

Wavelet Analysis of Nucleotide Genomic Sequences

Jianchang Ning^{1*}, Charles N. Moore², James C. Nelson³

¹Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711;

²Department of Mathematics, Kansas State University, Manhattan, KS 66506;

³Department of Plant Pathology, Kansas State University, Manhattan, KS 66506

Key words: wavelet, gene prediction, bioinformatics

Abstract

Wavelet algorithms are being developed as an alternative way to determine gene locations in genomic DNA sequences. The preliminary results from the development are presented. The data show the wavelet approach is feasible and better than knowledge-dependent approach based on a sample of sequences.

Introduction

Rapid and accurate determination of gene locations is imperative for genome sequencing projects. Computational approach is the fastest way so far to find genes in genomic DNA sequences. High-throughput genome sequencing projects have made urgent the development of accurate methods for annotation of DNA sequences, especially for identification of gene structures embedded in genomic sequences. Driven by this explosion of genomic data, computational gene prediction continues to be an active research field (Zhang 2002, Pertea and Salzberg 2002). Even though algorithms for *ab initio* gene prediction have been steadily improved in the past decade, the accuracy is still far from satisfactory: although, at the nucleotide level, up to 95% of genes can be accurately predicted, at the exon level only up to 75% are predicted, and at the whole-gene level only ~20% (Guigo et al. 2000, Claverie 1997). The most successful programs so far are based on Hidden Markov Models (HMM) (Pertea and Salzberg 2002, Guigo et al. 2000, Claverie 1997). With this algorithm, programs need be trained with data sets of well-characterized genes. However, the major limitation with HMM method is that we have a little knowledge of gene structures, especially, for new sequencing genomes. Furthermore, current set of known genes is limited and certainly does not represent all potential gene features or their organizational themes, which would lead to inevitable bias in the statistics and patterns extracted from the dataset. Thus, the present paper describes a method that adapts signal processing algorithms to predict genes in genomic DNA sequences.

A DNA sequence may be schematically represented as a non-branching string of nucleotides (bases), designated by their initials A, C, G and T. Certain regions of this string known as genes or coding regions are functioned as templates to guide synthesis of peptides by cellular machinery. Genes in the DNA of eukaryotic genomes are neither contiguous nor continuous. Large regions of non-coding DNA may intervene between genes and the genes themselves usually consist more of non-coding DNA in the form of

* Corresponding author: ning@dbi.udel.edu, phone: 302-8313229, fax: 302-831-3410

introns. Due to the nature of codon composition and codon usage, coding regions (CDS) tend to be three-based periodicity. On the other hand, however, non-coding regions (nCDS) have the random tendency or non-three-based periodicity (repeat portions) (Guigo 1997). Signal processing approach is the perfect way to detect periodicity of signals (Embree and Danieli 1999). But conventional Fourier analysis can only reveal “global” periodicity of “stationary” signals. Wavelets, on the contrast, provide multi-scale representation of signals. Basic idea of wavelets is to decompose a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. Coefficients at coarse scales capture gross and global features of the signal while coefficients at fine scales contain local details. There are many applications of wavelets in biology (see the review paper by Lio 2003 for the examples and references).

Wavelet method could be an alternative way to determine gene structures embedded in genomic DNA sequences. The present study employed discrete wavelets (DWT) to decompose genomic DNA sequences followed by data-dependent thresholding algorithms to remove the background. Then entropic segmentation method (Bernaola-Galvan et al. 2000) was applied to find boundaries between segments. Finally, we used biological information to validate the above results. Before the wavelet decomposition, genomic DNA sequences were digitized into numerical sequences based on their contents. This paper just presents our preliminary results since our algorithms are not finalized and implementation is not optimized yet.

Algorithms

A wavelet family with orthonormality is defined as $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)$, where j and k are dilation and translation parameters, respectively. A wavelet family includes a father wavelet $\phi(t)$ and a mother wavelet $\psi(t)$ and they satisfy the conditions of $\int_{-\infty}^{+\infty}\phi(t)dt = 1$, $\int_{-\infty}^{+\infty}\psi(t)dt = 0$, and $\phi(t) = \sqrt{2}\sum_k l_k\phi(2t - k)$, $\psi(t) = \sqrt{2}\sum_k h_k\phi(2t - k)$, where l_k and h_k are the low-pass and high-pass filters defined as $l_k = \sqrt{2}\int_{-\infty}^{+\infty}\phi(t)\phi(2t - k)dt$ and $h_k = \sqrt{2}\int_{-\infty}^{+\infty}\psi(t)\phi(2t - k)dt$ with the relationship of $h_k = (-1)^k l_{n-1-k}$, where n is the filter length. This family can be utilized to represent a signal $f(t)$ as $f(t) = \sum_{j,k} \hat{f}_{j,k} \psi(t)$, where $\hat{f}_{j,k}$ is the wavelet coefficients given by $\hat{f}_{j,k} = \int_{-\infty}^{+\infty} f(t)\psi_{j,k}^*(t)dt$. The coefficients $\hat{f}_{j,k}$ describe features of $f(t)$ at the time/spatial ($k2^j$) and frequency proportional to 2^j . Unlike classic Fourier, therefore, by wavelets signal components can be localized both in time/space domain and in frequency domain. As j increases (and the scale factor 2^j decreases), the oscillations in the mother wavelet increase and exhibit a “high frequency” behavior. On the other hand, we obtain a “low frequency” behavior while j decreases.

