This is a special issue on Computational Biology and Data Mining guest edited by professors Don Dong and Ji-Ping Wang:

Don Hong earned his Ph.D. in Mathematics from Texas A&M University and has held a postdoctoral position at the University of Texas-Austin. Before joining Middle Tennesse State University as a Professor of Mathematical Science in 2005, Professor Hong has held Professor/Visiting Professor positions at East Tennessee State University and Vanderbilt University, respectively .

Professor Hong's research areas include Approximation Theory and Computational Science, Splines and Wavelets with Applications, Medical Data Processing and Bioinformatics. He has published two books and more than 40 research articles in leading journals such as Transactions of the American Mathematical Society, Mathematics of Computation, SIAM Journal of Numerical Analysis, and Bioinformatics. He is currently serving on the editorial boards of Journal of Applied Functional Analysis, International Journal of Mathematics and Computer Science, the Current Development in Theory and Applications of Wavelets, Open Proteomics Journal, and the CAAI Transactions on Intelligent Systems. He has also served as a proposal reviewer for National Science Foundation (NSF) and a panel reviewer for National Institute of Health (NIH).

Ji-Ping Wang earned his Ph.D. in Statistics from the Pennsylvania State University right before he joined the Department of Statistics at Northwestern University in 2003 as an assistant professor and was promoted to associate professor in 2009.

Dr. Wang's research areas include mixture models theory and applications, species richness estimation, capture-recapture models for population size estimation, and bioinformatics and computational biology. He has published research articles in journals including Journal of American Statistical Association, Biometrika, Nature, Bioinformatics, and PLoS Computational Biology. He has served as panel reviewer and outside reviewer of the NSF/NIGMS joint program. He served as the chair of the organizing committee of the First Midwest Symposium for Bioinformatics and Computational Biology in 2007. He is now directing the Bioinformatics Core of the NCI funded Physical Sciences Oncology Center at Northwestern University.

International Journal of Mathematics and
Computer Science, 5(2010), no. 2, 65-85

Regularity of Irregularity: Testing for Monofractality by Multifractal Tools

Kichun Sky Lee[1], Jongphil Kim[2], Brani Vidakovic[3]

[1] Department of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive, NW
Atlanta, GA 30332-0205, USA

[2] H. Lee Moffitt Cancer Center & Research Institute
12902 Magnolia Drive Tampa
Tampa, FL 33612, USA

[3] Department of Biomedical Engineering
Georgia Institute of Technology
313 Ferst Drive
Atlanta, GA 30332-0535, USA

e-mail: skylee1020@gmail.com, Jongphil.Kim@moffitt.org, brani@bme.gatech.edu

(Received June 3, 2010, Revised Aug. 8, 2010, Accepted Nov. 1, 2010)

Abstract

Multifractality is a generalization of monofractality in which the regularity attribute has a distribution. Multifractals have become popular as flexible models in modeling real-life data of high frequency. We developed a method of testing whether the data of high frequency is consistent with monofractality using meaningful descriptors coming from a wavelet-generated multifractal spectrum. We discuss theoretical properties of the descriptors, their computational implementation, the use in data mining, the effectiveness in the context of simulations, an application in turbulence, and analysis of coding/noncoding regions in DNA sequences.

Key words and phrases: multifractality, monofractality, wavelets, Hurst exponent, Hölder exponent, bootstrapping, turbulence, DNA sequence.

AMS subject classification: 62J05, 62J07, 62H35, 62P10

ISSN 1814-0424 ©2010, http://ijmcs.future-in-tech.net

1. Introduction

This paper is concerned with assessing the deviation from monofractality in measured high-frequency signals. It has been observed that a wide range of complex structures in nature is characterized by seemingly irregular behavior. Examples of such irregular signals in both time and scale are abundant in medicine, physics, economics, and geosciences, to list a few. Although irregular, such signals can be well modeled by multifractal processes. Concepts of fractal dimension and self-similarity have been used to quantify the multifractal behavior. The key idea is to quantify statistical similarity of patterns at many different scales. The regularity index describes the strength of the similarity. The scaling is usually stochastically complex and may include inhomogeneity of patterns in both time and scale. Multifractal analysis has been developed in order to quantify the irregular scaling [25, 20].

The essence of multifractal analysis is to assess fractal dimensions of self-similar structures with varying regularities and to produce the distribution of indices of regularity, which constitutes the multifractal spectrum (MFS). The MFS describes the "richness" of local singularities in the signal. The multifractal formalism relates the MFS to the partition function measuring high-order dependencies in the data. In recent years, the multifractal formalism has been implemented with wavelets [1, 12]. This approach is very amenable to computation and estimation in practice. The advantages of using the wavelet-based MFS are availability of fast algorithms for wavelet transform, the locality of wavelet representations in both time and scale, and intrinsic dyadic self-similarity of basis functions.

Rigorous mathematical foundations of the multifractal process and wavelet-based approaches have been studied by several researchers [22, 12]. Many applications to dynamics of the multifractal processes [23, 15, 9], such as TCP/IP traffic data and financial data, can be found. In addition, the wavelet-based fractal analysis is a pervasive concept in the medical fields; many medical images, treated as signals, demonstrate a certain degree of self-similarity over a range of scales, driving the development of data mining algorithms based on fractal analysis of those images. A wavelet transform modulus maxima method combined with a multifractal analysis was used to detect tumors as well as microcalcifications [26]. A classification technique based on features derived from the fractal description of mammograms was used [11]. The wavelet-based multifractal discrimination model was proposed to determine ocular pathology based on the pupillary response behaviors exhibited by older adults with and without ocular disease during the performance of a computer-based task [8].

The presence of multifractality in real-life signals is difficult to assess due to finite signal sizes and numerical instability of assessing tools. Veitch and Abry [10] used a collection of regularities on blocks of the signal, and Veitch et al. [9] reviewed the evidence for multifractal behavior of aggregate TCP traffic using wavelet-based logscale diagram. In most of the approaches for the assessment, level-wise analysis of L^p norms of wavelet coefficients was utilized. However, the slopes in this scaling behavior could be misleading because multifractal signals may result in a perfect linear decay of energies. Also the slopes are sensitive to the exponent in the partition function. Extraction of meaningful multifractal characteristics for effectively assessing deviation from monofractality based on the MFS has not received much attention in the literature. The main contribution of this paper is the development of a test for monofractality of a signal based on relevant multifractal descriptors from the wavelet-based MFS. We demonstrate effectiveness of this test in simulations and real-life examples that include turbulence and DNA nucleotide sequences.

This paper is organized as follows: in Section 2, singularity and scaling are discussed in a wavelet context. Also we clarify the notion of deviation from monofractality. In Section 3, we discuss how to compute MFS by using discrete wavelet transform and describe multifractal descriptors in MFS. In Section 4, we propose a bootstrap-based testing procedure to detect deviations from monofractality. Conclusions are provided in Section 5.

2. Monofractality

In this section we examine monofractality of a process by using the properties of singularity and scaling in wavelet transforms. By inspecting decay of wavelet coefficients, we can detect singularity and scaling simultaneously. We will discuss possible deviations from monofractality at the end of this section.

2.1. Singularity and Scaling

A signal, or a process $Y(t)$ is regular if it can be locally approximated by a polynomial. The terms 'process' and 'signal' will be interchangeably used in referring to observed paths of a random process. An irregular signal features local singularities. The singularity behavior of a process $Y(t)$ at time t_0 is characterized by Hölder exponent H_{t_0} (Lipschitz exponent): H_{t_0} is defined as the largest h such that there exists a polynomial P satisfying $|Y(t) - P(t)| \leq C|t - t_0|^h$ for t sufficiently close to t_0. Roughly speaking, saying that $Y(t)$ has exponent h at t_0 means that, around t_0, the process Y is bounded by the curves of $Y(t_0) + C|t - t_0|^h$ and $Y(t_0) - C|t - t_0|^h$ (see Figure 1 for graphical interpretation). If h is close to 0, the wide boundary from the two curves allows for large variations. As H_t approaches 1, the process becomes regular or smooth at the point t.

A process scales if its distributional properties are intrinsically invariant to changes of a scale. A process $Y(t)$ is self-similar with self-similarity index $H > 0$ (H-ss) if $Y(at) \overset{d}{=} a^H Y(a)$. Here $\overset{d}{=}$ denotes equality in all finite-dimensional distributions. An H-ss process with stationary increments exhibits long range dependence (LRD) when $H > 1/2$. A zero mean Gaussian process $B_H(t)$ with stationary self-similar increments is called fractional Brownian motion (fBm) with Hurst exponent H if $B_H(t) \sim N(0, \sigma^2 |t|^{2H})$, and

$$B_H(t+\tau) - B_H(t) \overset{d}{=} B_H(\tau) - B_H(0) \overset{d}{=} \tau^H B_H(1). \tag{2.1}$$

As a fBm with Hurst exponent H, $B_H(t)$ is is sometimes referred as fBm$_H$. As a zero mean Gaussian process, $B_H(t)$ could be alternatively defined via its covariance structure:

$$\mathrm{E}\big[B_H(t)B_H(s)\big] = \frac{\sigma^2}{2}\Big[|t|^{2H} + |s|^{2H} - |t-s|^{2H}\Big]. \tag{2.2}$$

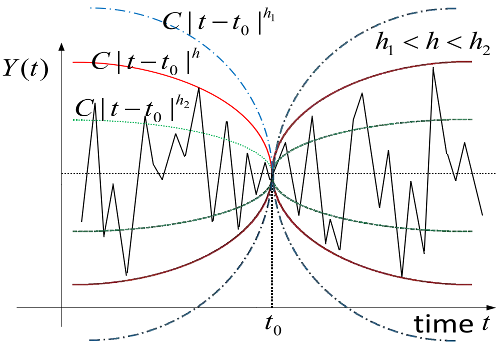


Figure 1: Graphical interpretation of Hölder exponent h of a process $Y(t)$ at a point t_0. Note that smaller h corresponds to a wider boundary within which the process is allowed to vary.

The scaling behavior of a signal is tightly related to singularity of wavelet coefficients [24, 18]. The singularity (Hölder exponent) and the self-similarity (Hurst exponent) are obtainable through multi-scale analysis of wavelet transforms. We will discuss wavelet transforms only at the level needed to introduce the concepts of singularity and self-similarity.

2.2. Wavelets: Detecting Singularity and Scaling

To detect the phenomena of singularity and self-similarity using wavelets, let us consider an L^1-normalized orthogonal wavelet basis comprised of $\psi_{j,k}(t) = 2^j\psi(2^jt-k)$. Wavelet functions $\psi_{j,k}(t)$ are generated from wavelet function $\psi(t)$ by dilation by a scale factor 2^{-j} and translation of $2^{-j}k$. We assume that the $\psi(t)$ has $\mathcal{R}$ vanishing moments: $\int t^r\psi(t)dt = 0, r = 0,\ldots,\mathcal{R}-1$. The coefficients of discrete wavelet transform of a process Y are defined by

$$d_{j,k} = \int_{-\infty}^{\infty} Y(t)\psi_{j,k}(t)dt, \tag{2.3}$$

which carries information on the local difference of the process near to the position k on a dyadic scale j. Let $k2^{-j} \to t$ means that $t \in [k2^{-j},(k+1)2^{-j}[$ and $j \to \infty$. The results of Jaffard [see 24, p. 291] and Gonçalves [19] concern detecting singularity of a signal: if $Y(t)$ is of Hölder exponent H, then

$$|d_{j,k}| = \mathcal{O}(2^{-jH}), \qquad \text{as} \;\; k2^{-j} \to t \tag{2.4}$$

for any wavelet with $\mathcal{R} > H$. This means that the decay of the local differences of a process is related to the singularity of the signal, provided that the decomposing wavelet is more regular than the process.

Wavelets also enable us to detect the self-similarity of a signal. For an H-ss process with stationary increments (H-sssi), it can be shown that

$$d_{j,k} \overset{d}{=} 2^{-jH}d_{0,k} \overset{d}{=} 2^{-jH}d_{0,0}, \qquad \forall k, \tag{2.5}$$

which leads to the same order of $|d_{j,k}|$ as in (2.4). Note that L^2 normalization is used in computations for the sake of computational simplicity, and L^1 normalization is selected in discussions to simplify the rate of the decay: for L^2-normalized wavelets, $d_{j,k} \overset{d}{=} 2^{-j(H+1/2)} d_{0,0}$. The equation (2.5) also serves as a basis for wavelet based estimation of H:

$$\log_2 \mathrm{E}|d_{j,k}|^q = -jqH + C_q, \tag{2.6}$$

where C_q is a constant depending on q, wavelet function ψ, and the magnitude of the signal. The partition function

$$T(q) = \lim_{j\to\infty} (-1/j) \log_2 \mathrm{E}|d_{j,k}|^q$$

measures the scaling of the higher order dependencies and the singularity structure of the process at the exponent q. Since index k is arbitrary, given $d_{j,k}$ within the level j, the partition function does not depend on k. In particular, for the H-sssi signal or the signal with Hölder exponent H, equation (2.6) rewrites as $\log_2 \mathrm{E}|d_{j,k}|^q = -jT(q) + C$, where C is a constant.

A practical estimation of H is based on empirical moments of the wavelet coefficients at dyadic scale j:

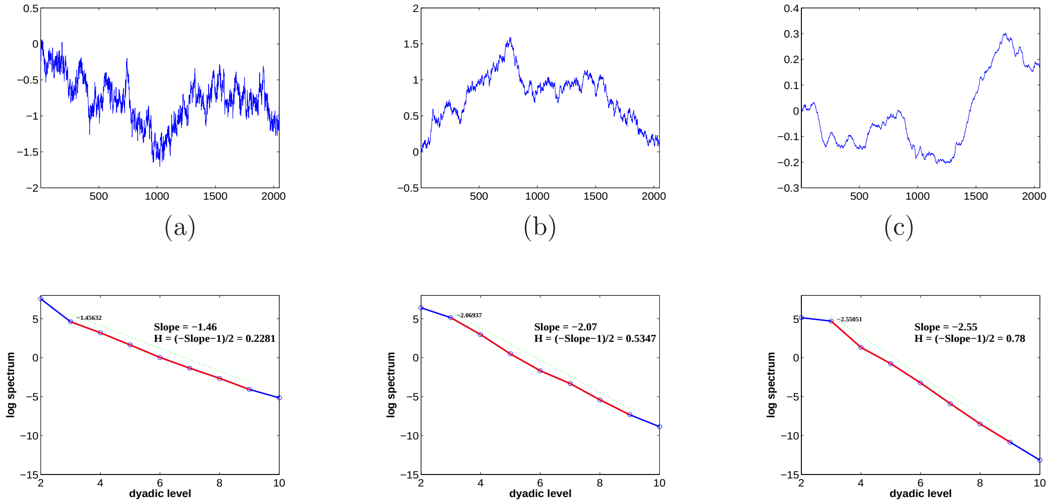
$$\hat{S}_j(q) = \frac{1}{n_j} \sum_k |d_{j,k}|^q,$$

where n_j is the number of $d_{j,k}$ available at dyadic scale j. We assume that the wavelet coefficients are uncorrelated, and hence independent, as has been approximately the case in various contexts (see [17] for a review and [16] for numerical simulations). A plot of the logarithm of the estimates $\hat{S}_j(q)$ against j, $\left(j, \log_2 \hat{S}_j(q)\right)$, is called qth order Logscale Diagram (q-LD): it is also a wavelet spectrum for q=2. These diagrams result in straight lines with slopes of $-qH$, or $-T(q)$, for the fBm$_H$ or signals with Hölder exponent H over the interval. Straight lines in q-LDs provide empirical evidence for monotone scaling. Partition function $T(q)$ is estimated as the slope in the following regression:

$$\log_2 \hat{S}_j(q) = -jT(q) + \varepsilon_j, \tag{2.7}$$

where the error term ε_j is introduced by the moment matching method when the true moments are replaced with the empirical ones. Simple ordinary least square (OLS) is the most convenient choice to estimate the partition function. It is convenient but not correct – according to Abry the regression has to be weighted since the variances of ε_j vary with the level j. Figure 2 shows wavelet spectra from three simulated fBm$_H$ with different slopes under L^2 normalized Haar wavelet. In what follows, the Haar wavelet was used unless mentioned otherwise.

Next, we will analyze the concept of deviation from monofractality by relating it to q-LDs.



(a) (b) (c)

(d) (e) (f)

Figure 2: Simulations of fBm with (a) $H = 0.33$, (b) $H = 0.50$, and (c) $H = 0.80$; in the lower, the corresponding wavelet spectra are shown; as H gets larger, the spectrum line gets steeper.

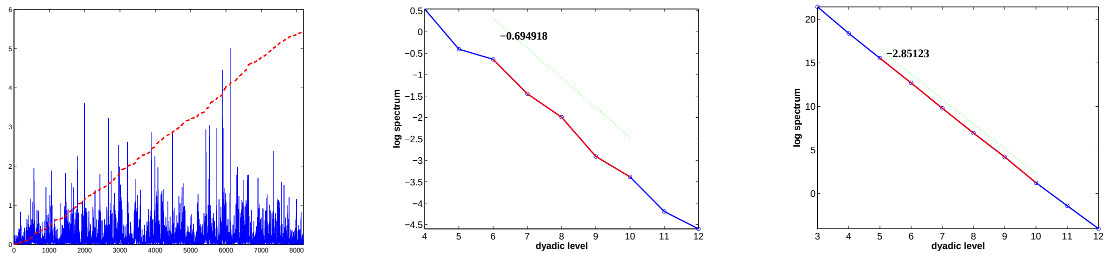
2.3. Deviation from Monofractality

We consider deviations from the linearity of $\log_2 \mathrm{E}|d_{j,k}|^q$ over dyadic scale j as evidence of deviation from monofractality. We make a distinction between deviation from monofractality and evidence for multifractality as the two are not synonymous: for multifractality is richer as a form of scaling behavior associated with randomness and distribution. Multifractal signals possess rich scaling behavior and deviation from linearity of the spectrum is not sufficient to characterize multifractal signals. For instance, multifractal signals can have perfectly linear 2nd order spectra as exemplified later. Distribution of local singularity is required to assess multifractality.

Linear scaling behavior at q-LDs does not necessarily provide evidence for monofractality of monofractality since multifractal signals can show linear scaling behaviors. Figure 3(a) shows a realization of the multifractal wavelet model (MWM) and its cumulative sum [23]. Since the signal, generated to be nonnegative, was regarded to be comparable to fractional Gaussian noise, we took the cumulative sum, which reveals a more stable scaling behavior. Indeed, the Hurst exponent, 0.9255, from the cumulative sum was reasonable. The wavelet spectra of the two signals are shown in Figure 3(b), which reveals that the spectra are linear while the signal is multifractal. MWM is a multifractal extension of traditional fBm models and the MWM synthesis is a multiplicative and coarse-to-fine construction of scaling coefficients for positive and stationary LRD signals. It models the wavelet coefficients of a signal as $d_{j,k} = a_{j,k} u_{j,k}$ with the multiplier $a_{j,k}$ being independent random variables on $[-1, 1]$ and $u_{j,k}$ being an approximation of the signal at dyadic scale j. The simulation in Figure 3(a) was done by β multifractal wavelet model using beta distribution as the multiplier $a_{j,k}$ and

$\mathrm{fBm}_{0.8}$ as the initial approximation of the signal $u_{j,k}$ in the coarsest level. This observation that the multifractal signal has linear spectrum indicates a weakness of spectral slopes in characterization of deviation from monofractality.

Moreover, scaling behavior is sensitive to the exponent q in q-LD. Figure 4 shows different scaling behavior over the exponent for simulated signals from $\mathrm{fBm}_{0.3} + \mathrm{fBm}_{0.7}$. The spectral slopes from 2-LD in 4(a) and 6-LD in 4(b) were -2.2013 and -6.659, respectively, which resulted in different Hurst exponents, 0.6626 and 0.7454. Figure 4(c) plots boxplots of $1,000$ estimators of H for different exponent q. Thus, looking at an isolated q will not be sufficient to confirm monofractality. This also emphasizes the shortcoming of spectral slopes and motivates the MFS to consider different scaling behaviors relevantly. Instead of making scaling inferences on the spectral slopes, one can adopt an empirical approximate of the MFS from wavelet-based partition functions that include information on the spectral slopes. Now, we propose a testing procedure to distinguish signals of monofractality from those that deviate from monofractality based on the MFS.



(a) (b) (c)

Figure 3: (a) One realization (in solid blue) of the multifractal wavelet model overlapping with its cumulative sum (in dashed red) scaled by 1/200; (b) the wavelet spectrum of the signal; (c) the wavelet spectrum of the cumulative sum, which shows a clearly linear scaling behavior.

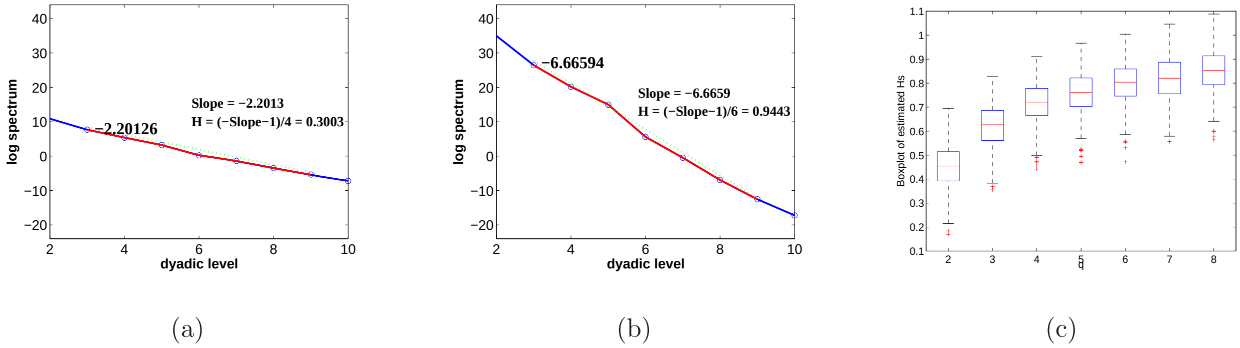
3. Multifractal Spectrum

MFS of a process is a summary of its scaling and singularity properties. Here we describe MFS and discuss how to apply it in measuring deviations from monofractality.

Let us consider the local singularity strength of wavelet coefficients as follows [19]:

$$\alpha(t) = \lim_{k2^{-j}\to t} -\frac{1}{j}\log_2 |d_{j,k}|. \tag{3.1}$$

The local singularity strength measure (3.1) converges to the local Hölder exponent of the process at time t. Small values of $\alpha(t)$ reflect the more irregular behavior at time t. Any



(a) (b) (c)

Figure 4: (a) 2th order logscale diagram for a $\mathrm{fBm}_{0.3} + \mathrm{fBm}_{0.7}$; (b) 6th order logscale diagram for the signal; (c) boxplots of estimated Hurst exponents over the scaling exponent q; the estimated Hurst exponents would not vary with all exponent q if the process was monofractal.

inhomogeneous process has a collection of local singularity strength measures and their distribution $f(\alpha)$ forms the MFS. A direct way to obtain this spectrum is to use the counting technique,

$$f(\alpha) = \lim_{\epsilon\to 0} \lim_{j\to\infty} \frac{1}{j} \log_2\Big(2^{-j}\#\{k \,:\, 2^{-j(\alpha+\epsilon)} < |d_{j,k}| < 2^{-j(\alpha-\epsilon)}\}\Big), \tag{3.2}$$

which captures the limiting frequency of occurrences of a given singularity α and ranges from -1 to 0. It relates to the distribution of the local singularities. Smaller $f(\alpha)$ implies fewer points in t behave with singularity α. If all points in t behave with singularity α^*, then $f(\alpha^*) = 0$.

Although it is feasible to estimate the MFS using (3.1) and (3.2), the method is not practicable due to the computational difficulty of approximating the limit. The multifractal formalism enables MFS f to be calculated by taking Legendre transform f_L of partition function T, $f_L(\alpha) := \inf_q\{q\alpha - T(q)\}$: using the theory of large deviations, one can show that $f_L(\alpha)$ converges to the true MFS $f(\alpha)$ [23, 22]. Because of the log-convex property of the moment generating function and concavity of $T(q)$, f_L is obtained as follows:

$$f_L(\alpha) = q\alpha - T(q) \quad \text{at} \quad \alpha = T'(q).$$

Using the estimator $\hat{T}(q)$ of $T(q)$ in (2.7), for equally spaced q_i with spacing $q_0 = q_i - q_{i-1}$, we estimate $f_L(\alpha)$ as follows [19]:

$$\begin{aligned} \hat{\alpha}_i &= [\hat{T}(q_{i+1}) - \hat{T}(q_i)]/q_0, \\ \hat{f}_L(\hat{\alpha}_i) &= q_i\hat{\alpha}_i - \hat{T}(q_i). \end{aligned} \tag{3.3}$$

Inspecting the MFS of monofractals is beneficial to build intuition on the variety of shapes of the Legendre transform based MFS of different signals.

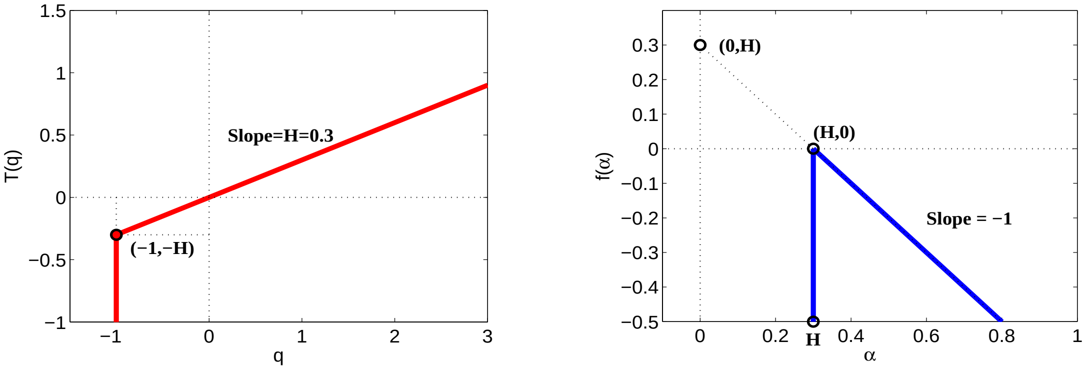
Example 1. For a fBm_H in (2.1), it is easy to show that L^1-normalized wavelet coefficients are

$$d_{j,k} \sim N(0, \sigma_\psi 2^{-2jH}),$$

where σ_ψ is a constant that depends on the wavelet function ψ and the magnitude of the signal, hence the partition function T and the MFS as Legendre transform f_L of T become

$$T(q) = \begin{cases} -\infty, & q \leq -1, \\ qH, & q > -1, \end{cases} \quad \text{and} \quad f_L(\alpha) = \begin{cases} -\infty, & \alpha < H, \\ 0, & \alpha = H, \\ H - \alpha, & \alpha > H. \end{cases}$$

Figure 5 depicts the theoretical partition function and the corresponding MFS for $\text{fBm}_{0.3}$. Note that both are the Legendre transform of each other: the slope -1 and the intercept $(0, H)$ in Figure 5(b) of MFS correspond to a point of $(-1, -H)$ in Figure 5(a) of partition function $T(q)$.



