Wavelet transform for denoising and quantification of microarray data.

Qazi M. Ali* & Omar Farooq

AMU Aligarh, India

1 Introduction

Wavelet transform is a tool that can process both the stationary as well as non-stationary signal and has got multi-resolution capabilities. Due to these advantages it has been used effectively in many application areas of bioinformatics such as detecting patterns in DNA sequences (Arneodo et al., 1996), protein structure classification (Mandell et al. 1997) and microarray data analysis (Myasnikova et al. 2001). In this paper we tried to exploit the denoising capabilities of the wavelet transform for removal of noise that is introduced during the preparation of microarray data. This helps in reducing the error during the quantification process, thereby, improving the interpretation of the microarray data. For evaluation of the gene expression data extensive computation is required if the raw data is used. This is because of the fact that the number of spots is very large and the data within each spot is also substantial. Here we proposed a quantification technique based on wavelet transform that will reduce the number of pixels for computation to one fourth without sacrificing the accuracy in interpretation of the microarray data.

2 Wavelet and Denoising

The Wavelet transform is a time frequency analysis tool that decomposes signals using wavelets as a basis function (ψ). A family of wavelets can be obtained by scaling by and translating it by Mallat (1998).

$$\psi_{u,s}(t) = s^{-\frac{1}{2}}\psi\left(\frac{t-u}{s}\right) \tag{1}$$

The Discrete Wavelet Transform (DWT) of a signal $f[n]$ with period N is computed as:

$$DWTf[n,a^j] = \sum_{m=0}^{N-1} f[m]a^{-j/2}\psi^*\left(\frac{m-n}{j}\right) \tag{2}$$

where m and n are integers. The value of a is equal to 2 for a dyadic transform. The signal representation is not complete if the wavelet decomposition is computed up to a scale a^j . The information corresponding to the scales larger then a^j is also required, and is computed by a scaling filter and is given by:

$$SFf[n,a^j] = \sum_{m=0}^{N-1} f[m]a^{-j/2}\phi^*\left(\frac{m-n}{a^j}\right) \tag{3}$$

where $\phi(n)$ is the discrete scaling filter. The output of the Equation 2 gives the detailed coefficients and the output of Equation 3 gives the approximate coefficients of a 1-D signal.

The denoising technique using wavelet analysis is based on the idea that the amplitude rather than the location of the spectrum of the signal to be different from the noise. Denoising by wavelet is quite different from traditional filtering approaches because it is non-linear, due to a thresholding step. Let a signal $x[n]$ be given as:

$$x[n] = f[n] + w[n] \quad 0 \leq n \leq N \tag{4}$$

where $f[n]$ is the original signal (which is assumed to be piecewise smooth) and $w[n]$ is the white Gaussian noise with zero mean. Denoising of the signal $x[n]$ by thresholding involves the following steps:

- Perform a suitable wavelet transform of the noisy data.
- Calculate the threshold d depending upon the noise variance.
- Perform thresholding of the wavelet coefficients.
- The coefficients obtained from the step above are then padded with zeros to produce a legitimate wavelet transform and this is inverted to obtain the signal estimate.

The two types of thresholding commonly used are hard and soft thresholding. Usually thresholding is applied on the detailed coefficients obtained after wavelet decomposition of the signal and the approximate coefficients are left untouched. In the hard thresholding all the coefficients with absolute value below a threshold δ are forced to zero while in soft thresholding, the detailed coefficients are modified as follows:

$$\tilde{d^s_{ij}} = \begin{cases} \text{sign}(d_{ij})(\mid d_{ij} \mid -\delta) & \text{if} \quad \mid d_{ij} \mid > \delta \\ 0 & \text{if} \quad \mid d_{ij} \mid \leq \delta \end{cases} \tag{5}$$

where sign(x) is $+1$ if x is positive and -1 if x is negative. Donoho and Johnson (1995) have shown that the optimal threshold for denoising is obtained by using the equation below:

$$\delta = \bar{\sigma}\sqrt{2\ln_e(N)} \tag{6}$$

where $\bar{\sigma}$ is the noise standard deviation estimate. The noise standard deviation estimate is carried out by performing one-level wavelet decomposition and calculating the median of the absolute values of detailed wavelet coefficients Mallat (1998).

3 Quantification

The main information to be extracted from microarrays is strength of expression in a target. The expression levels differ between the test and control mRNA populations. These two channels of a microarray are labeled with a different color. The total color intensity of a spot determines the expression strength. Through these intensities various statistical constants can be computed such as mean signal intensity and median signal intensity.

