Self-similarity in NMR spectra: An application in assessing the level of cysteine

Yoon Young Jung*, Youngja Park†, Dean P. Jones†‡, Thomas R. Ziegler‡, and Brani Vidakovic§

January 15, 2007

High resolution of NMR spectrocopic data of biosamples are a rich source of information on the metabolic response to physiological variation or pathological events. There are many advantages of NMR techniques such as the sample preparation is fast, simple and non-invasive. Statistical analysis of NMR spectra usually focuses on differential expression of large resonance intensity corresponding to abundant metabolites and involves several data preprocessing steps. In this paper we estimate functional components of spectra and test their significance using multiscale techniques. We also explore scaling in NMR spectra and use the systematic variability of scaling descriptors to predict the level of cysteine, an important precursor of glutathione, a control antioxidant in human body. This is motivated by high cost (in time and resources) of traditional methods for assessing cysteine level by high performance liquid chromatograph (HPLC).

1 Introduction

During the last decade, metabolomics has provided new opportunities to investigate complex dietary and nutritional questions by applying quantitative methodologies to information-rich profiles of dietary chemicals and their metabolites [11, 12]. NMR spectroscopy has been utilized in exploring

*Department of Statistics, Ewha Womans University, Seoul, Korea

†Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, GA 30322

‡Center for Clinical and Molecular Nutrition, Department of Medicine, Emory University, Atlanta, GA 30322

§Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30322

physiological variations in macronutrient metabolism and has shown to be a fast, simple, and non-invasive method for "fingerprinting" of metabolic compounds. These advantages, however, are offset by complex spectral representations. For example, ^{1}H NMR measures proton (hydrogen) signals from all plasma metabolites. However almost every molecule in plasma contains multiple protons which results in overlapped and complex spectra. For this reason advanced signal processing techniques are increasingly used to analyze the NMR spectra.

Statistical analysis of NMR spectra traditionally focuses on differential expression of large resonance intensity corresponding to abundant metabolites and involves several data preprocessing steps such as baseline correction, peak alignment and normalization. These preprocessing steps are not perfect and often lead to ambiguities and information loss. Researchers have developed statistical methods and multidimensional NMR techniques that identify important metabolites contributed to toxicological and pathophysiological conditions or treatments by comparing the spectra.

A previously unaddressed question is what is the interplay of metabolites with small "energies" in spectra, how they "communicate", and what is the position-lagged correlation of their spectral contents. In contrast to exploring a few large resonance intensity in the spectra after preprocessing of spectral curves, our analysis focuses on fractal properties of the output signals and regularities of their scalings. An advantage of the proposed method is that it does not require complicated preprocessing steps.

Formally speaking, we treat the spectra as functional data and employ functional data analysis (FDA) techniques [25, 26] for extracting spectral functional components characterized by treatments, subject blocking, and maybe some other factors of underlying experimental design. At the same time, we employ multiscale analysis that provides the tools for assessing the scaling of derived functional components which is an intrinsic property of functional observations and deriving descriptors that can be connected to energy activity of all metabolites in the spectrum.

Since wavelets and wavelet-based methodology offer domains in which the variation of a function can be explored at layers of nested scales, with the possibility of controlling the total energy allocated to each resolution level [21, 27, 28, 29, 36], we perform the multiscale analysis of spectral components in the wavelet domain.

Traditional applications of wavelets in NMR spectroscopy are for dimension and noise reduction. The statistical foundation of these methods is due to David Donoho and his coauthors. It is interesting that one of the first template functions to test performance of wavelet methods was a caricature

of an NMR spectrum, the function `bumps`, [6, 7]. More recent publications describe emerging methods in NMR data processing and some novel uses of wavelets in NMR processing [13, 14, 17, 35].

In the following, we suggest new methods to extract biologically significant information about the interactions of metabolites and their relationship with biological functions that is contained in NMR spectra by using scaling measures computed from wavelet coefficients. The method does not require preprocessing. As an application, we use the systematic variability scaling descriptors to predict cysteine concentrations from spectral data in which cysteine itself cannot be detected because its concentration is below detection limits. The measurement of plasma cysteine requires special blood collection techniques, and analysis by HPLC requires the long sample preparation time before actual HPLC running. On the other hand, NMR does not require any special blood collection technique or complicated sample preparation. NMR running time is much shorter than that of HPLC. The prediction of concentration of cysteine through multiscale analysis thereby could save the cost and time of analysis compared to other methods.

To focus on the effect of diurnal time on the scaling coefficient, we use functional repeated measure block design, a statistical design technique in which the observations are spectra. The influence of subjects on scaling index is not of interest and they serve as blocks. The scaling is assessed from the functional ANOVA components corresponding to the treatment effect of interest.

The paper is organized as follow. In Section 2 we describe the methodology of functional data analysis and wavelet-based assessment of scaling. The application of the methodology to assess the level of cysteine in blood plasma is provided in Section 3. Remarks and conclusions are given in Section 4.

2 Methodology

In this section we describe data and statistical methodology utilized in the analysis. Some technical details about the methods are deferred to Appendix. Our methodology is supported by two statistical techniques – (i) functional data analysis (FDA) and (ii) scaling assessment. Both techniques utilize multiresolution tools (wavelets) in their implementation.

