BMC Bioinformatics

BioMed Central

Proceedings

Time-dependent ARMA modeling of genomic sequences

Jerzy S Zielinski*[1], Nidhal Bouaynaya*[1], Dan Schonfeld[2] and William O'Neill[3]

Address: [1]Department of Systems Engineering, University of Arkansas at Little Rock, Little Rock, AR, USA, [2]Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL, USA and [3]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA

Email: Jerzy S Zielinski* - jszielinski@ualr.edu; Nidhal Bouaynaya* - nxbouaynaya@ualr.edu; Dan Schonfeld - dans@uic.edu; William O'Neill - woneill@uic.edu

* Corresponding authors

from Fifth Annual MCBIOS Conference. Systems Biology: Bridging the Omics
Oklahoma City, OK, USA. 23–24 February 2008

Published: 12 August 2008

BMC Bioinformatics 2008, **9**(Suppl 9):S14 doi:10.1186/1471-2105-9-S9-S14

This article is available from: http://www.biomedcentral.com/1471-2105/9/S9/S14

Abstract

Background: Over the past decade, many investigators have used sophisticated time series tools for the analysis of genomic sequences. Specifically, the correlation of the nucleotide chain has been studied by examining the properties of the power spectrum. The main limitation of the power spectrum is that it is restricted to stationary time series. However, it has been observed over the past decade that genomic sequences exhibit non-stationary statistical behavior. Standard statistical tests have been used to verify that the genomic sequences are indeed not stationary. More recent analysis of genomic data has relied on time-varying power spectral methods to capture the statistical characteristics of genomic sequences. Techniques such as the evolutionary spectrum and evolutionary periodogram have been successful in extracting the time-varying correlation structure. The main difficulty in using time-varying spectral methods is that they are extremely unstable. Large deviations in the correlation structure results from very minor perturbations in the genomic data and experimental procedure. A fundamental new approach is needed in order to provide a stable platform for the non-stationary statistical analysis of genomic sequences.

Results: In this paper, we propose to model non-stationary genomic sequences by a time-dependent autoregressive moving average (TD-ARMA) process. The model is based on a classical ARMA process whose coefficients are allowed to vary with time. A series expansion of the time-varying coefficients is used to form a generalized Yule-Walker-type system of equations. A recursive least-squares algorithm is subsequently used to estimate the time-dependent coefficients of the model. The non-stationary parameters estimated are used as a basis for statistical inference and biophysical interpretation of genomic data. In particular, we rely on the TD-ARMA model of genomic sequences to investigate the statistical properties and differentiate between coding and non-coding regions in the nucleotide chain. Specifically, we define a quantitative measure of randomness to assess how far a process deviates from white noise. Our simulation results on various gene sequences show that both the coding and non-coding regions are non-random.

However, coding sequences are "whiter" than non-coding sequences as attested by a higher index of randomness.

Conclusion: We demonstrate that the proposed TD-ARMA model can be used to provide a stable time series tool for the analysis of non-stationary genomic sequences. The estimated time-varying coefficients are used to define an index of randomness, in order to assess the statistical correlations in coding and non-coding DNA sequences. It turns out that the statistical differences between coding and non-coding sequences are more subtle than previously thought using stationary analysis tools: Both coding and non-coding sequences exhibit statistical correlations, with the coding regions being "whiter" than the non-coding regions. These results corroborate the evolutionary periodogram analysis of genomic sequences and revoke the stationary analysis' conclusion that coding DNA behaves like random sequences.

Background

Understanding the statistical properties of genomic sequences helps recreate the dynamical processes that led to the current DNA structure, and determine gene-related diseases like cancer and Alzheimer disease. For instance, based on the view that non-coding DNA exhibits long-range correlations [1-6], Li [7] proposed an expansion-modification model of gene evolution. The model incorporates the two basic features of DNA evolution: (i) sequence elongation due to gene duplication and (ii) mutations. It can be shown that the limiting sequence created by this dynamical process exhibits a long-range correlation structure, as attested by a $1/f^{\alpha}$spectrum, where the exponent α is a function of the probability of mutation. To understand the relationship between the DNA correlation structure and possible gene abberations, Dodin et al. [8] designed a simple correlation function intended to visualize the regular patterns encountered in DNA sequences. This function is used to revisit the intriguing question of triplet repeats with the aim of providing a visual estimate of the propensity of genes to be highly expressed and/or to lead to possible aberrant structures formed upon strand slippage.

