2006

Application of Spectral Analysis to DNA Sequences

Lan Zhao

Report Number:
06-003

APPLICATION OF SPECTRAL ANALYSIS TO DNA SEQUENCES

Lan Zhao

CSD TR #06-003
January 2006

Application of Spectral Analysis to DNA Sequences *†

Lan Zhao
Rosen Center for Advanced Computing
Purdue University, West Lafayette, IN 47907

Abstract

In this article, the spectral analysis of categorical time series is applied to the study of DNA sequences. By means of the study of the spectral envelope for a homologous DNA segment from five different primate species, some global criteria are proposed to measure their similarity, and some information concerning their evolutionary origin is obtained.

1 Introduction

A time series $X(t)$ or X_t is a sequence of data varying with time t, such that X_t is a random variable for each t and there exists correlativity between X_t and X_s for each pair (t, s). To represent one of its paths in a coordinate plane, we can plot a polygonal line by linking points (t, X_t) in order. In may situations, we are interested in the data which may vary with the location of points in space. In these cases, the sequence of data is also called as time series. The analysis and inference for time series has been extensively applied to may important fields such as the activity of sunspots, the movement of terrestrial poles, electroencephalogram(EEG), the operation of electrical machinery and so on, involving a great number of aspects of nature, industry, agriculture, economic, ecology and medicine. However it is only in recent years has this method been devoted to study of DNA sequences.

A DNA molecule consists of a long string of linked nucleotides of four kinds. Because a pair of strands are complementary, it is sufficient to represent a DNA molecule by a sequence of bases on a single strand, typically written in its 5' to 3' direction. Thus a strand of DNA can be represented as a sequence of letters, termed base pairs (BP), from the finite alphabet A, C, G, T. The genetic information is

* *AMS 2000 subject classifications.* Primary 62P10, 62M15.
† *Key words and phrases:* Categorical time series, DNA sequences, spectral enevelope.

believed to be stored in the particular order of the four kinds of nucleotides. From the view of statistics, such a sequence is a categorical time series.

One approach for exploring the nature of a categorical process is to assign numerical values to each of the possible states or categories followed by a spectral analysis of the resulting discrete-valued time series, which is the first method applied to DNA sequence analysis in this article. This statistical methodology is computationally simple, quick, as well as it could bring to light some internal properties of the sequence. But the resulting spectrum will inevitably depend on the value assigned to the state space. To eliminate such an influence of subjective scaling, we adopted here the spectral envelope approach introduced by Stoffer, Tyler and McDougall (1993), which could reflect the objective nature of a sequence to the greatest degree.

In the literature some criteria are proposed to measure and compare the homology of DNA sequences. In some papers, it is assumed that two species 1 and 2 to be compared are evolved from one ancestor 0, and the evolutionary processes from 0 to 1 and from 0 to 2 are both Markov processes. Here we have assigned A, C, G, T the values 1, 2, 3 and 4 respectively. Let (X_k, Y_k) denote the assigned values to the bases that occur at the kth position in species 1 and 2 respectively $(1 \leq k \leq n)$. Let P_{12} be the 4×4 transition matrix having (i, j) element $P_r(Y_k = j | X_k = i)$, which is independent of the position k. Barry and Hartigan (1987) proposed a distance of homology between DNA sequences of species 1 and 2 as follows:

$$d(1,2) = -\frac{1}{4} \log[\det(P_{12})], \tag{1}$$

where the measure $d(1,2)$ is well defined if $\det(P_{12}) > 0$, and we set $d(1,2) = \infty$ if $\det(P_{12}) \leq 0$. The authors also suggested to use $\hat{P}_{12} = (R_{ij})$ as an estimate of P_{12}, where R_{ij} is the proportion of times, given i in sequence 1, that j is obtained in sequence 2, i.e.,

$$R_{ij} = \#\{1 \leq k \leq n, X_k = i, Y_k = j\} / \#\{1 \leq k \leq n, X_k = i\}. \tag{2}$$

Of course the distance defined by (1) has many virtues. However, it also has obvious defects. In general, $d(1,2) \neq d(2,1)$, i.e., the distance d is not symmetric for species 1 and 2. In addition, even if $d(1,2)$ is very small, it does not follow that there must be great homology between the two DNA sequences. One artificial example can be given by putting

$$P_{12} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Here $\det(P_{12}) = 1$ and $d(1,2) = 0$. However, it is obvious that the two DNA sequences are completely different.

In this article, we proposed some new measures from an overall point of view on the ground of the spectrum analysis approach and used them in the study of homology among five DNA sequences. The measures we proposed here are still to be improved and perfected through further research, but we believe it is surely a feasible research line to take advantage of spectrum analysis methodology and to obtain some reasonable measures.

