Wavelet Variance and Entropy of DNA Sequences

En-Bing Lin and Mojisola Oyapero

Abstract—**To differentiate coding regions from noncoding regions is an important task in the field of bioinformatics. Our goal here is to analyze coding and noncoding regions and find similarities for coding and noncoding regions respectively. We use wavelet analysis to analyze coding and noncoding regions of DNA sequences by calculating wavelet variance and entropy which give rise to some distinctions between two regions. Based on the calculated results, we provide some similarities for coding and noncoding regions for variance and entropy respectively.**

***Index Terms*—Wavelets, wavelet transform, wavelet variance, entropy, coding and noncoding regions**

I. Introduction

A. Coding and Noncoding Regions of DNA Sequences

Biological data have generally a multi-scale structure, in particular, DNA sequences of most micro organisms have revealed some structures at different sequence scales. On the other hand, wavelet analysis provides many useful tools to perform multi-scale analysis and analyze the structures of given data. [4] Given the multi-scale structure of most biological data, wavelet methods appear to be a natural way to achieve improvements in the quality of mathematical or statistical analyses of such data. In a DNA strand, it is essential to find sequences which can be transcribed to complementary parts of the DNA strand. The DNA material in chromosomes is composed of coding and noncoding regions. The coding regions are known as genes and contain the information necessary for a cell to make proteins. [2] Different methods have been used to identify protein coding regions over the years which include Genscan algorithm [1] and MZFF method. [7] Another method explores the measure of spectral content in DNA sequences based on the fact that coding regions show a periodic organization of three bases, which are not found in noncoding regions. [6] In this paper, we use wavelet variance and entropy to analyze some similarities among coding and noncoding regions of several DNA sequences respectively. In what follows, we define wavelet transform which will be used to define wavelet variance in section 2. We will then define entropy and calculate wavelet variance and entropy, respectively, and compare the resulting data.

B. Wavelet Transform

The continuous wavelet transform of a continuous, square-integrable function x(t) is defined as:

$$W(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}x(t)\psi^{*}(\frac{t-b}{a})dt, \quad a>0, b\in R \tag{1.1}$$

where the symbol * represents the operation of complex conjugate. W(a, b) is the wavelet coefficient at space b and scale a; x(t) is the given signal (DNA sequence) at scale a and translational value b; ψ is the mother wavelet. [5]

II. Wavelet Tools

A. Wavelet Variance

It has been established that wavelets can break down the original signal into components of different scales; hence it provides a powerful tool to detect the pattern of variations across scales in observed data. The wavelet variance is thus calculated so that different data sets may be compared at different scales. It is defined as follows.

$$V(a)=\frac{1}{n}\sum_{j=1}^{n}W^{2}(a,x_{j}) \tag{2.1}$$

where $W^2(a,x_j)$ is the squared wavelet coefficient associated with scale *a* at data point x_j, and *n* is the number of data points. [3]

The formula (2.1) is used to calculate the wavelet variance for the various coding and noncoding regions. The results are compared in section 3.

B. Wavelet Entropy

Entropy is a measure of uncertainty. It is a measure of the disorderliness or randomness in a closed system. The definition of entropy given by Shannon is as follows:

$$WS^{(k)}=-\sum p_{j}^{(k)}\bullet\log_{2}p_{j}^{(k)} \tag{2.2}$$

where p_j is the wavelet coefficients for each level k. [5] The value of k depends on the level used for the different wavelets used. This is used to calculate the entropy values for the various coding and noncoding regions. The results are recorded in section 3.

C. Remark

From equations (2.1) and (2.2) wavelet variance and entropy are defined respectively.

Now using the Taylor series expansion for ln p in the neighborhood of a=1 in the entropy equation for each j is given below:

Manuscript received February 24, 2012; revised April 29, 2012.

The authors are with the Department of Mathematics, Central Michigan University, and Mount Pleasant, MI 48859, U.S.A (e-mail: enbing.lin@cmich.edu)

$$-\sum p \bullet \log_2 p = -\sum p \bullet \left(\frac{\ln p}{\ln 2}\right) = -\frac{1}{\ln 2}\sum p \bullet \ln p$$

$$= -\frac{1}{\ln 2}\sum p \bullet \left(-\frac{5}{3} + \frac{5p}{2} - p^2 + \frac{p^3}{6} + \cdots\right) = \sum\left(k_1 p - k_2 p^2 + k_3 p^3 - \cdots\right)$$

Comparing the above with the wavelet variance where p is the wavelet coefficients, we can write the wavelet variance as $c\sum p^2$, where c is a constant. For each DNA sequence, the difference and the ratio between the entropy and the wavelet variance have a consistent pattern since they both depend on p. These values may vary from one DNA sequence to another since the wavelet coefficients are different for each DNA sequence.

