

Preliminary Wavelet Analysis of Genomic Sequences

Jianchang Ning

*Delaware Biotechnology
Institute
University of Delaware
Newark, DE 19711, USA
ning@dbi.udel.edu*

Charles N. Moore

*Department of
Mathematics
Kansas State University
Manhattan, KS 66506,
USA
cnmoore@ksu.edu*

J. Clare Nelson

*Department of Plant
Pathology
Kansas State University
Manhattan, KS 66506,
USA
jcn@ksu.edu*

Abstract

Large genome-sequencing projects have made urgent the development of accurate methods for annotation of DNA sequences. Existing methods combine ab initio pattern searches with knowledge gathered from comparison with sequence databases or from training sets of known genes. However, the accuracy of these methods is still far from satisfactory. In the present study, wavelet algorithms in combination with entropy method are being developed as an alternative way to determine gene locations in genomic DNA sequences. Wavelet methods seek periodicity present in sequences. A promising advantage of wavelets is their adaptivity to varying lengths of coding/non-coding regions. Moreover, the wavelet methods integrated with entropy method just search the information contents of the sequences, which do not need to be trained. The preliminary results show that the wavelet approach is feasible and may be better than some knowledge-dependent approaches based on a sample of genomic DNA sequences.

1. 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. Even though algorithms for *ab initio* gene prediction have been steadily improved in the past decade, the accuracy is still far from satisfactory. Periodicity, due in general to nonrandomness of nucleotide usage associated with the triplet nature of codons, is an important target of most gene-parsing computer programs. The most successful programs so far are

based on Hidden Markov Models (HMM) [1-3]. 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.

Signal processing approach is the perfect way to detect periodicity of signals [4]. 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 [5-6].

2. 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 [7] 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.

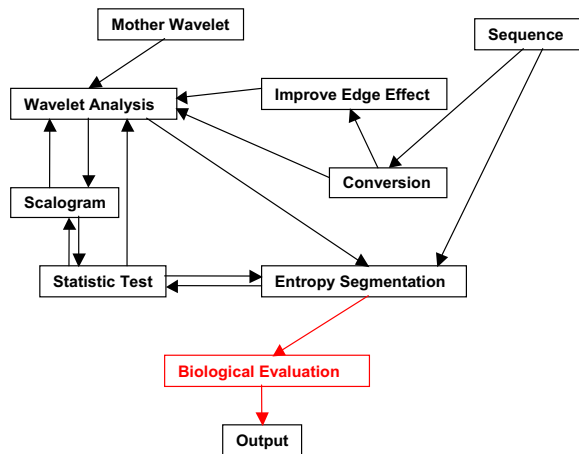


Figure 1. Schematic diagram of data flow

3. Preliminary Result

We tested our approach using Fickett & Tung [8] 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.

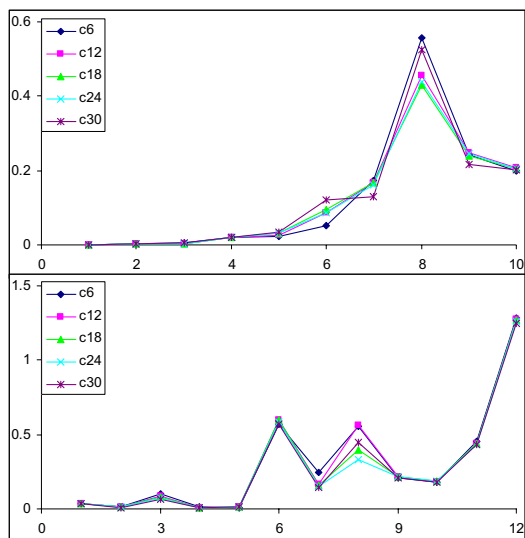


Figure 2. Scalograms of exons (up) and introns (down) from the benchmark datasets transformed by Coiflets (Energy vs. resolutions, cxx represent coiflets with different supports, the same in following figures)

Then, we applied the same algorithms to a real DNA sequence (GenBank accession #: AB009592) (Figure 3), 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.

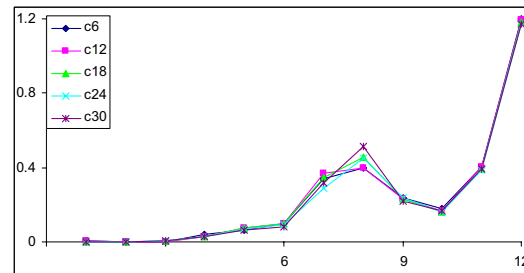


Figure 3. Scalogram of a real DNA sequence

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).

4. Reference

- [1] M. Pertea and S.L. Salberg, "Computational gene finding in plants", *Plant Molecular Biology* vol. 48, pp.39-48, 2002.
- [2] R. Guigo, P. Agarwal et al., "An assessment of gene prediction accuracy in large DNA sequences", *Genome Research*, vol. 10, pp. 1631-1642, 2000.
- [3] J.M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences", *Human Molecular Genetics*, vol. 6, pp. 1735-1744, 1997.
- [4] P.M. Embree and D. Danieli, *C++ Algorithms for Digital Signal Processing*, 2nd Edt, Prentice Hall PTR, New Jersey, 1999.
- [5] P. Lio, "Wavelets in bioinformatics and computational biology: State o art and perspectives", *Bioinformatics* vol 19, pp. 2-9, 2003.
- [6] C. Chiann and P.A. Morettin, "A wavelet analysis for time series", *J. Nonparametric Statistics*, vol. 10, pp. 1-46, 1998.
- [7] R.F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequence", *Physical Review Letters*, vol. 68, pp.3805-3808, 1992.
- [8] J.W. Fickett and C.S. Tung, "Assessment of protein coding measures", *Nucleic Acids Research*, vol. 20, pp. 6441-6450, 1992.