2 Some Conceptions

A. Stationary stochastic process. As previously pointed out, a stochastic sequence is a sequence of random variables $\{X(t)\}$ or $\{X_t\}$, varying with time t, with t being integer. The expectation function and covariance function of a stochastic sequence are defined by $EX(t)$ and

$$\gamma(t, t+h) = E((X(t) - E(X(t))(X(t+h) - EX(t+h))) \tag{3}$$

respectively. If $EX(t)$=constant, $\gamma(t, t+h) = \gamma(h)$, we call $\{X(t)\}$ a stationary sequence in wide sense or a stationary sequence for short. For simplicity and the application of spectrum analysis, we assumed the sequences studied are stationary.

B. Power spectrum and spectral density. Let $\tilde{X}(t)$ be a real function defined on $(-\infty, \infty)$ with period 2π, square-integral on $[-\pi, \pi]$. Then it has the following Fourier expansion:

$$\tilde{X}(t) = \sum_{n=-\infty}^{\infty} A(n)e^{int} = \frac{1}{2}A_1(0) + \sum_{n=1}^{\infty}[A_1(n)\cos nt + A_2(n)\sin nt] \tag{4}$$

where

$$\begin{cases} A_1(n) = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{X}\cos ntdt \\ A_2(n) = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{X}\sin ntdt, \end{cases}$$

that is, $\tilde{X}(t)$ is decomposed into a sum of a constant term $A_1(0)/2$ (the horizontal location of the function) and the waves of the form $\cos nt$ and $\sin nt$. The wave with $n = \pm 1$ and angular frequency $\omega = 1$ is called fundamental wave, and the rest waves are called homophonic waves. In this way, for each function with period 2π, there is a sequence of complex numbers $A(n)$ such that $|A(n)|$ is the amplitude of the corresponding fundamental or homophonic wave, and $\arg A(n) = -\arctan(A_2(n)/A_1(n))$ is the argument. The power of $\tilde{X}(t)$ is given by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{X}(t)^2 dt = \sum_{n=-\infty}^{+\infty} |A(n)|^2 = \frac{1}{2}\left[\frac{1}{2}A_1(0)^2 + \sum_{n=1}^{+\infty}(A_1(n)^2 + A_2(n)^2)\right], \tag{5}$$

$n = 0, 1, 2, \ldots$ are the angular frequencies of the relevant waves. We thus obtain the power spectrum of $\tilde{X}(t)$. Such a power spectrum is called discrete spectrum.

In general, for a stationary sequence, we define $F(\omega)$ as the sum of the power corresponding to those angular frequencies not exceeding ω, $-\pi \leq \omega \leq \pi$, and call $F(\omega)$ the power spectrum distribution function of this stationary sequence. In many situations, $F(\omega)$ has a density function $f(\omega)$, the spectral density, and $f(\omega)$ can be decomposed into the following form:

$$f(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{+\infty} \gamma(t) e^{-i\omega t}, \quad -\pi \leq \omega \leq \pi. \tag{6}$$

3 Analysis of Frequency Spectrum

Let X_t, $t = 0, \pm 1, \pm 2 \ldots$ be a categorical-valued time series with finite state space $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$. For DNA sequence we will typically have $k = 4, \Delta = \{A, C, G, T\}$. (Of course we can also have $k = 16, \Delta = \{AA, AC, AG, AT \ldots\}$, and so forth). It is convenient and reasonable to assume that X_t is a stationary sequence and the possibility of X_t to be each state among A, C, G, T is p_1, p_2, p_3, p_4 respectively. Now to A, C, G, T we assign the values $\beta_1, \beta_2, \beta_3, \beta_4$ respectively, that is, we put

$$X_t(\beta) = \begin{cases} \beta_1, & if \quad X_t = A \\ \beta_2, & if \quad X_t = C \\ \beta_3, & if \quad X_t = G \\ \beta_4, & if \quad X_t = T, \end{cases}$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$ is called the scaling. Obviously the resulting spectrum depends on the scaling β. According to (3), the covariance function of $\{X_t(\beta)\}$ is

$$\gamma_\beta(h) = E(X_t(\beta) - EX_t(\beta))(X_{t+h}(\beta) - EX_{t+h}(\beta)) \tag{7}$$

and the spectral density is

$$f_\beta(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_\beta(h) e^{-i\omega h}, \quad -\pi \leq \omega \leq \pi. \tag{8}$$

To work out the spectral density curve $f_\beta(\omega) \sim \omega$, we must give an estimate of $f_\beta(\omega)$. (the notation $\hat{}$ denote an estimate). Because $\{X_t\}$ is a stationary process, $EX_{t+h}(\beta) = EX_t(\beta) = \beta_1 p_1 + \beta_2 p_2 + \beta_3 p_3 + \beta_4 p_4$. For a given DNA sequence, p_1 can be estimated by the ratio of the number of positions in which the nucleotide is A to the number of total positions, and the rest may be deduced by analogy, i.e.,

$$\hat{p}_1 = n_A/n, \quad \hat{p}_2 = n_C/n, \quad \hat{p}_3 = n_G/n, \quad \hat{p}_4 = n_T/n.$$