Statistical analysis of genomic sequences has, however, relied, for a long time, on signal processing techniques for stationary signals (correlation and power spectrum) [2,4,9,10], and statistical tools for slowly-varying trends within stationary signals (Detrended Fluctuation Analysis or DFA) [1,5,6]. Stationarity can be argued as a valid assumption for time-series of short duration. However, such an assumption rapidly loses its credibility in the enormous databases maintained by various genome projects. Standard statistical tests (e.g., Priestley's test for stationarity) have been used to verify that genomic sequences are not stationary and the nature of their non-stationarity varies and is often more complex than a simple trend [11,12]. Subsequently, more recent analysis of genomic data [1] has relied on time-varying power spectral methods (the evolutionary spectrum and periodogram) to capture the statistical characteristics of genomic sequences [11,12]. The main difficulty in using time-varying spectral methods is that they are extremely unstable and very noisy. Typically, the power spectrum and the evolutionary spectrum are averaged over time in order to obtain smooth and less jittery curves. Moreover, as was pointed out in [13], the evolutionary spectrum is restricted to the class of oscillatory processes. A stochastic process, $X(t)$, is oscillatory if it has a representation of the form

$$X(t) = \int A(t, \lambda)e^{2i\pi\lambda t}\, dZ(\lambda), \qquad (1)$$

Where $Z(\lambda)$ is an orthogonal increment process, and the evolutionary power spectrum of the process is defined by $P(t, \lambda) = |A(t, \lambda)|^2$. This definition of the evolutionary power spectrum has the following disadvantages [13]:

(i) It is not uniquely defined for a given non-stationary process.

(ii) The estimation procedure for the evolutionary spectrum depends greatly on the nature of theamplitude function $A(t, \lambda)$, which is not known a priori.

(iii) An increase in the number of observations does not provide added information on the local behavior of the evolutionary spectrum, and thus does not improve estimation accuracy.

We propose to model non-stationary genomic sequences by a time-dependent autoregressive moving average (TD-ARMA) process. Cramer [14] showed that a non-stationary process still possesses a Wold decomposition in terms of its innovation and its generating system. However, the linear system generating the non-stationary signal, $x(t)$, when driven by the innovation, $w(t)$, is no longer shift-invariant; the parameters of the impulse response, h_u, of this system are time-dependent so that

$$x(t)=\sum_{u=0}^{\infty} h_u(t)w(t-u). \tag{2}$$

The existence of a time-varying ARMA representation of this process is ensured by two theorems due, independently, to Grenier [15] and Huang and Aggarwal [16]. The uniqueness of the TD-ARMA representation is obtained by constraining the ARMA model to be invertible, but this leads to conditions on the time-varying impulse response $\{h_u(t)\}$ and its inverse (namely to be absolutely summable at any time t), which cannot be easily expressed in terms of the time-dependent coefficients of the ARMA model. In this paper, we estimate the time-dependent coefficients of the general TD-ARMA model using mean-squares, least-squares and recursive least-squares algorithms. The mean-squares estimation leads to generalized Yule-Walker type equations [15]. Once the non-stationary parameters are estimated (as time series), we use them to provide a basis for statistical inference by defining an index of randomness, which quantitatively assesses how close the non-stationary signal is to white noise. Our simulation results on various gene sequences show that (i) both the coding and non-coding segments of a gene are not random, and (ii) the coding segments are "closer" to random sequences than non-coding segments. Our results support the view that both coding and non-coding sequences are not random [11,12,9,17-20], and revokes the stationary study that maintains that non-coding DNA sustains long-range correlations whereas coding DNA behaves like random sequences [1-3,5,6,10].