2.1 Data

Human plasma samples were collected hourly over a 24 hour period (from 8:30 am to 8:30 am) from nine healthy adults under a protocol approved

by the Emory University Institution Review Board. Subjects were given standardized, nutritionally balanced meals to provide caloric intake at estimated basal energy expenditure + 40% (derived from the Harris Benedict equation) and adequate protein at 15% of total energy intake. Total energy intake was provided as 15% protein (based on 0.8 gm protein/kg/day), 30% fat, and 55% carbohydrate. Subjects consumed each meal within 45 minutes (i.e., breakfast from 9:00–9:45 am, lunch from 1:00–1:45 pm and dinner from 5:00–5:45 pm) and the snack within 15 minutes (9:00–9:15 pm). Meals were provided as a percentage of total energy intake as breakfast (30%), lunch (30%), dinner (30%), and an evening snack (10%). Water was provided ad libitum throughout the admission. Activity (if desired) was confined to walking in the Emory General Clinical Research Center (GCRC) unit and only within the following time frames (after the hourly blood draw): 10:00–10:30 am, 12:00–12:30 pm, 14:00–14:30 pm, 16:00–16:30 pm, 18:00–18:30 pm and 20:00–20:30 pm. Otherwise, patients remained in their room, either lying in bed or sitting in a chair. Blood samples were collected via a heparinized butterfly needle and syringe. Tubes were spun in a microcentrifuge at 14,600 g for 30 seconds at room temperature to remove blood cells. The entire sampling procedure was less than 2 minutes for each hourly sample. Plasma samples were maintained on ice until convenient for transfer to a $-70°$C freezer.

Plasma samples were thawed and a 600 ml porions are mixed with 66 ml of deuterium oxide (D_2O) containing DSS [3–(trimethylsilyl)–1–propanesulfonic acid sodium salt ($C_6H_{15}NaO_3SSi$, 1% w/w)]. ^{1}H NMR spectra were measured at 600 MHz on a Varian INOVA600 spectrometer with water presaturation at 25°C. The samples were maintained at 25°C in the magnet at least 10 minutes before measurement in order to ensure temperature stability. NMR spectra were measured with 64 scans into 16,384 data points over a spectral width of 6600.7 Hz, which resulted in an acquisition time of 2.55s per sample (d1=0, pulse=5ms, presaturation=1s, acquisition=1.5s). To check the reproducibility of the NMR analysis, spectra were acquired on identical samples at multiple time points (1.5h, 3h, 4h and 6h). The correlation coefficients of spectra were 0.96, 0.93, 0.97, 0.97.

Figure 1 shows the ^{1}H FT(Fourier transform)-NMR spectra that measure physiologic variations in macronutrients in human plasma. The columns correspond to individuals while the rows represent time of sampling. For each subgraph the horizontal axis is expressed as ppm (part per million) and ranges between 10 and 0, while the vertical axis gives an artificial magnitude adopted for comparison. Although the range of spectra for all patients is the same, note that the individuals 5,8, and 9 have "richer" spectra which

can be attributed to varying rates of absorbtion, distribution, metabolism, and excretion.

The level of cysteine was measured by HPLC with fluorescence detection of dansyl derivatives, [18]. This method requires two days for processing and cysteine derivatition. Furthermore, HPLC running time took for 1 hour to evaluate the cysteine concentration. In this study, we extract the Hurst exponent from NMR spectrum to predict the level of concentration, although cysteine concentration of human plasma cannot be directly observed in the NMR spectrum. The acquisition time for one NMR spectrum is less than 15 minutes per sample. The preparation of sample for NMR is less than 5 minutes. Total time for NMR data collection per sample is less than 20 minutes. Comparing the NMR method to the HLPC method to extract the level of cysteine, the NMR approach of human plasma is much simpler and requires much less time than HPLC.

2.2 Assessing the Spectral Components via a Functional Design

Given that our observations are functions (spectra) observed under different conditions from different individuals, we employ functional data analysis (FDA) to estimate, separate, and test spectral components corresponding to different experimental factors.

FDA is a recent statistical methodology [25, 26] which treats functions, images, n-dimensional continuum objects as observations and performs standard statistical inference tasks (estimation, testing, classification) on such functional observations. Unlike the traditional statistical procedures that treat functional observations as multivariate data, the FDA makes inference on functions directly. For instance, estimating population mean function $\mu(\cdot)$ or testing that it is equal to 0, based on the sample of functional observations, are typical inferential tasks in FDA.

The traditional ANOVA statistical technique explores the scalar data which are obtained under one or more (fixed- or random-level) experimental treatments. It estimates the population treatment means and tests their equality. The functional ANOVA (FANOVA) assumes that observations are functions, in our case NMR spectra and performs equivalent statistical inference.

It is assumed that the experiment in which the NMR spectra are measured is performed under p different treatments. Let b represent the number of subjects observed under the treatment i, where $i = 1, 2, \ldots, p$. The total sample size is $n = pb$. It is of interest to estimate and test the functional