Components of a signal can be displayed by wavelet scalogram. The scalogram is defined as a plot of the sums of squares of the wavelet coefficients at different scales:

$I(j) = \sum_{k=0}^{2^j-1} \hat{f}_{j,k}^2$ for $j=0, \dots, n-1$. According to Chiann and Morettin (1998), $I(j)$ may be approximated to be a Chi-Squared distribution.

After decomposition of a signal, we followed the data-dependent method of Ogden and Parzen (1994) to remove the background noise. The fundamental idea of the data-dependent method is to kick out the largest element each time from a vector of coefficients at a particular scale until the element is no longer significantly different from the threshold value. Those elements whose values are below the threshold are deemed to be background noise. Then, inverse wavelet transform is applied on these shrunk vectors to reconstruct the signal components.

Delimitation of boundaries of segments is a challenge to signal processing algorithms since changes of a few positions will not result in significant changes in spectrum and/or scalogram. We modified the entropic segmentation algorithm (Bernaola-Galvan et al. 2000) to optimize the segmentation. The basic idea for the algorithm in the case of genomic DNA sequences is that the position with the largest joint entropy is the separate point.

Implementation

We first digitize a genomic DNA sequence using electron-ion interaction potentials of nucleotides with $A=0.1260$, $C=0.1340$, $G=0.0806$ and $T=0.1335$. The binary indicator sequence method (Voss 1992) was also implemented to convert genomic DNA sequences into numerical sequences but did not present the results here. We then applied Coiflets and Daubechies wavelets to decompose the sequences and reconstruct them. The final results are to be validated with biological information. But this part is to be finished yet. The scheme is present in Figure 1.

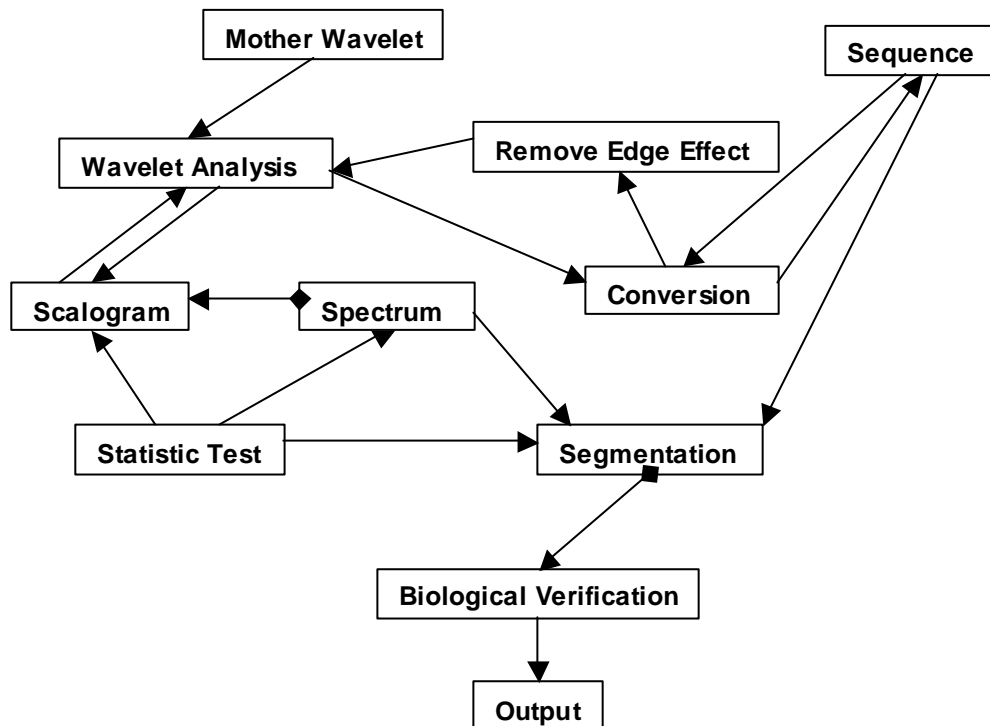


Fig. 1 Schematic Diagram of Data Flow

Results

We tested our approach using Fickett & Tung (1992) benchmark datasets. First of all, sequences of thirteen concatenated exons and twelve concatenated introns were digitized using electron-ion interaction potentials of nucleotides. Scalograms were calculated and plotted below (Figure 2) after the wavelet transforms were applied to these sequences (Because of the similarity between Coiflets and Daubechies, the results of Coiflets are showed only). The figures show that there is significant difference in scalograms between exons and introns.