(a) (b)

Figure 5: (a) Partition function $T(q)$ of $\text{fBm}_{0.3}$; (b) MFS $f(\alpha)$ of the signal.

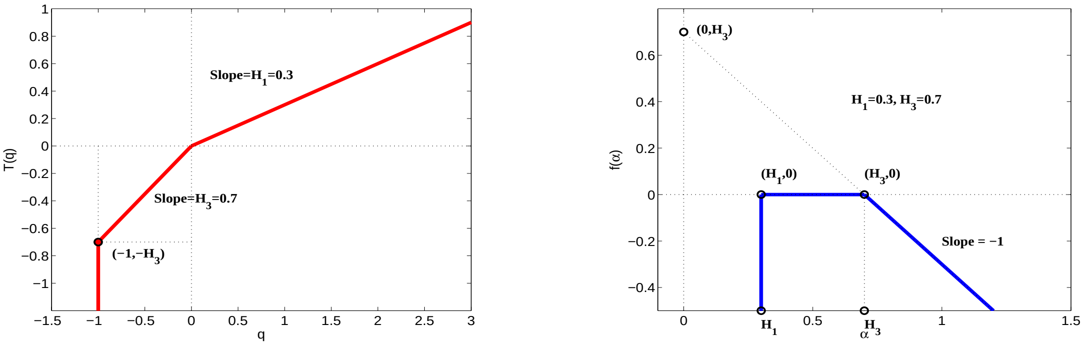
Example 2. Suppose we observe different fBm processes varying with time intervals: $X(t)$ is given by

$$X(t) = B_{H_k}(t), \qquad t \in [t_{k-1}, t_k[,$$

for $k = 1, 2, 3$, and $H_1 < H_2 < H_3$. Since $T(q)$ is determined by the minimum of Hurst exponents when $q > 0$ and by the maximum when $q < 0$ [14], we have

$$T(q) = \begin{cases} -\infty, & q \leq -1, \\ qH_3, & -1 < q < 0, \\ qH_1, & q \geq 0, \end{cases} \quad \text{and} \quad f_L(\alpha) = \begin{cases} -\infty, & \alpha < H_1, \\ 0, & H_1 \leq \alpha \leq H_3, \\ H_3 - \alpha, & \alpha > H_3. \end{cases} \qquad (3.4)$$

The illustration of (3.4) is shown in Figure 6 for $H_1 = 0.3$, $H_1 = 0.5$, and $H_3 = 0.7$. It is worth mentioning that the MFS is flat in the interval between $\min\{H_i\}$ and $\max\{H_i\}$ and that information on regularities between $\min\{H_i\}$ and $\max\{H_i\}$, which is H_2 in this example, is lost in $T(q)$ and $f_L(\alpha)$. We will see in the next section that some of the low-dimensional descriptors of MFS are consistent with the deviation from monofractality.



(a)

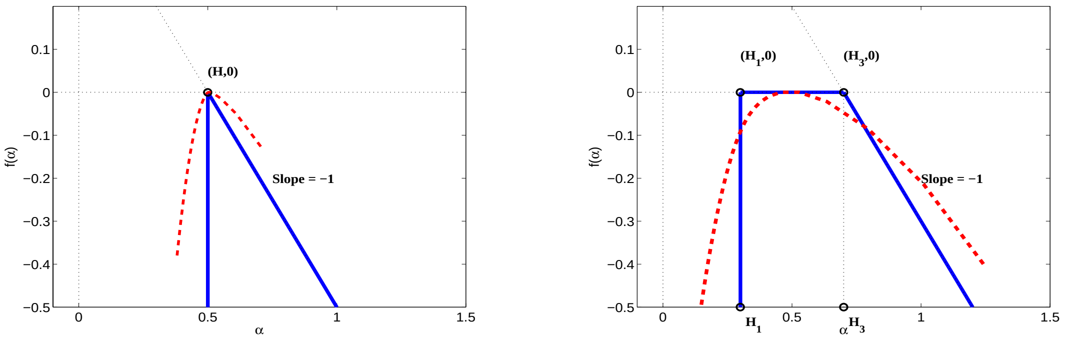
(b)

Figure 6: (a) Partition function $T(q)$ of $X(t)$; (b) MFS $f(\alpha)$ of the signal.

3.1. Multifractal Descriptors

Rather than operating with MFS as a function (density), we summarize it by a small number of meaningful descriptors. These descriptors are interpreted in terms of location and shape of MFS because they are calibrated by the counterpart of MFS of monofractal signals. Theoretically, the MFS of fBm (a representative of monofractal) consists of three geometric parts: the vertical line, the maximum point, and the right slope, as is shown in Figure 5(b). However, it is rare to obtain such a perfect spectrum in practice. Even for the well simulated fBm, due to error of estimation (most of them are due to the partition function estimation and derivative calculation as presented above), the MFS deviates from the theoretical form, as shown in Figure 7. Panels (a) and (b) of Figure 7 show theoretical MFS as a solid blue line and empirical MFS as a dashed red line for fBm$_{0.5}$ and $X(t)$ in the example 2, respectively. Notice that the maximum or mode is well approximated, but the slope exhibits discrepancy between theoretical and empirical MFS due to numerical instability.

Despite the existence of estimation error, the MFS can be approximately summarized by 3 canonical descriptors (multifractal descriptors) without a loss of the discriminant information. The proposed summaries are (1) the spectral mode (Hurst exponent, H), (2) left slope (LS) or left tangent (LT) and (3) width spread (broadness, B) or right slope (RS) or right tangent (RT). A typical MFS can be quantitatively described as shown in Figure 8(a). There are many ways to define the broadness (B). These descriptors have been successfully

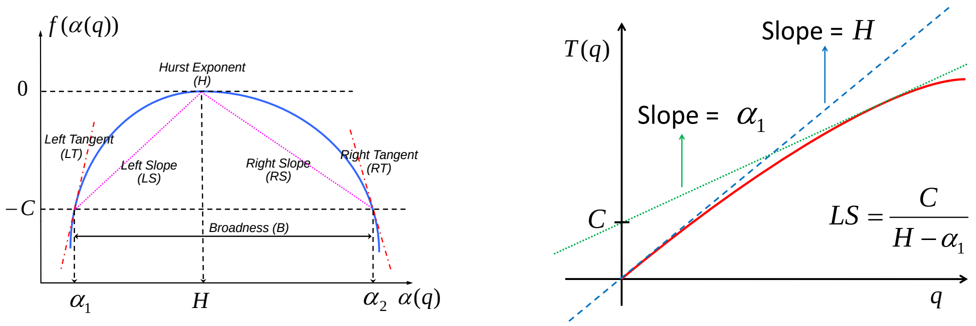


(a) (b)

Figure 7: Theoretical MFS (solid blue line) and an empirical MFS (dashed red line); (a) for $\mathrm{fBm}_{0.5}$; (b) for $X(t)$ in the example 2; empirical MFSs deviate from theoretical ones.

used in classification procedures as in Derado et al. [11] and Shi et al. [8] In this paper, we select the following definition [8].

Definition 3. Suppose that α_1 and α_2 are two roots which satisfy the equation $f(\alpha)+C=0$ and $\alpha_1<\alpha_2$. The broadness (B) of MFS is defined as $B=\alpha_2-\alpha_1$.



(a) (b)

Figure 8: (a) Illustration of geometric descriptors of MFS. Note that the horizontal axis represents values of Hölder exponent $\alpha(q)$, while the vertical axis represents values proportional to the relative frequency of these indices, $f\big(\alpha(q)\big)$; (b) interpretation of left slope (LS) with partition function $T(q)$; LS is obtained by the two slopes (H and α_1) of the two tangent lines; LS is adopted as a measure of deviation from the straight line passing through the origin.

The suggested multifractal descriptors are graphically presented in Figure 8(a). Figure 8(b) shows the link of descriptor LS to the configuration of partition function $T(q)$. Note that LS is determined by two points $(-C, \alpha_1)$ and $(H, 0)$ in Figure 8(a), which correspond to the tangent line passing through $(0, C)$ in dotted green and the tangent line passing through the origin in dashed blue in Figure 8(b). It should be also noted that the threshold value C in the definition could be adjusted empirically in the practical analysis to ensure that this measure is well computed for signals under analysis. In practical implementations, we use $C = 0.15$ or 0.2.

Another difficulty in computation is caused by the discreteness of $\alpha(q)$. The problem is that it may be approximate to find the exact roots of the equation $f(\alpha) + C = 0$ among the discrete values of α's. To circumvent this issue, we first find the two closest points $(\alpha_i^l,\ f(\alpha_i^l))$ and $(\alpha_i^u,\ f(\alpha_i^u))$ for each i such that

$$f(\alpha_i^l) < -C \quad \text{and} \quad f(\alpha_i^u) > -C, \quad i = 1, 2,$$

and then obtain the two solutions α_1, α_2 by interpolation. The slopes LS and RS and tangents LT and RT can be obtained using the interpolation technique, as computed by

$$\begin{aligned} LS = C/(H - \alpha_1) \quad &\text{and} \quad RS = -C/(\alpha_2 - H), \\ LT = (f(\alpha_1^u) - f(\alpha_1^l))/(\alpha_1^u - \alpha_1^l) \quad &\text{and} \quad RT = (f(\alpha_2^u) - f(\alpha_2^l))/(\alpha_2^u - \alpha_2^l). \end{aligned} \tag{3.5}$$

Interpretation of H and LS (or LT) is straightforward. The apex of the spectrum or the most common Hölder exponent α found within the signal represents the Hurst exponent H. The slope of the distribution produced by the collection of Hölder exponents α with smaller values of the mode (H) represents LS (or LT).

In this study, we selected the LS as the multifractal characteristic for measuring deviation from monofractality because the monofractality theoretically corresponds to a vertical line at H, that is, infinite LS, in MFS. We related the extent of deviation from the vertical line to the characterization of monofractality, which is explored more in the following section.

4. Test for Deviation from Monofractality

In this section, we analyze the MFS summaries as possible statistics for assessing deviation from monofractality. For this goal, the LS turns out to be an informative index.

4.1. Left Slope as a Measure of Deviation from Monofractality

We start with intuitive interpretation of LS, connecting it with the partition function $T(q)$. Geometrically, α_1 in Figure 8(a) is the slope of the tangent line whose intercept is C in Figure 8(b). In addition, H in Figure 8(a) is the slope of the tangent line that passes through the origin. Theoretically, the expectation of $T(q)$ is linear in an ideal case of fBm, which leads to a perfect vertical line at the Hurst exponent as in Figure 5(b) and thus the infinite LS. Empirically, the wavelet-based estimator $\hat{T}(q)$ of $T(q)$ in (3.3) deviates from the straight line because of the finite approximation to the moments and numerical instabilities. This causes LS to be finite for empirical fBms. As a result, LS incorporates information on the shape of the partition function. It reflects deviation from the straight line passing through the origin: the more linear the partition function, the larger LS. The process $X(t)$, which is a synthetic superposition of the three fBms, in Example 2, lead the flat segment between $\min\{H_i\}$ and $\max\{H_i\}$ (as the solid blue line in Figure 7(b)) and much wider breadth compared to MFS of the individual fBm (as the solid blue line in Figure 7(a)). The theoretical MFS has a wide breadth (as the solid blue line in Figure 7(b)) leading to small LSs for its empirical processes (as the dotted red line in Figure 7(b)) in comparison with those for empirical fBms (as the dotted red line in Figure 7(a)). We attribute this decrease to the deviation from monofractality of the signal.

To emphasize this point, consider a multifractional Brownian motion (mBm) with time varying Hurst exponent. A mBm with $H(t)$ ($\text{mBm}_{H(t)}$) is a zero mean Gaussian process defined as in (2.2) by replacing H with a time-varying $H(t)$ [3, 13]. Specifically, we consider a $\text{mBm}_{H(t)}$ with $H(t)$ given as, for $T = 2^{11}$,

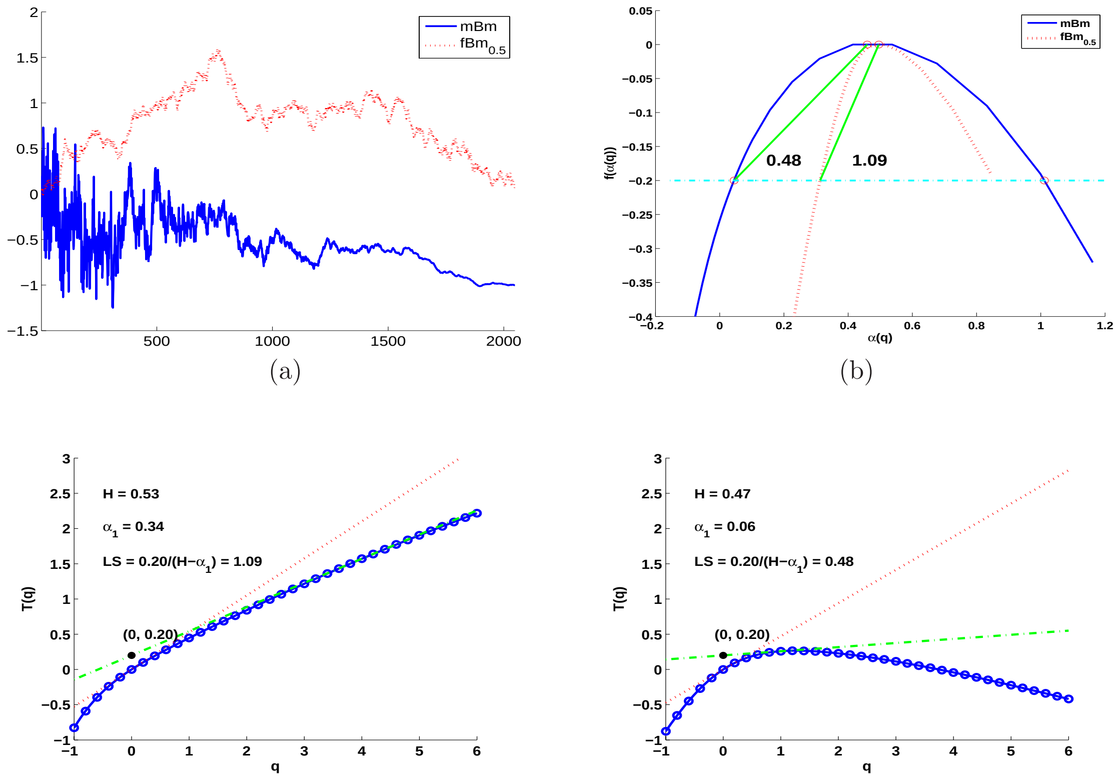
$$H(t) = \frac{0.6}{T}t + 0.2, \qquad t \in [\,0 \;\; T\,].$$

Next we compare this $\text{mBm}_{H(t)}$ with a standard Brownian motion, $\text{fBm}_{0.5}$. In Figure 9(a) and Figure 9(b), simulated signals of the two processes and the corresponding MFS are shown. The LS of $\text{mBm}_{H(t)}$ was smaller than that of $\text{fBm}_{0.5}$ (0.48 compared to 1.09). In Figure 9(c) and 9(d), the partition functions from the two signals are shown, respectively. The shapes of the two partition function that carry information equivalent to MFS are strikingly different. The LS reflects the difference of the two tangent lines (dotted red and dashed green) for each partition function: the larger the difference of the two slopes, the smaller the LS. We observe that the partition function in 9(d) deviates from the theoretical partition function (the straight and dotted red line) more severely than that in 9(c). The slopes of the two tangent lines (the dashed green lines) of the two partition functions that pass through the point $(0, 0.2)$ are different due to dissimilar shapes of the two empirical partition functions.

This behavior of LS is consistent whenever the monofractality of the signal is violated. Using this observation, we propose a testing procedure described in the next section, which tests the monofractality of a signal based on LS. The details of the testing procedure and its applications are provided.

4.2. Parametric Bootstrap Test

Bootstrapping is a computer-based method for assigning measures of accuracy to statistical estimates with sampling from an approximating distribution [7]. The advantage of bootstrapping is that it is straightforward to simulate empirical null-distributions of complex statistics such as percentile points, proportions, odds ratios, or correlation coefficients. The bootstrap method may also be used for constructing hypothesis tests as an alternative to inference based on parametric assumptions. In the case in which exact distributions are



(a) (b)

(c) (d)

Figure 9: (a) Simulated signals of fBm$_{0.5}$ in a dotted red line, mBm with straight line $H(t) = \frac{0.6}{T}t + 0.2$ in a solid blue line, $T = 2^{11}$; (b) MFS of the fBm$_{0.5}$ in a dotted red line and of the mBm in a blue solid line; (c) partition function for the fBm$_{0.5}$; (d) partition function for the mBm.

unknown or analytic procedures are too complex to obtain, even an approximation to the distribution, the bootstrap techniques are employed. In our case, the distribution of LS for monofractality of fixed size, wavelet basis, and precision settings of MFS calculation are overly complex.

With LS as a measure of deviation from monofractality, we propose a new testing procedure to check if a signal is monofractal; H_0: the signal is monofractal vs. H_1: not H_0. This type of hypothesis is a goodness-of-fit type. Not rejecting H_0 leads to the conclusion that the signal is consistent with assumption of monofractality. Rejecting H_0 does not indicate multifractality, but just a violation of monofractality or inconsistency of the monofractality assumption. The proposed test is conducted with parametric bootstrap which is outlined in Figure 10. We start with an observed signal and a wavelet basis with a sufficient number of vanishing moments; and also fix C as in Definition 3 and q_i in (3.3). The following steps describe the testing algorithm:

[1] Calculate $\hat{LS}$ and $\hat{H}$ as estimators of LS and H, respectively, for an input.

(a) Calculate wavelet coefficients $d_{j,k}$ as in (2.3).

  (b) Estimate the partition function $T(q)$ with $d_{j,k}$ as in (2.7).
  (c) Estimate the MFS $f(\alpha)$ with $T(q)$ as in (3.3).
  (d) Estimate H as the maximizer of $f(\alpha)$ and find LS as in (3.5).

[2] Generate B copies of $\mathrm{fBm}_{\hat{H}}$ and for each copy (realization) find LS^{*b}, $b = 1, \ldots, B$; this is the parametric bootstrap step.

[3] Construct a bootstrap distribution of $\hat{\mathrm{LS}}$ using bootstrap replicates; find empirical 0.05 quantile ($q_{0.05}$).

[4] If the LS is less than $q_{0.05}$, reject H_0.

To simulate a sample path from a fBm, we used the method of Wood and Chan, which is based on Fourier transform [6]. We construct an empirical distribution of $\hat{LS}$ as a surrogate of the true distribution of the true LS from B number of replicates of $\mathrm{fBm}_{\hat{LS}}$. Since signals of monofractality have high LS values, the achieved significance level (ASL) of the test is the proportion of the number of replicates for which the left slope (LS^{*b}) is less than $\hat{LS}$ to the total number of replicates (B). In hypothesis testing with bootstrapping, ASL is the counterpart of the p-value in the classical hypothesis testing. We can also adjust the quantile of $q_{0.05}$ to be different from that of 0.05.

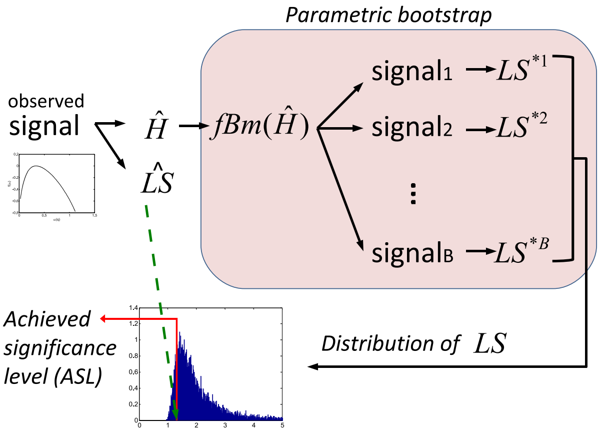


Figure 10: Parametric bootstrapping for testing whether a signal observed is monofractal; the achieved significant level (ASL) of the test is the area of the bootstrap distribution enclosed with the solid red line since monofractal signals have high values of LS.

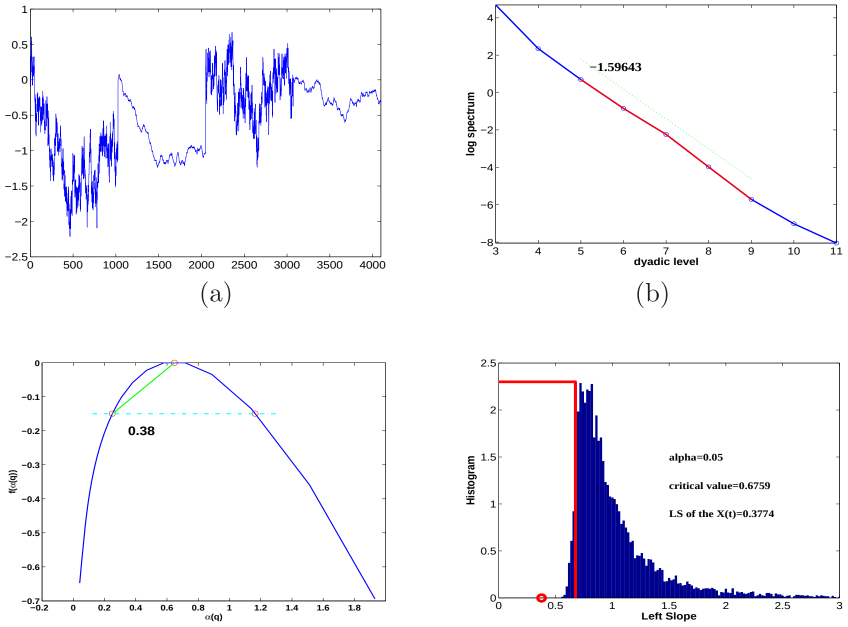
4.3. Experimental Result

We perform a simulation experiment to test the following non-monofractal signal $X(t)$ for monofractality:

$$X(t) = B_{H_k}(t), \qquad t \in [t_k + 1, t_k + 2^{10}[,$$

for $t_k = (k-1)2^{10}$, $k = 1, \ldots, 4$, and $H_1 = H_3 = 0.3, H_2 = H_4 = 0.7$. We chose B as 5000, C as 0.15, and q_i as equi-spaced with size 0.2 on $(-1\ 6]$; and tested 3000 samples of

$X(t)$. Obviously, the signal is not monofractal since regularity is not constant across the time. We want to test if H_0: $X(t)$ is monofractal vs. H_1: not H_0. An illustration of $X(t)$, its wavelet spectrum, its MFS, and the empirical distribution of LSs are shown in Figure 11. The wavelet spectrum in Figure 11(b) shows a monotone decay across the dyadic scales, and yet the LS was 0.38 in 11(c), which is indicating irregular scaling. We show the normalized histogram (bootstrap distribution) of LS^{*b} from the 5000 bootstrap fBm simulations and the critical region of level 0.05 in Figure 11(d). The LS clearly falls in the rejection region and H_0 is rejected: The ASL of this test was 0. Out of the 3000 tested signals, 2838 signals were concluded not to be monofractal: The rate, or ability to recognize the true non-monofractal signals, was 0.946. Next we apply this to real-life examples.



(a) (b)

(c) (d)

Figure 11: (a) An illustration of $X(t)$; (b) its wavelet spectrum; (c) its MFS; (d) bootstrap distribution of LS^{*b} with the LS (0.38) for the signal (in red circle) and a rejection region of 95% achieved significance level (within the solid red line).

4.4. Turbulence and DNA Examples

To illustrate the test procedure in a real-world example, we compared a turbulence signal with a fBm$_{1/3}$. Understanding the properties of turbulence is a major problem of modern physics, which remains mostly open despite intense research efforts from 1941 when Kolmogorov formulated a statistical theory of turbulence [4]. Kolmogorov introduced his theory, often referred to as K41 theory, for locally isotropic turbulence. The velocity field is modeled as a process $U(x)$ with increments having the following structure function of order p:

$$E\left[|U(x+r)-U(x)|^p\right] \propto (\epsilon r)^{\frac{p}{3}}.$$

Parameter ϵ, energy per unit of fluid mass per unit time, describes the energy transmission from large eddies, where the energy is injected, to small eddies, where the energy is converted to heat by viscosity. The K41 theory states that a one-dimensional longitudinal trace of a three-dimensional velocity field is a fractal noise process with constant Hurst exponent 1/3 and models turbulence as a monofractal. Though the theory was verified in many empirical observations possessing the property of monotone spectral decay, it does not take into account the existence of coherent structures such as vortices and helicity. Kolmogorov [5] refined the homogeneity assumption of ϵ to be a location-varying dissipation rate $\epsilon(x)$, which leads to the model of multifractional Brownian motion. This turbulence model is not monofractal.

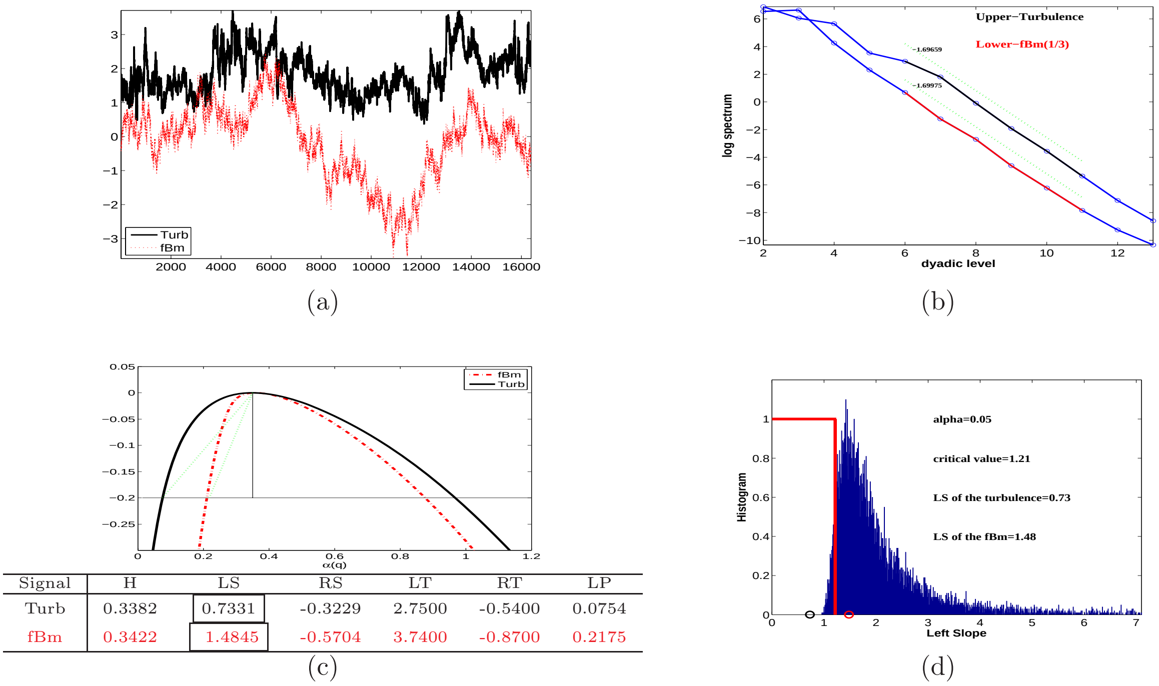
We tested a turbulence signal of length 2^{14} from velocity measurements on July 12, 1997 at 5.2 m above the ground surface over an Alta Fescue grass site at the Blackwood division of the Duke Forest in Durham, North Carolina to check if the turbulence is monofractal. In Figure 12(a), the turbulence signal and $\mathrm{fBm}_{1/3}$ are in a solid black line and a dotted red line, respectively. The average log spectrum of squared wavelet coefficients for the two signals are shown in Figure 12(b). The spectral slopes are indistinguishable; the two signals do not differ with respect to their second order properties. The wavelet spectrum of the turbulence signal is shifted upwards from that of $\mathrm{fBm}_{1/3}$ because of difference in their energies.