4 Dataset

DNA microarray is a powerful tool for parallel detection of multiple target genes in biological systems. In this study, a low-density DNA microarray data on differential expression of Pseudomonas stutzeri strain KC ORF's have been analysed using wavelet transform. The data, the raw data, were originally obtained from scanning of the microarray glass slide on GenePix 4000A microarray scanner (AxonTM Instruments) for simultaneous detection of Cy3 and Cy5 fluorescence using GenePix Pro 3.0 software (Musarrat et al. 2003). A grid of circles was aligned to the image using the known dimensions of the array and signal intensities of each spot calculated. Total intensity of all the pixels within each circle was averaged and data normalized for constant spot size of 100 mm diameter. Ratios for differences in dye-incorporation efficiency determined following the default computed normalization procedure obtained from Stanford microarray database. Finally the log mean of the intensities was obtained - termed as log mean (raw data). The same dataset was validated by directly computing the above statistical parameters on the image data. This precluded the involvement of expensive instrumentation and technical expertise in image analysis and makes it a simple and cost-effective approach.

5 Data Analysis and Results

During the process of preparation of the microarray data there are various possibilities of noise being added to the data. In this paper we propose denoising of microarray data using discrete wavelet transform. In order to demonstrate the capabilities of wavelet transform to remove noise we first generated an additive white Gaussian noise with different power. The generated noise was added to the clean microarray data. This gave different values of Signal to Noise Ratios (SNR), which is defined as:

$$\mathrm{SNR} = 10 * \log_{10}\left(\frac{\text{Signal Power}}{\text{Noise Power}}\right) \tag{1}$$

The denoising was carried out by having one level wavelet decomposition of the noisy data. Finally the inverse transform was applied to recover the signal with reduced level of noise. It is found that when the noise power is large the wavelet based denoising technique can effectively remove the noise. Thus, the denoising technique can be effectively used to remove the noise introduced during the process of creating the microarray data.

Under the procedure of quantification of data, an image file is used instead of raw data. Here we want to justify that the interpretation from our analysis is the same as that of the raw data analysis. However, the analysis and quantification of the image file is much simpler as compared to raw data because in an image file the RGB components are stored in one byte only. Thus the proposed technique is computationally efficient as compared to the previous.

For the quantification of the expression data a rectangular grid inscribing the circular spot was first placed to get the pixel information in the desired spot area. Once the rectangular grid was identified, mean of the red and green intensities was calculated. Finally, the log of this mean was computed. The results are obtained by doing the same calculation on the raw data and on the image data. The difference in the results is due to the fact that resolution of the raw data is much higher as compared to the actual image data. Due to this fact, the results obtained on raw data show higher values as compared to that of the image data. However, the correlation coefficient between the log of mean values of raw data and that of image data is found to be 0.7719, a significant correlation. This means that a very high degree of similarity exists between these two results. In other words the interpretation of the results obtained by these two datasets will be the same.

Furthermore, wavelet based quantification of the data was also carried out. In the case of clean image, the spot intensity of a color will be approximately uniform. The variation may occur due to the presence of noise being added during the process of preparation of microarray data, which is undesirable. A DWT splits the image into four sub-bands of low frequency, which has global information, and three other high frequency bands, which carry information about the vertical horizontal and diagonal edges present in the image. Two images were obtained: the image of a microarray data and the output of four wavelet filtered images giving the global view of the image, vertical, horizontal and diagonal edges. Thus the output band in the low frequency region (*i.e.* the LL band) is suitable for extraction of average intensity of the spot area.

The correlation coefficient between the log of mean values of raw data and that of wavelet transformed data is found to be 0.7741. This shows, a slightly higher correlation as compared to the previous case.

6 Conclusions

The quantification results computed by using the raw image data show high degree of correlation to the quantification results obtained using the original data. There is a further increase in correlation coefficient if wavelet based quantification is carried out. This implies that the statistical results obtained by using the images gives the similar interpretation as one derived from the full data analysis. Thus the proposed technique is not only fast and cost effective but also gives the same interpretation of the microarray data.

References

Mallat, S. (1998). *A Wavelet Tour of Signal Processing,* Academic Press, San Diego.

Donoho, D. L. and Johnston, I. M. (1995). De-noising by soft-thresholding,” *IEEE Transactions Information Theory,* **41**, 613-627.

Musarrat J. and Hashsham S. A. (2003). Customized cDNA for expression profiling of environmentally important genes of Pseudomonas stutzeri strain KC, *Teratogenesis, Carcinogenesis and Mutagenesis Supplement,* **1**, 1-12.

Arneodo, A., d”Aubenton-Carafa, Y. Barcy, E., Graves, P. V., Muzy, J. F and Thermes, C. (1996). Wavelet based fractal analysis of DNA sequences. *Physica D,* **1328**, 1-30.

Mandell, A. J., Selz, K. A., and Shlesinger, M. F. (1997). Wavelet transform of protein hydrophobicity sequences suggests their membership of structural families. *Physica A,* **244**, 254-262.

Myasnikova, E. Samsonova, A., Kozlov, K., Samsonova, M. and Reinit, J. (2001). Registration of the expression patterns of Drosophila segmentation gene by two independent methods. *Bioinformatics,***17**, 3-12.