

Wavelet Transforms for the Analysis of Microarray Experiments

Taku A. Tokuyasu, Donna Albertson, Dan Pinkel, Ajay Jain
UCSF Cancer Center, Box 0128
San Francisco, CA 94143-0128
tokuyasu@cc.ucsf.edu

Abstract

Array comparative genomic hybridization (cgh) is a microarray technology for measuring the relative copy number of thousands of genomic regions. Visual examination of cgh profiles shows that genomic changes occur on a variety of length scales. Such changes may be characteristic of phenotypic variables such as tumor type and gene mutational status. To aid in identifying such features and exploring their relationship with phenotypic outcomes, we are applying wavelet transforms to the analysis of such profiles. This allows us to decompose a cgh signal into components on different length scales, even when the genome is severely aberrated, providing a convenient basis for exploring their behavior. Wavelet transforms may also be useful in the realm of gene expression. The expression signal given by genes in clustered order can be wavelet transformed, which compresses the signal from many genes into a few components, possibly aiding in the development of new tumor classifiers.

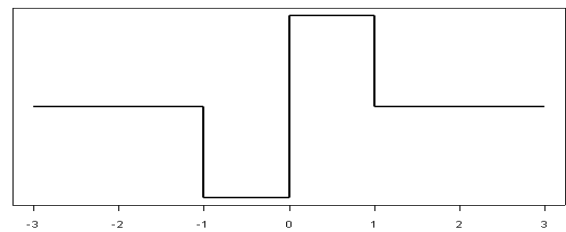
1. Introduction

Microarray technologies query the state of cells over thousands of variables in a single experiment. The expectation that such information will provide rich insights into biology and the nature of disease has been tempered somewhat by the sheer complexity of the datasets. This has spurred the rapid development of statistical methods designed to address this issue [1]. Before turning to such methods, it may be fruitful to exploit coherence in the data over multiple genes or clones. Here we propose the use of wavelet transforms as a tool to characterize structure at multiple positions and length scales. A basic assumption in the following is that the data can be ordered in one dimension.

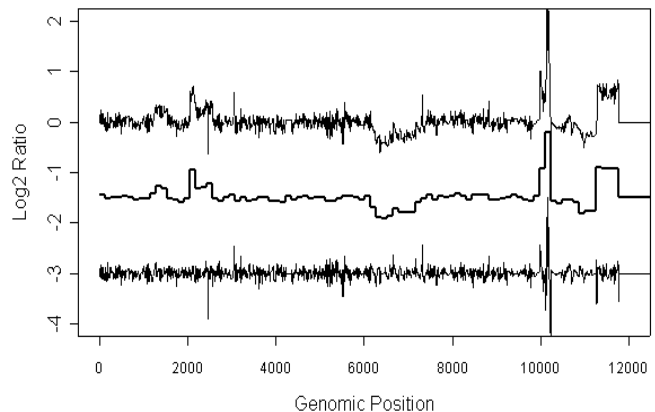
2. Wavelet transforms

A wavelet transform is a lossless linear transformation of a signal into coefficients on a basis of wavelet functions, akin to a Fourier transform [2]. We use Haar wavelets (Fig. 1a) for ease in interpreting the results. The region of support where the Haar wavelet is nonzero

defines both a characteristic position and length scale. The complete set of basis functions consists of all possible copies of this function, shifted to cover the domain of interest (e.g., the genome) and expanded by powers of two to cover length scales from two (as in the figure) to the size of the genome (plus an additional function that captures the overall signal average). We are using Wavelab software [3] for this purpose.



(a)



(b)

Figure 1: (a) Haar wavelet. (b) Sample breast tumor cgh profile (top), and its long (middle) and short (bottom) length-scale reconstructions.

3. Application to cgh profiles

We focus on array comparative genomic hybridization (CGH) [4], which measures the copy number of genomic regions represented by clones spotted on a microscope slide. It is now possible to measure copy number across the genome at a resolution of about 1 Mb [5], with higher resolution expected shortly. The genome can exhibit significant deviations from the normal copy number of two in cancer and various inherited genetic diseases.

The top signal in Fig. 1b is a sample breast tumor cgh profile [5], where 0 on the vertical axis corresponds to a normal genome. This is a discrete signal arising from 2300 clones arranged in genome order, beginning with chromosome 1 on the left and ending at chromosome X on the right. Note the presence of both extended features that correspond to the gain or loss of a substantial part of a chromosome, and focal deletions and amplifications ("amplicons") that may correspond to specific genes selected for in the evolution of the tumor [6].

As suggested previously, the wavelet transform has a bias for locations and lengths scales that are a power of two. One method that we use to take advantage of this characteristic is to scale each chromosome to the same power of two. This has in fact been done in Fig. 1b, where each chromosome corresponds to 512 units (roughly Mb) along the x-axis. While this stretches smaller chromosomes considerably, this (reversible) rescaling does not appear to impair our ability to detect focal events (see below). It also may help separate the effects of mechanisms that operate on the length scale of a chromosome from more focal mechanisms. We linearly interpolate the signal between the observations from actual clones, and pad the overall signal with zeros to a power of two. The wavelet transform then produces a number of wavelet coefficients equal to the signal length.

For illustrative purposes, below the cgh profile in Fig. 1b, we plot its long-length and short-length scale reconstructions (displaced vertically for clarity), which result from alternately zeroing the wavelet coefficients above and below the length scale of 128 units and inverse transforming. In the long-length reconstruction, each chromosome corresponds to four horizontal segments, so e.g. the loss of a chromosome arm would appear in this component. The short length reconstruction describes the remaining deviations in the signal away from the local average given by the long-length component, and thus tends to be centered on zero. This clearly captures focal events in the original profile and could be used to define such features. There are occasional artifacts such as overshooting and strong signals at dramatic changes in profile value, such as appear at either end of the X chromosome in Fig. 1b. These can be adjusted for with reference to the original signal if required.

The wavelet coefficients themselves allow a new set of statistical questions to be more easily formulated, such as the correlation of chromosomal regions at different positions and length scales. Note that the decomposition into short- and long-length scale components can be made even when the genome is considerably more aberrated than is the case in Fig. 1b.

4. Gene expression

The basic intuition for applying wavelets in the case of gene expression is that genes are often co-expressed in groups. It would be useful to treat the group as a single variable, akin to the motivation behind methods such as principal component analysis. This is particularly important in tasks such as feature selection for classifiers. We begin by hierarchically clustering the genes, and then wavelet transform the per-sample expression profiles for the genes in clustered order. Our initial results are encouraging, with wavelets on a variety of "length" scales being highly correlated with outcome in a number of standard datasets. It remains to be seen whether the promise of additional leverage can be realized.

5. References

- [1] T. P. Speed (ed), *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC, Boca Raton, 2003.
- [2] Stephane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, January 1998.
- [3] David Donoho et al, *WAVELAB 802*, <http://www-stat.stanford.edu/~wavelab/>.
- [4] D. Pinkel et al, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nat Genet.* 1998 Oct;20(2):207-11.
- [5] A. M. Snijders et al, "Assembly of microarrays for genome-wide measurement of DNA copy number," *Nat Genet.* 2001 Nov;29(3):263-4.
- [6] D. G. Albertson, et al, "Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene," *Nat Genet.* 2000 Jun;25(2):144-6.