The two MFS along with their descriptors are shown in Figure 12(c): the MFS of the turbulence signal is wider than that of the $\mathrm{fBm}_{1/3}$ signal. To quantify the degree of deviation of the turbulence signal from monofractality, $10,000$ samples of the $\mathrm{fBm}_{1/3}$ and the empirical distribution of estimated LSs were obtained as is shown in Figure 12(d). Two circles represent LSs of the turbulence (left in black) and the $\mathrm{fBm}_{1/3}$ (right in red). The critical point at 95% is highlighted with the solid red line in Figure 12(d). The left black circle, corresponding to the turbulence signal falls in the critical region, leading to rejection of the null hypothesis. We concluded the turbulence signal was not monofractal.

Next, we demonstrate our method in an analysis of DNA sequences. In the analysis of DNA sequences, one of the most important tasks is to study whether two sequences are related. This is studied by using a scoring system to rank the possible relations between the sequences and by considering statistical methods to evaluate the significance of such relations [21]. Often the sequences of nucleotides (A, C, G, and T) are coded as functions or DNA walks, and fractal properties of these associated functions can be informative for functional properties of DNA segments [2].

The analysis of DNA walks is influenced by the presence of a global linear trend induced by the excess of purines over pyramidines. In all eukaryotic species, a DNA molecule consists of a long complementary double helix of purine nucleotides (denoted as A and G) and pyrimidine nucleotides (denoted as C and T). A single strain of this DNA can be represented as a long word that corresponds to a random walk. Depending on the letter at position i in the word, the random walk gets a cumulative sum of increments of $x(i) = 1$ for A and G, and $x(i) = -1$ for C and T. Hence, the corresponding random walk is defined as $s(n) = \sum_{i=1}^{n} x(i)$, in which n is an index smaller than the length of the sequence.

In Figure 13(a), we show an 8196-long DNA random walk for a spider monkey from the EMBL Nucleotide Sequence Database, which is also known as the EMBL-Bank. The wavelet spectrum and MFS of the signal are shown in Figure 13(a) and 13(b), respectively. The estimated Hurst exponent was 0.648 and the left slope was 1.47. We noticed that the MFS yielded only the left part from the mode because the partition function was flat for negative exponents and made the right part of the MFS computationally unobtainable. The empirical distribution of LS^{*b} from $10,000$ simulations of $\text{fBm}_{0.648}$ is shown in Figure 13(d). The ASL of the observed LS (1.47) was greater than 0.05, by which we conclude that the signal is monofractal. This conclusion is in accordance with the observation made by Arneodo that the DNA sequences are the most perfect monofractals found in nature (personal communication).



(a) (b)

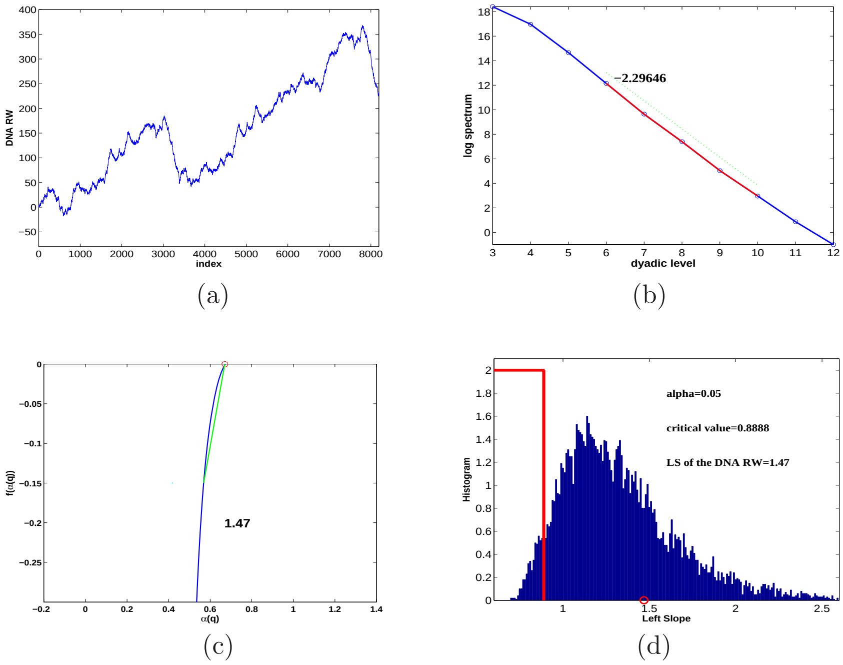
Signal	H	LS	RS	LT	RT	LP
Turb	0.3382	0.7331	-0.3229	2.7500	-0.5400	0.0754
fBm	0.3422	1.4845	-0.5704	3.7400	-0.8700	0.2175

(c) (d)

Figure 12: Comparison of the turbulence and $\text{fBm}_{1/3}$ signals; (a) turbulence in a solid black line and $\text{fBm}_{1/3}$ in a dotted red line are indistinguishable with respect to their second order properties; (b) log spectra for the two signals with two spectral slopes produced identical slopes; (c) MFS and the descriptors for (a); (d) the bootstrap distribution of LS^{*b} along with two circles (left in black for LS, 0.73, of the turbulence signal, right in red for LS, 1.43, of the $\text{fBm}_{1/3}$) signal and a rejection region of 95% achieved significance level (within the solid red line).

5. Conclusion

Evidence for deviation from monofractality, which is contained in the partition function and MFS, can be used to develop a paradigm for formal testing for deviation from monofractality. In this paper, we introduced a measure of deviation from monofractality by using left



(a) (b)

(c) (d)

Figure 13: Demonstration of the test of monofractality to a DNA random walk: (a) 8196-long DNA random walk for a spider monkey from the EMBL Nucleotide Sequence Database; (b) wavelet scaling with slope −2.296 and estimated Hurst exponent 0.648; (c) MFS with left slope 1.47; only the left part from the mode was computationally available due to a straight line in the partition function of negative exponents; (d) the distribution of LS^{*b} with the LS (1.47) in a red circle.

slope as one of the descriptors of wavelet-based MFS of a signal. We constructed a test procedure based on parametric bootstrapping of sampled fBm signals which are monofractals. This produces a distribution of left slopes consistent with the assumption of monofractality. Our simulation results indicate that the testing procedure effectively separates multifractal Brownian motion signals from fBm signals. Its effectiveness is also shown in the real life example of turbulence and DNA sequences, first as an example of multifractal, and second as an example of monofractal.

6. Acknowledgments

The authors thank editors of the special issue for their kind invitation to contribute and their support and encouragement. They also thank Kiersten Peterson and an anonymous referee for useful comments. They are also grateful to Gabriel Katul for providing the turbulence data set. This study was supported by the National Science Foundation grant NFS-ATM 0724524 at Georgia Institute of Technology.

References

[1] A. Arneodo, E. Bacry, S. Jaffard, and J.F. Muzy, Singularity spectrum of multifractal functions involving oscillating singularities, Journal of Fourier Analysis and Applications, 4 (1998), 159–174.

[2] A. Arneodo, Y. d'Aubenton Carafa, E. Bacry, P.V. Graves, J.F. Muzy, and C. Thermes, Wavelet based fractal analysis of DNA sequences, Physica D: Nonlinear Phenomena, 96 (1996), 291–320.

[3] A. Benassi, S. Cohen, and J. Istas, Identifying the multifractional function of a gaussian process, Statistics & Probability Letters, 39 (1998), 337–345.

[4] A.N. Kolmogorov, The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers, Proceedings: Mathematical and Physical Sciences, 434 (1991), 9–13.

[5] A.N. Kolmogorov, A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number, Journal of Fluid Mechanics, 13 (1962), 82–85.

[6] A.T.A. Wood, G. Chan, Simulation of stationary Gaussian processes in [0, 1] d, Journal of Computational and Graphical Statistics, 3 (1993), 409–432.

[7] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1997.

[8] B. Shi, K. Moloney, K. V. Leonard, J. Jacko, and F. Sainfort, Multifractal discrimination model (mdm) of high-frequency pupil diameter measurements for human-computer interaction, In: Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques, (Don Hong and Yu Shyr Eds.), World Scientific Publications, New Jersey, 2007, 333–350.

[9] D. Veitch, N. Hohn, and P. Abry, Multifractality in TCP/IP traffic: the case against, Computer Networks, 48 (2005), 293–313.

[10] D. Veitch, P. Abry, A statistical test for the time constancy of scaling exponents, IEEE Transactions on Signal Processing, 49 (2001), 2325–2334.

[11] G. Derado, K. Lee, O. Nicolis, F. Bowman, M. Newell, F. Rugger, and B. Vidakovic, Wavelet-based 3-D Multifractal Spectrum with Applications in Breast MRI Images, In: Bioinformatics Research and Applications, (Ion Mandoiu, Rajshekhar Sunderraman, and Alexander Zelikovsky Eds.), Springer Berlin, Heidelberg, 2008, 281–292.

[12] H. Wendt, S.G. Roux, S. Jaffard, and P. Abry, Wavelet leaders and bootstrap for multifractal analysis of images, Signal Processing, 89 (2009), 1100–1114.

[13] J. Coeurjolly, Identification of multifractional brownian motion, Bernoulli, 11 (2005), 987–1008.

[14] J.L. Vehel, R. Riedi, Fractional Brownian motion and data traffic modeling: The other end of the spectrum, In: Fractals in Engineering, (Jacques Levy Vehel, Evelyne Lutton, and Claude Tricot Eds.), Springer, New York, 2007, 185–202.

[15] L. Calvet, A. Fisher, Multifractality in asset returns: theory and evidence, Review of Economics and Statistics, 84 (2002), 381–406.

[16] P. Abry, D. Veitch, Wavelet analysis of long-range-dependent traffic, IEEE transactions on information theory, 44 (1998), 2–15.

[17] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, Wavelets for the analysis, estimation and synthesis of scaling data, In: Self-Similar Network Traffic and Performance Evaluation, (Kihong Park and Walter Willinger Eds.), John Wiley & Sons, New York, 2000, 39–88.

[18] P. Flandrin, On the spectrum of fractional Brownian motions, IEEE Transactions on Information Theory, 35 (1989), 197–199.

[19] P. Gonçalves, R. Riedi, and R. Baraniuk, A simple statistical analysis of wavelet-based multifractal spectrum estimation, Proceedings 32nd Asilomar Conference on Signals, Systems and Computers, 1 (1998), 287–291.

[20] P.C. Ivanov, L.A.N. Amaral, A.L. Goldberger, S. Havlin, M.G. Rosenblum, Z. Struzik, and H.E. Stanley, Multifractality in human heartbeat dynamics, Letters to Nature, 399 (1999), 461–465.

[21] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, Cambridge, 1998.

[22] R.H. Riedi, Multifractal processes, In: Theory and applications of long-range dependence, (Paul Doukhan, George Oppenheim, and Murad S. Taqqu Eds.), Birkhauser, Boston, 2003, 625–716.

[23] R.H. Riedi, M.S. Crouse, V.J. Ribeiro, and R.G. Baraniuk, A multifractal wavelet model with application to network traffic, IEEE Transactions on Information Theory, 45 (1999), 992–1018.

[24] S. Jaffard, Local behavior of Riemann's function, Contemporary Mathematics, 189 (1995), 287–287.

[25] U. Frisch, Fully developed turbulence and intermittency, Annals of the New York Academy of Sciences, 357 (1980), 359–367.

[26] W.S. Kuklinski, Utilization of fractal image models in medical image processing, Fractals, 2 (1994), 363–369.

International Journal of Mathematics and
Computer Science, 5(2010), no. 2, 87-100

An Empirical Bayes Approach for Methylation Differentiation at the Single Nucleotide Resolution

Kenneth McCallum, Wenxin Jiang, Ji-Ping Wang

Department of Statistics
Northwestern University
Evanston, IL 60208, USA

e-mail: kennethmccallum2013@u.northwestern.edu, jzwang@northwestern.edu

Abstract

DNA methylation is an important epigenetic phenomenon that is associated with a variety of diseases, particularly cancers. Recent development of high throughput sequencing technology has enabled researchers to investigate the methylation rate at a single nucleotide resolution for any given sample. Testing for methylation rate equality or difference between two samples, however, is challenged by the small sample size observed at many sites across the genome. Fisher's exact test is typically used in this situation; however, it is conservative and it cannot be used to test for specific difference in methylation rate between two samples. In this paper, we propose an empirical Bayes approach that utilizes the genome-wide data as prior information for methylation differentiation between two samples. We show that this new approach is more powerful than Fisher's exact test. In addition, it can be used to test for any specific methylation difference while controlling the false discovery rate (FDR). The new method is applied to a real data set from a colon tumor study.

Key words: 'Empirical Bayes', 'DNA Methylation', Single-nucleotide
AMS subject classification: 62J05, 62J07, 62H35, 62P10
ISSN 1814-0424 ©2010, http://ijmcs.future-in-tech.net

1. Introduction

The epigenetic phenomenon of DNA methylation, in which cytosines in CpG dinucleotides are chemically modified by the addition of a methyl group, plays an important role in genetic regulation [1, 2]. Methylation rates are known to change throughout the genome during development in mammals [3, 4]. Furthermore, differential methylation rates are associated with a variety of diseases, including neurodevelopmental disorders [5], and numerous cancers [6, 7, 8, 9].

Most research up to now has focused on methylation rates over large regions of the genome; however, an increasing number of studies attempt to quantify and analyze methylation rates at specific sites [1, 2]. Methods making use of universal bead arrays have been able to detect differential methylation rates at the single nucleotide resolution and have demonstrated that these differences can be used to distinguish normal and cancer tissues [10]. However, the array based methods are limited in terms of the number of sites that can be examined; for example, only 1536 sites were included in the study by Bibikova et al. [10]. Another approach, the one which will be the focus of this study, makes use of high throughput bisulfite sequencing. This method is increasingly common and has the potential to examine hundreds of thousands or millions of sites simultaneously. For example, Laurent et al. [11] and Gu et al. [12] both made use of bisulfite sequencing to generate maps of methylation with single-nucleotide resolution, Han et al. [9] tested for site specific differential methylation in samples taken from subjects with and without lung cancer using bisulfite sequencing, and Houseman et al. [13] used clustering methods to differentiate methylation rates.

The data set produced by Gu et al. [12] is used for illustration in this study. The data was for two tissue samples, a colon tumor and normal colon tissue, both taken from the same donor. Bisulfite sequencing was used to determine methylation status at targeted CpG sites across the genome. At each site (corresponding to the C in a CpG dinucleotide), the data included the number of reads (number of sequenced DNA fragments) that covered the given site, and the number of reads that were positive for methylation at the given site. Figure 1 illustrates the format of the data. A total of 920,441 sites had at least one read for both tissue samples and only these sites were included in the present study. Although a few sites had very large numbers of reads, some with more than 1400, the majority were small, with a median of 10 or fewer across both samples. Summary statistics for the number of reads are given in Table 1.

The central goal of methylation studies is to identify CpG sites or regions that show differential methylation rates between disease or cancer tissue and normal tissue. Given the small number of reads (or sample size) at a given site, Fisher's exact test is often the only choice for testing the equality of methylation rates between two samples. Fisher's exact test, however, is conservative in power. Furthermore, it cannot be used to test for specific difference in the methylation rates between two samples. The latter is particularly important, as in practice, a meaningful difference in methylation is often called only if the methylation rate in one sample is higher/lower than the other by a predefined threshold value (see details

Figure 1: Format of data.

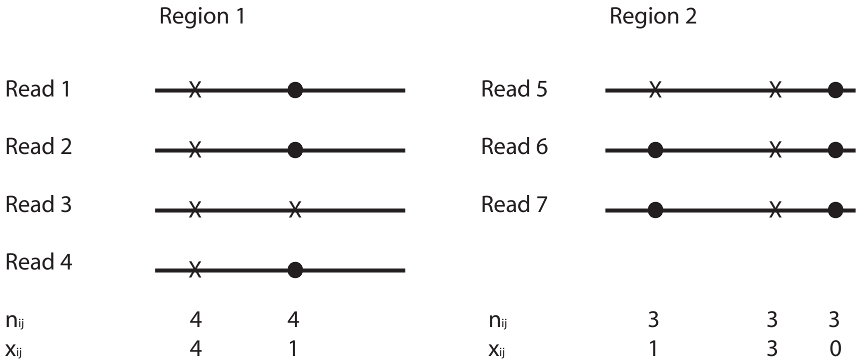


Figure 1 shows two regions of equal length that contain the targeted CpG sites for methylation examination. Bisulfite sequencing generated two groups of short reads of identical length, each of which covered one of these regions. For example, in region 1, four reads were generated. Within each read, the methylation status at the sites were identified as positive or negative. In the figure, n is used to denote the total number of reads observed at a given site, and x the number of methylations (positives) observed across reads.

Table 1: Quantiles for number of reads per site (n_{ij}).

	Minimum	Q1	Median	Q2	Maximum
Normal	1	2	6	14	14043
Tumor	1	3	10	25	14361

below). These two limitations motivate us to seek an alternative approach.

In this paper, we proposed an empirical Bayes (eB) approach, in which we utilize the large amount of data observed from the entire genome to construct a prior distribution for the methylation rate. Based on the posterior distribution of the methylation rate at each site, we test for difference of methylation rate while controlling false discovery rate. We show this new approach has improved power compared to Fisher's exact test. In addition, it can be used to test any specific difference of methylation rates between two samples.

2. Methods & Results

2.1. The Model

Let x_{ij} be the observed methylations out of a total of n_{ij} reads at site i from sample j for $i = 1, ..., M$ and $j = 1, 2$, and θ_{ij} be the true, unobserved, methylation rate. Let

$\mathbf{N} = (\mathbf{N}_1, \mathbf{N}_2)$, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{N}_j = \{n_{ij} : i = 1, ..., M\}$ and $\mathbf{X}_j = \{x_{ij} : i = 1, ..., M\}$. We assume that the methylation rates are independent across samples and sites, and have a common distribution within each sample,

$$\theta_{ij} \sim Beta(\gamma_j, \lambda_j).$$

Note that although the model allows for the possibility that the samples may differ with respect to the hyper-parameters γ and λ, results given below show that in practice a single set of hyper-parameters can be used if the samples are similar. We further assume that each read is a random observation from the population, i.e., the entire underlying tissue. Then x_{ij} follows a binomial distribution

$$x_{ij}|(n_{ij}, \theta_{ij}) \sim Binomial(n_{ij}, \theta_{ij}).$$

The posterior probability for the methylation rate given the reads data is then

$$\theta_{ij}|(n_{ij}, x_{ij}) \sim Beta(\gamma_j + x_{ij}, \lambda_j + n_{ij} - x_{ij}).$$

To estimate the hyper-parameters, observe that the likelihood function is

$$L(\gamma_j, \lambda_j; \mathbf{N}_j, \mathbf{X}_j) = \Pi_{i=1}^{M} \int_0^1 P[X_{ij} = x_{ij}|\theta_{ij}, n_{ij}]p(\theta_{ij}|\gamma_j, \lambda_j)d\theta_{ij}.$$

Under the model,

$$P[X_{ij} = x_{ij}|\theta_{ij}, n_{ij}] = \binom{n_{ij}}{x_{ij}} \theta_{ij}^{x_{ij}}(1 - \theta_{ij})^{n_{ij}-x_{ij}}$$

and

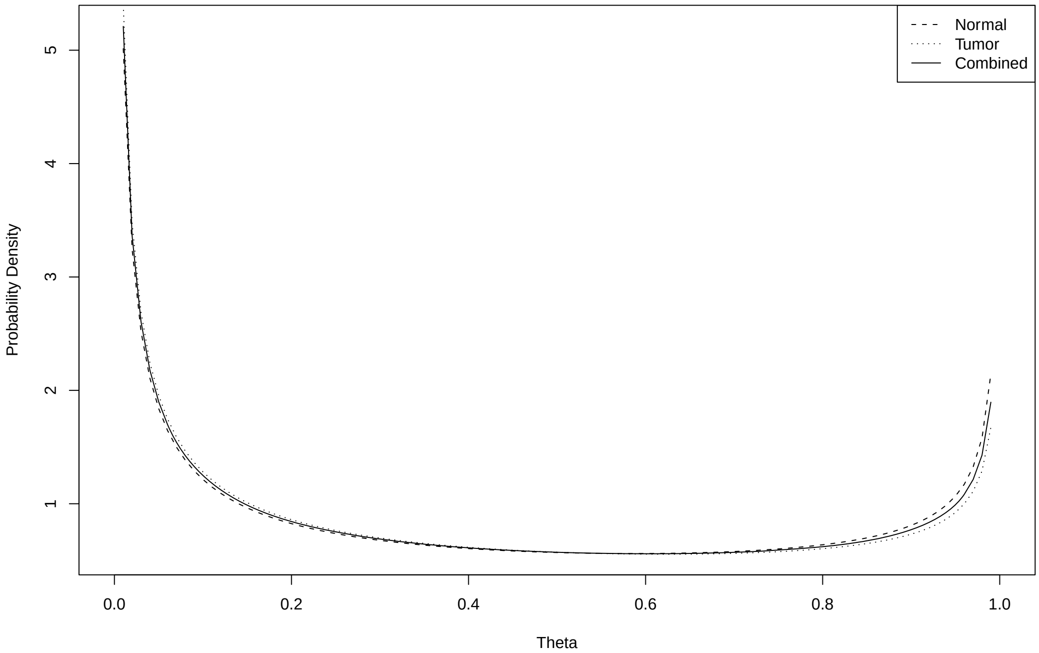
$$p(\theta_{ij}|\gamma_j, \lambda_j) = B^{-1}(\gamma_j, \lambda_j)\theta_{ij}^{\gamma_j-1}(1 - \theta_{ij})^{\lambda_j-1},$$

where B stands for the Beta function. Therefore,

$$\begin{aligned} L(\gamma_j, \lambda_j; \mathbf{N}_j, \mathbf{X}_j) &= \Pi_{i=1}^{M} \int_0^1 \binom{n_{ij}}{x_{ij}} B^{-1}(\gamma_j, \lambda_j)\theta_{ij}^{\gamma_j+x_{ij}-1}(1 - \theta_{ij})^{\lambda_j+n_{ij}-x_{ij}-1}d\theta_{ij} \\ &= \Pi_{i=1}^{M} \binom{n_{ij}}{x_{ij}} B^{-1}(\gamma_j, \lambda_j)B(\gamma_j + x_{ij}, \lambda_j + n_{ij} - x_{ij}), \end{aligned}$$

The maximum likelihood estimates of γ_j and λ_j, denoted $\hat{\gamma}_j$ and $\hat{\lambda}_j$, can easily be found using a method such as the Newton-Raphson algorithm. For the tumor and normal colon tissue data from [12], we fitted the Beta-binomial model for each sample separately, and then for the combined data. Results are summarized in Table 2. The MLEs for γ are approximately equal across samples while the MLEs for λ show a greater difference, with the tumor tissue having a λ value approximately 10% greater than the normal tissue. Despite this small discrepancy in λ, the density curves, shown in Figure 2, appear almost identical. This suggests that little would be gained by specifying separate priors for the two samples in this case.

Figure 2: Prior Distribution Densities.



The densities of the fitted prior distributions are given. These priors are assumed to be i.i.d. across all sites in the data set used to fit them.

To verify the fit of the model, based on the empirical distribution of n_{ij}, we calculated the expected empirical distribution of the observed methylation rates, defined as x_{ij}/n_{ij}, based on the fitted beta model from the joint data, treating x_{ij} as a random variable. For a given n_{ij}

$$P(X_{ij} = x_{ij}|n_{ij}, \hat{\gamma}, \hat{\lambda}) = \int_0^1 P[X_{ij} = x_{ij}|n_{ij}, \theta_{ij}]p(\theta_{ij}|\hat{\gamma}, \hat{\lambda})d\theta.$$

The probability for observing methylation rate $q \equiv x_{ij}/n_{ij}$ is then found by taking the weighted average over all pairs (x, n) such that $x/n = q$. That is, if $S_q = \{(x, n) : x/n = q\}$, then the expected probability to observe q in the sample given the empirical distribution of n_{ij} is

$$\Sigma_{S_q} P(X_{ij} = x_{ij}|n_{ij}, \hat{\gamma}, \hat{\lambda})P(N_{ij} = n_{ij})$$

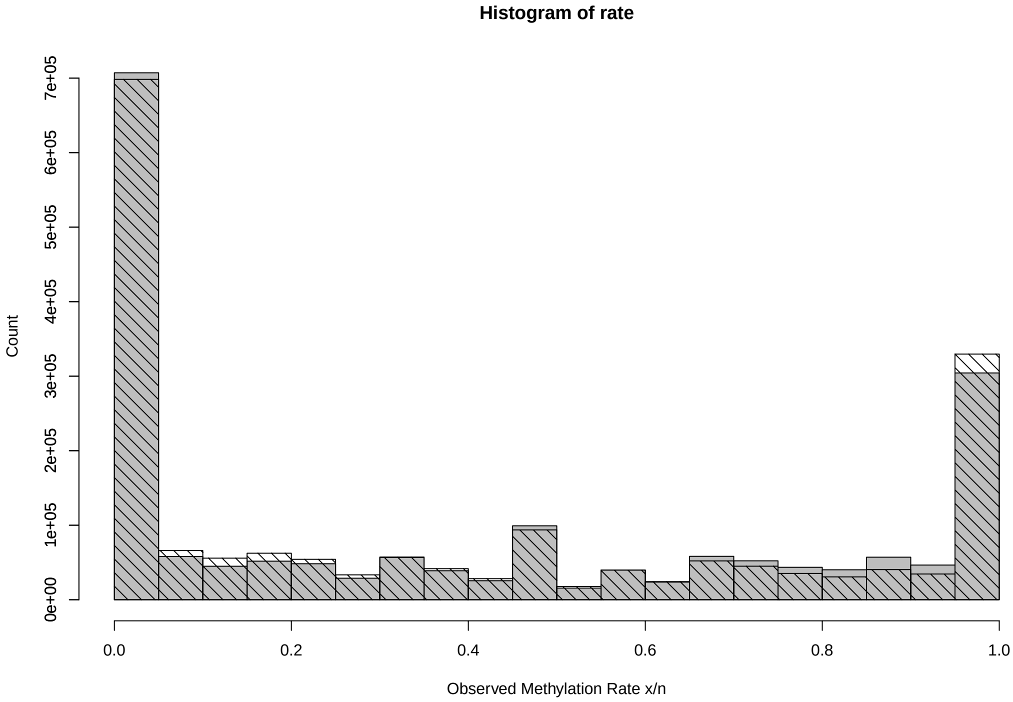
where $P(N_{ij} = n_{ij})$ is the probability of a site having n_{ij} reads based on the empirical distribution. This is then compared to the observed distribution of methylation rates. Figure 3 gives the shape of the distributions. The shape is similar to that of the curves in Figure

Table 2: Maximum likelihood estimates of hyperparameter values.