So for given β, we can estimate $EX_t(\beta)$ by

$$\bar{X}(\beta) = \sum_{i=1}^{4} \beta_i \hat{p}_i \tag{9}$$

and estimate $\gamma_\beta(h)$ by

$$\hat{\gamma}_\beta(h) = \frac{1}{n-h}\sum_{j=1}^{n-h}(X_j(\beta) - \bar{X}(\beta))(X_{j+h}(\beta) - \bar{X}(\beta)). \tag{10}$$

Especially,

$$\hat{\gamma}_\beta(0) = \frac{1}{h}\sum_{j=1}^{n}(X_j(\beta) - \bar{X}(\beta))^2. \tag{11}$$

There are many ways to estimate the spectral density, see for example Hannan(1970), p.273-288. The estimation of spectral density we adopted here is given by

$$\hat{f}_\beta(\omega_j) = \frac{1}{2\pi}\sum_{h=-r}^{r}\hat{\gamma}_\beta(h)(1-|h|/n)e^{-i\omega_j h}, \quad \omega_j = 2j\pi/n. \tag{12}$$

Here r is defined as window width, which satisfy $r \to \infty$ and $n/r \to \infty$, as $n \to \infty$, i.e., $r \gg 1$ and $n \gg r$ (in this article $n = 896, r = 20$). Thus the spectral density estimate for DNA sequence assigned given values could be obtained.

Time series is a scientific description for stochastic sequence. So the power spectrum could reflect some internal laws and relation between DNA sequence to a certain degree. To compare the similarity or homology of two DNA sequence quantitatively, we proposed here a global measure of distance as follows:

$$\int_{-\pi}^{\pi}|\hat{f}_1(\omega) - \hat{f}_2(\omega)|d\omega. \tag{13}$$

It means to integrate over the sections not overlapping between the two spectral densities, so it is to measure the overall difference form the point of energy distribution in power spectrum.

The greatest defection of this method lies in its dependence upon the assigned value β.

4 The Spectral Envelope

To remove the effect of assigned value on spectral analysis, it is proposed to study DNA sequence by using spectral envelope approach, which has been studied by Stoffer et al.(1993). Its main principle is to represent the categories in terms of a set of $k \times 1$ vectors and then we obtain a new k-dimensional stationary time series $\{Y_t\}$. In the case of DNA sequence,

$$Y_t = \begin{cases} (\,1,\,0,\,0,\,0\,)' & if \quad X_t = A \\ (\,0,\,1,\,0,\,0\,)' & if \quad X_t = C \\ (\,0,\,0,\,1,\,0\,)' & if \quad X_t = G \\ (\,0,\,0,\,0,\,1\,)' & if \quad X_t = T. \end{cases}$$

The time series $\{X_t(\beta)\}$ can be obtained from the $\{Y_t\}$ by $X_t(\beta) = \beta' Y_t$. And for Y_t, the expectation function EY_t, spectral density $g(\omega)$ and covariance function $\delta(h)$ are

$$EY_t = (p_1, p_2, p_3, p_4)', \tag{14}$$

$$g(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \delta(h) e^{-i\omega h}, \quad -\pi \leq \omega \leq \pi, \tag{15}$$

$$\delta(h) = E(Y_t - EY_t)(Y_{t+h} - EY_{t+h})' = \delta(-h)'. \tag{16}$$

Specially, covariance matrix

$$V = \delta(0) = E(Y_t Y_t') - EY_t EY_t'. \tag{17}$$

By the definition it is easy to see $g(\omega)$ is a 4×4 Hermite matrix, i.e. $g(\omega) = g^*(\omega)$ ($*$ represents conjugate transpose of a matrix). Since $X_t(\beta) = \beta' Y_t$, the spectral density of X_t is $f_\beta(\omega) = \beta' g(\omega) \beta = \beta' g^{Re}(\omega) \beta$, $g^{Re}(\omega)$ represent the real part of $g(\omega)$. Let $\beta = ae$, a is any real number, $e = (1, 1, 1, 1)'$ is the 4×4 vector of ones, then there is only one state a for $X_t(\beta)$. By (7), $\gamma_\beta(h) = 0$, and $f_\beta(\omega) = 0$, that is

$$e' g(\omega) e = e' g^{Re}(\omega) e = 0. \tag{18}$$

So $g(\omega)$ is not of full rank. Similarly from (17), we can get

$$e' V e = 0. \tag{19}$$

The rank of V is only 3, it is not of full rank too. To eliminate the influence of β on spectral analysis, define $\lambda(\omega)$ as follows:

$$\lambda(\omega) = \sup_{\beta \neq ae} \frac{\beta' g^{Re}(\omega) \beta}{\beta' V \beta} \tag{20}$$

where a represents any real number. Obviously $\lambda(\omega) = \lambda(-\omega)$. Please note $X_t(\beta) = \beta' Y_t$ means $f_\beta(\omega) = \beta' g^{Re}(\omega) \beta$, so the meaning of $\lambda(\omega)$ is to find β to maximize the spectral density after standardization. so we could remove the influence of the scaling and the spectral envelope of a stationary categorical time series $X_t (t = 0, \pm 1, \pm 2, \ldots)$ is defined to be $\lambda(\omega) (-\pi \leq \omega \leq \pi)$. Apparently $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. for any particular scaled process $\{X_t(\beta)\}$, $\lambda(\omega) d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies $\omega d\omega$. Spectral envelope can be used to study the nature of a categorical time series more objectively.