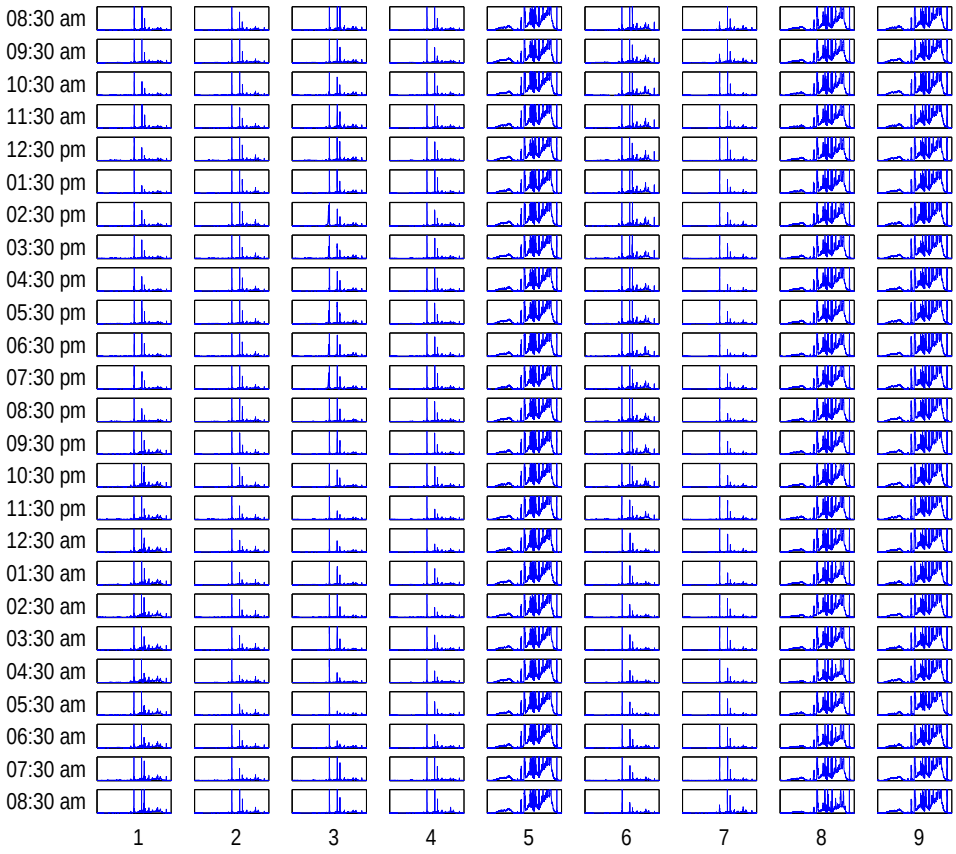

Figure 1: The ^{1}H FT-NMR spectra of human plasma samples of nine patients for 25 time points. The columns correspond to individuals while the rows represent time instants. For each subgraph the horizontal axis is chemical shift expressed as *ppm* unit and ranges between 10 and 0, while the vertical axis gives NMR spectral intensity.

contributions of the treatments to the spectral output. In the FANOVA jargon, the observed spectra $s_{i\ell}(\delta)$ can be represented as superposition of 4 functions, $\mu(\delta)$ which is a common part, $\alpha_i(\delta)$ which is the contribution from the treatment i, $\beta_\ell(\delta)$ which is the contribution of the subject ℓ (blocking variable), and the error term $\epsilon_{i\ell}(\delta)$. This can be expressed as

$$s_{i\ell}(\delta) = \mu(\delta) + \alpha_i(\delta) + \beta_l(\delta) + \epsilon_{i\ell}(\delta),\ i = 1, \ldots, p,\ \ell = 1, \ldots, b. \quad (1)$$

Here the variable δ represents chemical shift expressed in *ppm* unit. It is assumed that for each fixed δ, $\epsilon_{i\ell}(\delta)$ are independent normal random variables with mean zero and common variance σ^2. A rigorous way to introduce (1) involves random fields and is provided in the Appendix. In simple terms, each observed spectra is a sum of the mean spectra, treatment effect component, subject effect component, and an error attributed to the measurement procedure and uncontrollable fluctuations. The validity of this analysis is contingent on precise alignment of spectra across times and subjects since the estimators involve averaging the observed functions.

In the context of our data, the repeated measures are calibrated so that measure 1 corresponds to 8:30 am. The each subsequent measure is 1 hour apart from the previous one, so that 25th measurement corresponds to 8:30 am of the following day, i.e., $p = 25$. A total of nine individuals are followed through all the treatment times. This study is not interested in differences among the individuals; thus, the subjects are considered as a blocking factor.

Our major interest is the hourly variation of nutritional metabolomics. We first separate the observed spectra as the sum of the mean spectra $\hat{\mu}$, time effects $\hat{\alpha}_i$ and the subject effects $\hat{\beta}_j,\ j = 1, \ldots, 9$. The estimates of the time effects are shown in Figure 2. The mean hourly contributions to the spectra are estimated as in the Appendix. Note that $\hat{\alpha}_1$ and $\hat{\alpha}_{25}$ (upper left and lower right panels numbered as panels 1 and 25 respectively) are similar in size, as expected. Note also that at some hours there is increased expression of dominant metabolites compared to the average (panels 9:30 am, 3:30 pm, for example), while for some other times (panels 11:30 pm, 2:30 am, for example) the expression decreases.

The estimators of the block effects, i.e., the mean contributions to the spectra by each subject, are given in Figure 3. Although these estimators are not of interest in assessing the treatment means, their inequality is desirable since it shows that our model accounts for the variability among the subjects contributing to the precision of the assessment of the differences between the treatment means. This is a universal benefit of blocking in all experimental designs where blocking is possible. As evident in Figure 3, the

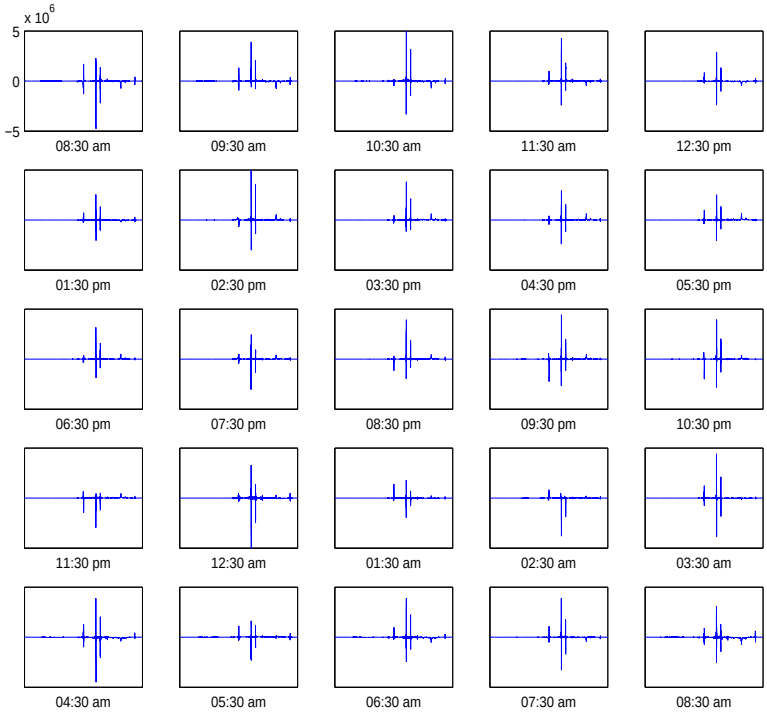

Figure 2: Estimators of the time effects $\hat{\alpha}_i$ for 25 times. The upper left panel shows $\hat{\alpha}_1$ while the lower right panel shows $\hat{\alpha}_{25}$.

mean contribution of each subject shows a different pattern. For example, subjects 5, 8, 9 show increased expression of dominant features compared to the average and subjects 1, 3, 4, 6, 7 show a decrease in the expression.