Parameter	Normal	Tumor	Combined
γ	0.365518	0.359081	0.362389
λ	0.550387	0.614117	0.582426

Estimates of the parameters for the prior distribution are given for normal colon tissue data, tumor colon tissue data, and the combined data set.

Figure 3: Theoretical and Observed Methylation Rates.



Proportion of positive reads out of total reads (x_{ij}/n_{ij}) is given on the x-axis. Number of sites matching a given proportion is shown on the y-axis. The grey bars represent the observed data while the bars with black diagonal stripes indicate the theoretical number given the prior distribution for the underlying methylation rate and the empirical distribution for the number of reads.

2, though it reflects the fact that the distribution of observed rates is discrete. The spikes that occur near 0 and 1 in the plot are partially due to the large number of sites with small numbers of reads, which are highly constrained in terms of values they can take on. Overall, the evidence shows that the model is a very good fit for the data.

2.2. Hypothesis Tests

Two different sets of hypotheses are considered. The first is a simple test of equality

$$H_0 : \theta_{i1} = \theta_{i2} \text{ vs. } H_A : \theta_{i1} \neq \theta_{i2}.$$

The second is a test of difference of rates given by

$$H_0' : |\theta_{i1} - \theta_{i2}| \leq c \text{ vs. } H_A' : |\theta_{i1} - \theta_{i2}| > c$$

for some constant c. The second hypothesis is particularly interesting, as in practice, differential methylation is often called when the difference is substantial, e.g., c=0.2 [11].

Given $\hat{\gamma}_j$, $\hat{\lambda}_j$ the posterior distribution of the methylation rate is

$$\theta_{ij}|(x_{ij}, n_{ij}) \sim Beta(\hat{\gamma}_j + x_{ij}, \hat{\lambda}_j + n_{ij} - x_{ij}).$$

For convenience, we shall denote the posterior distribution as $\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{ij})$ in the following context. For testing $H_0 : \theta_{i1} = \theta_{i2}$ versus $H_1 : \theta_{i1} \neq \theta_{i2}$, we define the posterior log odds as follows:

$$\Delta_i \equiv log\Big[\frac{\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{i1} > \theta_{i2})}{1 - \pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(\theta_{i1} > \theta_{i2})}\Big].$$

We reject H_0 if $|\Delta_i| > \delta_\alpha$, where δ_α is the cutoff value corresponding to level α.

Given the prior distribution and the number of reads at a site for each sample, it is possible to calculate the level (α) and power of the test for a given critical value (δ_α) analytically. However, doing so for every combination of number of reads appearing in the sample would be extremely computationally intensive. Here we estimate it using Monte Carlo simulations. We first generate Monte-Carlo samples as follows:

1. Sample (n_{i1}, n_{i2}) pairs with replacement from the observed data. We sample the pairs instead of individual n_{ij}'s to account for possible dependence of reads count between samples due to various factors including DNA sequence features.

2. Sample θ_{ij} values for each site in each sample from the fitted prior distribution, i.e., $Beta(\hat{\gamma}, \hat{\lambda})$ from the combined data or $Beta(\hat{\gamma}_j, \hat{\lambda}_j)$ from separate samples.

3. Generate x_{ij} from $Binomial(n_{ij}, \theta_{ij})$

Two simulated data sets of size equal to the original data are generated. In one set, we use $Beta(\hat{\gamma}, \hat{\lambda})$ to generate the θ_{ij} values for both samples. For the second set, $Beta(\hat{\gamma}, \hat{\lambda})$ is used only for the sites with equal θ_{ij} values while the remaining sites are simulated using the separate estimates from the normal and tumor tissues (i.e., $Beta(\hat{\gamma}_j, \hat{\lambda}_j)$). In both cases, the first 100,000 sites are set so that $\theta_{i1} = \theta_{i2}$ while the remaining ones are allowed to vary.

Table 3: Test of Equality

	Simulation 1			Simulation 2		
	Critical	level	power	Critical	level	power
1 prior	2.5	0.1	0.518	2.5	0.1	0.522
1 prior	3.15	0.05	0.438	3.15	0.05	0.442
2 priors	2.5	0.1	0.524	2.5	0.1	0.522
2 priors	3.15	0.05	0.452	3.21	0.05	0.441
Fisher's Exact		.1	0.356		0.1	0.355
Fisher's Exact		0.05	0.317		0.05	0.316

Simulation 1 used a single prior from the combined data set to generate the methylation rates. Simulation 2 used the prior from the combined data set to generate methylation rates for the subset of the simulated data where the rates were set equal across tissue samples, and used each sample's individually calculated prior for the remaining data points. The designations of 1 prior and 2 prior refer to whether the combined data estimates of the parameters or the individual tissue sample estimates were used in calculating the log odds.

After the simulated data set is complete, the posterior log odds can be calculated for each site. A suitable critical value can then be selected for a level α test by setting δ_α equal to the $100(1-\alpha)th$ percentile of the absolute values of the log odds for the subset of sites with $\theta_{i1} = \theta_{i2}$. Similarly, power can be estimated by taking the proportion of sites with $\theta_{i1} \neq \theta_{i2}$ with posterior log odds that have absolute values less than δ_α. In implementing this test for the simulated data sets, rather than refitting the values of $\hat{\gamma}$ and $\hat{\lambda}$, the values used in the simulation were reused in calculating the posteriors. This is justified by the large size of the data sets and the resultant accuracy and precision of the MLE.

Two versions of the test are conducted on each of the simulated data sets. The first uses the combined estimates of the hyper-parameters, while the second uses the separate estimates of the hyper-parameters for each of the two data sets. The results in this case indicate that it makes little or no difference which way the priors are specified. This is not unexpected since the two prior distributions were so similar. However, this might not generalize to all cases if two specimens are markedly different in their methylation patterns. Table 3 shows the approximate critical values for level 0.1 and 0.05 tests, and estimated power. Results for Fisher's exact test are also given for comparison. It should be cautioned that the critical values given here depends on both the hyper-parameters and the distribution of read counts, and hence are specific to this data set and should not be taken to be generally applicable.

For the test of difference, we define the posterior log odds that $|\theta_{i1} - \theta_{i2}| > c$ as follows,

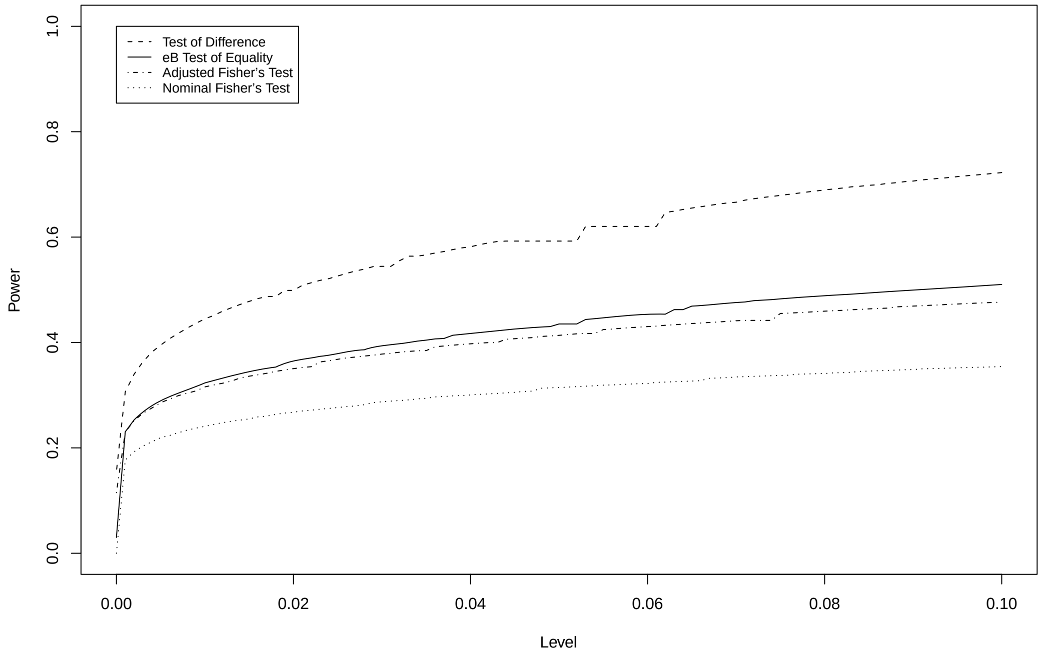
$$\Delta_i^c = log\Big[\frac{\pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(|\theta_{i1} - \theta_{i2}| > c)}{1 - \pi_{\boldsymbol{\theta}|\mathbf{X},\mathbf{N}}(|\theta_{i1} - \theta_{i2}| > c)}\Big].$$

Table 4: Test of Differences

Critical	Level	Power
0.857	0.1	0.741
1.658	0.05	0.620

Critical values, level, and power for the test of differences of methylation rates are reported with a $c = 0.2$ as the null hypothesized largest absolute difference. Test were done with a single prior on simulated data using the prior fitted to the combined data.

Figure 4: Power versus Level.



Power is shown on the y-axis and level on the x-axis. Values are estimates based on simulated data.

As with the test of equality, a critical value for a given level α, and the corresponding power, can be determined by simulations. A difference threshold of $c = 0.2$ was chosen for the test, which corresponds to the bin width for categorizing methylation rates used in other studies (eg. Laurent et al., 2010, [11]). Simulations indicate greater power for the test of differences than for the test of equality at a given level. Results are summarized in Table 4. The power of all three tests is plotted against the level in Figure 4.

Table 5: Hypotheses and True Values

	Test Negative	Test Positive	Total
H_0	M_{00}	M_{01}	$M_{0\bullet}$
H_1	M_{10}	M_{11}	$M_{1\bullet}$
Total	$M_{\bullet 0}$	$M_{\bullet 1}$	$M_{\bullet\bullet}$

The number of true negative, $M_{0\bullet}$, and true positives, $M_{1\bullet}$, compared to the number testing as negative, $M_{\bullet 0}$, and testing as positive, $M_{\bullet 1}$. Only $M_{\bullet 0}$, $M_{\bullet 1}$, and $M_{\bullet\bullet}$ are directly observed.

2.3. False Discovery Rate

Using the estimate of level, α, and power, β, from the simulations, the false discovery rate (FDR) can be estimated for the original data set. Let $M_{\bullet\bullet}$ be the number of sites, $M_{0\bullet}$ and $M_{1\bullet}$ be the total number of true null and alternative hypotheses respectively. Let $M_{\bullet 0}$ and $M_{\bullet 1}$ be the numbers of claimed negatives and positives. Table 5 tabulates four different incidents incurred in hypothesis testing: true negatives (M_{00}), false natives (M_{10}), true positives (M_{11}), and false positives (M_{01}). Then

$$E[M_{\bullet 1}] = \alpha M_{0\bullet} + \beta M_{1\bullet} = \alpha M_{0\bullet} + \beta(M_{\bullet\bullet} - M_{0\bullet}).$$

This implies that $M_{0\bullet}$ can be estimated by

$$\hat{M}_{0\bullet} = [M_{\bullet 1} - \beta M_{\bullet\bullet}]/[\alpha - \beta]$$

The FDR is then estimated by

$$FDR = \frac{\alpha \hat{M}_{0\bullet}}{M_{\bullet 1}}.$$

Since M_{00}, M_{01}, M_{10}, and M_{11} are all functions of the specified level α, estimation of $M_{0\bullet}$ and FDR requires an appropriate choice of α. For the real data, using the estimated α and β from the simulation studies presented in Figure 4, we calculated $\hat{M}_{0\bullet}$ for α ranging from 0.1 to 0.0001. Interestingly, $\hat{M}_{0\bullet}$ increased monotonically from around 740,000 to over 870,000 as the type I error level decreased from 0.1 to less than 0.0001. To determine which value of α can lead to a most accurate estimate of $M_{0\bullet}$, we simulated data sets containing 800,000 true nulls and 192,000 true alternatives where the methylation rate θ_{ij} followed the prior distribution fitted from the eB approach. The monotonicity, however, was not observed; and $M_{0\bullet}$ was estimated very accurately for any α value used in the same range. This likely indicates some violations of model assumptions in the real data. We leave this as an open question for future investigation.

In the absence of a reliable estimate of $M_{0\bullet}$, a precise estimate of FDR cannot be calculated. The most conservative estimate of FDR can be obtained by substituting $M_{\bullet\bullet}$ for $\hat{M}_{0\bullet}$ into the FDR formula. An FDR of 0.05 is achieved by setting the level as $\alpha = 0.00092$, at which $M_{\bullet 1} = 16,976$ sites were identified as differentially methylated. In contrast, only

5,003 sites were identified as differentially methylated at the same FDR using Fisher's exact test (the FDR was controlled by requiring the q-value of each individual hypothesis to be ≤ 0.05 using the QVALUE R package downloaded from http://www.bioconductor.org). The eB test clearly shows improved power over Fisher's test, however, the majority of the true positive sites remain un-identified due to the limitation of small sample size (n_{ij}).

The same method was applied to the test of difference ($H'_0 : |\theta_{i1} - \theta_{i2}| \leq c$ vs. $H'_A : |\theta_{i1} - \theta_{i2}| > c$ at $c = 0.20$). An FDR≤ 0.05 was achieved at level 0.000088. A total of 1,630 sites were identified to have significantly pronounced difference (≥ 0.2) in methylation rates between the two samples.

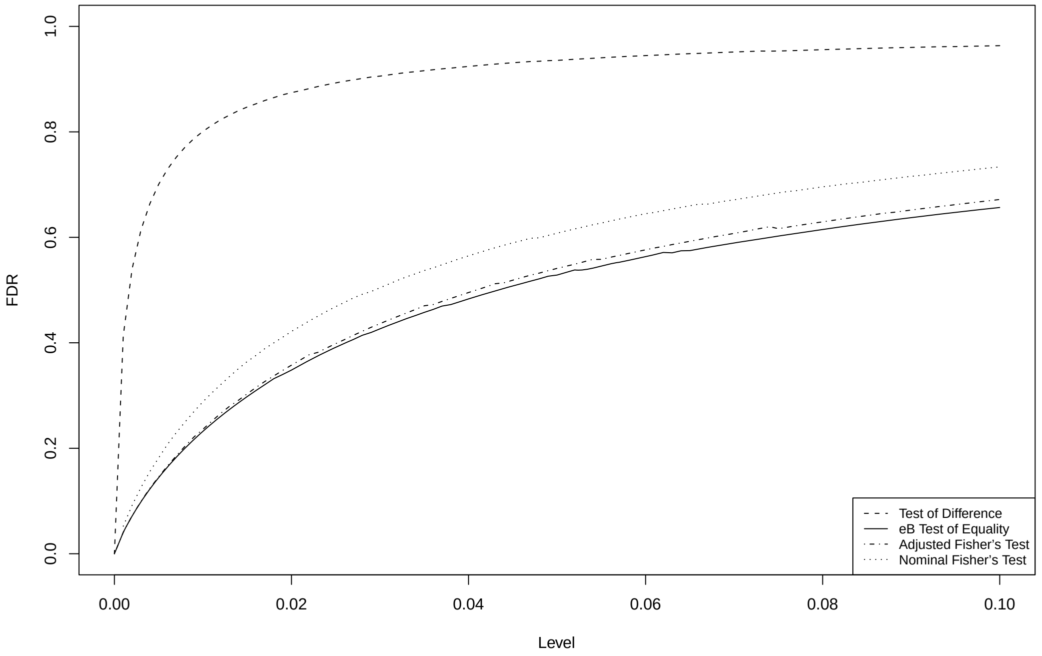
To gain further insights into the FDR behavior, in Figure 5, we plotted the FDR as a function of the level summarized from the simulation studies described above. The proposed eB method has a lower FDR than Fisher's exact test at all levels less than 0.1. Since the actual level of Fisher's exact test is typically lower than the nominal level, we also plotted the FDR vs actual level of Fisher's exact test. The actual level was assessed in the same way as for the eB approach, by finding the actual type I error rate in the simulated data at each given nominal level. The posterior log odds test has a uniformly lower FDR than the Fisher's exact test even after the adjustment for the difference in nominal and actual levels. In practice, as Fisher's test is always performed under the nominal level while the true level is never known, a comparison of the power or FDR under the nominal level is more meaningful. A spreadsheet with the locations on the genome that tested as positive is available at http://bioinfo.stats.northwestern.edu/∼jzwang/.

3. Discussion

In this paper, we showed the two advantages of proposed empirical Bayes approach over Fisher's exact test in methylation differentiation studies. This method is particularly useful when the number of reads at each site (or sample size) is small, while genome-wide data can provide rich information regarding the methylation rates across sites. Indeed, as shown in Figure 2, the fitted beta distribution has majority of probability mass concentrated around 0 and 1. This suggests that most sites have either very high or very low methylation rates. This prior information tends to shrink the posterior distribution of θ_{ij} towards the two ends. For example, if the observed $x_{ij} = 0$ and $n_{ij} = 2$, then it is highly likely that this site has a low methylation rate regardless of the small sample size, and vice versa. This strong prior information forms the basis for the power improvement when using posterior log odds as the test statistics.

Several possibilities exist for generalizations or refinements of this approach. Firstly, the bias issue in estimation of $\hat{M}_{0.}$ needs further investigation. It is not clear to us whether there is a causal relationship between the type I error level α and the bias. Secondly, currently all sites are treated as independent. In a real genome, it is possible that sites nearby may be correlated in methylation rate. Characterizing such dependence may help further improve the power of the eB approach. Finally, only two tissue samples were used to generate the data for this study; however, it will often be desirable to incorporate multiple specimens for each

Figure 5: FDR versus Level.



FDR is shown on the y-axis and level on the x-axis. The two curves for Fisher's exact test differ due to the overly conservative nature of the test. Values are estimates based on simulated data. The abbreviation "eB" stands for empirical Bayes.

condition. If methylation rates across specimens can be considered to be independent, then the density of the vector of methylation rates will be a product of beta densities. Similarly, the vector of positive reads will have probability mass given by the product of independent binomial pmfs. Because of independence, the posterior density for the vector of methylation rates will then be the product of beta densities, with the beta densities being the same as the posteriors would be if each specimen were treated separately. Once this distribution is known, the distribution of weighted averages of the methylation rates can be easily obtained.

Acknowledgements

This work is supported by NCI grant U54CA143869.

References

[1] J. Tost, DNA Methylation: An Introduction to the Biology and Disease-Associated Changes of a Promising Biomarker, Molecular Biotechnology, 44 (2010), 71-81.

[2] A. Bird, DNA methylation patterns in epigenetic memory, Genes and Development, 16 (2002), 6-21.

[3] S. Guibert, T. Forne, and M. Weber, Dynamic regulation of DNA methylation during mammalian development, Epigenetics, 1 (2009), 81-98.

[4] W. Reik, W. Dean, and J. Walter, Epigenetic reprogramming in mammalian development, Science, 293 (2001), 1089-1093.

[5] K. D. Robertson, DNA methylation and human disease, Nature Reviews Genetics, 6 (2005), 597-610.

[6] G. Heller, C. C. Zielinski, and S. Zöchbauer-Müller, Lung Cancer: From single-gene methylation to methylome profiling, Cancer Metastasis Review, 29 (2010), 95-107.

[7] B. C. Christensen, C. J. Marsit, A. E. Houseman, J. J. Godleski, J. L. Longacker, S. Zeng, R.-F. Yeh, M. R. Wrensch, J. L. Wiemels, M. R. Karagas, R. Bueno, D. J. Sugarbaker, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, Differentiation of Lung Adenocarcinoma, Pleural Mesothelioma, and Nonmalignant Pulmonary Tissues Using DNA Methylation Profiles, Cancer Research, 69 (2009), 6315-6321.

[8] D. T. Hsiung, C. J. Marsit, E. A. Houseman, K. Eddy, C. S. Furniss, M. D. McClean, and K. T. Kelsey, Global DNA Methylation Level in Whole Blood as a Biomarker in Head and Neck Squamous Cell Carcinoma, Cancer Epidemiology, Biomarkers and Prevention, 16 (2007), 108-114.

[9] W. Han, T. Wang, A. A. Reilly, S. M. Keller, and S. D. Spivack, Gene promoter methylation assayed in exhaled breath, with differences in smokers and lung cancer patients, Respiratory Research, 10 (2009), 86.

[10] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. W. Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, and Jian-Bing Fan, High-throughput DNA methylation profiling using universal bead arrays, Genome Research, 16 (2006), 383-393.

[11] L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. K. Sung, I. Rigoutsos, J. Loring, and C. L. Wei, Dynamic changes in the human methylome during differentiation, Genome Research, 20 (2010), 320-331.

[12] H. Gu, C. Bock, T. S. Mikkelsen, N. Jäger, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander, and A. Meissner, Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution, Nature Methods, 7 (2010), 133-136.

[13] E. A. Houseman, B. C. Christensen, R.-F. Yeh, C. J. Marsit, M. R. Karagas, M. Wrensch, H. H. Nelson, J. Wiemels, S. Zeng, J. K. Wiencke, and K. T. Kelsey, Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions, BMC Bioinformatics, 9 (2008), 365.

International Journal of Mathematics and
Computer Science, 5(2010), no. 2, 101-122

Bayesian Methods in Multi-Color Optical Mapping

Liping Tong[1,2]

[1] Department of Mathematics and Statistics
Loyola University Chicago
1032 W. Sheridan Road
Chicago, IL 60660, USA

[2] Department of Preventive Medicine and Epidemiology
Loyola University Medical Center
2160 S. First Ave.
Maywood, IL 60153, USA

e-mail: ltong@luc.edu

(Received July 7, 2010, Revised Oct. 27, 2010, Accepted Nov. 1, 2010)

Abstract

In this study design, data consist of noisy observations of multiple copies of a DNA molecule of interest. The main goal is to construct a physical map with lists of colors and positions. The secondary goal is to estimate error rates and assess uncertainty in parameter estimations. The existing maximum likelihood estimation (MLE) method works well when reasonably large error rates were introduced. However, besides the difficulties due to incomplete and unbounded likelihood, the MLE method does not provide an easily interpretable assessment of uncertainty in the discrete parameter space of sequence of colors. In this paper, we propose the Markov chain Monte Carlo (MCMC) and reversible jump MCMC methods to overcome the difficulties in an MLE procedure and to search and evaluate the space of color sequences. These methods are not only useful for this particular study design, but also important in general to show how to combine discrete and continuous distributions and effectively apply MCMC methods in a complicated situation.

Key words and phrases: Optical mapping, Hidden Markov model, Markov chain Monte Carlo, Bayesian factor, Reversible jump MCMC

AMS subject classification: 62J05, 62J07, 62H35, 62P10

ISSN 1814-0424 ©2010,http://ijmcs.future-in-tech.net

1. Introduction

Optical restriction mapping is a non-electrophoretic approach first developed by David Schwartz and his colleagues in 1993 [18]. This method offers a strategy for restriction mapping that overcomes many of the limitations of the conventional method and has been widely accepted and used since then [17]. However, as it is pointed out by Laurie Mets et al. (1999) [13], since the only information output of the original optical mapping method is ordered fragment sizes, for the method to be useful, fragment sizes must be determined with high accuracy, which is difficult to achieve and therefore limits the applications of optical mapping. In contrast, the multi-color optical mapping procedure developing by Laurie Mets and colleagues not only provides relative positions, but also distinguishes different recognition sites using distinct colors [15]. Therefore, more fuzzy fragment size measurement is allowed since the additional color information greatly improves the accuracy in a physical map construction [19].

The multi-color optical mapping method switches away from restriction sites as mapping landmarks to more general DNA binding probes. Some distinct recognition sites with length of 6 or 7 letters, such as CCTCTTT, are chosen. Fluorochromes (i.e. colored fluorescent beads) are bound to probes to mark and locate the recognition sites along multiple copies of the DNA molecule of interest, with each color corresponding to a different recognition site or DNA word. Multiple copies of the DNA molecule of interest, with probes attached, are spread and viewed with an optical microscope. The relative order and positions of colors are determined and measured within the limits of resolution of the microscope. The main goal is to construct a physical map with lists of recognition sites and positions from these noisy observations.

The statistical problem of map estimation for the multi-color optical mapping method can be thought either as a "multi-color" version of the map construction problem that has been considered by Anantharaman et al. (1997) [1], Lee et al. (1998) [10], Parida (1998) [14], Karp et al. (2000) [7] [8] and others, or as a continuous position version of the multiple sequence alignment problem that has been considered by Feng & Doolittle (1987) [3], Lipman et al. (1989) [11], Waterman (1995) [20] and many other researchers. However, as we explain below, the major problems and concerns in a multi-color optical mapping are very different from both cases.

One major issue in a "one-color" optical mapping design is the orientation uncertainty. When the molecule is laid out on a surface, the left-to-right or right-to-left order is lost. Though some elaborate biochemical methods can be used to determine the orientation, there is still uncertainty. This issue also occurs in a multi-color optical mapping process. But, it becomes much easier to handle due to the additional color information in the observations [19]. However, the sequence of colors in a multi-color procedure also complicates the problem and results in a very large discrete parameter space. Let c be the number of distinct colors (recognition sites), B be the sequence of colors and N be the number of sites on the underlying true map. Then the sequence of colors B has c^N possible values. Taking into account that N is also unknown results in an even larger space. Thus, some of the major considerations in the analysis of the multi-color optical mapping data are completely different from those in the one-color case.

In a traditional multiple sequence alignment problem, the distance between two adjacent positions is usually unknown and therefore cannot affect an alignment. In addition, since the sequences in comparison usually come from different sources, the final map is usually constructed based on the assumption of evolution history. However, in a multi-color optical mapping design, observations are multiple copies of the same DNA molecule of interest, and the distances between adjacent colors can be determined under an optical microscope. The additional distance information is important and must be considered in an optical mapping since it can vary from several hundred base pairs (BP) to hundreds of thousands BPs. Adding distance information becomes even more challenging when (unknown) scaling of distance measurements varies from molecule to molecule.