Since both $g^{Re}(\omega)$ and V are not of full rank, we use $\tilde{g}^{Re}(\omega)$ and $\tilde{V}$ to represent the upper 3×3 block of $g^{Re}(\omega)$ and V respectively. According to Stoffer et al.(1993),

it can be proved that $\lambda(\omega)$ is the largest eigenvalue of $\tilde{g}^{Re}(\omega)$ relative to $\tilde{V}$, i.e. the largest real root of equation

$$\det(\tilde{g}^{Re}(\omega) - \lambda\tilde{V}) = 0. \tag{21}$$

The corresponding eigenvector $b(\omega)$ is a 3×1 vector satisfying

$$\tilde{g}^{Re}(\omega)b(\omega) = \lambda(\omega)\tilde{V}b(\omega). \tag{22}$$

For the use in the sequel, we put

$$\mathcal{B}(\omega) = \begin{pmatrix} b(\omega) \\ 0 \end{pmatrix},$$

and define ω^*, the frequency with energy focused by

$$\lambda(\omega^*) = \sup\{\lambda(\omega), 0 \leq \omega \leq \pi\}.$$

The estimates of spectral density and covariance function we adopted here are

$$\hat{g}(\omega_j) = \frac{1}{2\pi}\sum_{h=-r}^{r}\hat{\delta}(h)(1 - \frac{|h|}{n})e^{-i\omega_j h}, \quad \omega_j = \frac{2\pi j}{n}, \tag{23}$$

$$\hat{\delta}(h) = \frac{1}{n-h}\sum_{j=1}^{n-h}(Y_j - \bar{Y})(Y_{j+h} - \bar{Y})' = \hat{\delta}(-h)', \tag{24}$$

$$\hat{V} = \hat{\delta}(0), \tag{25}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Using (21), we can obtain the spectral envelope $\hat{\lambda}(\omega) \sim \omega$ as well as the related estimates of ω^*, $\lambda(\omega^*)$ and $\mathcal{B}(\omega^*)$. Similar to (13), we define

$$\int_{-\pi}^{\pi} |\hat{\lambda}_1(\omega) - \hat{\lambda}_2(\omega)| d\omega, \tag{26}$$

as a global measure to compare the similarity of spectral envelope of two DNA sequences and the homology between them. In addition, another measure is proposed here according to spectrum analysis approach:

$$\sup_{||\beta||=1} \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}'(g_1^{Re}(\omega) - g_2^{Re}(\omega))^2 \mathcal{B} d\omega. \tag{27}$$

5 DNA Sequence Analysis

Here we will study a homologous DNA segment from five different primate species, and compare the distance of their homology quantitatively using the spectrum analysis approach of time series. The data adopted here are a segment of

mitochondrial DNA from human(H), chimpanzee(C), gorilla(G), orangutan(O), and gibbon(GB). This segment is 896 bp in length. It contains the genes for three transfer RNAs and parts of two proteins. The nucleotide sequences for this mtDNA segment appear in Fig 1, which is from Brown et al.(1982) Fig 3, and the original caption under the sequences is as follows:

Sequences of 896 bp fragments of primate mitochondrial DNAs. Common and pygmy chimpanzee are considered as on entity(see text). The published (Anderson et al. 1981)human mtDNA sequence is shown in the uppermost line in small capital letters. Differences from the published human sequence are indicated with large capital letters. Since our human mtDNA differs in sequence from the published one only at positions 40 and 569, the human is listed only once, with the changes observed here indicated as just described. Sequence differences exist at 284 positions among the primates studied. At 177 positions the changes are unique - i.e., there is a different base in only one species. The distribution of these unique changes is human, 17; chimpanzee, 19; gorilla, 22; orangutan, 55; gibbon, 64. At the remaining 107 sites the changes are non-unique -i.e., in at least two lineages there is a base which is different from that existing in the remaining lineages. The only deletion was at position 560 in the orangutan, and there were no additions. The 5 genes present encompass the following nucleotides: Protein 4, 1-458; histdine tRNA, 459-527; serine tRNA, 528-586; leucine tRNA, 587-657; Protein 5, 658-896.

It is thought that the mt genome of animals is suitable material to examine the tempo and mode of molecular evolution. But to choose a series of species for comparison there are two points which should be noted: Their divergence times should

- be different.
- lines within a time range not too long on the other hand.