III. Discrete Analysis

A. Wavelet Variance and Entropy

This section contains results obtained when continuous wavelet analysis is carried out on the following DNA sequences.

The coding and noncoding regions are divided into sections for the three model organisms as follows:

Ateles geoffroyi

Coding region S2: 3951:3955 S3: 5434:5516 S4: 5810:5911 S5: 6665:6739 S6: 7637:8415

Noncoding regions: B2: 3956:5433 B3: 5517:5809 B4: 5912:6664 B5: 6740:7636
B6: 8416:10894

Anolis carolinensis

Coding regions: K2: 1341-1701 K3: 3898-4066 K4: 5207-5372 K5: 6003-6242 K6: 9361:9482

Noncoding regions: J2: 1702-3897 J3: 4067-5206 J4: 5373-6002 J5: 6243-9360
J6: 9483-9923

Bos taurus

Coding regions: M2: 80:97 M3: 1261:1350 M4: 3248:3332 M5: 4914:4972

Noncoding regions: L2: 98:1260 L3: 1351:3247 L4: 3333:4913 L5: 4973:5116

The numerical values of the wavelet variance and entropy are given in the following tables for the various coding and noncoding regions for the different model organisms:

TABLE 3.1 Wavelet Variance and Entropy for Ateles Geoffoyi

	Coif5		db10		bior6.8			
Ateles	Var	Entropy	Var	Entropy	Var	Entropy	ratio	diff
S2	0.9774	2.8888	1.0134	2.6298	0.994	2.3894	2.955596	1.9114
S3	1.0134	2.6335	0.9972	2.6478	0.9802	2.8019	2.598678	1.6201
S4	1.2231	2.8759	1.3787	3.0292	1.1359	2.7318	2.35132	1.6528
S5	1.2349	2.3252	1.062	2.2748	1.1983	2.4345	1.882905	1.0903
S6	1.1106	3.643	1.1912	3.3238	1.0387	3.6761	3.280209	2.5324
B2	1.2257	3.4773	1.2123	3.58	1.1578	3.5003	2.836991	2.2516
B3	1.1785	3.1459	1.1977	3.6887	1.1363	3.0741	2.66941	1.9674
B4	1.0938	3.6591	1.663	3.6449	1.0317	3.6672	3.34531	2.5653
B5	1.2005	3.216	1.2005	3.4472	1.1179	3.3028	2.678884	2.0155
B6	1.1406	3.9318	1.1847	3.8749	1.0788	3.9533	3.447133	2.7912

TABLE 3.2 Wavelet Variance and Entropy for Anolis Carolinensis

	coif5		db10		bior6.8			
Anolis	Var	Entropy	Var	Entropy	Var	Entropy	ratio	diff
K2	1.2296	2.6946	1.1903	3.1635	1.1552	2.8714	2.48563	1.7162
K3	1.1934	3.6714	1.096	3.4391	1.1862	3.3786	2.848255	2.1924
K4	1.2349	3.0473	1.22	2.7592	1.1662	2.9179	2.502058	1.7517
K5	1.1265	3.238	1.2309	3.1331	1.0194	3.4793	3.413086	2.4599
K6	1.097	3.0852	1.1049	2.6658	1.0228	3.0394	2.971646	2.0166
J2	1.28	3.8112	1.2949	3.9617	1.2029	3.8894	3.233353	2.6865
J3	1.3264	3.4625	1.2569	3.5337	1.243	3.4094	2.74288	2.1664
J4	1.2796	3.1788	1.2551	3.4518	1.1925	3.5529	2.979371	2.3604
J5	1.2737	3.7613	1.2972	3.9411	1.2056	3.7642	3.122263	2.5586
J6	1.1914	3.2403	1.2247	3.604	1.1354	3.3083	2.913775	2.1729

TABLE 3.3 WAVELET VARIANCE AND ENTROPY FOR BOS TAURUS

	Coif5		db10		bior6.8			
Bos taurus	Var	Entropy	Var	Entropy	Var	Entropy	ratio	diff
M2	1.089	2.9078	0.987	2.678	0.9016	2.348	2.713273	1.691
M3	1.2003	2.8498	0.9969	2.3241	1.1571	2.4702	2.331327	1.3272
M4	1.016	2.7167	1.016	2.829	1.2707	2.785	2.784449	1.813
M5	1.2079	2.9165	1.1625	3.1405	0.9939	2.9022	2.701505	1.978
L2	1.2281	3.4624	1.2707	3.6569	1.1532	3.4931	2.877863	2.3862
L3	1.4273	3.4616	1.4611	3.6513	1.3386	3.5001	2.499008	2.1902
L4	1.4022	3.2897	1.3954	3.6903	1.3108	3.4112	2.644618	2.2949
L5	1.2106	2.6141	1.28	3.1873	1.1641	3.331	2.490078	1.9073

We have calculated variance and entropy for each region of different organisms by using different wavelets. In what follows, we will do the comparisons by observing the last two columns of the above tables.