The MLE method proposed by Tong et al. (2007) [19] has described a way for map estimation given certain types of errors, which worked well to estimate a map with acceptable error rates when reasonably large noises were introduced. However, the MLE does not provide an easily interpretable assessment of uncertainty in the discrete parameter space of sequence of colors, which can be easily addressed by a Bayesian model. In addition, the Bayesian model provides an alternative of MLE for parameter estimation and can work better in certain situations, especially when the unbounded likelihood becomes an issue. In fact, the MLE procedure breaks down frequently when the error rates are as high as $f_n = 0.3$, $f_p = 0.02$, and $\sigma^2 = 2$, while the MCMC procedures are pretty stable.

This paper is structured as follows. In Methods section, the statistical model is described and the prior probabilities on parameters and hidden variables of alignments are proposed. Then the updating procedures for the MCMC and reversible jump MCMC are introduced. In Results section, these two MCMC procedures are tested and compared with the MLE method using simulated data sets based on the bacteriophage lambda genome, which contains about 50k nucleotide pairs along the linear double-stranded DNA molecule.

2. Methods

Two situations are considered to perform Bayesian inference using the MCMC method. In the first situation, the number and sequence of colors, (N, B), on the candidate map is fixed and assumed to be the true value of the underlying map, which might or might not be correct. The other parameters are then sampled from their posterior distributions conditional on (N, B) and observed data $\mathcal{D}$. In this situation, estimations of the other parameters can be calculated and compared with those from some other methods such as MLE. In addition, Bayes factors can be computed to compare candidate maps, which can be those inferred from some other procedures, such as those that are found to give the highest several maximized likelihoods using the MLE method. In the second situation, a reversible jump MCMC procedure is constructed to allow updating the values of (N, B). Comparing to the MLE method, the Bayesian model not only provides natural assessment of uncertainty in parameter estimations, but eliminates computational issues due to incomplete likelihood in the expectation step and the difficulties due to the unbounded likelihood problem in the maximization step.

2.1. Notation

We use the same error models as those in Tong et al. (2007) [19]. To make this paper concise but still easy for reading, only necessary notations and assumptions are introduced here.

The true (unknown) map is denoted by $\mathcal{H} = (h_1, b_1; h_2, b_2; \cdots ; h_N, b_N; L)$, where $N \geq 1$ is the number of sites, $L > 0$ is the length of the map, and H and B are the corresponding position and color sequences with $H = (h_1, h_2, \cdots , h_N)$, $B = (b_1, b_2, \cdots , b_N)$, where $0 \leq h_1 < h_2 < \cdots < h_N \leq L$ and $b_j \in \mathbf{C}$. Here h_t is the position of the t^{th} site on the map relative to one end of the DNA molecule (where the map orientation is chosen so that position values increase in the same direction on the map as on the oriented molecules), b_t is the color of the t^{th} site on the true map, and $\mathbf{C}$ is the set of possible colors. Assume there are M observations $\mathcal{D}_1, \mathcal{D}_2, \cdots , \mathcal{D}_M$, which are obtained from M independent copies of the DNA molecule with true map $\mathcal{H}$. Let m_j be the number of observed sites on the j^{th} observation $\mathcal{D}_j$, where $\mathcal{D}_j = (s_{j1}, c_{j1}; s_{j2}, c_{j2}; \cdots ; s_{jm_j}, c_{jm_j})$ includes position sequence $S_j = (s_{j1}, s_{j2}, \cdots , s_{jm_j})$, $0 = s_{j1} < s_{j2} < \cdots < s_{jm_j} < \infty$, and color sequence $C_j = (c_{j1}, c_{j2}, \cdots , c_{jm_j})$, $c_{jk} \in \mathbf{C}$. Here s_{ji} and c_{ji} represent the observed position (relative to s_{j1}) and color, respectively, of the i^{th} site on the j^{th} observation. In addition, assume that all the observations $\mathcal{D}_1, \mathcal{D}_2, \cdots , \mathcal{D}_M$ can be oriented relative to one another without error since it is not a major concern here [19]. Use $0 < \alpha_j < \infty$ and $0 \leq \beta_j \leq L - \alpha_j s_{jm_j}$ to represent the scale and shift of the jth observation $\mathcal{D}_j$. Then the standardized position $s^*_{jt} = \alpha_j s_{jt} + \beta_j$ would allow positions from different observations to be comparable.

2.2. Error Model

Assume that each site on $\mathcal{D}_j$ aligns to at most one site on $\mathcal{H}$, and each site on $\mathcal{H}$ aligns to at most one site on $\mathcal{D}_j$. In addition, assume that (1) the false negative sites relative to the true DNA molecule map are determined by independent Bernoulli trials, with the probability of missing a site being $0 < f_n < 1$; (2) the standardized positions of false positive sites appearing on the observed DNA molecule follow a Poisson process with rate $f_p > 0$ and the colors of false positives are independently and uniformly chosen from $\mathbf{C}$; (3) the joint density of the standardized positions is $f\left(s^*_{jt_1}, s^*_{jt_2}, \cdots , s^*_{jt_{d_j}}\right) \propto \exp\left\{-\sum_{i=1}^{d_j}\left(s^*_{jt_i} - h_{r_i}\right)^2/2\sigma^2\right\} \cdot 1\left\{0 \leq s^*_{jt_1} \leq s^*_{jt_2} \leq \cdots \leq s^*_{jt_{d_j}} \leq L\right\}$, where d_j is the number true positive sites on $\mathcal{D}_j$, $t_1 < t_2 < \cdots < t_{d_j}$ are the indices for the set of true positive sites on D_j, $r_1 < r_2 < \cdots < r_{d_j}$ are the indices for their corresponding sites on the true map $\mathcal{H}$ and σ^2 is the common variance.

There are two types of sites in each observation $\mathcal{D}_j$: either a true positive site meaning that it is the manifestation of a site on the true map, or a false positive site meaning that it is not the manifestation of any site on the true map. Given an underlying map, true positive, as well as false positive sites with standardized positions can be simulated according to the above error model. Let the list of positions is denoted by $s^*_{j1}, \cdots, s^*_{jm_j}$. The observed positions on $\mathcal{D}_j$ are then calculated by $s_{jt} = (s^*_{jt} - \beta_j)/\alpha_j$, where $\beta_j = s^*_{j1}$ and $\log(\alpha_j) \sim N(0, 0.35^2)$. For simplicity and easy computation, allow $-\beta_j/\alpha_j < s_{j1} < s_{j2} < ... < s_{jm_j} < (L - \beta_j)/\alpha_j$ in the model although in the data we always have $0 = s_{j1} < s_{j2} < ... < s_{jm_j} < (L - \beta_j)/\alpha_j$.

The "aliasing" error (a color is misspecified as another) is not modeled because it is expected to occur at a low enough rate that each instance can be accounted for as a combination of a false positive and a false negative. There are two other scenarios that might happen for a long entangled genome. First, the observed sequence of colors may not be in correct order. Second, the observed sequence might be a fragment that is much shorter than the true map. These two scenarios are not considered here since they are not likely to happen for a short simple genome, such as the bacteriophage lambda genome we use in the simulation study.

2.3. Hidden Variables

The (hidden) alignment variables are defined as follows. For $j = 1, 2, \cdots, M$, $t = 1, 2, \cdots, N$, let $W_{jt} = 1$, if the t^{th} site on the map $\mathcal{H}$ is observed on the j^{th} observation $\mathcal{D}_j$; $W_{jt} = 0$, otherwise. If $W_{jt} = 1$, define Q_{jt} as the observed position (before standardization) on $\mathcal{D}_j$ for the t^{th} site of $\mathcal{H}$; if $W_{jt} = 0$, define $Q_{jt} = Q_{jt-1}$. For simplicity, we have the convention: $Q_{j0} = -\beta_j/\alpha_j$. Since, under the model, the observed sites are in correct order, we must have $-\beta_j/\alpha_j \leq Q_{j1} \leq Q_{j2} \leq \cdots \leq Q_{jN} \leq (L - \beta_j)/\alpha_j$, for each $j = 1, 2, \cdots, M$. From the definition of W and Q and the modelling assumptions above, it can be verified that for each j, $\{(W_{jt}, Q_{jt})\}_{1 \leq t \leq N}$ is a Markov Chain.

2.4. Prior Distributions

The complete set of model parameters includes L, f_p, f_n, σ^2, $\alpha_1, \cdots, \alpha_M$, $\beta_1, \cdots, \beta_M$, N, $b_1, \cdots$, b_N, $h_1, \cdots, h_N$. In order to obtain an identifiable parameterization, we must fix one of the following parameters: map length L, false positive rate f_p, variance σ^2, and scale parameters α_1, $\cdots$, α_M. For computational reasons, we choose to fix L and re-scale all the positions on the j^{th} observation by multiplying by the factor L/s_{m_j}, which ensures the new scale $\alpha_j \in (0, 1)$, $1 \leq j \leq M$. Consider the following priors.

1. The false positive rate f_p follows Gamma (a_p, b_p). In practice, a convenient choice is to set $a_p = b_p = \sqrt{m_{min}/L}$, where $m_{min} = \min_j\{m_j\}$ is the minimum number of sites on an observation.

2. The false negative rate f_n follows Beta (a_n, b_n). In practice, if a rough estimate for f_n is available and denoted by $\hat{f}_n$, let $a_n = 1, b_n = (1 - \hat{f}_n)/\hat{f}_n$. Otherwise, let $a_n = b_n = 1$.

3. The common variance σ^2 follows Inverse Gamma (a_σ, b_σ). In practice, use $a_\sigma = b_\sigma = 1/(5\hat{\sigma})$, where $\hat{\sigma}$ is an estimate (not necessarily accurate) for σ. From our MCMC practice, it seems robust when choosing different a_σ and b_σ as long as they are not too large.

4. The scale parameters $\alpha_1, \alpha_2, \cdots, \alpha_M$ are i.i.d. Beta (a_α, b_α). The choice for a_α and b_α is similar to a_n and b_n.

5. The shift parameters $\beta_1, \beta_2, \cdots, \beta_M$ are i.i.d. Uniform $(0, L)$. Note that the likelihood will be 0 when $\beta_j > L - \alpha_j s_{jm_j}$.

6. The number of sites N on map $\mathcal{H}$ follows a Poison distribution with mean λ_N. In practice, let λ_N be the average number of sites on all observations divided by the map length L.

7. The positions $(h_1, h_2, \cdots, h_N)$ are the order statistics of i.i.d. Uniform $(0, L)$ random variables.

8. The color b_t at the t^{th} site, $t = 1, \cdots, N$, is uniformly chosen from $\mathbf{C}$, the set of colors.

Note that the choice of those prior distributions is mainly for computational convenience and assumes no knowledge about the true underlying model. These priors should be reasonable but do not have to be consistent with the true underlying model. In fact, though we assume a prior distribution for each of the parameter in the estimation procedure, but in the true model used for simulation, some parameters are simply unknown constants, such as f_p, f_n, and σ^2; some are obtained from real data with unknown distributions, such as N, $h_1, \cdots, h_N$, $b_1, \cdots, b_N$; some have completely different distributions, for example, $\alpha_1, \cdots, \alpha_M$ are generated from a lognormal distribution in the simulation but are assumed a beta distribution in the estimation.

2.5. MCMC Updating Procedures, Given $\mathbf{N}$ and $\mathbf{B}$

We fix N and $B = (b_1, b_2, \cdots, b_N)$ throughout this subsection. Given initial values for all the other parameters $\theta = (f_p, f_n, \sigma^2, \alpha_1, \cdots, \alpha_M,\ \beta_1, \cdots, \beta_M, h_1, \cdots, h_N)$ and hidden variables w and q, update the following sequentially.

1. The alignments w and q by a Metropolis-Hastings sampling strategy [12]. the idea here is to propose new values of w and q based on the approximation of the transition probability functions that are easy to compute [19]. Then adjust using acceptance ratio for a correct distribution. Since the approximation is really close to the exact ones, the acceptance ratio is high in general.

2. The false positive rate f_p and false negative rate f_n by the full conditional distributions [4].

3. The variance σ^2 by a Metropolis-Hastings step since the full conditional distribution of σ^2 is hard to obtain directly.

4. The j^{th} scale α_j and shift β_j by a small or big step. In this step, we first propose a probability distribution to decide which observation to be updated. Then, to update the scale α_j, we propose a Beta distribution according to the deviations of aligned sites on $\mathcal{D}_j$, where the deviation is defined to be the difference between the standardized observed position and the corresponding true position. If most of the deviations are positive, we will have more probability to propose a smaller scale, and vise versa. Finally, to update β_j, in a small step we use the current available alignment information to propose a new value of β_j, while in a big step we propose a new value of β_j according to a uniform distribution.

5. The t^{th} position h_t by a small or big step. Similarly to the updating of shift parameters, in a small step, the new value of h_t is proposed according to the current available alignment information, while in a big step when the current alignment is not reliable the new value of h_t is sampled from a uniform distribution.

The details of the above updating process are described in Appendix (section 2).

2.6. The Bayes Factor

Consider two different sequences of colors (N_1, B_1) and (N_2, B_2). If $\mathcal{D}$ denotes the actual observations and $Pr\{\mathcal{D}|N_i, B_i\}$ denotes the conditional probability distribution of observations under color sequence (N_i, B_i), $i = 1, 2$, the Bayes factor BF $= Pr\{\mathcal{D}|N_1, B_1\}/Pr\{\mathcal{D}|N_2, B_2\}$ provides the relative weight of evidence for the sequence (N_1, B_1) compared to the sequence (N_2, B_2) [5]. Since $Pr\{\mathcal{D}|N_i, B_i\}$ is not available explicitly, it must be computed as a marginalization by integrating over parameters θ. Specifically, since

$$\begin{aligned}
\frac{1}{Pr\{\mathcal{D}|\ N_i, B_i\}} &= \frac{Pr\{\theta, W, Q|\mathcal{D}, N_i, B_i\}}{Pr\{\theta, W, Q, \mathcal{D}|N_i, B_i\}} \\
&= \int_{\theta\in\Theta} \frac{Pr\{\theta, W, Q|\mathcal{D}, N_i, B_i\}Pr\{\theta|N_i, B_i\}}{Pr\{\theta, W, Q, \mathcal{D}|N_i, B_i\}} d\theta \\
&= \int_{\theta\in\Theta} \frac{Pr\{\theta, W, Q|\mathcal{D}, N_i, B_i\}}{Pr\{W, Q, \mathcal{D}|\theta, N_i, B_i\}} d\theta,
\end{aligned}$$

and we are able to sample $\theta_{(k)}, W_{(k)}$ and $Q_{(k)}$ from distribution $Pr\{\theta, W, Q|\mathcal{D}, N_i, B_i\}$ using MCMC (see Appendix, section 2), then

$$Pr\{\mathcal{D}|N_i, B_i\} \approx \frac{n \cdot N_{WQ}}{\sum_{k=1}^{n} 1/L(\theta(k); \mathcal{D}, w_{(k)}, q_{(k)})} \qquad (2.1)$$

where $L(\theta(k); \mathcal{D}, w_{(k)}, q_{(k)})$ is the complete data likelihood and can be calculated using formula (1) in Appendix (section 1), and N_{WQ} is the number of all possible alignments, given

$\mathcal{D}$ and (N_i, B_i). The calculation of N_{WQ} is similar to the calculation of likelihood with backward and forward variables, except that now the complete data likelihood is the constant 1.

2.7. Reversible Jump MCMC to Update **N** and **B**

When variation in N and B is allowed, the parameter space is no longer constant. In this situation, the reversible jump MCMC strategy can be used to account for discrepancies in parameter space [6]. In this subsection, we propose such a reversible jump MCMC procedure to directly estimate the posterior probabilities for N and B.

Consider the step to update a position. Some possible transitions are: (1) Type U: update h_t to h_t^*, and keep N and b_t unchanged; (2) Type B: "birth" of a randomly chosen color at a randomly chosen site; (3) Type D: "death" of a randomly chosen site. Let u'_k, b'_k and d'_k be the probabilities for types U, B and D transition from the k^{th} to the $k+1^{st}$ steps of the sampler. Suppose N_{max} and N_{min} are the maximum and minimum values for N. Let $0 < \delta < 1$ and $0 < \rho < 1$ be constants. Use N_k to denote the number of sites on a candidate map at step k. Let $u'_k = \delta$. In addition, if $N_{min} < N_k < N_{max}$, let $b'_k = (1-\delta)\rho$ and $d'_k = (1-\delta)(1-\rho)$; if $N_k = N_{min}$, let $b'_k = 1-\delta$ and $d'_k = 0$; if $N_k = N_{max}$, let $b'_k = 0$ and $d'_k = 1-\delta$. Randomly select the transition type using u'_k, b'_k and d'_k to update from $(N_k, B_k, H_k, W_k, Q_k)$ to $(N_{k+1}, B_{k+1}, H_{k+1}, W_{k+1}, Q_{k+1})$.

If a transition type U is chosen, then the updating procedure for a specific position and the related alignments has been described in Appendix (subsection 2.5). If a transition type B is chosen, we describe in Appendix (section 3) how to decide the color, position, and corresponding alignments, and how to calculate the acceptance ratio. If a transition type D is chosen, the site (b_t, h_t) and corresponding alignments (w_{jt}, q_{jt}) are simply deleted. The acceptance ratio has the same form as the one for a type B transition, with appropriate change of labelling of the variables, and the ratio terms inverted.

3. Results

3.1. Data Simulation

Bacteriophage lambda was originally discovered from E. coli in 1953 by E. Lederberg and J. Lederberg [9]. It has been a very important tool in the study of molecular biology since then. The bacteriophage lambda genome has a linear genetic and physical map, sometimes presented in circular representation because the molecule circularizes at the cohesive ends during some stages of its life circle. The total genome size is 48,502 base pairs. The genome sequence is obtained from Gen Bank, with accession number # NC_001416. The recognition sites and their corresponding colors are defined in Table 1. The total number of occurrences summed across all these colors in the lambda genome is 26.

Table 1: Recognition Sites and Colors

Name	Recognition Site			Shape
I	TTTTCCC	or	CCCTTTT	circle
J	TTTCTCC	or	CCTCTTT	diamond
K	TTCTTCC	or	CCTTCTT	triangle
M	TTCTCTC	or	CTCTCTT	cross

The underlying true map is obtained by by considering the bacteriophage lambda genome sequence from base pairs 1 to 28,502 (out of 48,502 total base pairs), which includes 9 recognition sites. We re-scale the positions and fix the map length to be $L = 100$ throughout this paper. Twenty observations are generated using the error model described in subsection 2.2, with $f_n = 0.3$, $f_p = 0.02$, and $\sigma = 1.0$. With these values, the expected number of true positive sites per observation is 6.3, the expected number of false positive sites per observation is 2, and the expected number of false negative sites per observation is 2.7. We also did simulation studies for some other choices of parameter values. The conclusion is similar. Since in practice the above parameter values are expected to be about the same, only results for the above values are listed and discussed in the following.

3.2. Map Position, Scale, Shift and Error Parameter Estimates

In this subsection, consider the Bayesian approach in which the number of sites N and the sequence of colors B are fixed to be its true value. All the other parameters are updated according to the Gibbs and M-H prodecures in subsection 2.5. The run length of the MCMC is 10,000 iterations and the first 2000 are discarded for the analysis. The actual acceptance ratio for σ updating is 0.987, for the position updating is 0.500, and for the scale and shift updating is 0.112.

Figure 1 shows the sequence of sampled parameter values across the MCMC iterations in the left-hand column and a histogram of the posterior sample distributions in the right-hand column, for the error parameters f_n, f_p and σ. The plots suggest that this MCMC mixes reasonably well at least for the error parameters. In addition, we see that the posterior sample distributions for these parameters are approximately symmetric, with center close to the true values. Compared to the MLE estimates $(0.03, 0.36, 0.46)$ for (f_p, f_n, σ), the posterior sample means $(0.022, 0.301, 0.940)$ are much closer to the true parameter values $(0.02, 0.3, 1.0)$.

The results for posterior positions are showed in the MCMC columns in Tables 2. Because the ends of the molecule are poorly estimated, to make the estimates for positions and shifts comparable, consider estimates of h_1, $\triangle h_t = h_t - h_{t-1}$ for $2 \leq t \leq N = 9$ and $L - h_9$ in Table 2. To make them comparable, the j^{th} row (α_j) in the "true" and "MLE" columns are adjusted by multiplying L/s_{jm_j}, for $1 \leq j \leq M$, as what is done in the MCMC procedures.

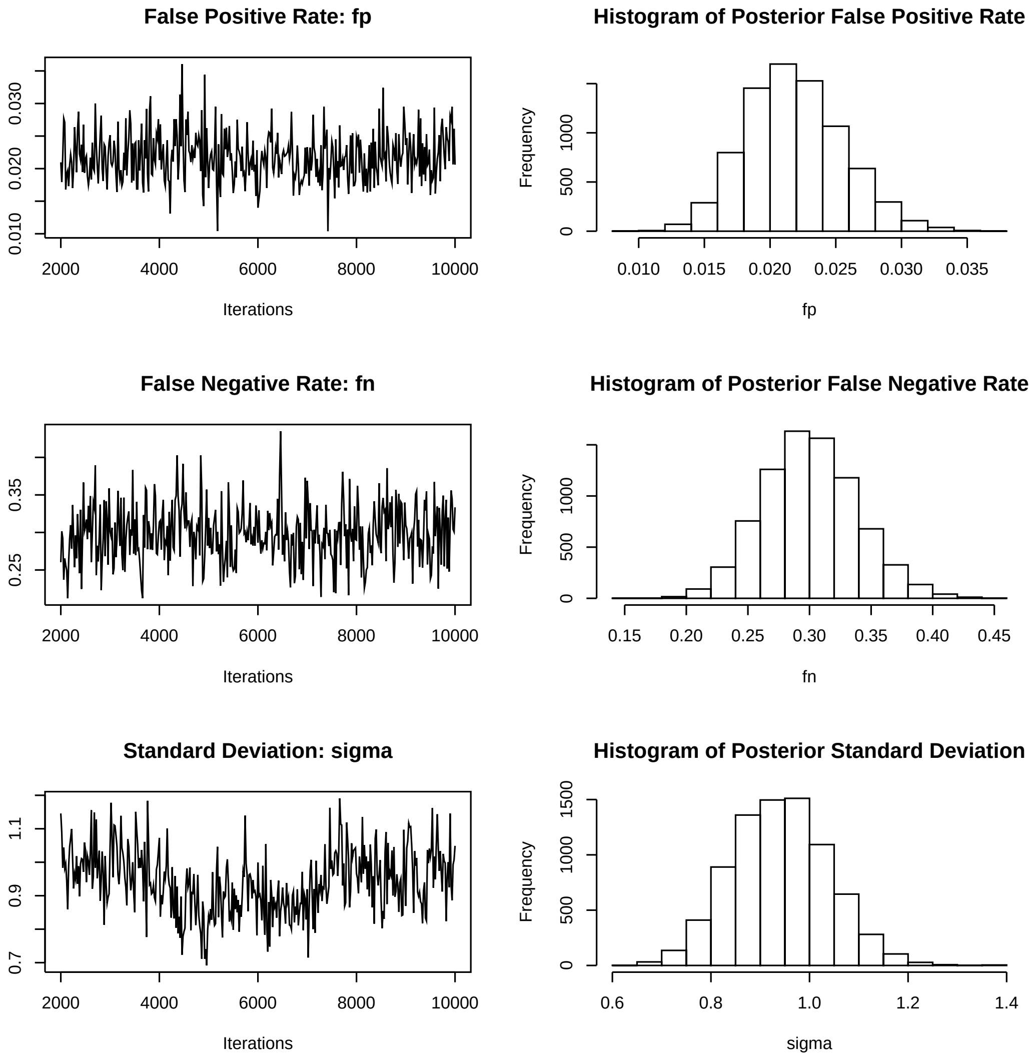


Figure 1: MCMC Iterations and Posterior Distributions of Error Parameters. The number of sites N and sequence of colors B on the map $\mathcal{H}$ are fixed. The true values of the error parameters are: $f_p = 0.02$, $f_n = 0.3$, $\sigma = 1$; the posterior means are: $\tilde{f}_p = 0.022$, $\tilde{f}_n = 0.301$,$\tilde{\sigma} = 0.940$; and the MLE are: $\hat{f}_p = 0.03$,$\hat{f}_n = 0.36$, $\hat{\sigma} = 0.46$.

From Table 2, we see that the MCMC estimates for positions tend to be closer to the true values than the MLEs do, especially when the distances between two consecutive sites are close (e.g. $\triangle h_4$ and $\triangle h_8$). In addition to reflecting genuine differences between the MLE and Bayesian approaches, this could also be related to approximations, valid for σ small relative to $\triangle h_t$, that were used in the MLE but not the Bayesian procedure. The posterior means for scales and relative shifts and are also close to their true values (results not shown).

Table 2: Position Comparisons of True, MLE, MCMC and Jump MCMCs

			MCMC	Jump I	Jump II	Jump III
	True	MLE	mean (s.d.)	mean (s.d.)	mean (s.d.)	mean (s.d.)
h_1	11.07	12.06	9.77 (0.70)	8.50 (0.76)	10.55 (0.46)	11.01 (0.63)
$\triangle h_2$	28.14	27.89	28.85 (0.47)	29.47 (0.50)	28.60 (0.46)	28.26 (0.61)
$\triangle h_3$	14.10	13.34	14.09 (0.43)	14.21 (0.51)	14.01 (0.42)	13.90 (0.43)
$\triangle h_4$	0.37	1.35	0.65 (0.42)	0.74 (0.40)	0.61 (0.40)	0.61 (0.36)
$\triangle h_5$	10.69	10.60	10.90 (0.39)	11.05 (0.47)	10.82 (0.41)	10.75 (0.39)
$\triangle h_6$	23.18	23.19	23.58 (0.48)	23.86 (0.39)	23.45 (0.40)	23.22 (0.45)
$\triangle h_7$	7.01	5.96	6.86 (0.48)	6.94 (0.34)	6.72 (0.37)	6.65 (0.39)
$\triangle h_8$	0.30	0.77	0.27 (0.25)	0.33 (0.26)	0.27 (0.23)	0.25 (0.23)
$\triangle h_9$	3.66	2.65	3.84 (0.47)	3.94 (0.50)	3.95 (0.45)	3.88 (0.45)
$L - h_9$	1.48	2.19	1.18 (0.63)	0.90 (0.47)	0.88 (0.54)	1.36 (0.62)

3.3. Bayes Factors for the Candidate Sequences of Colors

To assess the relative weight of evidence for different sequences of colors (N, B) using the Bayes factors, choose the sequences of colors with the top 4 profile likelihood values from the maximum likelihood procedure described by Tong et al., 2007 [19]. The run length of the MCMC is 10000 iterations and the first 2000 are discarded for the analysis. Table 3 lists the values of the log Bayes factors (BF), the posterior means and s.d.'s of the error parameters f_n, f_p and σ (with sequence of colors fixed). The log(BF) column are the values of $\log Pr\{\mathcal{D}|N_i, B_i\} - \log Pr\{\mathcal{D}|N, B\}$, where (N_i, B_i) is the top i^{th} candidate sequence of colors, and N, B is the true sequence of colors.