Methods

Numerical representation of genomic sequences

Converting the DNA sequence into a digital signal offers the opportunity to apply powerful signal processing methods for the handling and analysis of genomic information. This is, however, not an easy task as the analysis results might depend on the chosen map. Various numerical mappings have been adopted in the literature. To cite few, Peng et al. [1] construct a one-dimensional map of nucleotide sequences onto a walk, $u(i)$, which they termed "DNA walk". The DNA walk is defined by the rule that the walker steps up ($u(i) = +1$) if a pyrimidine resides at position i, and steps down ($u(i) = -1$) otherwise. Voss [9] represents a DNA sequence by four binary indicator sequences, which indicate the locations of the four nucleotides A, T, C and G. Berthelsen et al. [21] proposed a two-dimensional representation of DNA sequences, characterized by a Hausdorff dimension (also called fractal dimension) that is invariant under (i) complementarity, (ii) reflection symmetry, (iii) compatibility and (iv) substitution symmetry of AT and C↔G. The corresponding embedding assignment is given by A = (-1; 0), T = (1; 0), C = (0; -1) and G = (0; 1). In this paper, since we are interested in time-dependent ARMA modeling of one-dimensional non-stationary genomic sequences, we adopt the widely used "DNA walk" map proposed by Peng et al [1]. The DNA walk provides a nice graphical representation for each gene. For instance, Figure 1 shows the structure of the Human gene 276 located in chromosome 1, and its DNA walk is displayed in Fig. 2.

Time-dependent ARMA model

Grenier [22] showed that a discrete non-stationary signal, $\{x\,[n]\}$, can be represented by finite-order time-varying ARMA processes of the form

$$x[n]+\sum_{i=1}^{p} a_i[n-i]\,x[n-i]=w[n]+\sum_{i=1}^{q} b_i[n-i]\,w[n-i], \quad n=0,\cdots,N-1, \tag{3}$$

where N is the length of the sequence $x\,[n]$, $a_i\,[n]$ and $b_i\,[n]$ are the time-dependent model parameters, p and q are the model orders and $w\,[n]$ is a white noise process. Observe that the coefficients $a_i\,[n]$ and $b_i\,[n]$ appear with an argument $n - i$ depending on i. This is purely arbitrary since any time origin can be chosen, without restraining the generality of the model. We assume that the time-dependent

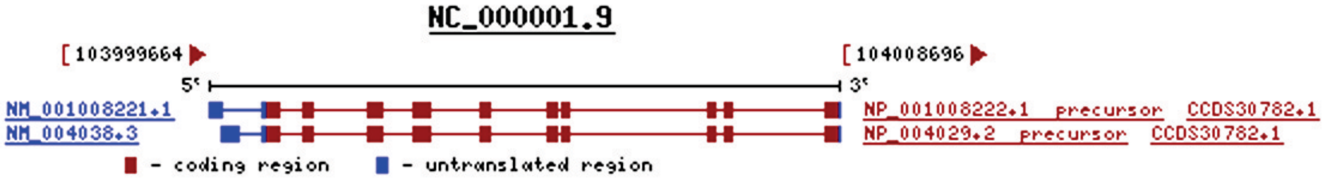

Figure 1

Gene Structure. Gene structure of the Human gene 276 located in chromosome 1: The boxes correspond to the exons (coding regions), and the lines between them represent the introns (non-coding regions). The total length of the gene is N = 8208 bases, including 1536 coding bases and 6672 non-coding bases.