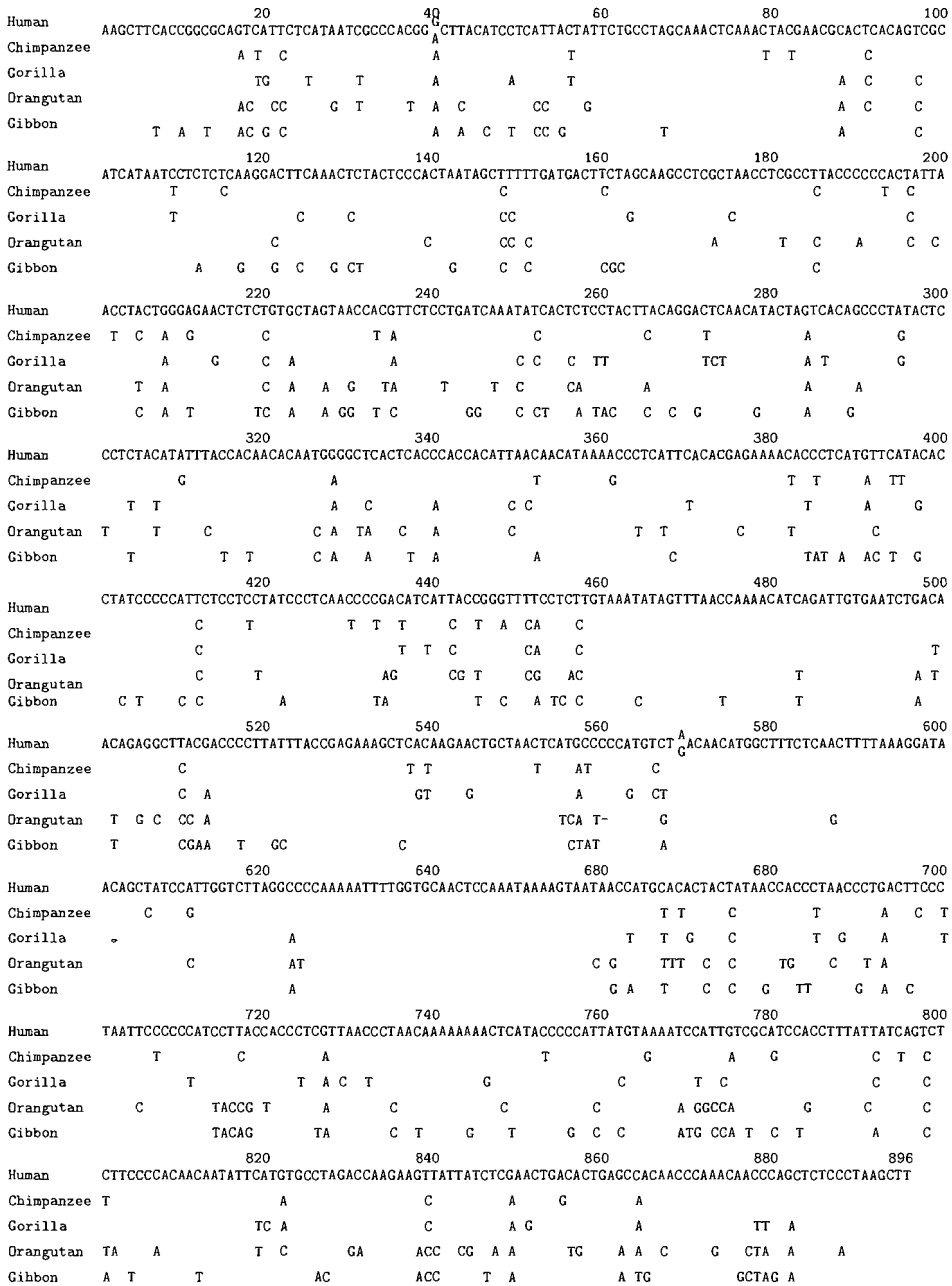

Fig 1: DNA Sequence Map

Because according to Brown et al.(1979), the readily-substituted positions in the mtRNA will have become "saturated" in a time range which is too long, i.e., although the average mutation rate is not changed, more and more mutations will be multiple substitutions at the same nucleotide position which causes the dependence

of difference between DNA sequences upon their divergence time distance to become nonlinear.

The divergence time distances among the five species we studied here are within 10 million year which satisfies above requirements.

5.1

From the primary analysis of the five groups of data, we can find:

1. The five DNA sequences differ from one another at 284 positions. Only the difference at position 560 in the orangutan is caused by deletion while all the others are caused by base substitution.

2. To count up the number of positions at which there exists difference between any two DNA sequences, we obtain the following table in which the number of sequence differences between any two mtDNAs appears in the upper right-hand section and the percentage of sequence difference is given in the lower left-hand section.

	H	C	G	O	Gb
H	-	79	92	144	162
C	8.8	-	95	154	169
G	10.3	10.6	-	150	169
O	16.1	17.2	16.7	-	169
Gb	18.1	18.9	18.9	18.9	-

3. At all the positions which are different between two mtDNAs, the ratio of transition to transversion is high. This is probably related to the selective pressure in evolutionary process and the structure interrelation of bases.