The first sequence of colors is the same as the true sequence of colors. Therefore, we should have $\log Pr\{\mathcal{D}|N_1, B_1\} - \log Pr\{\mathcal{D}|N, B\} = 0$. Our actual value in Table 3 for this quantity is -0.11, which is a little bit different from 0. This could be an effect of small number of MCMC iterations. The values of Bayes factor for these four maps decrease and the posterior means for error parameters (f_p, f_n) increase. From more detailed check of these maps, note that the number of sites for these four candidate maps are 9, 10, 11, 12 (from the best to the fourth best). Therefore, higher posterior means for f_n are expected when taking them as candidate maps. Meanwhile, this high probability for false negatives could

Table 3: Log Bayes Factors for the Top 4 Sequences of Colors

	$\log L(\theta; \mathcal{D})$	log(BF)	f_n (s.d.)	f_p (s.d.)	σ (s.d.)
1	-424.23	-0.11	0.31 (0.040)	0.022 (0.0038)	0.93 (0.092)
2	-437.85	-11.78	0.39 (0.041)	0.023 (0.0042)	0.88 (0.096)
3	-439.64	-13.10	0.40 (0.045)	0.031 (0.0082)	0.82 (0.101)
4	-441.04	-38.09	0.49 (0.053)	0.094 (0.0104)	0.89 (0.300)

introduce more variability in the posterior samples of other parameters.

3.4. Reversible Jump MCMC

In this subsection, all the parameters, including number of sites N and sequence of colors B, are subject to updating. The reversible jump MCMC is able to give posterior distributions for the parameters that take into account the uncertainty in (N,B). Thus, it would be of interest to compare those with the posterior distributions of parameters from MCMC with (N,B) fixed and with the sampling distributions from MLE.

We run three separate reversible jump chains, in which the starting values of (N, B, H) are based on the first (Jump I), second (Jump II) and third observation (Jump III) respectively. All the staring values for the other parameters are obtained from the pairwise comparison method described in Tong et al. (2007) [19]. Then the map, observations and related parameters are re-scaled to make $L = 100$. The length of the first jump MCMC is $10,000$ iterations, and the length for each of the other two is $20,000$ iterations. The first 20% are discarded for the analysis. In all three of these jump MCMCs, the true sequence of colors is found quickly (within 2000 iterations), and the chain stays close to the true sequence of colors stably thereafter. The three maps used as starting points for the 3 jump chains (after adjustment for scale and shift obtained from pairwise comparison) are shown and compared with the true map in Figure 2. Note that all three starting maps deviate from the true map in both sequence of colors and positions.

The reversible jump MCMC turns out to be surprisingly successful in estimating parameters, especially sequence of colors. The posterior probabilities for the true sequence of colors in these 3 reversible jump MCMC are 93.2%, 98.0% and 97.0%, respectively. The posterior sample means and standard deviations for the error parameters f_p, f_n and σ are listed in Table 4, where the "partial" column considers the iterations with true sequence of colors only and the "all" column considers all the iterations. Table 4 shows very similar posterior means and standard deviations for error parameters using MCMC with N and B fixed to their true values, and using any one of the three jump chains, both when all iterations are used and when only those having the most likely sequence of colors (N, B) are used. This suggests that the posterior distributions of (f_p, f_n, σ) are stable. Notice that the standard deviations for the partial set of iterations are uniformly smaller than the standard

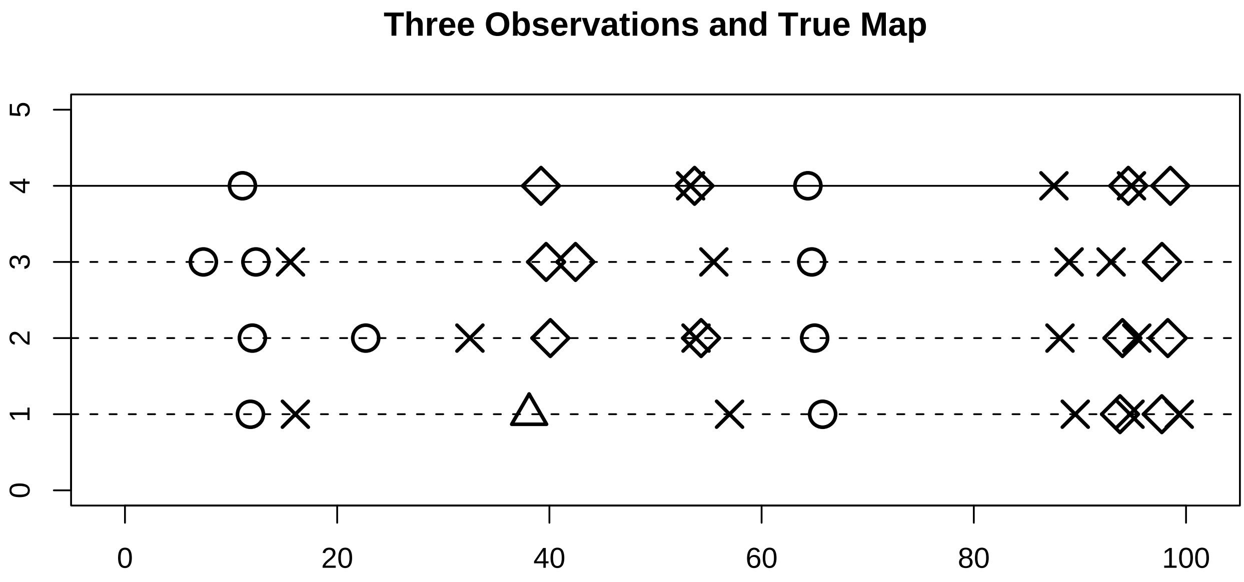


Figure 2: The Three Starting Maps, Compared with True Map. The first line (solid) shows the true map. The remaining lines (dotted) show the first 3 observations (after adjustment for scale and shift that are obtained from pairwise comparison), which are used as the starting maps in jump MCMC I, II and III, respectively.

deviations for the full set of iterations, which is reasonable because the uncertainty in N and B would increase the uncertainty in the posterior distributions of (f_p, f_n, σ). Table 2 also gives the posterior means and standard deviations of positions using partial iterations of reversible jump MCMC, which have very similar results to the posterior means and standard deviations using iterations of MCMC when N and B are fixed.

Table 4: Estimates of Error Parameters in Reversible Jump MCMC

True		MCMC	Jump I		Jump II		Jump III	
			all	partial	all	partial	all	partial
$f_p = 0.02$	mean	0.022	0.022	0.022	0.022	0.022	0.022	0.022
	s.d.	0.004	0.004	0.004	0.004	0.004	0.004	0.004
$f_n = 0.3$	mean	0.301	0.311	0.307	0.299	0.297	0.301	0.299
	s.d.	0.038	0.042	0.039	0.039	0.038	0.041	0.039
$\sigma = 1.0$	mean	0.924	0.977	0.968	0.947	0.946	0.942	0.939
	s.d.	0.098	0.100	0.095	0.087	0.087	0.104	0.103

4. Discussion

In this paper, we have proposed the Bayesian method to estimate the underlying physical map of a DNA molecule and the corresponding error rates, using the multi-color optical mapping data.

There are two situations when the MCMC strategies can be used. First, if there are candidate maps already, which might be obtained from MLE or some other methods, the regular MCMC method assuming fixed N and B can be used to refined these candidate maps, to compare them and to estimate error rates. Second, if a candidate map shall be inferred from observed molecules, the reversible jump MCMC can be used to search for a map and to estimate other parameters.

The maximum likelihood inference works well with reasonably high error rates, and mild violations to model assumptions. However, there are some considerations and concerns still left. First, the MLE method may not be able to find a non-singular solution when the true error rates are high. Second, the MLE method can not assess uncertainty in some of the parameter estimates since the parameter space itself changes.

The Bayesian method uses the exact log likelihood (although the MCMC method itself is of course, always approximate) and is able to assess the uncertainty of N and B (number of sites and sequence of colors on the map $\mathcal{H}$) using the marginal posterior distribution or Bayes factors. However, like all the other Bayesian methods [5], this one also requires intensive computation, which makes it practically difficult when M (number of observations) and m_j (number of sites on $\mathcal{D}_j$) are large.

Therefore, one realistic way is to first use likelihood based inference when M and m_js are large. Then use the MCMC methods for a subset of the whole map, such as our analysis only considering the first 9 sites on the lambda genome, to refine and verify the parameter values.

In the case when the observed sequences of colors may not be in the correct order, our hidden Markov and MCMC models should be adjusted to allow for mild crossover (within the limit of resolution). One more difficult problem is to construct a physical map from observations of multiple overlapping fragments, instead of complete genome copies, with orientations and amount of overlap not completely known. In this situation, the indicators for alignment variables do not necessarily start from 1 or end at N. We could in principle think the starting and ending indicators as parameters or hidden variables in the model. This could be a topic of future work.

Acknowledgements

This work was supported by grants Burroughs Wellcome Fund Interfaces #1001774 and Packard Foundation. The author is grateful to many useful suggestions from Professor Mary Sara McPeek. The author also appreciates a referee whose many detailed and constructive comments have much improved the paper.

References

[1] Thomas S. Anantharaman, Bud Mishra, and David C. Schwartz. Genomics via optical mapping II: Ordered restriction maps. J Comput Biol, 1997. 4: p.91-118.

[2] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistics functions of Markov chains. Ann. Math. Stat., 1970. 41: p.164-171.

[3] D.F. Feng, and R.F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution, 1987. 25: p.351-360.

[4] S. Geman, D. Geman. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. IEE Transactions on Pattern Analysis and Machine Intelligence, 1984. 6: 721741.

[5] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Markov chain Monte Carlo in practice. Chapman & Hall (CRC), 1st ed., 1996.

[6] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 1995. 82: p.711-732.

[7] Richard M. Karp, and Ron Shamir. Algorithms for optical mapping. J of Comput Biol, 2000. 7: p.303-316.

[8] Richard M. Karp, Itsik Pe'er, and Ron Shamir. An algorithm combining discrete and continuous methods for optical mapping. J Comput Biol, 2000. 7: p.745-760.

[9] E.M. Lederberg, and J. Lederberg. Genetic studies of lysogenicity in Escherichia coli. Genetics, 1953. 38: p.51-64.

[10] Jae K. Lee, Vlado Dancik, and Michael S. Waterman. Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. J Comput Biol, 1998. 5: p.505-515.

[11] D.J. Lipman, S.F. Altschul, and J.D. Kececioglu. A tool for multiple sequence alignment. Proceedings of the National Academy of Sciences of the USA, 1989. 86: p.4412-4415.

[12] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of static calculations by fast computing machines. Journal of Chemical Physics, 1953. 21: p.1087-1092.

[13] Laurie Mets. Direct multi-feature mapping of single DNA molecules. Proposal submitted, on March 31, 2000, to the research announcement RA00-14 "Fundamental research at the [BIO:INFO:MICRO] interface" published in the Commerce Business Daily 23, December, 1999.

[14] Laxmi Parida. A uniform framework for ordered restriction map problems. J Comput Biol, 1998. 5: p.725-739.

[15] Xiaohui Qu, David Wu, Laurens Mets, and Norbert F. Scherer Nanometer-localized multiple single-molecule fluorescence microscopy. PNAS, 2004. 101: p. 11298-11303.

[16] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 1989. 77: p257-286.

[17] Akhtar Samad, Edward J. Huff, Weiwen Cai, and David C. Schwartz. Optical mapping: a novel, single-molecule approach to genomic analysis. Genome Res, 1995. 5: p.1-4.

[18] D.C. Schwartz, X. Li, L.I. Hernandez, S.P. Ramnarain, E.J. Huff, and Y.K. Wang. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science, 1993. 262: p.110-114.

[19] L. Tong, L. Mets, M.S. McPeek. Likelihood-based Inference for Multi-Color Optical Mapping Data. Statistical Applications in Genetics and Molecular Biology, 2007. Vol. 6: Iss. 1, Article 5.

[20] M.S. Waterman. Introduction to computational biology. Chapman & Hall, 1995.

Appendix

1. Complete-Data Log Likelihood Function

Let $(\mathcal{D}, w, q)$ denote the complete data, where $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_M)$, $w = (w_1, \cdots, w_M)$ and $q = (q_1, \cdots, q_M)$, with $w_j = (w_{j1}, \cdots, w_{jN})$, $q_j = (q_{j1}, \cdots, q_{jN})$, where w_{jt} is a realization of W_{jt} and q_{jt} is a realization of Q_{jt}. The complete-data log likelihood function is $\log L(\theta; \mathcal{D}, w, q) = \sum_{j=1}^{M} \log L(\theta; \mathcal{D}_j, w_j, q_j)$, where

$$\begin{aligned}\log L(\theta; \mathcal{D}_j, w_j, q_j) \;=\;& (m_j - \sum_{t=1}^{N} w_{jt}) \log f_p - f_p L + \sum_{t=1}^{N} w_{jt} \log(1 - f_n) + \sum_{t=1}^{N} (1 - w_{jt}) \log f_n \\ &+ m_j \log \alpha_j - \sum_{t=1}^{N} w_{jt} \log \sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} w_{jt} (\alpha_j q_{jt} + \beta_j - h_t)^2 \\ &- \sum_{t=1}^{N} w_{jt} \log(\sqrt{2\pi}) + \log C(H, L, w_j, \sigma) \qquad (1)\end{aligned}$$

provided all the following hold: (1) $w_{jt} \in \{0, 1\}$ for each $1 \le t \le N$; (2) $-\beta_j/\alpha_j \le q_{j1} \le \cdots \le q_{jN} \le \frac{L-\beta_j}{\alpha_j}$; (3) $q_{j1} = -\beta_j/\alpha_j$ if $w_{j1} = 0$, and $q_{jt} = q_{jt-1}$ if $w_{jt} = 0$ and $2 \le t \le N$; (4) $q_{jt} \in \{-\beta_j/\alpha_j, s_{j1}, \cdots, s_{jm_j}\}$ for each $1 \le t \le N$; (5) $-\beta/\alpha \le s_{j1} < s_{j2} < \ldots < s_{jm_j} \le (L - \beta_j)/\alpha_j$. If any of these conditions fail to hold then the likelihood is 0. Here, $C(H, L, w_j, \sigma)$ is a normalizing factor and can be calculated as $C(H, L, w_j, \sigma) = [P\{0 \le X_1 \le X_2 \le \cdots \le X_{d_j} \le L\}]^{-1}$ where $X_i \sim N(h_{r_i}, \sigma^2)$ independently, $1 \le i \le d_j$, where $d_j = \sum_{t=1}^{N} w_{jt}$, and $r_1, r_2, \cdots, r_{d_j}$ are the values of t when $w_{jt} = 1$. If $d_j = 0$, then $C(H, L, w_j, \sigma) = 1$.

2. MCMC Updating Procedures, Given $\mathbf{N}$ and $\mathbf{B}$

Use $P(A)$ to denote the probability measure evaluated at event A and $Pr(\cdot)$, here and after, as a notation for probability measure functions of discrete random variables, for probability densities of continuous random variables and for products of these. At each step, write the current parameter and alignment values to be θ, w and q, and the proposed ones to be θ^*, w^* and q^*, which might or might not be accepted according to the acceptance ratio.

2.1. Update alignment variables

We first construct an observation sequence $\{Y_{jt}\}_{1 \le t \le N+1}$, $j = 1, \cdots, M$, of the latent-variable Markov chain purely for computational convenience [19]. Then define the observation distributions $b_{jt}(u,v;z) = Pr(Y_j|Q_{j,t-1} = u, Q_{jt} = v, W_{jt} = z)$, the transition probability distributions $a_{jt}(u,v;z) = P(Q_{j,t+1} = v, W_{j,t+1} = z|Q_{jt} = u)$ and the initial distribution $\pi_j(u,z) = P(Q_{j1} = u, W_{j1} = z)$. Following the approach due to Baum et al. (1970) [2] and described in Rabiner (1989) [16], we are able to find formulas to calculate the forward variable $\eta_{jt}(u,z) = Pr(Y_{j1},\cdots,Y_{jt}; Q_{jt} = u|W_{jt} = z)$ and backward variable $\tau_{jt}(u,z) = Pr(Y_{j,t+1},\cdots,Y_{j,N+1}|W_{jt} = z, Q_{jt} = u)$.

For each $j = 1,2,\cdots,M$, propose new alignment variables (W^*_{j1}, Q^*_{j1}), $\cdots$, (W^*_{jN}, Q^*_{jN}) as follows. When $t = 1$ and $u \in \{s_{j1},\cdots,s_{jm_j}\}$,

$$\begin{aligned} P\{Q^*_{j1} = u, W^*_{j1} = 1|\mathcal{D}_j,\theta\} &= \frac{\eta_{j1}(u,1)\tau_{j1}(u)}{\sum_{i=0}^{m_j}(\eta_{j1}(u,0)+\eta_{j1}(u,1))\tau_{j1}(u)} \\ P\{Q^*_{j1} = u, W^*_{j1} = 0|\mathcal{D}_j,\theta\} &= \frac{1\{u = s_{j0}\}\eta_{j1}(s_{j0},0)\tau_{j1}(s_{j0})}{\sum_{i=0}^{m_j}(\eta_{j1}(u,0)+\eta_{j1}(u,1))\tau_{j1}(u)}. \end{aligned} \tag{2}$$

When $t = 2,\cdots,N$ and $u,v \in \{s_{j1},\cdots,s_{jm_j}\}$,

$$\begin{aligned} P\{Q^*_{jt} = v, W^*_{jt} = 1|Q^*_{jt-1} = u,\mathcal{D}_j,\theta\} &= a_{j,t-1}(u,v;1)\cdot b_{jt}(u,v;1)\cdot\frac{\tau_{jt}(v)}{\tau_{j,t-1}(u)} \\ P\{Q^*_{jt} = v, W^*_{jt} = 0|Q^*_{jt-1} = u,\mathcal{D}_j,\theta\} &= 1\{v = u\}\cdot f_n\cdot\frac{\tau_{jt}(v)}{\tau_{j,t-1}(u)} \end{aligned} \tag{3}$$

Assume that the proposed values are $q^*_j = (q^*_{j1},\cdots,q^*_{jN})$ and $w^*_j = (w^*_{j1},\cdots,w^*_{jn})$. The acceptance ratio is

$$r_j(w,q) = \min\left\{1, \frac{L\{\theta;\mathcal{D}_j,w^*_j,q^*_j,\}}{L\{\theta;\mathcal{D}_j,w_j,q_j\}}\cdot\frac{P\{Q^*_j = q_j, W^*_j = w_j|\mathcal{D}_j,\theta\}}{P\{Q^*_j = q^*_j, W^*_j = w^*_j|\mathcal{D}_j,\theta\}}\right\}$$

where the first ratio is the complete data likelihood ratio for the j^{th} observation, which can be calculated using (1), and the second ratio becomes

$$\frac{P\{Q^*_{j1} = q_{j1}, W^*_{j1} = w_{j1}|\mathcal{D}_j,\theta\}}{P\{Q^*_{j1} = q^*_{j1}, W^*_{j1} = w^*_{j1}|\mathcal{D}_j,\theta\}}\cdot\prod_{t=2}^{N}\frac{P\{Q^*_{jt} = q_{jt}, W^*_{jt} = w_{jt}|Q^*_{jt-1} = q_{jt-1},\mathcal{D}_j,\theta\}}{P\{Q^*_{jt} = q^*_{jt}, W^*_{jt} = w^*_{jt}|Q^*_{jt-1} = q^*_{jt-1},\mathcal{D}_j,\theta\}},$$

which can be calculated using (2)-(3). Note that $r_j(w,q) = 1$ if $a_{jt}(u,v;z)$ is the exact transition probability function.

2.2. Update false positive and false negative rates

The new values f^*_p and f^*_n are sampled independently from their corresponding full conditional distribution and accepted with probability 1, where

$$f^*_p \sim \text{ Gamma }\left(\sum_j m_j - \sum_j\sum_t w_{jt} + a_p, b_p/(MLb_p+1)\right)$$

and

$$f^*_n \sim \text{ Beta }\left(\sum_j\sum_t(1-w_{jt}) + a_n, \sum_j\sum_t w_{jt} + b_n\right).$$

2.3. Update common variance

Sample the new variance σ^{*2} according to Inverse Gamma (g_1, g_2), where $g_1 = \sum_j \sum_t w_{jt}/2 + a_\sigma$, and $g_2 = \sum_j \sum_t w_{jt}(\alpha_j q_{jt} + \beta_j - h_t)^2/2 + b_\sigma)$, and accept with probability

$$r(\sigma) = \min\left\{1, \prod_{j=1}^{M} C(H, L, w_j, \sigma^*)/C(H, L, w_j, \sigma)\right\}.$$

2.4. Update scale, shift and related alignments

Since we want the probability to update α_j and β_j to be high when the alignment between the j^{th} observation and candidate map is bad, that is, there are few aligned sites, and vice versa, then let

$$P\{\text{select the } j^{th} \text{ scale and shift to update}\} \;=\; \frac{\sum_t (1 - w_{jt}) + 1}{\sum_j \sum_t (1 - w_{jt}) + M}. \tag{4}$$

We correct the probability by adding number 1 both on the numerator and denominator because we want to have a non-zero probability to update the scale and shift when all the sites on the candidate map find matched sites on an observation.

Let $K = \lfloor \frac{N+1}{2} \rfloor$ and define $\tilde{\beta}_j = \sum_t w_{jt}(h_t - \alpha_j q_{jt})/\sum_t w_{jt}$ if $\sum_t w_{jt} > 0$, $\tilde{\beta}_j = 0$ otherwise; $I_{jt} = 1$ if $w_{jt}(\alpha_j q_{jt} + \tilde{\beta}_j - h_t) > 0$, $I_{jt} = 0$ if $w_{jt}(\alpha_j q_{jt} + \tilde{\beta}_j - h_t) = 0$, $I_{jt} = -1$ if $w_{jt}(\alpha_j q_{jt} + \tilde{\beta}_j - h_t) < 0$; $a_1 = \alpha_j \left(N + \sum_{1 \le t \le K} I_{jt} - \sum_{K \le t \le N} I_{jt}\right) \left(\sum_{t=1}^{N} |I_{jt}|\right)/N$, and $a_2 = (1 - \alpha_j) \left(N - \sum_{1 \le t \le K} I_{jt} + \sum_{K \le t \le N} I_{jt}\right) \left(\sum_{t=1}^{N} |I_{jt}|\right)/N$. Let α_j and α_j^* represent the current and proposed scale respectively. Consider the simple situation when $a_\alpha = b_\alpha = 1$ (uniform prior). From intuition, we want $\alpha_j^* \simeq \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} \simeq \sum_{K \le t \le N} I_{jt}$; $\alpha_j^* < \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} - \sum_{K \le t \le N} I_{jt} \ll 0$; $\alpha_j^* > \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} - \sum_{K \le t \le N} I_{jt} \gg 0$. At the same time, we believe the larger the $\sum_t w_{jt}$, the more reliable the current α_j. Then the following proposal is suggested to obtain α_j^*,

$$\begin{aligned} \text{when } \alpha_j \le 0.5 \;&:\; \alpha_j^* \sim \text{ Beta } \left(a_1 + a_\alpha, \sum_t |I_{jt}| - a_1 + b_\alpha\right), \\ \text{when } \alpha_j > 0.5 \;&:\; \alpha_j^* \sim \text{ Beta } \left(\sum_t |I_{jt}| - a_2 + a_\alpha, a_2 + b_\alpha\right). \end{aligned} \tag{5}$$

Let M_o denote the mode of this distribution. Then $M_o = \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} = \sum_{K \le t \le N} I_{jt}$; $M_o < \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} - \sum_{K \le t \le N} I_{jt} < 0$; and $M_o > \alpha_j$ when $\sum_{1 \le t \le K} I_{jt} - \sum_{K \le t \le N} I_{jt} > 0$.

To update the current shift β_j to a new shift β_j^* conditional on new scale α_j^*, consider two cases. When most of the sites on the candidate map find matched sites on the j^{th} observation, that is, $\sum_t w_{jt}$ is large, one tends to use the current available alignment information to update shift, which is called a small step; otherwise, when $\sum_t w_{jt}$ is small, the current alignment information would not be reliable, then it would be better off trying a uniform distribution within the region defined by the constraint $0 \leq \beta_j^* \leq L - \alpha_j^* s_{jm_j}$ for all j, which is called a big step. When $\sum_t w_{jt} > C$ (C is some positive constant), propose a small step

$$\beta_j^* \sim N\left(\frac{\sum_t w_{jt}(h_t - \alpha_j^* q_{jt})}{\sum_t w_{jt}}, \frac{\sigma^2}{\sum_t w_{jt}}\right) \tag{6}$$

truncated to have $0 \leq \beta_j^* \leq L - \alpha_j^* s_{jm_j}$. When $\sum_t w_{jt} \leq C$, we need a big step. Let $T = \{t : 0 \leq h_t - \alpha_j^* q_{jt} \leq L - \alpha_j^* s_{jm_j}, 1 \leq t \leq N\}$. Suppose there are n_T elements in T. Then sort $h_t - \alpha_j^* q_{jt}$, $t \in T$, in ascending order. We get $(n_T + 1)$ intervals with left end 0 and right end $L - \alpha_j^* s_{jm_j}$. Let the probability to choose one specific interval be $1/(n_T + 1)$, and use uniform distribution on this interval to obtain β_j^*.

