HMMSeg wavelet methods

Robert E. Thurman
Division of Medical Genetics
Department of Genome Sciences
University of Washington

February 5, 2007

Wavelets provide a framework for multi-scale analysis. By decomposing a given data type into increasingly coarser scales, wavelet analysis allows broader and broader trends in the data to reveal themselves. As opposed to Fourier analysis, which also provides a decomposition of a given signal in terms of multiple scales (frequencies), wavelet analysis localizes behavior in both frequency and "time" (genomic position, in our case), and is thus better suited to pick up transient behavior.

The basis for understanding wavelet transforms is the continuous wavelet transform (CWT) [1, 2]. Mathematically, for a given time series $x(t)$, the CWT wavelet coefficient $W(a, s)$ for given scale a and time s is given by

$$W(a,s) \equiv \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(u)\psi\left(\frac{u-s}{a}\right) du,$$

where $\psi(t)$ is the wavelet function of choice, satisfying the basic properties $\int_{-\infty}^{\infty} \psi(u), du = 0$ and $\int_{-\infty}^{\infty} \psi^2(u), du = 1$. Simple examples of $\psi(t)$ include the Haar wavelet and the so-called "Mexican hat" wavelet. The wavelet coefficient $W(a, s)$ captures information about the local behavior of x at scale a near time (genomic position, in our case) s.

The discrete wavelet transform (DWT) can be thought of as a discretization of the CWT across evenly-spaced values of t and dyadic scales $a_j = 2^j\delta$, where δ is the resolution of $x(t)$. For fixed level J (scale $2^J\delta$), the DWT allows for a decomposition

$$x = \sum_{j=1}^{J} \mathcal{D}_j + \mathcal{S}_J, \tag{1}$$

of x into a sum of the wavelet *smooth* $\mathcal{S}_J$ and wavelet *details* $\mathcal{D}_j$, each of which is a time series the same length as x. Each $\mathcal{D}_j$ represents the local variation in x at scale $2^j\delta$ while $\mathcal{S}_j = x - \sum_{j=1}^{J} \mathcal{D}_j$ can be thought of as a smoothed version of x, with the details at lower scales removed. The MODWT (maximal overlap DWT) [1] is a modification of the DWT that also gives rise to the multiresolution

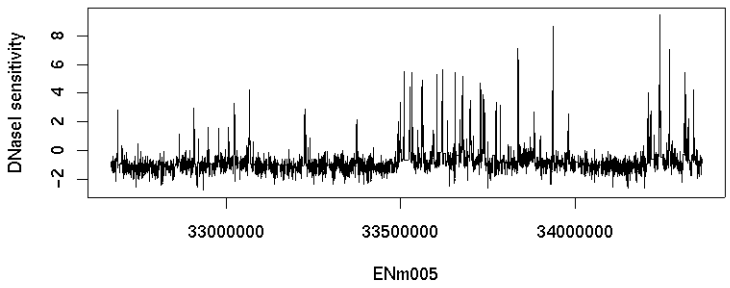

Figure 1: DNaseI sensitivity, ENCODE region ENm005 on chr21.

analysis (1) but which, among other things, allows input sequences x of arbitrary length (the DWT requires length of x to be a power of two), at the cost of requiring more, redundant, intermediary wavelet coefficients. HMMSeg uses the Daubechies "least asymetric" LA(8) wavelet filter (discrete analog of the wavelet function ψ), with reflection boundary conditions. The LA(8) wavelet is considered a good general-purpose wavelet whose "width" (8) strikes a balance between providing smooth approximations with few artifacts, and minimal edge effects at the boundaries of the data [1].

As an example, Figure 1 contains a plot of DNaseI sensitivity measured as part of the ENCODE project [3] in ENCODE region ENm005 on chromosome 21. The data have been interpolated to give equally-spaced values at every 50bp. Figure 2 shows the MRA decomposition of the same signal at the 6.4kb scale (level $J = 7$), using the LA(8) wavelet. Figure 3 shows an enlargement of the 6.4kb scale smooth. Only the smooth portion of the MRA is output by HMMSeg.

References

[1] Donald B. Percival and Andrew T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[2] C. Torrence and G.P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

[3] ENCODE Consortium. The ENCODE pilot project: functional annotation of 1% of the human genome. Submitted, 2006.