The FANOVA tests (details in Appendix) showed that both null hypotheses $H_0' : \alpha_1(\delta) = \alpha_2(\delta) = \cdots = \alpha_{25}(\delta) = 0$ and $H_0'' : \beta_1(\delta) = \beta_2(\delta) = \cdots = \beta_9(\delta) = 0$ were rejected with p-values of 0.0001 and 10^{-6}, strongly suggesting that the mean functional contributions to the spectra are non-zero functions and vary significantly with δ, time and subjects.

Although these results are important, their practicality is limited. Other relevant but exogenous parameters influence the functional estimators. This motivated us to summarize the functional components of spectra via scalar descriptors with realistic physical interpretation, as described in the following Section.

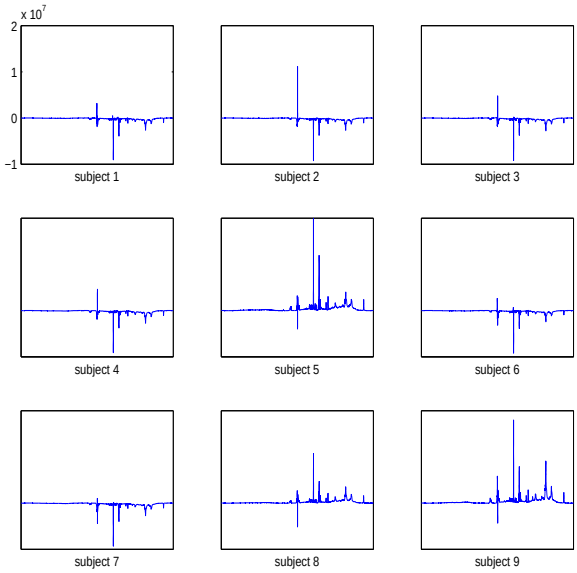

Figure 3: Estimators of the block effects for the 9 individuals.

2.3 Scaling of Spectral Components

Most high frequency biomedical measurements exhibit scaling. The regular scaling of high frequency data has been used in statistical modeling tasks involving regression, classification, and experimental design [23, 30]. The scaling is described as regular decay of the energy in signals when this energy is progressively measured at scales for which the resolution is increasing. More precisely, the regular scaling is described by a linear relationship between the log-scale (scale defined as reciprocal of the frequency) and log-average-energy within the scale. The slope of this linear relationship uniquely determines the Hurst exponent, H, a constant between 0 and 1 that characterizes the scaling. For example, white noise is characterized by $H = 1/2$, all turbulent signals have $H = 1/3$, and "random DNA walk" corresponding to non-coding parts of human DNA have $H \approx 0.6$. Most neural, ocular, and many other physiological high-frequency measurements scale and this scaling has been used as a statistical summary of the outputs. Theoretical details describing the estimation of the Hurst exponent are given in the Appendix.

Next, we briefly discuss the rationale for use of scaling to summarize NMR spectra. When trends in data are irrelevant and when smoothing does not make sense, scaling analysis of row noisy measurements may yield

useful information. For example, in the study on links between dynamics of change of pupil diameter and ocular pathologies, Shi et al. [30] argue that trends in high frequency measurements (> 200 Hz) are irrelevant since they could be affected by the change of environmental light intensity, clearly not related to the pathologies. However, the scaling in these measurements assessed by the Hurst exponent carries discriminatory information about the eye pathologies. Similarly, traditional analysis of ^{1}H NMR spectra of human plasma can be considered irrelevant to the plasma cysteine concentration because the dominant spectral measurements are insensitive to directly detect cysteine.

Another important property of scaling is that it is invariant with respect to shift/scale of the spectra, and does not require data preprocessing steps such as baseline correction, peak alignment and normalization, unless performed on one of the FANOVA components. The consequence is that the estimator of the Hurst exponent is robust with respect to changes in a few dominant resonance intensities corresponding to expressed metabolites or marker chemicals.

If the signal has high Hurst exponent, the autocorrelations (correlations between the signal and its shifts) are strong, signifying considerable internal regularity. On the other hand, the signals with low Hurst exponent exhibit intrinsic irregularity and antipersistency. In terms of NMR spectra, spectra with a larger Hurst exponent would possess more internal regularity and autocorrelation. This informally means that metabolites communicate more when the Hurst exponent is higher and that they are more "co-expressed."

The signature of scaling in the NMR spectral data is visible in a logscale diagram (Figure 4). The horizontal axis represent diadic scales in which the largest number (13 in Figure 4) corresponds to the Nyquist frequency i.e., the finest discernable scale. Note that the slope of the graph in the logscale diagram corresponding to scales 10, 11, 12, and 13 differ from the slope corresponding to scales that are below 10. This is an artifact of preprocessing of spectra. The low scales of logscale diagram (2-5) are not of interest in assessing the scaling since their values are affected by global energy of the spectra and a few energetic peaks. The region with fairly constant slope in the middle of the diagram is used to calculate the Hurst exponent.