The new alignments (W_{jt}^*, Q_{jt}^*), $t = 1, 2, \cdots, N$, between the j^{th} observation and the candidate map conditional on α_j^* and β_j^* are proposed by distributions (2) and (3). The acceptance ratio for the proposed value $\alpha_j^*, \beta_j^*, w_j^*$ and q_j^* is

$$r_j(\alpha, \beta, w, q) = \min\{1, f_1 \cdot f_2 \cdot f_3\},$$

where f_1 is the complete date likelihood ratio with

$$f_1 = \frac{L(\alpha_j^*, \beta_j^*, \theta_-; \mathcal{D}_j, w_j^*, q_j^*)}{L(\alpha_j, \beta_j, \theta_-; \mathcal{D}_j, w_j, q_j)},$$

f_2 is the prior ratio with

$$f_2 = \left(\frac{\alpha_j^*}{\alpha_j}\right)^{a_\alpha - 1} \left(\frac{1 - \alpha_j^*}{1 - \alpha_j}\right)^{b_\alpha - 1},$$

and f_3 is the proposal ratio with

$$\begin{aligned} f_3 \;=\; & \frac{\sum_t (1 - w_{jt}^*) + 1}{\sum_j \sum_t (1 - w_{jt}^*) + M} \cdot \frac{\sum_j \sum_t (1 - w_{jt}) + M}{\sum_t (1 - w_{jt}) + 1} \cdot \frac{Pr(\alpha_j | \alpha_j^*, \theta_-, w_j^*, q_j^*)}{Pr(\alpha_j^* | \alpha_j, \theta_-, w_j, q_j)} \\ & \cdot \frac{Pr(\beta_j | \alpha_j, \theta_-, w_j^*, q_j^*)}{Pr(\beta_j^* | \alpha_j^*, \theta_-, w_j, q_j)} \cdot \frac{P(Q_j^* = q_j, W_j^* = w_j | \mathcal{D}_j, \alpha_j, \beta_j, \theta_-)}{P(Q_j^* = q_j^*, W_j^* = w_j^* | \mathcal{D}_j, \alpha_j^*, \beta_j^*, \theta_-)}. \end{aligned}$$

Here θ_- denotes parameters other than α_j and β_j, $Pr(\alpha_j^* | \alpha_j, \theta_-, w_j, q_j)$ is one of the densities in (5) depending on the value of α_j, and $Pr(\beta_j^* | \alpha_j^*, \theta_-, w_j, q_j)$ is either the density function of (6) or 1 (uniform). Likewise, the values of $Pr(\alpha_j | \alpha_j^*, \theta_-, w_j^*, q_j^*)$ and $Pr(\beta_j | \alpha_j, \theta_-, w_j^*, q_j^*)$ can be calculated.

2.5. Update position

First, choose the t^{th} position to update according to alignment information. Let

$$P\{\text{select the } t^{th} \text{ position to update}\} = \frac{\sum_j (1 - w_{jt}) + 1}{\sum_j \sum_t (1 - w_{jt}) + N} \tag{7}$$

To update the current position h_t to a new position h_t^*, consider two cases again. When the t^{th} site on the candidate map finds a matched site on most of the observations, that is, $\sum_j w_{jt}$ is large, one tends to use the current available alignment information to update position, which is a small step; otherwise, when $\sum_j w_{jt}$ is small, the current alignment information would be not reliable, then it would be better off trying a uniform distribution within the support of position h_t, which is a big step. When $\sum_j w_{jt} > C$ (C is some positive constant), we have a small step

$$h_t^* \sim N\left(\frac{\sum_j w_{jt}(\alpha_j q_{jt} + \beta_j)}{\sum_j w_{jt}}, \frac{\sigma^2}{\sum_j w_{jt}}\right) \tag{8}$$

truncated to have $0 \leq h_{t-1} \leq h_t^* \leq h_{t+1} \leq L$. When $\sum_j w_{jt} \leq C$, we need a big step. For the j^{th} observation, consider those positions with indicator i, $i = 1, \cdots, m_j$, that satisfy $s_{ji} \in [q_{jt-1}, q_{jt'}]$, where $t' = \min\{k : k \geq t+1 \bigcap w_{jk} = 1\}$. Standardize those positions to $s_{ji}^* = \alpha_j s_{ji} + \beta_j$. Then further choose those that satisfy $s_{ji}^* \in [h_{t-1}, h_{t+1}]$ (define $h_0 = 0$ and $h_{N+1} = L$). Suppose there are k positions. Then $h_{t-1}, s_{j,i_1}, \cdots, s_{j,i_1+k-1}, h_{t+1}$ form $(k+1)$ consecutive intervals. Let the probability to choose a specific interval be $1/(k+1)$, and use uniform distribution on the chosen interval to obtain h_t. The acceptance ratio for the new position h_t^* is

$$r(h_t) = \min\left\{\prod_{j=1}^{M} \frac{C(H^*, L, w_j, \sigma)}{C(H, L, w_j, \sigma)} \cdot \frac{\exp\left\{-w_{jt}(\alpha_j q_{jt} + \beta_j - h_t^*)^2/2\sigma^2\right\}}{\exp\left\{-w_{jt}(\alpha_j q_{jt} + \beta_j - h_t)^2/2\sigma^2\right\}} \cdot \frac{Pr(h_t/h_t^*, w, q, \mathcal{D}, \theta_-)}{Pr(h_t^*/h_t, w, q, \mathcal{D}, \theta_-)}\right\}.$$

Here θ_- denotes the parameters other than h_t. The value of $Pr(h_t^*/h_t, w, q, \mathcal{D}, \theta_-)$ is the density of (8) or of the uniform distribution in the big step. Likewise, we may calculate $Pr(h_t/h_t^*, w, q, \mathcal{D}, \theta_-)$.

3. Type B Transition in the Reversible Jump MCMC

Choose to add a new color before the t^{th} site with probability

$$p_t = \frac{\sum_{j=1}^{M} V_{jt} + 1}{\sum_{t=1}^{N}\sum_{j=1}^{M} V_{jt} + N}$$

$t = 1, 2, \cdots, N+1$, where artificially define the $N + 1^{st}$ site on map $\mathcal{H}$ to be at position L and $Q_{jN+1} = (L - \beta_j)/\alpha_j$. Here V_{jt} is defined as the number of observed sites between Q_{jt-1} and Q_{jt} (ends excluded) on observation $\mathcal{D}_j$. To decide which color to be added between h_{t-1} and h_t, let $V_{jt}(c)$, $c \in \mathbf{C}$, be the number of observed sites with color c between Q_{jt-1} and Q_{jt} (ends excluded) and the probability to choose a color c is

$$p_c = \sum_{j=1}^{M} \frac{V_{jt}(c) + 1}{\sum_{c\in\mathbf{C}}\left(\sum_{j=1}^{M} V_{jt}(c) + 1\right)}.$$

To decide the position h_t^* of this new site with color $b_t^* = c$, use the big step strategy described in the subsubsection 2.5.5. Let us clarify notations: $N_{k+1} = N_k + 1$, $B_{k+1} = (b_1, \cdots, b_{t-1}, b_t^*, b_t, \cdots, b_{N_k})$, $H_{k+1} = (h_1, \cdots, h_{t-1}, h_t^*, h_t, \cdots, h_{N_k})$, $(w_j)_{k+1} = (w_{j1}, \cdots, w_{jt-1}, w_{jt}^*, w_{jt}, \cdots, w_{jN_k})$, and $(q_j)_{k+1} = (q_{j1}, \cdots, q_{jt-1}, q_{jt}^*, q_{jt}, \cdots, q_{jN_k})$. To propose alignment w_{jt}^* and q_{jt}^* for each $j = 1, \cdots, M$, define $t'_j = \min\{r : r \geq t \bigcap w_{jr} = 1\}$. Suppose that all the sites with color c between q_{jt-1} and $q_{jt'_j}$ on the j^{th} observation $\mathcal{D}_j$ are represented by $\{s_{ji_1}, \cdots, s_{ji_k}\}$. For $u \in \{s_{ji_1}, \cdots, s_{ji_k}\}$, let

$$P\{W_{jt}^* = 0\} = \frac{f_n}{\text{denom}} \tag{9}$$

$$P\{W_{jt}^* = 1, Q_{jt}^* = u\} = \frac{(1-f_n)\phi\left((\alpha_j u + \beta_j - h_t^*)/\sigma\right)/\sigma}{\text{denom}} \tag{10}$$

where denom $= f_n + (1-f_n)\sum_{l=1}^k \phi\left((\alpha_j s_{ji_l} + \beta_j - h_t^*)/\sigma\right)/\sigma$. Then the acceptance ratio for this type B transition is

$$\begin{aligned} r_b(b,h,w,q) = \min\Bigg\{1, &\frac{C(H_{k+1}, L, w_{k+1}, \sigma)}{C(H_k, L, w_k, \sigma)}(f_p f_n)^{-\sum_j w_{jt}^*}(1-f_n)^{\sum_j w_{jt}^*} \\ &\prod_j \left(\frac{1}{\sigma}\phi\left(\frac{\alpha_j q_{jt}^* + \beta_j - h_t^*}{\sigma}\right)\right)^{w_{jt}^*} \frac{\lambda_n}{N+1}\frac{d'_{k+1}}{b'_k} \\ &\prod_j \frac{Pr(N_k, B_k, H_k, w_k, q_k | \mathcal{D}_j, N_{k+1}, B_{k+1}, H_{k+1}, w_{k+1}, q_{k+1})}{Pr(N_{k+1}, B_{k+1}, H_{k+1}, w_{k+1}, q_{k+1} | \mathcal{D}_j, N_k, B_k, H_k, w_k, q_k)}\Bigg\}, \end{aligned}$$

where $Pr(N_{k+1}, B_{k+1}, H_{k+1}, w_{k+1}, q_{k+1} | \mathcal{D}_j, N_k, B_k, H_k, w_k, q_k)$ is the product of p_t, p_c, density of h_t^* from (8) and probability from either (9) or (10), while $Pr(N_k, B_k, H_k, w_k, q_k | \mathcal{D}_j, N_{k+1}, B_{k+1}, H_{k+1}, w_{k+1}, q_{k+1}) = (\sum_j (1 - w_{jt}) + 1)/(\sum_t \sum_j (1 - w_{jt}) + N)$, which is the probability to delete the tth color.

International Journal of Mathematics and
Computer Science, 5(2010), no. 2, 123-140

Mathematical Tools and Statistical Techniques for Proteomic Data Mining

Don Hong[1,2], Shi Yin Qin[3], and Fengqing (Zoe) Zhang[4]

[1] Department of Mathematical Sciences
Program of Computational Sciences
Middle Tennessee State University
Murfreesboro, TN, TN 37132, USA

[2] College of Science
Ningbo University
Ningbo, Zhejiang, China

[3] School of Automation Science and Electrical Engineering
Beihang University
Beijing, China

[4] Department of Statistics
Northwestern University
Evanston, IL, 60208, USA

e-mail: dhong@mtsu.edu, qsy@buaa.edu.cn, FengqingZhang2015@u.northwestern.edu

(Received June 30, 2010, Revised Aug. 15, 2010, Accepted Nov. 1, 2010)

Abstract

Proteomics is the study of and the search for information about proteins. The development of mass spectrometry (MS) such as matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) MS and imaging mass spectrometry (IMS), greatly speeds up proteomics studies. At the same time, the MS and IMS applications in medical science give rise to many challenges in mathematics and statistics regarding to the MS and IMS data analysis including data preprocessing, classification, and biomarker discovery. In this paper, we give a review of recent development of mathematical techniques and statistical tools for MS and IMS based proteomic data mining including wavelet based MS data preprocessing and multivariate statistical methods for IMS data classification and biomarker discovery.

Key words and phrases: biomarker discovery, classification, data analysis, elastic net, mass spectrometry, image MS, proteomics, wavelet denoising

AMS subject classification: 62-07, 62P10, 65T50, 92C55

ISSN 1814-0424 ©2010, http://ijmcs.future-in-tech.net

1. Introduction

The widespread adoption of matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) MS for protein identification in proteomics studies is driven by the sensitivity, unlimited mass range capabilities, versatility and intact protein analysis (see [1], [11], [30] for example). One promising area of MS based proteomics is that the information hidden in the noisy mass spectral data can help people to detect cancer even in early stage. MS has already been widely used to find disease related proteomic patterns in complex mixtures of proteins. These new techniques have made proteomics possible, especially when involving large molecules. Indeed, the Nobel prize in chemistry in 2002 recognized MALDI's ability to analyze intact biological macromolecules. MALDI IMS has emerged as a powerful technique for analyzing the spatial distribution of proteins directly in tissue specimens. IMS as a platform has shown great potential and is very promising for rapid mapping of protein localization and the detection of sizeable differences in protein expression (see [11], [28], [31], and references therein). However, the complexity and high dimensionality of the MS and IMS data pose great challenges for data processing.

Usually, a raw MS spectrum consists of three components: true peaks, baseline, and noise. Disentangling these three components is a complex task. Concerning to IMS, the data processing becomes even more difficult. IMS data has two spatial dimensions ($x-$ and $y-$ dimensions) and the mass-over-charge (m/z) dimension. Each MALDI IMS data set is multidimensional and has hundreds of pixels. Each pixel is associated with a complete mass spectrum. This contrasts with regular images where for each pixel there is a set of RGB values. Each mass spectrum contains mass-to-charge (m/z) values ranging from $2k$ to $70k$ Daltons and ion intensity values which are associated with each pixel. There are hundreds of mass spectra represented in a single MS image. To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels, but also to study correlation and distribution using the spatial information for the entire image cube. Another important distinction that should be made is determining if the m/z values selected as potential biomarkers are caused by the biological structure of the tissues or by the disease state being investigated.

Generally, the mathematical processing of MS signals can be roughly divided into two steps. First, in the preprocessing step, we attempt to recover true signals from the raw data, as accurately as possible. This step includes calibration, denoising, baseline correction, data alignment cross samples, and peak selection. The second type processing is to involve operations such as dimension reduction, feature selection, clustering, and pattern recognition for classification. Preprocessing is of great importance and can improve the performance of classifiers to separate cancer and non-cancer samples. It has been studied that ineffective or inadequate algorithms in preprocessing will introduce substantial biases and thus prevent to extract valuable biomarkers from raw data.

Wavelet is an important method in signal processing and has very broad applications in image processing and statistical data analysis. The characteristics of wavelet analysis such as localized representation, orthogonal decomposition and multi-resolution analysis (MRA) guarantee that wavelets are very suitable for MS data analysis ([6], [7], [9], [12], [13], [18], [19] for example). Wavelet could reveal more information than other conventional methods. Combing the zoom-in and pan-out properties, wavelets, as building blocks of models, are well localized in both time and frequency scale. The MRA enables us to analyze the signal in different frequency bands and thus enables us to observe any transient in time domain as well as in frequency domain. Furthermore, wavelet is very useful in analyzing data with gradual frequency changes. In addition, wavelets select widths of time slices according to the local frequency in the signal. This adaptive property of wavelets certainly can help us to determine the location of peak differences of MS protein expressions between cancer and non-cancer samples.

In this paper, we present an overview of wavelet applications in the MALDI MS and IMS proteomics data processing and multivariate statistical tools for IMS data biomarker discovery. We explore the characteristics of MALDI MS and IMS data as well as wavelet application in this area, both theoretically and practically. The application includes not only in preprocessing steps but also in the feature selection. We provide some guidelines to algorithm development and parameter selection in different data processing stages of both the conventional MALDI MS data and IMS data. After analyzing the MS and IMS data from the view of wavelet transform, we suggest integrating all MS and IMS data processing steps by using wavelet transform. Biomarker selection from IMS data is a problem of global optimization. A recently developed regularization and variable selection method, elastic-net (EN), produces a sparse model with admirable prediction accuracy and can be an effective tool for IMS data processing. Very recently, we have incorporated a spatial penalty term into the EN model and developed a new tool for IMS data biomarker selection and classification ([20], [35]). The remainder of the paper is organized as follows. The characteristics of MS and IMS based proteomics data and wavelet applications in this area are discussed in the next section. In Section 3, the preprocessing steps and wavelet applications are presented in detail. Wavelet-based procedure for feature selection and classification algorithms are also discussed this section. Newly developed elastic net based biomarker discovery tools used for IMS data are discussed in Section 4.

2. Characteristics of MALDI MS and IMS Data and Wavelet Applications

In this section, we describe in detail the characteristics of MALDI TOF MS and IMS data and apply wavelets in the data (pre)processing.

2.1. Mathematical Model for MALDI MS and IMS Proteomics Data

A commonly used model ([6], [9], [18], [25]) for MS data analysis is that each raw spectrum data can be represented in three parts: $f(t) = B(t) + N * S(t) + \epsilon(t)$, where $f(t)$ is the observed signal. $B(t)$ stands for baseline, a systematic artifact commonly seen in mass spectrometry data. $S(t)$ is the true signal, which consists of a sum of possibly overlapping peaks, each corresponding to a particular biological molecule such as a protein or a peptide. N is the normalization factor, a constant multiplicative factor to adjust for spectrum specific variability. $\epsilon(t)$ stands for the noise function. In general, the preprocessing steps include calibration, denoising, baseline correction, normalization, peak alignment, and peak detection and quantification. The difficulty in processing MS data stems from the mixture of true peaks, baseline and noise. Separating these three components from each other is very complex.

Concerning the MADLI IMS data, many of the characteristics in mass spectrum preprocessing stage are the same as those for the MS data. However, the main difference is that the IMS data has two spatial dimensions (both x- and y- dimensions) plus a mass-over-charge (m/z) dimension. The combination of spatial and mass resolution results in large and complex data sets, which gives a great challenge to the quantitative analysis and interpretation tools. Figure 1 shows a mouse brain IMS data set, and the darker region in Figure 1(a) indicates the presence of the tumor. The grid in Figure 1(a) forms a matrix of points of the sample surface. Individual mass spectra are acquired for every point (pixel) of the sample surface and stored digitally. Behind each pixel, it is an entire mass spectrum with a very large range of m/z values. Figure 1(b) displays three mass spectra corresponding to three different pixels of the mouse brain tissue section. Specially designed software enables the election of an analyte signal (m/z value) from the mass range and plots the intensity of the signal for each individual point in a matrix. If the signal intensity is plotted by a color scale, the matrix can be represented by an ion image of the analyte distribution, which is shown in Figure 1(c) and (d). Ion images can show the spatial spread of a particular peak's intensity over the tissue and the mass spectral peak represented by the amount of a particular ion that was measured. From Figure 1(c), we can see that the m/z value 5442.704 is differentially expressed between the cancer region and the normal region, which maybe due to the latent biological function of this m/z value and its effect to cancer growth. Figure 1(c) and Figure 1(d) have clearly different intensity distributions over the mouse brain. If one has already known a particular m/z value having biological meaning and plans to know the spatial distribution of a particular molecule, ion image is very informative. However, a more important application should be the determination of unknown variants for metabolite and protein profiling in both clinical and disease studies. Noticing that we have huge number of ion images per data set, it is necessary to have statistical models to do biomaker selection instead of doing visually checking. In addition, these conventional images, derived from a specific analyte mass do not identify the spatially localized correlations between analytes that are latent in IMS data processing. Although it is difficult to make full utilization of both spatial information and spectrum information in IMS data, it is very necessary. For IMS data processing, multi-scale representation and global analysis of spatial and protein information of the biological samples will be the key for data processing. Therefore, the multivariate data analysis methods can be applied in IMS data for identifying both spatial and mass trends and merit further investigation.

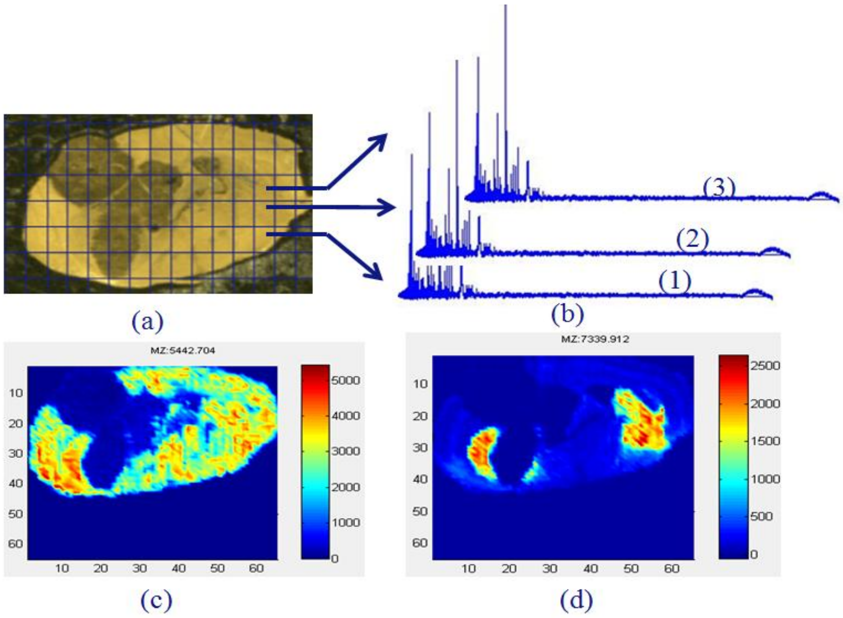


Figure 1: Mouse brain IMS Data. (a) photomicrograph of a mouse brain tissue section, implanted with a GL26 glioma cell line and tumor growth; (b) three mass spectra from three different pixels; (c) the ion image of $m/z = 5442.704$; and (d) the ion image of $m/z = 7339.912$.

2.2. Wavelets for MALDI MS and IMS Data

To extract the true signal from MS/IMS data, we need to remove the noise and the incoherent signal from the observed data. Wavelet analysis serves as an efficient mathematical tool that can be utilized to extract or encode the feature signal. Wavelet theory represents data signals by breaking them down into many interrelated component pieces, similar to the pieces of a jig-saw puzzle. When pieces are scaled and translated by wavelets, the breaking down process is called a wavelet decomposition or wavelet transform. A discrete wavelet transform (DWT) decomposes a signal into several vectors of wavelet coefficients. Different coefficient vectors contain information about the signal function at different scales. Coefficients at coarse scale capture gross and global features of the signal while coefficients at fine scale contain detailed information. Applying wavelet transform to MALDI-TOF MS data, the protein expression difference can be measured at different resolution scales based on a molecular weight-scale analysis. It may reveal more information than other conventional methods. Wavelets, as building blocks of models, are well localized in both time and scale (frequency). In wavelet analysis, a function is approximated by a weighted sum over the scaled and translated mother wavelets. Each weighted wavelet acts as a building block, and when all the blocks are summed together, an approximation is found. Wavelet is very useful in analyzing data with gradual frequency changes. Signals with rapid local changes (signals with discontinuities, cusps, sharp spikes, etc) can be precisely represented with just a few wavelet coefficients.

The wavelet approximation to a signal function f is built up over multiple scales and many localized positions. The fundamental concept involved in multi-resolution analysis (MRA) is to find the average features and the details of the signal via scalar products with scaling signals and wavelets. The MRA enables us to analyze the signal in different frequency bands and thus enables us to observe any transient in time domain as well as in frequency domain. The high frequency band output is viewed as the wavelet transform coefficients for a fine scale and the low frequency band output is decimated by a factor of 2. This low frequency band is then split into a high and low band again. The band splitting and decimation process continues and produces an octave band representation of the signal. The high pass filter output wavelet coefficients represent the signal's characteristics and energy at a particular scale. The output of the final low pass filter is the residual namely the most coarse signal.

Since true signal $S(x)$, the baseline $B(x)$ and the machine noise $\epsilon(x)$ have different time-frequency attributes, it is then possible to separate them in wavelet coefficients. In the wavelet representation, the noise $\epsilon(x)$ is concentrated in the fine scale wavelet coefficients and the incoherent signal can be approximated by the projection onto the coarse space. In contrast to Fourier transforms, wavelets select widths of time slices according to the local frequency in the signal. This adaptive property of wavelets certainly can help us to determine the location and intensity (peak) difference(s) of MALDI-TOF MS protein expressions between cancerous and normal tissues in term of molecular weights. Most peak signals can be represented by a small number of wavelet coefficients while white noise is distributed equally over all wavelet coefficients. However the separation of noise and peaks is not so straightforward since they both have fast changing parts. A variety of threshold strategies can be used to remove the machine noise from the data including feedback strategies from MS data information[7]. But inappropriate thresholding may cause peak attenuation. Baseline correction is an important step in MS data preprocessing. Through wavelet transforms, baseline $B(x)$ can be considered as a coarse approximation and a component with small coefficients in a wavelet space[12]. As for the peak selection step, peaks of MS data can be viewed as the singularities of the MS output signal. Singularities of a signal can be represented by the modulus maxima of their wavelet transforms (see [15], [27], [34], for example). Concerning the 3D IMS data, wavelet is also suitable for data preprocessing as well as feature selection.

3. Wavelet Applications for MALDI MS Data Preprocessing

In this section, we discuss in detail the wavelet applications for MALDI MS data preprocessing with an emphasis on denoising, baseline correction, and feature (peak) selection. Figure 2 illustrates the effects of the preprocessing steps visually. Recall that each raw spectra data can be represented in three parts: $f(t) = B(t) + N * S(t) + \epsilon(t)$. Figure 2 displays the MALDI MS data preprocessing framework in terms of denoising, baseline correction and peak selection.

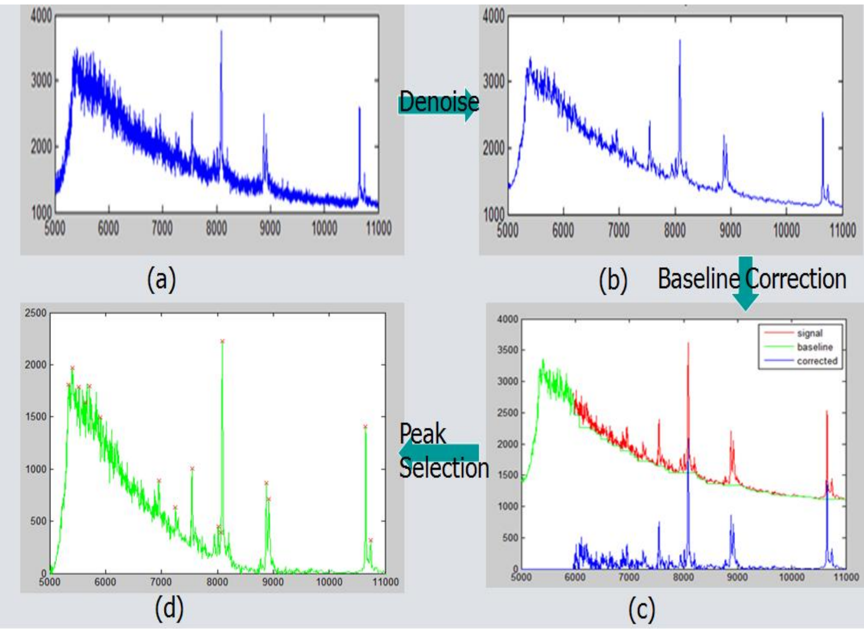


Figure 2: MALDI MS data preprocessing framework. (a) a raw mass spectrum; (b) the denoised mass spectrum; (c) the mass spectrum after baseline correction; and (d) the mass spectrum marked with selected peaks.

3.1. Denoising

Yasui et al.[33] and Coombes et al.[9] independently proposed the application of wavelet transformation in proteomics. Recent applications of wavelets for MS data processing can be found, for example in [6], [7], [12], [18], and [22].

The general wavelet denoising procedure is as follows:

1. Apply wavelet transform to the noisy signal to produce the noisy wavelet coefficients to the level which we can properly distinguish the peaks.

2. Select an appropriate threshold limit at each level and a threshold method (hard or soft thresholding) to best remove the noises.