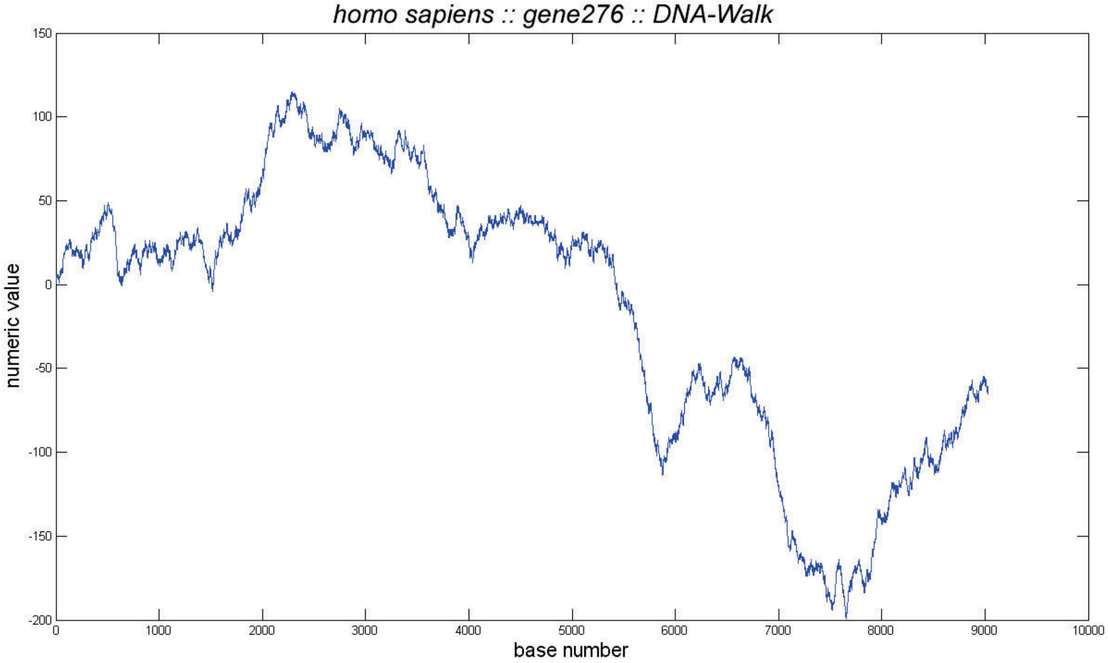

Figure 2
DNA Walk. DNA walk of the Human gene 276.

coefficients $a_i\,[n]$ and $b_i\,[n]$ can be expressed as linear combinations of some basis functions $\{f_k[n]\}_{k=0}^{m}$,

$$a_i[n] = \sum_{k=0}^{m} c_{i,k} f_k[n] \qquad (4)$$

$$b_i[n] = \sum_{k=0}^{m} d_{i,k} f_k[n] \qquad (5)$$

The advantage of the basis parametrization is clear from the fact that the identification of the time-dependent coefficients $a_i\,[n]$ and $b_i\,[n]$ reduces to the identification of the constant coefficients $\{c_{i,k}\}_{k=0}^{m}$ and $\{d_{i,k}\}_{k=0}^{m}$, and therefore the linear non-stationary problem reduces to a linear time-invariant problem. The basis functions $\{f_k[n]\}_{k=0}^{m}$ do not have to be limited to the standard choices of Legendre, Fourier, or the prolate spheroidal basis but can also take advantage of any prior information, such as the presence of a jump in the coefficients at a known instant [22].

Estimation of the time-dependent ARMA coefficients

From Eqs. (4) and (5), the unknown parameters of the TD-ARMA model are the weights of the linear combinations defining each time-varying parameter. The linearity is the key to the algorithms proposed in this paper. We will derive mean-squares, least-squares and recursive least-squares solutions to estimate the time-dependent coefficients $\{a_i[n]\}_{i=1}^{p}$ and $\{b_j[n]\}_{j=1}^{q}$.

Mean-squares estimation

Define the process

$$v[n] = x[n] + \sum_{i=1}^{p} a_i[n-i]\; x[n-i] = w[n] + \sum_{i=1}^{q} b_i[n-i]\; w[n-i], \quad n = 0,\cdots N-1, \qquad (6)$$

and define the vector

$$X\,[n] = [f_0[n]x[n],\, \cup,\, f_m\,[n]x[n]]^t, \qquad (7)$$

where the symbol ^{t} stands for the transpose of a vector or a matrix. It is possible to derive for this process orthogonality conditions similar to the stationary ARMA model conditions [23]. Observe that the process $v\,[n]$, defined in Eq. (6), is orthogonal to $[w[n - q - 1],\, w\,[n - q - 2],\, \cup]$; hence, it is orthogonal to $x\,[n - q - i]$ for all $i > 0$, and orthogonal to $X\,[n - q - i]$ for all $i > 0$ [22]. This orthogonality condition leads to a generalized Yule-Walker equation [22]

$$E\left(\begin{bmatrix} X[n-q-1] \\ \vdots \\ X[n-q-p] \end{bmatrix}[X[n-1]^t \cdots X[n-p]^t]^t\right)\theta = -E\left(\begin{bmatrix} X[n-q-1] \\ \vdots \\ X[n-q-p] \end{bmatrix}\cdot x[n]\right) \tag{8}$$

Although the process $x\,[n]$ is non-stationary, the stationarity and ergodicity of the process $w\,[n]$, together with the linearity of the model, allow us to replace in Eq. (8) the expectation by a summation. However, although consistent, the above estimator is often considered a poor one [22].