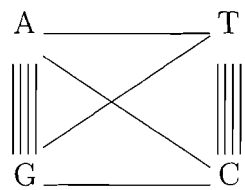

Fig 2: Structure interrelation

4. The mutation rate of mtDNA far exceeds that of nuclear DNA: the mtDNAs coding for proteins evolve 10 times faster than single copy nuclear DNA and the tRNA genes evolve 100 times faster in mitochondrial than in nuclear. There are several points which may be related to this high mutation rate:

(a) the influence of greater exposure to oxidative environment in mitochondrial.

(b) a more error-prone system of replication.

(c) less efficient editing or repair functions.

(d) a higher rate of turnover.

(e) the nuclear tRNAs engage in numerous protein synthesis while the mt genome only code for no more than 20 peptides. So both the functional constraints and selective pressure are smaller for mt tRNA than for nuclear tRNA. Also the size of mt tRNA genes is small and the degree of base modification is low. All of these are consistent with the specificity of mt DNA's genetic code.

5.2

Here we proceed to make the spectral analysis of the five DNA sequences:

1. According to the methods presented in section 3, we assign a value for each of the four states, for example: A=1, C=2, G=3, T=4, then we can plot the DNA power spectrum of five species (Fig 3). The vertical axis represents the spectral density and the horizontal axis represents angle frequency ω. Because $f(-\omega) = f(\omega)$, we only plot the section of $0 \leq \omega \leq \pi$.

   (a) Fig 3 shows the major peaks in five plots at approximately $\omega = 2\pi \times 300/896 \doteq 2\pi/3$ unanimously, which is $T = 2\pi/\omega \doteq 3bp/cycle$ when converted into period. That means the strongest periodicity with concentrated energy is 3bp per cycle, which is very consistent with the triplet characteristic of genetic code.

   (b) Using (13) to measure the homology between the four species C, G, O, Gb and H respectively, the following relationship can be deduced: Distance with H: Gb(0.20)>O(0.15)>C(0.12)>G(0.09). As we have pointed out previously, this distance reflects the difference of energy distribution in a global sense.

2. Consider the following sequence: $\{a, b, c, b, a, b, c, b \ldots\}$, clearly there is not just one cyclic pattern here. For example, the sequence can be viewed as $\{\bar{b}, b, \bar{b}, b\}$ with period 2, where $\bar{b}$ means "not b", however it can also be viewed as $\{\underline{a, b, c, b}, \underline{a, b, c, b}\}$ with period 4. If we put $a = c = 0, b = 1$, the former periodity is emphasized, while if we put $a = 0, b = 1, c = 2$, the latter is emphasized. This example shows the inevitable limitation of subjective scaling.

The spectral envelope approach can eliminate such an affect to the greatest degree, putting forward the essential and objective laws.

The spectral envelope is plotted in Fig 4, in which solid lines represent spectral envelope and dotted lines represent standardized power spectrum density with scaling A=1, C=2, G=3, T=4.

(a) It can be seen that the power peaks also appear at 3bp per cycle, consistent with the triplet characteristic of genetic code.

(b) Computing the difference between H and the other species by using (26) and (27), we obtain consistent results. Distances with H are: O(0.19 and 2.0×10^{-4}), Gb(0.17 and 1.7×10^{-4}), G(0.11 and 1.2×10^{-4}), C(0.09 and 0.9×10^{-4})). So the order of distance to H is:

$$O > Gb \gg G > C.$$

But it is different from the order in 1 (Gb>O>C>G), and in 1 the distances between H and each of the other four species are distributed with nearly equal internal, while in the latter result G and C is more closer to H than Gb and O. According to the earlier study such as that based on cleavage map comparisons of the mitochondrial genome (Ferris et al. 1981), the possible evolutionary relationships are as follows:

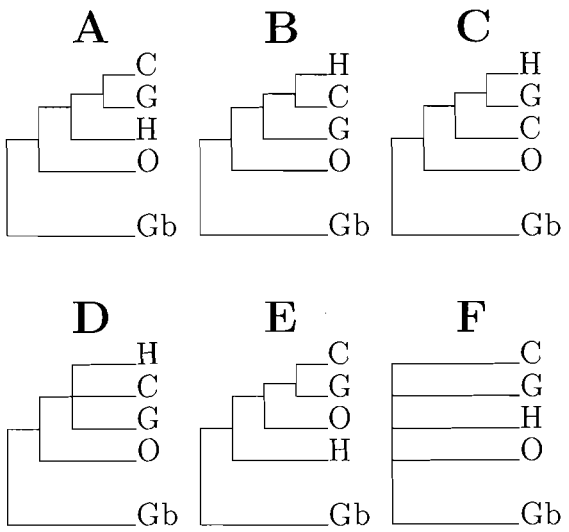

Fig 5: Evolutionary relation plot

In addition, research using different methods appears to rule out tree E and suggests that trees C and D are far less likely than trees A and B. such a general trend is confirmed once again by our results, i.e., the relationship between H and C or G is much more closer than that between H and O or Gb. However, the results obtained from spectral envelope appears to support the following model more accurately:

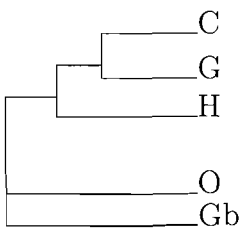

Fig 6: A possible evolutionary relation plot

Of course the final statement can be made only when more research has been done using more data and from more aspects.