3. Inverse wavelet transform of the thresholded wavelet coefficients to obtain a denoised signal.

In denoising step, the most important issues are the selection of a suitable wavelet, the decomposition levels, and the coefficients of the denoising threshold. Selecting the mother wavelet is of great importance for wavelet transform. There is no strict rule for selection. However, the analysis becomes more precise if the wavelet shape is adapted to the signal. DWT is sufficient for exact reconstruction, and the discrete forms are necessary for most computer implementations. Du at al.[13] proposed an improved DWT smoothing algorithm which utilizes the cross-level DWT coefficients information during smoothing. It estimates the noise distribution based on the first DWT decomposition level, and then infers the threshold at other levels. In order to reduce the peak attenuation in smoothing, the related DWT coefficients of the detected peak-related DWT coefficient are also used for reconstruction. Coombes et al[9] proposed to use undecimated discrete wavelet transform (UDWT). They suggested when using the DWT for denoising, it tends to create significant artifacts near the ends of the signal. With the DWT, there is usually a trade-off between the smoothness of the denoised signal and its squared-error performance ([8], [23]). While DWT denoising effect will change drastically if the starting position of the signal is shifted, UDWT is shift-invariant. Comparing with DWT, it is reported that UDWT gives better visual and qualitative denoising, with a small added cost in computational complexity[9].

For the choice of a decomposition level, the maximum level to apply the wavelet transform depends on how many data points are contained in a data set, since there is a down-sampling by 2 operation from one level to the next one. One factor that affects the noise removal results is the signal-to-noise ratio (SNR) in the original signal. Fewer levels of wavelet transform are needed to remove most of the noise if the signals have higher SNR. The signals with lower SNR should be decomposed by relatively more levels of wavelet transform.

For the denoising threshold, different kinds of methods have been tried to extract and preserve the desired signals as much and accurately as possible. The commonly used universal threshold is computed as $\lambda = \sigma\sqrt{2\log N}$, where σ stands for the estimation of the variation of the coefficients on the standard deviation scale. N represents the number of data points (wavelet coefficients). However, local threshold performs better than universal threshold when applied to MS data. The noise distribution of MS data over the m/z value is heterogeneous[13]. The most important issue in a denoising procedure is the wavelet coefficient threshold selection. In [9], UDWT is used to denoise the spectra and Daubechies wavelet of degree 8 is applied. The denoising procedure starts by transforming observed signal from time domain to the wavelet domain, then computes the median absolute deviation (MAD) of the wavelet coefficients, sets coefficients to zero with a hard thresholding (some threshold expressed as a multiple of 0.67 MAD) and finally transforms the signal back to the time domain. However, the denoising parameters are usually chosen by experience. Chen et al [7] introduced feed-back concepts to the MS data denoising in order to target the optimal parameters setup as objectively as possible. It is reported in the paper that the elevated baseline and the height of such baseline can be also associated with the proportion of falsely detected peaks. Thus an adaptive threshold selection algorithm is suggested by utilizing the proportion of the baseline as an index to adjust the thresholds. There are many other useful and practical principles for adaptive threshold selection such as Stein's unbiased risk estimate (SURE), minimal mean squared error, generalized cross validation and etc.. In addition, Du et al.[13] proposed to utilize the cross-level DWT coefficients information for denoising. It estimates the noise distribution based on the first DWT decomposition level, and then infers the threshold at other levels. There is an approximate linear relationship of the noise component distribution at different levels. It is known that wavelet transform modulus maxima of signal and noise have different transmission properties across different scales. The method of utilizing wavelet coefficients relativity to tell desired signal from noise is called spatially selective noise filtration. This method can provide more steady denoising results but requires higher computational cost. In addition, Li et al [24] proposed Bayesian wavelet shrinkage and thresholding estimators which outperform the classical data adaptive wavelet thresholding estimators in terms of mean squared error with finite samples. It is proposed to use block threshold strategy in [18] because the high frequency components decrease as the mass weight increases. Block threshold is to threshold the wavelet coefficients in groups (blocks) rather than individually to increase estimation accuracy by utilizing information about neighboring coefficients.

Two rules are generally used for thresholding the wavelet coefficients (soft/hard thresholding). Hard thresholding sets zeros for all wavelet coefficients whose absolute value is less than the specified threshold limit. Generally hard thresholding provides an improved signal to noise ratio but the reconstructed signal may have additional oscillation. Soft thresholding only reduces these wavelet coefficients less than the specified threshold limit instead of setting them to be zeros. Soft thresholding can preserve the smoothness but not as good as hard thresholding in the sense of mean squared error. However the selection of hard or soft thresholding should refer to the principles for thresholding selection. With an appropriate threshold, noise can be removed without biasing the signal too much, since the wavelet coefficients greater than the particular threshold level still remain unaltered. However, if the threshold is too large, then the signal will be altered. If the threshold is too small, then the level of denoising is not enough. This causes a crucial problem of peak attenuation. An ideal transform can project signal to a domain where the signal energy is concentrated in a small number of coefficients. On the other hand, if the noise is evenly distributed across this domain, this domain will be a good place to do denoising, due to the fact that the SNR is significantly increased in some important coefficients, or we can say the signal is highlighted in this domain while the noise is not. However, the signal peaks also have fast changing components just like noise and will be attenuated by wavelet transform as well. All these algorithms discussed above aim to extract noise from desired signal as accurately as possible.

3.2. Baseline Correction

Another important issue that needs to be addressed in MS data preprocessing is the baseline correction. Baselines of different spectra can have large variation. Usually baseline is viewed as a very low frequency component of the observed signal. The region of the spectrum below 950 m/z is typically dominated by noise from the matrix molecules and may also contain extensive areas of saturation where the number of ions hitting the detector exceeds its ability to count them. Denoising also plays a critical role in baseline correction. Without it, the extremes of the noise (on the low end) will tend to drive the estimated baseline below the actual baseline, and the baseline-corrected spectra will tend to drift upward to the right [2].

Baselines are corrected by fitting a monotone local minimum curve to the denoised spectra in [9]. However, wavelet transform can also serve as an effective tool in baseline correction. The discrete wavelet transform decomposes the MS signal into an approximate component and several detail components. It has been mentioned in [12] that the approximation component at a certain higher level of DWT is a good estimation of the smoothly decaying baseline.

But later, they point out that the baseline removal does not perform well when large peak regions exist in the spectrum. However this does not mean baseline correction in wavelet domain is not useful. If the spectrum has large peak regions, the approximate component at a certain higher level of DWT still contains information of peaks. Thus the method to remove the approximate component as baseline will, of course, fail in this situation. It will badly affect the peak selection. If we perform baseline correction by fitting a monotone local minimum curve in the approximate component, it would be more effective and reasonable. In addition, the empirical mode decomposition (EMD) introduced by Huang et al [21] could also serve as a useful method in MS data baseline correction. EMD is adaptive and therefore highly efficient method for analyzing nonlinear and non-stationary data.

3.3. Feature Peak Selection

One of the critical problems in the MS data analysis is to select meaningful feature peaks. Therefore, the final goal of the MS data preprocessing is to identify the locations and intensities of important peaks which can be used for biomarker discovery. Actually, the peak selection procedure can be considered as two parts: peak detection that is to find the m/z values of peaks and peak quantification that is to quantify the intensities of peaks. Peak selection procedure can reduce data dimension significantly. Most of the current methods use the height of the local maximum to quantify the peak within estimated boundaries. Peaks are selected in [9] using all local maxima of a spectrum after denoising, baseline correction, and normalization. The height of the peak is used to quantify peaks. For this purpose, a local maximum is defined as a point where the intensities change from increasing to decreasing (allowing for flat plateaus when the tops of peaks are more than one clock tick in width). The signal-to-noise ratio (SNR) of a peak is estimated as the height above baseline divided by a median-smoothed version of the wavelet-defined noise. Also the local maximum points need to be larger than certain intensity threshold and SNR threshold in order to be considered as peaks. However, point measurement may be subject to high variation form various resources. Also, the height may not be a good index of the total amount of ions for a specific feature.

As we know, high amplitudes do not always guarantee real peaks: some sources of noise can result in high amplitude spikes. Conversely, low amplitude peaks can still be real. Measuring a small region or bounded neighborhood around that peak would be more robust and informative. It is pointed in [12] that the estimated peak strength is proportional to the area under the curve (AUC) of the peak in simple situations. Therefore, they suggested to study AUC estimation in spectra with multiple overlapping peaks for possibility to improve peak detection. A new peak selection algorithm for MS data analysis is proposed in [23] by using asymmetric Lorentzian and Sech2 functions to fit peak shape. A Bayesian wavelet-based functional mixed model is used to represent mass spectra as functions in [29]. This flexible framework in modeling nonparametric fixed and random effect functions enables it to simultaneously model the effects of multiple factors. From the model output, they identify spectral regions that are differentially expressed across experimental conditions, while controlling the Bayesian FDR, in a way that takes both statistical and clinical significance into account.

All these peak selection algorithms discussed above require denoising, baseline correction, and normalization beforehand. A peak detection algorithm called MassSpecWavelet, by applying CWT-based pattern matching and wavelet transform modulus maxima, is introduced in [12]. This method can be directly applied to the raw data and requires no baseline removal or peak smoothing preprocessing steps before this peak detection. As mentioned earlier, peaks of MS data can be viewed as the singularities of the MS output signal. Singularities of a signal can be represented by the modulus maxima of their wavelet transforms ([27], [34], [15]). For the CWT, a symmetric Mexican Hat wavelet is used in [12]. The Mexican Hat wavelet is proportional to the second derivative of the Gaussian probability density function. The symmetric property of Mexican Hat wavelet can help to remove the baseline component after continuous wavelet transform. In the peak identification process, instead of using Lipschitz exponent to check the singularity, the SNR is used. However, as we know Lipschitz exponent is a very reliable and popular way to measure the singularity of signal. Thus by taking advantage of Lipschitz exponent, this peak identification algorithm may be further improved.

The peak quantification algorithms can be measured in terms of reproducibility while the evaluation of peak detection algorithms can base on sensitivity and false discovery rate. Guerra et al. [26] use ANOVA and F-tests to measure the reproducibility of peak quantification. Their results show for peak quantification, among five peak selection algorithms, MassSpecWavelet has the best performance. This wavelet-based direct peak detection method shows its advantage with sensitivities above 0.95 with a FDR of 0.1 in their experiment results. It is possible, as well as quite meaningful, to integrate all MS data preprocessing steps by wavelet transform. Another wavelet-based peak selection algorithm for MS data analysis, which is available in an open source framework OpenMS, can be found in [23]. This algorithm is a three-step technique including determining the positions of putative peaks in the wavelet-transformed signal, fitting an analytically given peak function to the data in that region, and optionally improving the resulting fit by using nonlinear optimization.

Algorithms above are mainly about peak selection in individual spectrum. In real application, however, the positions of peaks in each spectrum around the same m/z value may be slightly different from each other. Thus a processing step that determines which peaks found in individual spectrum should be identified as representing the same biochemical substance across spectra is necessary. Coombes et al. [9] started selecting the set of peaks with $S/N > 10$ first, then coalesced two peaks if they differed in location by at most 7 clock ticks or if they differed in relative mass by at most 0.3%. These parameters were determined empirically by visually checking the spectra. Then they considered the peaks with $2 < S/N < 10$, and added these to the list if they fell within the same distance limits of a previously identified peak. A new algorithm called project spectrum binning (PSB) for the cross sample peak alignment was introduced by Hong et al in [19]. Averaging is a fundamental principle underlying many statistical methods. Peak detection on average spectrum is a direct and simple way for spectrum alignment. Using mean spectrum for peak extraction and quantification can be found in [25].

4. Classification and Biomarker Discovery

After preprocessing procedures, the MS data is ready for biological feature extraction or called biomarker discovery associated with certain diseases. Machine learning or pattern recognition methods can be applied to extract features from wavelet coefficients of the MS data after wavelet transform. In this section, we emphasize on image mass spectrometry data classification and biomarker discovery using multivariate statistical analysis. For IMS data, while heat map plots of individual ion intensities make pretty pictures, viewing the data one ion at a time is tedious and relies on the expertise and interpretation of the operator. There has been good progress in applying multivariate statistical methods such as PCA, LDA and SVM to IMS data analysis. However, these methods for IMS data processing are inadequate due to their limited use of spatial information and the advantages of IMS technology [35].

MALDI-Imaging is an emerging and very promising new technique for protein analysis from intact biological tissues [11]. It measures a large collection of mass spectra spreading out over an organic tissue section and retains the absolute spatial information of the measurements for analysis and imaging. The current interest in IMS lies in its unique advantage: the ability to correlate anatomical information provided by histology with the spatially resolved biochemical information provided by the imaging mass spectrometry experiments. Compared with MALDI-MS, IMS, by automatic spotting of matrices on the tissue in an array format, results in comprehensive structural analysis at a higher spatial resolution, saves more time, and provides hundreds of identical independent spectra which address the measurements repeatability. However, each MALDI imaging data set is multidimensional, with hundreds of pixels covering the tissue section and an entire mass spectrum in which mass-over-charge (m/z) values can range from 2k to 70k Dalton associated to each pixel. In this case, the number of predictors (m/z values) is greatly larger than the number of observations. To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels but also to study correlation and distribution using the spatial information for the entire image cube. Another important issue is to distinguish the selected feature m/z values according to the differences caused by biological structure of the tissue or purely by cancer. All these difficulties compounded together, pose great challenges to IMS data processing and are yet to be well solved.

The application of MVA methods has opened new doors for the exploration of IMS data. Very recently, two statistical models are presented in [20] and [35], respectively, for biomarker selection and classification of the high dimensional and complex IMS data. The aim is to extract as much useful information as possible from IMS data, by not only utilizing the spectrum information within individual pixels but also studying correlation and distribution using the spatial information. Compared with other currently popular methods, these models work efficiently and effectively for IMS data processing in terms of confirming new biomarkers, producing more precise peak list by including significant peaks and reducing the number of side peaks, and providing more accurate classification results.

4.1. EN4IMS Model for IMS Data Processing

Two fundamental criteria for evaluating the quality of a model in statistical modeling are high prediction accuracy and discovering relevant predictive variables. In the practice of statistical modeling, variable selection is especially important; it is often desirable to have an accuracy predictive model with a sparse representation since modern data sets are usually high dimensional with a large number of predictors. One would like to have a simple model to enlighten the relationship between the response and covariates and also to predict future data as accurate as possible. Ordinary least squares (OLS) estimates are obtained by minimizing the residual sum square (RSS). It is well known that OLS does poorly in both prediction and variable selection. Penalized methods have been proposed to improve OLS, starting with Ridge regression [17], followed by Bridge regression [16], the Garotte [4], the Lasso [32], LARS [14], and very recently the elastic net [36]. The Dantzig selector method was proposed in [5] by using sparse approximation and compressive sensing.

The newly developed variable selection method, elastic net (EN), can simultaneously perform automatic variable selection and continuous shrinkage [36]. That is, it can continuously shrink the coefficients toward zero as its regularization parameters increase; some coefficients are shrunk to exactly zero if the regularization parameters are sufficiently large. The shrinkage often improves the prediction accuracy due to the bias-variance trade-off. Thus, the EN model simultaneously achieves accuracy and sparsity. The achievement of sparsity is particularly useful when the number of variables (p) is much larger than the number of observations (n). In addition, the EN model encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. Compared with other current commonly used analysis methods, the EN model is much more suitable for IMS data processing. In [35], a spatial penalty term is incorporated into the EN model in order to develop a new tool for IMS data biomarker selection and classification. The motivation of this new model EN4IMS is to fully utilize not only the spectral information within individual pixels, but also the spatial information for the entire IMS data cube. A software package for comprehensive IMS data processing, called IMSmining, is developed based on this new model. By incorporating the spatial penalty term, this package helps to distinguish the IMS feature peaks caused by biological structural differences from those truly associated with cancer.

The EN4IMS method has been tested on extensive simulation studies, and the algorithm has also been applied to real IMS data sets provided by Vanderbilt University Mass Spectrometry Research Center (VUMSRC). The analysis results of both simulation studies and real data examples show that the EN4IMS algorithm works efficiently and effectively for IMS data processing: producing a more precise listing of feature peaks, helping to discover new potential biomarkers, and providing more accurate classification results.

4.2. Weighted Elastic Net Model

In order to better consider the spatial information for more precise biomarker selection, a more general model called weighted elastic net (WEN), which incorporates the spatial penalty directly into the EN model equation, is developed in [20]. Theoretical properties of the WEN model such as the variable selection accuracy are discussed there as well.

In IMS data analysis, if a biomarker in terms of an m/z value in the MS spectrum is truly related to a cancer disease, then it is reasonable to expect that the ion intensity values at this m/z from different pixel locations in a cancer area are approximate the same. Therefore, the standard deviation of the intensities at the m/z should be small. In comparison, if the biomarker selected by the statistical model based on differentiation mainly caused by the tissue structure, then the ion intensities at the m/z point vary significantly from pixel to pixel. Therefore, the standard deviation of intensities at such an m/z point should be relatively large. Thus, it is proper to associate its standard deviation at each predictor with the corresponding coefficient in the model to enforce penalty on predictors caused by structure differences.

To better consider the spatial information for more precise biomarker selection, we propose the following weighted elastic net (WEN) model [20]:

$$\mathrm{argmin}_{\beta}\frac{1}{2}\|\mathbf{y}-\sum_{\mathbf{j=1}}^{\mathbf{p}}\mathbf{x_j}\beta_\mathbf{j}\|_\mathbf{2}^\mathbf{2}+\mathbf{n}\lambda_\mathbf{1}\sum_{\mathbf{j=1}}^{\mathbf{p}}\mathbf{w_j}|\beta_\mathbf{j}|+\frac{\mathbf{n}}{\mathbf{2}}\lambda_\mathbf{2}\sum_{\mathbf{j=1}}^{\mathbf{p}}|\mathbf{w_j}\beta_\mathbf{j}|^\mathbf{2}, \tag{4.1}$$

where $w_j > 0$, $j = 1, \cdots, p$ are weighted penalty coefficients. Let $\mathbf{W} = \mathrm{diag}[\mathrm{w_1}, \cdots, \mathrm{w_p}]$. Then the WEN model can be rewritten as

$$\mathrm{argmin}_{\beta}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_\mathbf{2}^\mathbf{2}+\mathbf{n}\lambda_\mathbf{1}\|\mathbf{W}\beta\|_\mathbf{1}+\frac{\mathbf{n}}{\mathbf{2}}\lambda_\mathbf{2}\|\mathbf{W}\beta\|_\mathbf{2}^\mathbf{2}. \tag{4.2}$$

The weighted elastic net model (4.1) puts the weights associated with ion intensity spreading information directly into the elastic net model and thus enforces larger penalty on the coefficients of predictors caused by structure differences. This model inherits good properties from the EN model including sparse representation, ability to deal with $p \gg n$ problem and grouping effect. In addition, compared with the EN model, it is more suitable for IMS data analysis since it makes good use of the spatial information and thus it helps to distinguish the selected feature m/z values according to the differences caused by biological structure of the tissue or purely by cancer.

By an algebraic simplification, we can see that WEN also enjoys the computational advantage of the Lasso. Thus, an algorithm for the WEN method based on the algorithm LARS [14] can be developed [20]. The WEN algorithm is applied to IMS data sets for predictor selection and classification, and results show that the WEN method works effectively and efficiently for IMS data processing.

The WEN algorithm together with EN4IMS plus many other functions are integrated into a software package called IMSmining. Classification results of using the EN4IMS and WEN models are compared with those of other current popular methods used in the IMS community. In a real data set of two mouse brain tissue sections, one is used for model training and the other section is used for model testing. 110 pixels are selected from the cancer area to be used as the training cancer data set, and 110 pixels are selected from the normal area to be used as the training noncancer data set. Similarly, 110 cancer pixels and 110 noncancer pixels are selected from the second mouse brain tissue section as test data. Classification rates show that the EN4IMS and WEN models outperform the other methods.

Since both the EN4IMS and the WEN models are based on linear regression, it would be interesting to consider piecewise linear spline regression classifiers for IMS data analysis. However, due to the nonlinearity and the mixed ℓ_1 and ℓ_2 constrains, we expect that such a study is non-trivial at all. It would be also very interesting to incorporate wavelet transform of IMS data into the study of classification and biomarker discovery.

Acknowledgements

The authors are grateful to an anonymous referee for constructive comments. This research was partially supported by Middle Tennessee State University Faculty Research Grant #2-21519 and Yue-Kong Pao Endowment Fund at Ningbo University, China..

References

[1] R. Aebersold and M. Mann M, Mass spectrometry-based proteomics, Nature, 422 (2003), 198-207.

[2] K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C. Xiao, and K.R. Coombes, A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples, Proteomics, 3 (2003), 1667-1672.

[3] K.A. Baggerly, J.S. Morris, and K.R. Coombes, Reproducibility of SELDI- TOF protein patterns in serum: comparing datasets from different experiments, Bioinformatics, 20 (2004), 777-785.

[4] L. Breiman, Better subset regression using the nonnegative garrote, Technometrics, 37 (1995), pp. 373-384.

[5] E. Candes and T. Tao, The dantzig selector: statistical estimation when p is much larger than n, Annals of Statistics, 35 (2007), 2313-2351.

[6] S. Chen, D. Hong, and S. Yu, Wavelet-based procedures for proteomic mass spectrometry data processing, Computational Statistics & Data Analysis, 52 (2007), 211-220.

[7] S. Chen, M. Li, D. Billheimer, D. Hong, B. Xu, and Y. Shyr, A Novel Comprehensive Wave-form MS Data Processing Method, Bioinformatics, 25 (2009), 808-814.

[8] R. Coifman and D. Donoho, Translation invariant de-noising, In: Wavelets and Statistics, pp. 125-150, New York. Springer-Verlag, 1995.

[9] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.-C. Hung, and H.M. Kuerer, Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, Proteomics, 5 (2007), 4107-4117.

[10] Coombes, K.R., et al. Preprocessing mass spectrometry data, In: Fundamentals of Data Mining in Genomics and Proteomics, pp. 79-99, Kluwer, Boston, 2007.

[11] D.S. Cornett, M.L. Reyzer, P. Chaurand, and R.M. Caprioli, MALDI imaging mass spectrometry: molecular snapshots of biochemical systems, Nat.Methods, 4 (2007), 828-833.

[12] Pan Du, Warren A. Kibbe, and Simon M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, 22 (2006), 2059-2065.

[13] P. Du, S.M. Lin, W.A. Kibbe, and H.H. Wang, Application of wavelet transform to the MS-based proteomics data preprocessing, Bioinformatics and Bioengineering, BIBE. Proceedings of the 7th IEEE International Conference, pp. 680 - 686, Boston, MA, 2007.

[14] B. Efron, T. Hastie, R. Tibshirani, Least angle regression, Annals of Statistics, 32 (2004), 407-499.

[15] A. Faghfouri and W. Kinsner, Local and global analysis of multifractal singularity spectrum through wavelets, IEEE CCECE/CCGEI, pp.2163-2169, Saskatoon, 2005.

[16] I. Frank and J. Friedman, A statistical view of some chemometrics regression tools, Technometrics, 35 (1993), 109-148.

[17] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12 (1970), 55-67.

[18] D. Hong and Y. Shyr, Mathematical framework and wavelets applications in proteomics for cancer study. In: Handbook of Cancer Models with Applications to Cancer Screening, Cancer Treatment and Risk Assessment (W.Y. Tan and L. Hannin Eds.), pp. 471-499, World Scientific, Singapore, 2008.

[19] D. Hong, H.M. Li, M. Li, and Y. Shyr, Wavelets and Projecting Spectrum Binning for Proteomic Data Processing, In: Quantitative Medical Data Analysis Using Math Tools and Statistical Techniques (Hong and Shyr Eds.), pp. 159-178, World Scientific Publications, LLC., Singapore, 2007.

[20] D. Hong and F. Zhang, Weighted Elastic Net Model for Mass Spectrometry Imaging processing, Math. Model. Nat. Phenom., 5(2010), 115-133.

[21] N.E. Huang, S. Zheng, S.R. Long, M.C. Wu4, H.H. Shih, Q. Zheng, N.-C. Yen7, C.C. Tung, and H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. Lond. A., 454 (1998), 903-995.

[22] D. Kwon, M. Vannucci, J.J. Song, J. Jeong, and R.M. Pfeiffer, A novel wavelet-based thresholding method for the preprocessing of mass spectrometry data that accounts for heterogeneous noise, Proteomics, 8 (2008), 30193029.

[23] E. Lange, C. Gropl, K. Reinert, High-accuracy peak picking of proteomics data using wavelet techniques, Biocomputing, 11 (2006), 243-254.

[24] X.L. Li, J. Li, and X. Yao, A wavelet-based data preprocessing analysis approach in mass spectrometry, Computers in Biology and Medicine 37 (2007), 509-516.

[25] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, and R. Kobayashi, Feature extraction and quantification for mass spectrometry in biomedical application using the mean spectrum, Bioinformatics, 21 (2005), 1764-1775.

[26] R. Guerra, M. Vannucci, Y. Li, C.C. Lau, T.K. Man, and A.C. Marcelo, Comparison of algorithms for preprocessing of SELDI-TOF mass spectrometry data, Bioinformatics, 24 (2008), 2129-2136.

[27] S. Mallat and W.L. Hwang, Singularity detection and processing with wavelets. IEEE transactions on information theory, 38 (1992), 617-643.

[28] H. Meistermann, J.L. Norris, H.R. Aerni, and et al., Biomarker discovery by imaging mass spectrometry, Molecular & Cellular Proteomics, 5 (2006), 1876-1886.

[29] J.S. Morris, P.J. Brown, R.C. Herrick, K.A. Baggerly, and K.R. Coobes, Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models, Biometrics, 64 (2008), 479-489.

[30] S.A. Schwartz, R.J. Weil, M.D. Johnson, and et al., Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression, Clin Cancer Res. 10 (2004), 981-987.

[31] M. Stoeckli, P. Chaurand, D.E. Hallahan, R.M. Caprioli, Imaging massspectrometry: a new technology for the analysis of of protein expression in mammalian tissues, Nat. Med. , 7 (2001), 493-496.

[32] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc., Series B., 58 (1996), 267-288.

[33] Y. Yasui, M. Pepe, M.L. Thompson, B.L. Adam, G.L. Wright, Y. Qu Jr., J.D. Potter, M. Winget, M. Thornquist, and Z. Feng, A data analytic strategy for protein biomarker discovery: profiling of high dimensional proteomic data for cancer detection, Biostatistics, 4 (2003), 449-463.

[34] C.L. Tu, W.L. Hwang, and J. Ho, Analysis of singularities from modulus maxima of complex wavelets, IEEE transactions on information theory, 51 (2005), 1049-1062.

[35] F. Zhang and D. Hong, Elastic Net Based Framework for Imaging Mass Spectrometry Data Biomarker Selection and Classification, Stat. in Medicine, in press.

[36] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, J. R. Statist. Soc., B. 67 (2005), Part 2, 301-320.

International Journal of Mathematics and Computer Science

Author Index
Vol. 5, 2010