We estimated the Hurst exponent from each of the spectra normalized by subtracting the mean estimator, $\hat{\mu}(\delta)$. The rationale is to inspect the scaling of the functional contributions for time and subject only. From $s^*_{i\ell}(\delta) = s_{i\ell}(\delta) - \hat{\mu}(\delta),\ i = 1, \ldots, 25,\ \ell = 1, \ldots, 9$, the matrix of Hurst exponents, $\{H_{i\ell}\}$ is obtained. Assume that each $H_{i\ell}$ can be decomposed to a "grand mean" H', effect of time H''_i, effect of subject H'''_ℓ, and an error

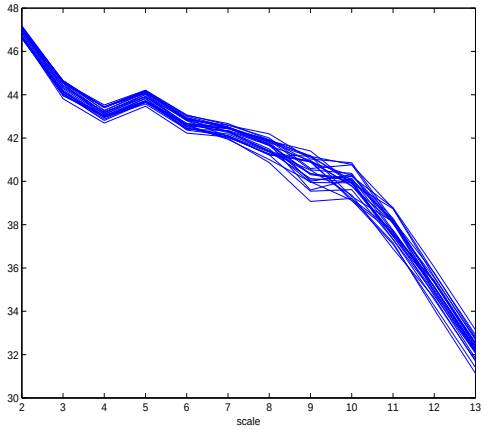

Figure 4: An average logscale diagram for each of 25 times.

$\epsilon_{i\ell}$ in the form of a block-design model

$$H_{i\ell} = H' + H_i'' + H_\ell''' + \epsilon_{i\ell},\ i = 1, \ldots, 25,\ \ell = 1, \ldots, 9.$$

A standard analysis of this model yielded that the hypothesis $H_0 :\ H_i'' = 0, i = 1, \ldots, 25$ was rejected (p-value 0.0013); that is, there is a significant difference in scaling with respect to times. The hypothesis $H_0 :\ H_\ell''' = 0, \ell = 1, \ldots, 9$ was rejected as well (p-value < 0.0001), and a significant difference in scaling is attributed to subjects. This is expected and justifies the blocking. We note that if this blocking was omitted, i.e., if $H_{i\ell}$'s are analyzed by one way ANOVA,

$$H_{i\ell} = H' + H_i'' + \epsilon_{i\ell},\ i = 1, \ldots, 25,\ \ell = 1, \ldots, 9,$$

the hypothesis $H_0 :\ H_i'' = 0, i = 1, \ldots, 25$ was not rejected, in fact unaccounted variabilities among the subjects masked the variability in times.

Figure 5 shows the hourly variations of Hurst exponent, estimated from the FANOVA components corresponding to the time effects α_i, as in Figure 2). Since α_is are obtained by manipulating spectra, the alignment is necessary (e.g., common average spectra is subtracted). We argue that even if the alignment is not perfect and a few big peaks result from a misalignment, the scaling is not affected if robust measures of average level energies are used, as proposed in [33].

The left panel shows the average Hurst exponent by the hour, while the right panel shows a compass-plot of the truncated average Hurst exponent. It is noticeable that H values tend to be higher in the afternoon/evening and

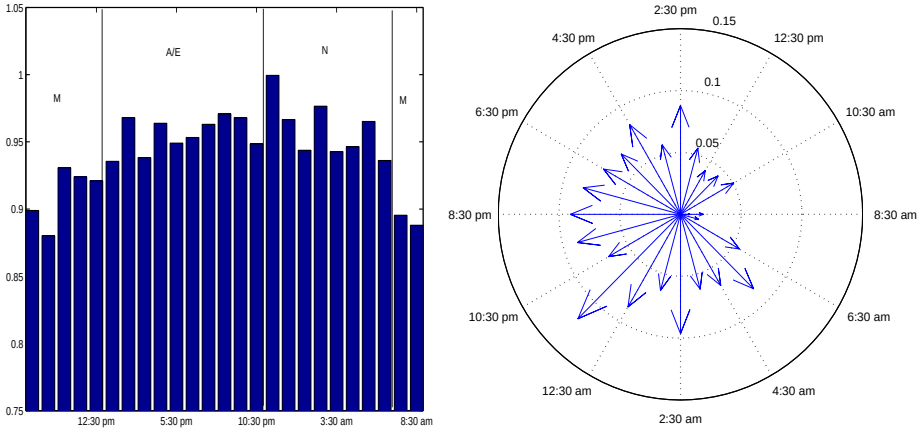

Figure 5: Hourly variations of Hurst exponent as bar plot (left) and as compass plot (right).

tend to be lower in the night to morning. This indicates that the metabolites have more tendency to be co-expressed in the late afternoon than in the morning. The three classes of time of day (morning, afternoon/evening, night) we used are from the previous PCA (Principal Components Analysis) results of the data [22].

2.4 Assessing the level of Cysteine

Cysteine (Cys) is an amino acid used for protein synthesis as well as many other metabolic functions. Therefore, metabolic changes could potentially serve as a biological response indicator of plasma cystaine. This suggests that scaling measure of NMR spectra of human plasma could be useful to assess the level of cysteine.

Cysteine is obtained directly from the diet and also from the essential amino acid, methionine (Met), which is metabolized in individuals by the transulfuration pathway to form Cys [16]. In addition to use in the primary sequence of most proteins, both Met and Cys are required for other metabolic functions. Met is converted to S-adenosylmethionine, which is used for methylation reactions [4] for structural and functional modifications of proteins, RNA and DNA, as well as synthesis of phospholipids and signaling molecules. The carbon skeleton of Met is also used for biosynthesis of polyamines, which are required for cell division and cell growth [37]. Cys is used for biosynthesis of glutathione (GSH), coenzyme A, taurine and sulfate [32]. GSH functions in redox regulation [18] and detoxification of

oxidants and reactive electrophiles [19]. Coenzyme A is central to fatty acid metabolism and the citric acid cycle; taurine is utilized for bile acid synthesis and osmotic regulation [15]; sulfate is used as a structural component of oligosaccharides [34], transport of steroid hormones [31] and detoxification of foreign compounds [20]. Both are required for physiologic processes in addition to maintenance of protein synthesis and nitrogen balance.