Summary: Spectral envelope method is easily understood, parsimonious and quantitative tool for DNA sequence study with a minimal loss of useful information. There is extensive application space in Biology for such an approach, especially in the aspects of molecular evolution, species formation, gene structure and function etc.. Although there is still many things not clear at present, we will surely be able to gain more and more information from spectral analysis of DNA sequences through further and deeper research.

References

[1] Anderson,S., Bankier,A.R., Barrell,B.G., de Bruijn, M.H.L, Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J.H., Staden.R., Young,I.G. (1981). Sequence and organization of the human mitochondrial genome. Nature. **290**, 457-465.

[2] Barry,D., Hartigan,J.A. (1987). Asynchronous distance between homologous DNA sequences. Biometrics, **43**, 261-276.

[3] Brown,W.M., Prager,E.M., Wang,A., Wilson,A.C. (1982). Mitochondrial DNA sequences of Primates: Tempo and mode of evolution. J. Mol. Evol., **18**, 225-239.

[4] Chen Zhaoguo (1988). Time Series and Its Spectral Analysis. Science Press, (in Chinese), Beijing.

[5] Ferris,S.D., Wilson A.C., Brown W.M (1981). Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Ntal Acad. Sci. USA. **78**, 2433-2436.

[6] Hannan,E.J. (1970). Multiple Time Series. John Wiley, New York.

[7] Stoffer,D.S., Tyler,D.E., McDougall,A.J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. Biometrika, **80**, 611-622.

[8] Stoffer,D.S., Tyler,D.E., McDougall,A.J. (1993). Spectral analysis of DNA sequences. Proceedings of the ISI 49th Session, Firenze. 345-361.