B. Remark

To analyze the data in the above tables, we draw the wavelet variance and entropy in the following Bar Chart.

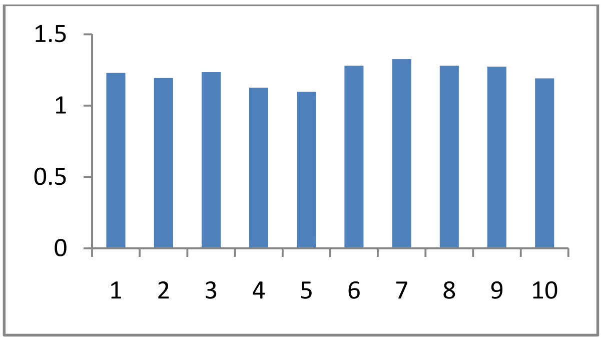

Fig. 3.1. Bar chart of the wavelet variance of the coding regions (1-5) and non-coding regions (6-10) 0f the Anolis carolinensis using coiflets. This shows some similarities within coding or noncoding regions and some distinctions between coding and noncoding regions.

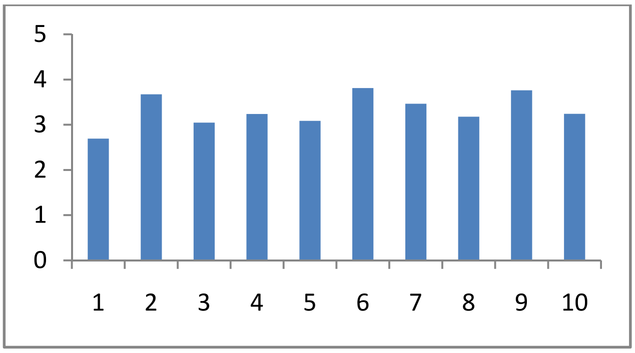

Fig. 3.2. Bar chart of the entropy of the coding regions (1-5) and non-coding regions (6-10) 0f the Anolis carolinensis using coiflets. To some extent, there appears some similarities within coding or noncoding regions and some distinctions between coding and noncoding regions.

The results from Tables 3.1, 3.2 and 3.3 show that there is a relationship between the wavelet variance and the entropy. The three DNA sequences analyzed show that the ratio and the difference between the wavelet variance and the entropy is close to a constant for a given DNA sequence; from the cases analyzed the ratio is close to 3 while the difference is close to 2.

The bar charts for the wavelet variance and entropy in figures 3.1 and 3.2 respectively show that the noncoding regions have relatively higher values than the coding regions. In this paper, the discrete wavelet transform is used to decompose DNA sequences with respect to a set of basis functions, which is associated with a particular scale. The properties of a DNA sequence at each scale are obtained by the wavelet variance. On the other hand, wavelet entropy gives useful criterion for analyzing and comparing probability distribution that provides a measure of the information of a given DNA sequence. More precisely, the wavelet entropy basically appears as a measure of the degree of order or disorder of the DNA sequence, so it can provide useful information about the underlying dynamical property associated with the DNA sequence. Several further studies can be obtained such as mutual information-based gene or feature selection method where features are wavelet-based; the bootstrap techniques employed to obtain an accurate estimate of the mutual information and other new methods to analyze data. [8] Entropy and variance are essentially used to measure uncertainty, information and risk and other needs. Variance has been prominent but the use of entropy is growing rapidly. This paper has reflected the relationship between them to some extent.

IV. ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Felix Famoye and anonymous referees' helpful comments.

REFERENCES

[1] C. Burge, and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of molecular biology*, vol. 26, no. 1, pp. 78-94, 1997.

[2] J. M. Butler, "Forensic DNA Typing: Biology," *Technology and Genetics of STR Marker. Elsevier Academic Pres*s, 2005.

[3] G. A. Bradshaw and T. A. Spies, "Characterizing canopy gap structure in forests using wavelet analysis," *Journal of Ecology*, pp. 205-215, 1992.

[4] E. B. Lin and P. C. Liu, *A discrete Wavelet Analysis of freak Waves in the ocean. Journal of Applied Mathematics*, 2004, No 5, 379-394.

[5] S. A. Mallat, *Wavelet Tour of Signal Processing*, Academic Press, 1998.

[6] S. Tiwari, et al, *Prediction of probable genes by Fourier analysis of genomic sequences*, pp. 263-270. 1997.

[7] M. Q. Zhang, "Identification of protein coding regions in the human genome by quadratic discriminant analysis," in *Proc. Natl Acad. Sci*, pp. 565–568. 1997.

[8] X. Zhou, X. Wang, and E. R. Dougherty, "Nonlinear probit gene classification using mutual information and wavelet-based feature selection," *Journal of Biological Systems*, no. 3, pp. 371-386. 2004.