Accordingly, Cys could have a central role in controlling metabolism. Consequently we tested the association of the Hurst exponents of NMR spectra with a quantitative measures of Cys in simultaneously collected samples to determine whether a useful estimate of plasma Cys could be derived from the metabolic spectrum.

Figure 6 shows the plot of the hourly variation of the average Cys level with the average Hurst exponent and the associated scatter plot. The biological implication of co-behavior pattern of the Cys level and the scaling measure reveals that we can make predictions of Cys level based on the Hurst exponents of ^{1}H NMR spectra. This means that, in principle, we can use ^{1}H NMR spectra for nutritional assessment, i.e., we can assess Cys levels even though Cys is not directly detected in the sample.

The rationale is the following. When Cys level is high, the major metabolic pathways producing different metabolites are well regulated. The links between metabolites are strong in the sense that there is required coordination of metabolism of lipids, carbohydrates, and proteins. On the ^{1}H NMR spectra, this well regulated link results in a more regular appearance. Some portion of this regularity is likely to be due to multiple signals arising from the same chemicals, especially among the metabolites not so distant in the chemical shift. This regularity is properly sensed and assessed by wavelet spectra and is measured by Hurst exponent. The higher exponent corresponds to more regulated spectra which is linked to the increased level of Cys.

3 Conclusions

NMR spectroscopy of human plasma and urine is attractive because it requires minimal sample preparation, has a short run time and provides quantitative spectral information that depends upon intrinsic properties of the biologic molecules. In this study, we performed FDA and scaling assessment of NMR spectra and proposed a means to predict Cys concentration using the scaling in the ^{1}H NMR data.

Such a wavelet-based global spectral analysis can be extended to local

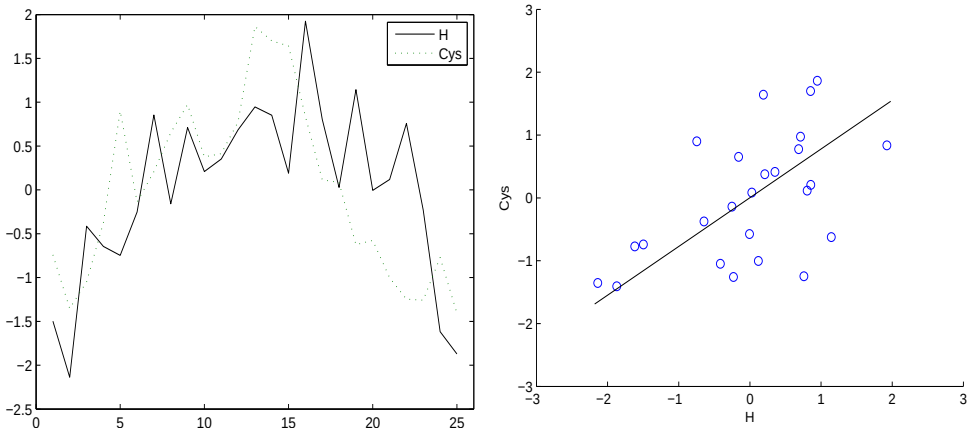

Figure 6: Hourly variation of cysteine level with Hurst exponent and associated scatter plot.

analysis that will identify neighborhoods of metabolites close in chemical shift sense, responsible for particular changes. This analytic approach may be useful for single, high-throughput analysis for chemical assessment of cysteine as well as other key nutrients.

Acknowledgements

This research was supported by NIH Grants DK066008, ES012929 and M01RR00039 at Emory University and NSF Grant 0505490 at Georgia Institute of Technology.

Technical Appendix

In this Technical Appendix we give some details concerning the functional ANOVA and wavelet-based assessment of scaling.

The functional ANOVA (FANOVA) model has been utilized by several authors. For example, Ramsay and his team use the FANOVA to model lip motion from acoustical data [24] and Fan and Lin apply it to test longitudinal effects of business advertisement [10], while Abramovich et al [1] apply a functional block design on the data coming from sport medicine.

In the FANOVA, the observations $\mathbf{y}$ are modeled as

$$dy_{i\ell}(\mathbf{t}) = (\mu(\mathbf{t}) + \alpha_i(\mathbf{t}))\, d\mathbf{t} + \sigma\, dW_{i\ell}(\mathbf{t}),$$
$$i = 1, \ldots, p;\ \ell = 1, \ldots, n_i;\ \sum_{i=1}^{p} n_i = n, \quad \mathbf{t} \in \mathbf{T} \subset \boldsymbol{R}^{\mathbf{s}},$$

where $\sigma > 0$ is the diffusion coefficient, p and s are finite integers, $\mu(\mathbf{t})$ and $\alpha_i(\mathbf{t})$ are (unknown) s-dimensional mean and treatment effect functions and $W_{i\ell}(\mathbf{t})$ are independent s-dimensional standard Wiener processes. To ensure identifiability of treatment effect functions α_i, it is standardly imposed:

$$\int \left| \sum_i n_i \alpha_i(\mathbf{t}) \right| d\mathbf{t} = 0. \tag{2}$$

It is understood that the observations $\mathbf{y}$ are taken at a regular grid in s-dimensional space $\mathbf{t}_m = (t_{1,m}, \ldots, t_{s,m})$,

$$t_{i,m} = m/N,\ 1 \le i \le s, 1 \le m \le N,$$

and that N is the discretization size.