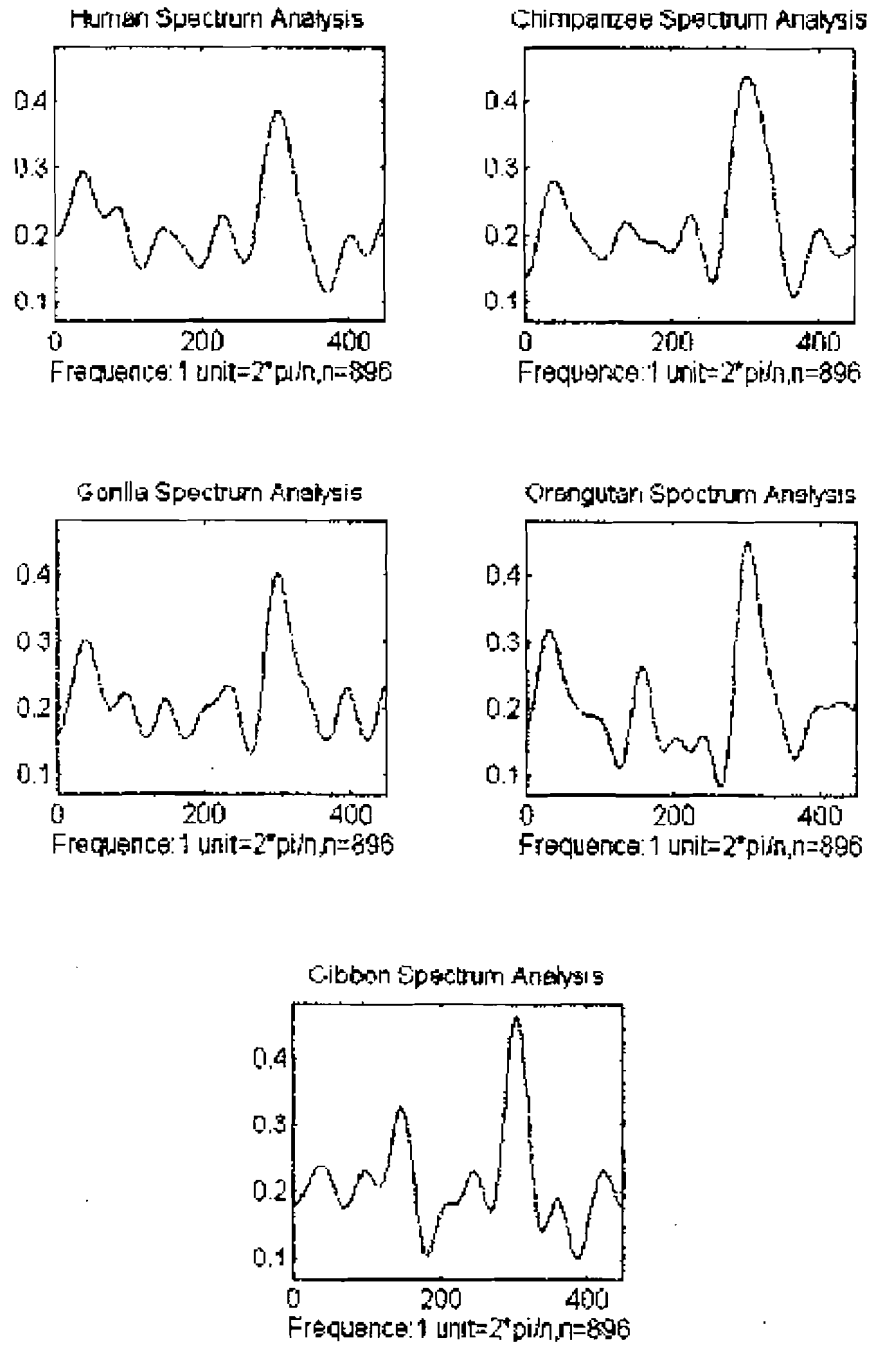

Figure 1: Nonstandardized power spectrum densities with A=1, C=2, G=3, T=4

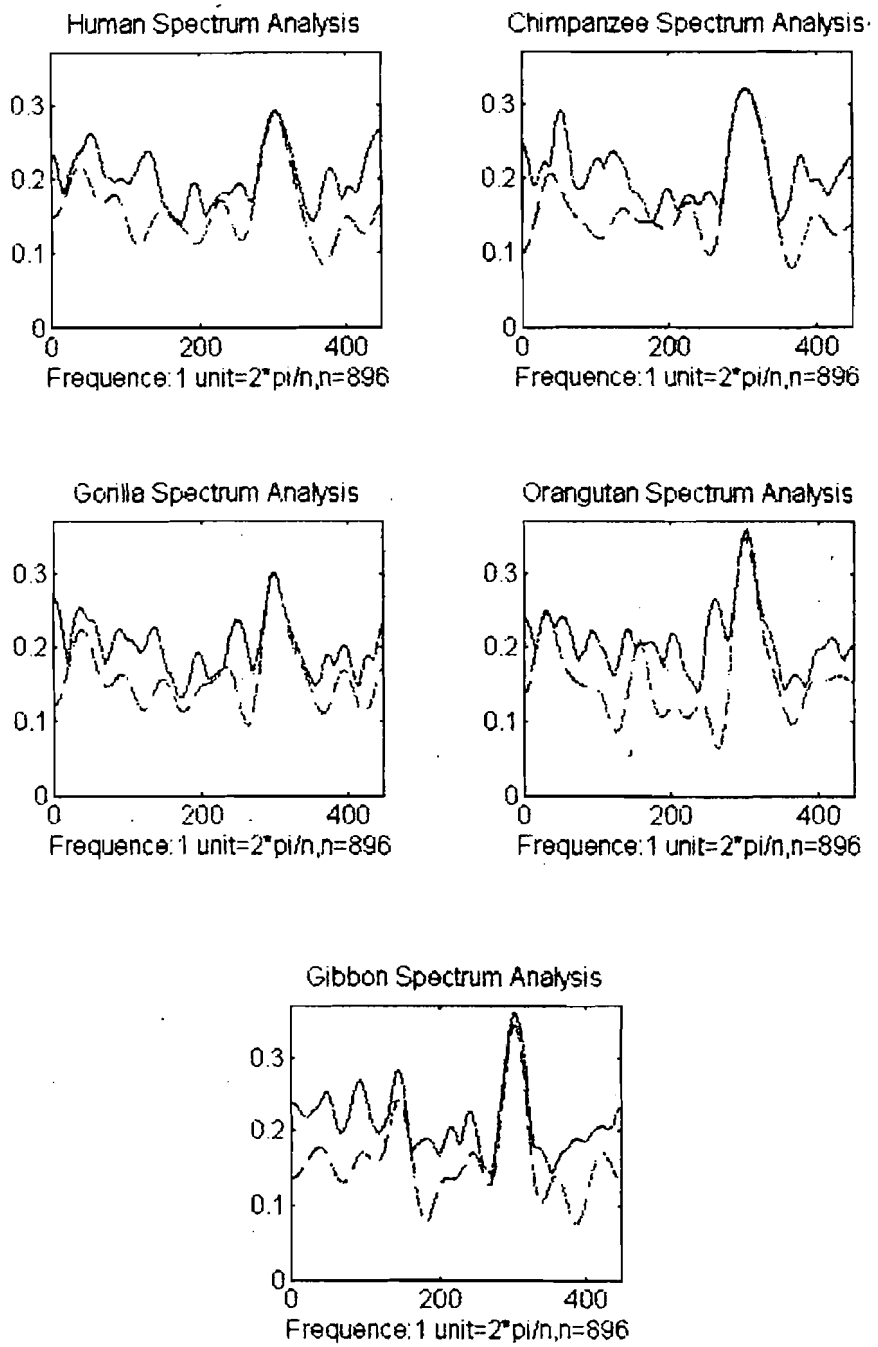

Figure 2: Spectral envolopes and standardized power spectrum densities (Solid lines represent spectral envolopes and dotted lines represent standardized power spectrum densities with scaling A=1, C=2, G=3, T=4.)