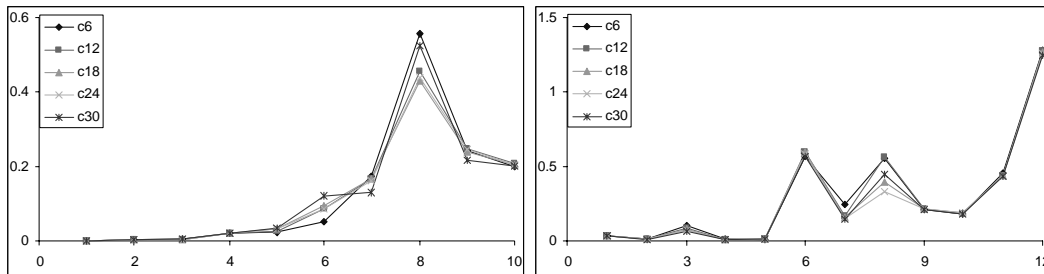


Figure 2. Scalograms of exons (left) and introns (right) from the benchmark datasets transformed by Coiflets (Energy vs. resolutions, the same in following figures)

Then, we applied the same algorithms to a real DNA sequence (GenBank accession #: AB009592) (Figure 4), which contains a single gene consisting of intervened seventeen exons and sixteen introns. Total length of the exons is about 20 % of the whole sequence length. So, its scalogram reflects more intron features.

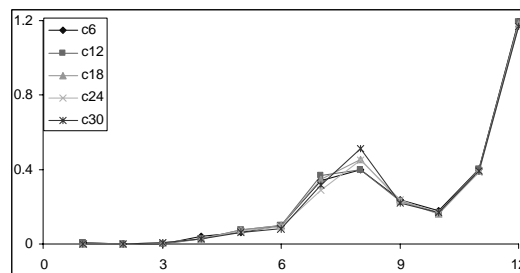


Figure 3. Scalogram of a real DNA sequence (accession # AB009592) transformed by Coiflets

Our results of a sample of sequences (Accession # AB009592 and AC078840) with the wavelet algorithms showed that this approach has higher sensitivity (on average, 0.52 vs. 0.21 for bases and 0.47 vs. 0.18 for exons) and similar specificity (on average, 0.29 vs. 0.25 for bases and 0.43 vs. 0.40 for exons) in comparison with the most popular gene-finding programs, GENSCAN (1997) and GLIMMER (1999) (Detail, see the table).

Table 1. Comparison of my results (GENELET) with others

<i>GenBank ID: AB009592</i>		Specificity1		Sensitivity2	
		Base	Exon	Base	Exon
GENSCAN	Maize	0.20	0.40	0.07	0.12
	Arab	0.33	0.30	0.30	0.18
GLIMMER	Rice	0.23	0.43	0.12	0.18
	Arab	0.23	0.40	0.25	0.24
GENELET		0.29 (0.085)	0.43 (0.105)	0.52 (0.185)	0.47 (0.159)

¹ *Specificity = (right predictions)/(positive predictions)*

² *Sensitivity = (right predictions)/(real number)*

Discussions

Precise prediction of gene structure by purely computational means is notoriously difficult, mostly because we have not completely understood the mechanisms of gene transcription and translation in the cellular machinery. With current technology and experimental approaches, we have no much hope to solve the mystery soon. Computational determination of gene location in a high-throughput fashion will be still the main approach. The present study is just to provide an alternative way to achieve higher accuracy of gene prediction. The data present in this paper is our preliminary results. This results show that the wavelet approach is feasible and better than the knowledge-based methods in these cases.

In our ongoing research, we are optimizing the algorithms, comparing different wavelets, and refining the biological model. We are even evaluating different digitizing methods and statistic models. We will intensively test our approach with the benchmark dataset and our own developing dataset.

Reference

- Bernaola-Galvan P, Grosses I, Carpena P, Oliver JL, Roman-Roldan R, and Stanley HE. 2000. Finding borders between coding and noncoding DNA regions by an entropic segmentation method. **Physical Review Letters** **85(6)**: 1342-1345.
- Chian, C. and Amoretin, P. A. 1998. A wavelet analysis for time series. **J. Nonparametric Statistics** **10**: 1-46.
- Claverie JM. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. **Human Molecular Genetics** **6(10)**: 1735-1744.
- Fickett, J, W. and Tung, C. S. 1992. Assessment of protein coding measures. **Nucleic Acids Research** **20**: 6441-6450.
- Guigo, R, Agarwal, P. et al. 2000. An assessment of gene prediction accuracy in large DNA sequences. **Genome Research** **10**: 1631-1642.
- Guigo R. 1997. Computational gene identification. **J. Mol. Med.** **75**: 389-393.
- Lio, P. 2003. Wavelets in bioinformatics and computational biology: State o art and perspectives. **Bioinformatics** **19(1)**: 2-9.
- Pertea, M. and Salberg, S. L. 2002. Computational gene finding in plants. **Plant Molecular Biology** **48**: 39-48.
- Voss RF. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequence. **Physical Review Letters** **68**: 3805-3808.
- Zhang, M. Q. 2002. Computational prediction of eukaryotic protein-coding genes. **Nature Review Genetics** **3(9)**: 698-709.