The standard least square estimators for $\mu(\mathbf{t})$ and $\alpha_i(\mathbf{t})$

$$\begin{aligned} \hat{\mu}(\mathbf{t}) &= \bar{y}_{..}(\mathbf{t}) = \frac{1}{n} \sum_{i,\ell} y_{i\ell}(\mathbf{t}), \\ \hat{\alpha_i}(\mathbf{t}) &= \bar{y}_{i.}(\mathbf{t}) - \bar{y}_{..}(\mathbf{t}), \end{aligned}$$

where $\bar{y}_{i.}(\mathbf{t}) = \frac{1}{n_i} \sum_\ell y_{i\ell}(\mathbf{t})$, are obtained by minimizing the discrete version of LMSSE ([25], p. 141),

$$LMSSE = \sum_{\mathbf{t}} \sum_{i,\ell} [y_{i\ell}(\mathbf{t}) - (\mu(\mathbf{t}) + \alpha_i(\mathbf{t}))]^2,$$

subject to discretized version of constraint (2), $(\forall \mathbf{t})\ \sum_i n_i \alpha_i(\mathbf{t}) = 0$.

The fundamental ANOVA identity becomes functional identity,

$$\mathrm{SST}(\mathbf{t}) = \mathrm{SSTr}(\mathbf{t}) + \mathrm{SSE}(\mathbf{t}),$$

with $\mathrm{SST}(\mathbf{t}) = \sum_{i,\ell} [y_{il}(\mathbf{t}) - \bar{y}_{..}(\mathbf{t})]^2$, $\mathrm{SSTr}(\mathbf{t}) = \sum_i n_i [y_{i.}(\mathbf{t}) - \bar{y}_{..}(\mathbf{t})]^2$, and $\mathrm{SSE}(\mathbf{t}) = \sum_{i,\ell} [y_{i\ell}(\mathbf{t}) - \bar{y}_{i.}(\mathbf{t})]^2$. If $\mathrm{MSE}(\mathbf{t}) = \mathrm{SSE}(\mathbf{t})/(n-p)$ and $\mathrm{MSTr}(\mathbf{t}) = \mathrm{SSTr}(\mathbf{t})/(p-1)$, then for each $\mathbf{t}$, the function

$$F(\mathbf{t}) = \frac{\mathrm{MSTr}(\mathbf{t})}{\mathrm{MSE}(\mathbf{t})}$$

is distributed as non-central $F_{p-1,n-p}\left(\frac{\sum_i n_i \alpha_i^2(\mathbf{t})}{\sigma^2}\right)$. For more on functional statistical designs, use of decorrelating transformations (wavelets), and estimation, regularization and testing of design components, see [5, 9, 10, 27, 36].

The self-similarity is an inherent property of many high-frequency functional responses. If the data are self-similar, that is, scale in a regular fashion, then a single descriptor in the form of a Hurst exponent, fully describes the scaling.

There are many ways to assess the self-similarity and to estimate the Hurst exponent. We mention the methods based on contrasting estimators of variability, on various aspects of Fourier and wavelet spectra, methods based on level-crossings, filtering, etc. The literature on this methodology is rich and the monograph [8] provides a comprehensive overview.

We utilized the wavelet-based estimation of the Hurst exponent because of its locality and robustness. A brief description of wavelet spectra follows.

Assume that the signal (^{1}H NMR data) is wavelet-transformed to a range of scales $j_0 \leq j \leq j_1$, where the j_0 scale contains wavelet coefficients corresponding to the coarsest details while the j_1 scale corresponds to the details in the highest resolution. A complete wavelet transformation contains in addition the scaling coefficients, but they play no role in determining the Hurst exponent. The structure of decomposition (details of various scales and scaling exponents) is the embodiment of the multiresolution analysis performed by wavelets. The Hurst exponent quantifies scaling behavior in the data, and classifies these intrinsic autocorrelations as persistent ($H > 0.5$), antiperspirant ($0 < H < 0.5$), or white noise ($H = 0.5$). Researchers realized the practical importance of scaling descriptors and utilized them in the statistical inference tasks, see for instance [30] and references therein. Persistent signals show more visual regularity while the antiperspirant signals exhibit irregular, almost a zig-zag appearance.

The magnitudes of the detail coefficients over all scales are second order descriptors of the process and, in total, constitute a wavelet spectrum of the signal. Formally, within the scale j, averages of squared wavelet coefficients (energies) are found. We denote these averages by $E(j)$. The logarithms of such average energies are proportional to the scale index j and this proportionality is directly linked to the Hurst exponent; that is,

$$\log_2 E(j) = aj + C, \tag{3}$$

where a is the slope, and C is an intercept. The slope a can be expressed in terms of the Hurst exponent H as $a = 2H - 1$, which provides a practical

approach to Hurst exponent estimation. For more information, consult [2, 3, 33].

References

[1] Abramovich, F., Antoniadis, A., Sapatinas, T., and Vidakovic, B. (2004) Optimal testing in functional analysis of variance models. *Int. J. Wavelets, Multiresolution Info. Processing*, 2, 323–349.

[2] Abry, P., Flandrim, P., Taqqu, M., and Veitch, D. (2003) Self-similarity and long-range dependence through the wavelet lens. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, Theory and Applications of Long-Range Dependence. Birkhaüser.

[3] Abry, P., Veitch, D. and Flandrim, P. (1998) Long-range dependence: revisiting aggregation with wavelets. *Journal of Time Series Analysis*, 19(3), 256–266.

[4] Bottiglieri, T. (2002) S-Adenosyl-L-methionine (SAMe): from the bench to the bedside–molecular basis of a pleiotrophic molecule. *American Journal of Clinical Nutrition.*, 76, 1151S–1517S.

[5] Brown, P.J., Fearn, T. and Vannucci, M. (2001) Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96, 398–408.

[6] Donoho, D.L. and Johnstone, I.M. (1994) Ideal special adaptation via wavelet shrinkage. *Biometrika*, 81, 425–455.

[7] Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. of Amer. Stat. Assoc.*, 90, 1200–1224.

[8] Doukhan, P., Oppenheim, G. and Taqqu M.S., editors (2002) Theory and Applications of Long-range Dependence. *Birkhaüser, Boston.*

[9] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* 91, 674–688.

[10] Fan, J. and Lin, S. K. (1998) Test of significance when data are corves, *J. Amer. Statist. Assoc.* 93, 1007–1021.