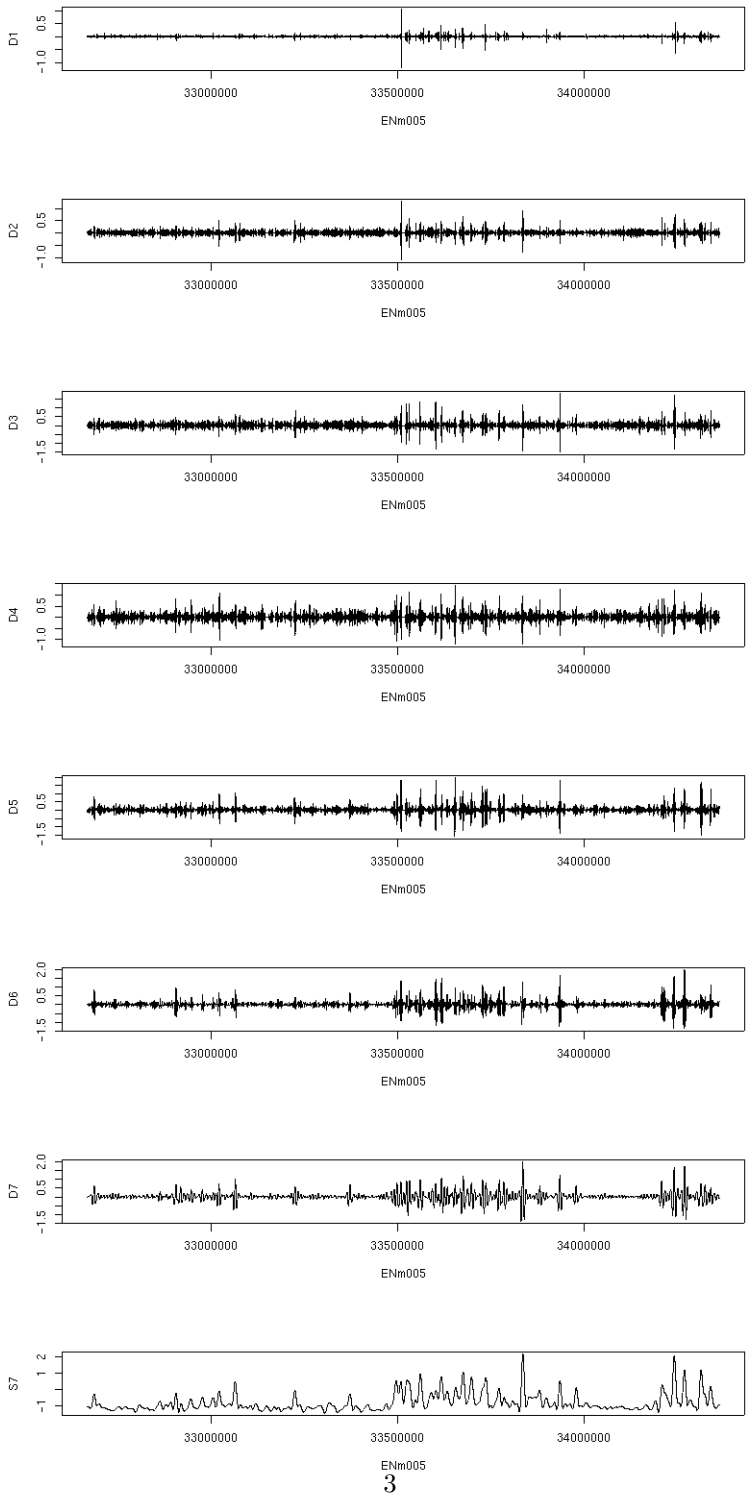

Figure 2: MRA for the DNaseI data. Top to bottom, details $\mathcal{D}_1 \ldots, \mathcal{D}_7$; smooth $\mathcal{S}_7$.

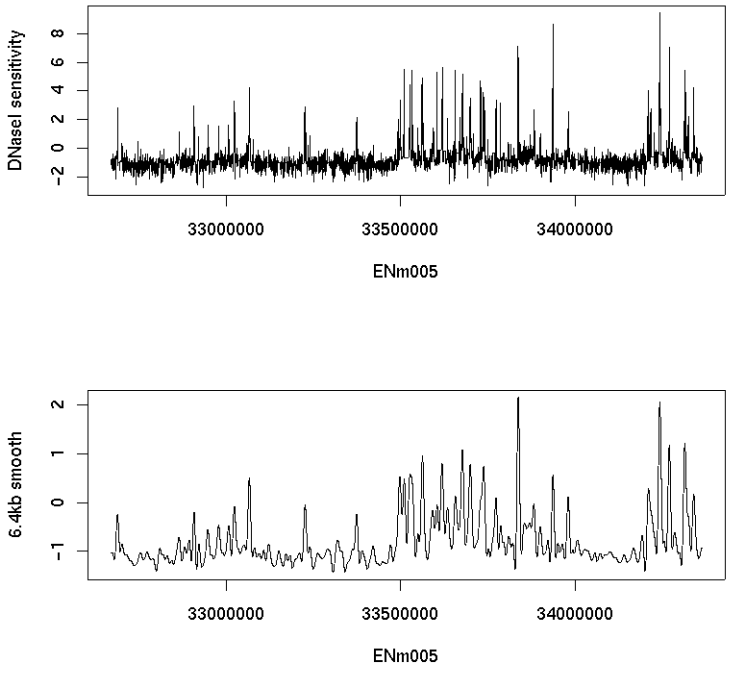

Figure 3: DNaseI original signal, top; 6.4kb wavelet smooth, bottom