectively grouped by patterns produced and compared with known metabolites. Our protocol showed agreement between the results and the metabolites known to be affected by the toxin.

6 Acknowledgment

This work was supported in part by grants from the Air Force Research Laboratory, ManTech Environmental Technology, Inc. (Contract No. ManTech/WBI-002), Alion Science and Technology (Contract No. SUB1174146RB), and The Henry M Jackson Foundation for Advancement of Military Medicine, Inc. (Contract No. 132633).

References