[11] German, J. B., Roberts, M. A. and Watkins, S. M. (2003) Genomics and metabolomics as markers for the interaction of diet and health: lessons from lipids. *J. Nutr.*, 133, 2078S–2083S.

[12] German, J. B., Bauman, D. E., Burrin, D. G., Failla, M. L., Freake, H. C., King, J. C., Klein, S., Milner, J. A., Pelto, G. H. et al. (2004) Metabolomics in the opening decade of the 21st century: building the roads to individualized health. *J. Nutr.* 134, 2729–2732.

[13] Gunther U.L., Ludwig, C. and Ruterjans H. (2001) Improved automatic structure calculation using wavelet denoised data. In preparation, 2001.

[14] Gunther U.L., Ludwig, C. and Ruterjans H. (2001) WAVEWAT-Imporoved solvent suppression in NMR spectra employing wavelet transforms. Submitted to *J. Magn. Reson.*

[15] Hansen, S.H. (2001) The role of taurine in diabetes and the development of diabetic complications. *Diabetes/Metabolism Research Reviews.*, 17, 330–46.

[16] Hoffer, L.J. (2002) Methods for measuring sulfur amino acid metabolism. *Curr. Opin. Clin. Nutr. Metabol. Care.*, 5, 511–517.

[17] Jeffrey C.H. and Alan S.S. (1996) NMR data processing.

[18] Jones, D.P. (2002) Redox state of GSH/GSSG couple: Assay and biological significance. *Meth. Enzymol.*, 348, 93–112.

[19] Jones, D.P., Brown, L.A. and Sternberg, P. (1995) Variability in glutathione-dependent detoxication in vivo and its relevance to detoxication of chemical mixtures. *Toxicology*, 105, 267–274.

[20] McCarver D.G. and Hines R.N. (2002) The ontogeny of human drug-metabolizing enzymes: phase II conjugation enzymes and regulatory mechanisms. *J. Pharmacol. Exp. Therap.*, 300, 361–366.

[21] Morris, J.S., Brown, P.J., Herrick, R.C., Baggerly K.A., and Coombes K.R. (2006). Bayesian Analysis of Mass Spectrometry Proteomics Data Using Wavelet Based Functional Mixed Models. University of Texas, MD Anderson Cancer Center Department of Biostatistics and Applied Mathematics Working Paper Series.

[22] Park, Y., Kim, S.B., Wang, B., Blanco, R., Le, N. Wu, S., Jonas, C.R., Ziegler, T.R. and Jones, D.P. (2006) Nutritional Metabolomics: Statistical pattern recognition of diurnal variation of macronutrients in human plasma by high-resolution ^{1}H NMR spectroscopy. Technical report at Department of Medicine, Emory University.

[23] Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons M., and Stanley H.E. (1992) Long-Range Correlation in Nucleotide Sequences. *Nature*, 356, 168–170.

[24] Ramsay, J.O., Munhall, K.G., Gracco, V.L. and Ostry, D.J. (1996) Functional data analysis of lip motion. *Journal of the Acoustical Society of America*, 99, 3718–3727.

[25] Ramsay, J.O. and Silverman, B.W. (1997) Functional Data Analysis, Springer-Verlag, New York.

[26] Ramsay, J.O. and Silverman, B.W. (2002) Applied Functional Data Analysis Methods and Case Studies, Springer-Verlag, New York.

[27] Raz J. and Turetsky B. (1999) Wavelet ANOVA and fMRI. In Proceedings of SPIE: Wavelet Applications in Signal and Image Processing VII, 3813, 561–570.

[28] Ruttimann, U.E., Unser, M., Rawlings, R.R., Rio, D., Ramsey, N.F., Mattay, V.S., Hommer, D.W., Frank, J.A. and Weinberger, D.R. (1998) Statistical Analysis of Functional MRI Data in the Wavelet Domain. *IEEE Transactions on Medical Imaging*, 17(2),142–154.

[29] Sajda, P., Laine, A., and Zeevi, Y. (2002) Multi-resolution and wavelet representations for identifying signatures of disease. *Disease Markers*, 18, 339–363.

[30] Shi, B., Moloney, K.P., Pan, Y., Leonard, V. K., Vidakovic, B., Jacko, J., and Sainfort, F. (2006) Classification of High Frequency Pupillary Responses Using Schur Monotone Descriptors in Multiscale Domains. *Journal of Statistical Computation and Simulation*, 76, 431–446.

[31] Song, W.C. (2001) Biochemistry and reproductive endocrinology of estrogen sulfotransferase. *Annals of the New York Academy of Sciences.*, 948, 43–50.

[32] Stipanuk, M.H. and Watford, M. Amino acid metabolism. In Biochemical and Physiological Aspects of Human Nutrition (M.H. Stipanuk, Ed). W.B. Saunders, Philadelphia, 233–286.

[33] Stoev,S., Taqqu, M., Park, C., and and Marron, J.S. (2004) Strengths and Limitations of the Wavelet Spectrum Method in the Analysis of Internet Traffic, SAMSI, Technical Report #2004-8

[34] Sugahara, K. and Kitagawa, H. (2000) Recent advances in the study of the biosynthesis and functions of sulfated glycosaminoglycans. *Current Opinion in Structural Biology*, 10, 518–27.

[35] Vannucci, M., Sha, N. and Brown, P.J. (2005). NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection. *Chemometrics and Intelligent Laboratory Systems*, 77, 139-148.

[36] Vidakovic, B. (2001) Wavelet-Based Functional Data Analysis: Theory, Applications and Ramifications. F3399, Proceedings of The 3rd Pacific Symposium on Flow Visualization and Image Processing, Editor T. Kobayashi, ISBN 1-930746-01-6.

[37] Wallace, H.M. Caslake, R. (2001) Polyamines and colon cancer. *Eur. J. Gastroenterol Hepatol.*, 13, 1033–1039.