Least-squares estimation

Equations (4) and (5) can be written in vector format as follows

$$a_i\,[n] = \mathbf{f}^t\,[n]\,\mathbf{c}_i, \quad \text{and} \quad b_i\,[n] = \mathbf{f}^t\,[n]\,\mathbf{d}_i,$$

where

$$\mathbf{f}[n] = \begin{bmatrix} f_0[n] \\ \vdots \\ f_m[n] \end{bmatrix}, \quad \mathbf{c}_\mathrm{i} = \begin{bmatrix} c_{i,0} \\ \vdots \\ c_{i,m} \end{bmatrix}, \quad \mathbf{d}_\mathrm{i} = \begin{bmatrix} d_{i,0} \\ \vdots \\ d_{i,m} \end{bmatrix}.$$

Define

$$\mathbf{u}^t\,[n] = x\,[n]\,\mathbf{f}^t\,[n], \quad \text{and} \quad \mathbf{v}^t\,[n] = w\,[n]\,\mathbf{f}^t\,[n].$$

Then, we have

$$a_i[n-i]\;x[n-i] = \mathbf{u}^t[n-i]\;\mathbf{c}_\mathrm{i}$$
$$b_i[n-i]\;w[n-i] = \mathbf{v}^t[n-i]\;\mathbf{d}_\mathrm{i}$$

Using this vector notation, Eq. (3) can be written as

$$\begin{aligned} &x\,[n] + \mathbf{u}^t\,[n-1]\,\mathbf{c}_1 + \cup + \mathbf{u}^t\,[n-p]\,\mathbf{c}_\mathbf{p} = \\ &w\,[n] + \mathbf{v}^t\,[n-1]\,\mathbf{d}_1 + \cup + \mathbf{v}^t\,[n-q]\,\mathbf{d}_\mathrm{q} \end{aligned} \tag{9}$$

Or equivalently

$$x\,[n] + \phi^t\,[n]\;\theta = w\,[n], \tag{10}$$

where $\phi^t\,[n]$ is the row vector

$$\phi^t\,[n] = [\mathbf{u}^t\,[n-1], \cup, \mathbf{u}^t\,[n-p], -\mathbf{v}^t\,[n-1], \cup, \mathbf{v}^t\,[n-q]],$$

and

$$\theta = [\mathbf{c}_1, \cup, \mathbf{c}_\mathbf{p}, \mathbf{d}_1, \cup, \mathbf{d}_\mathbf{q}]^t.$$

Observe that the vector θ contains all the unknowns of the TD-ARMA model. Writing Eq. (10) for $n = 0, 1, \cup, N - 1$ leads to

$$\mathbf{x} = \Phi\;\theta + \mathbf{w}, \tag{11}$$

where

$$\Phi = \begin{bmatrix} -\phi^t[0] \\ \vdots \\ -\phi^t[N-1] \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x[0] \\ \vdots \\ x[N-1] \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w[0] \\ \vdots \\ w[N-1] \end{bmatrix}.$$

The least-squares solution of Eq. (11) is given by

$$\theta = (\Phi^t\,\Phi)^{-1}\,\Phi^t\,\mathbf{x} \tag{12}$$

To overcome the computational complexity associated with the least-squares estimation (involving in particular the inversion of a square $(m + 1)(p + q) \times (m + 1)(p + q)$ matrix), we opted for a recursive least-squares estimation as follows.

Recursive least-squares estimation

The recursive least squares algorithm is summarized as [24]

$$\hat{\theta}[n] = \hat{\theta}[n-1] + L[n]\;\{x[n] + \phi^t[n]\hat{\theta}[n-1]\} \tag{13}$$

$$L[n] = -\frac{P[n-1]\;\phi[n]}{1+\phi^t[n]\;P[n-1]\;\phi[n]} \tag{14}$$

$$P[n] = P[n-1] - \frac{P[n-1]\;\phi[n]\;\phi^t[n]\;P[n-1]}{1+\phi^t[n]\;P[n-1]\;\phi[n]} \tag{15}$$

The initial conditions can be chosen arbitrarily.

Index of randomness

Over the past decade, there has been a flow of conflicting papers about the long-range power-law correlations detected in eukaryotic DNA [1-3,5,6,9-12,17-20]. The controversy is generated by conflicting views that either advocate that non-coding DNA sustains long-range correlations whereas coding DNA behaves like random sequences [1,10,2,3,5,6], or maintains that both coding and non-coding DNA exhibit long-range power-law correlations [11,12,9,17-20]. Based on the analysis of the time-dependent power spectrum of genomic sequences, Bouaynaya and Schonfeld [11,12] showed that the statistical differences between coding and non-coding sequences are more subtle than previously concluded using stationary analysis tools. In fact they found that both coding and non-coding sequences are non-random. However, coding sequences are "whiter" than non-coding sequences.

We propose to qualitatively assess the degree of randomness of both coding and non-coding sequences using the time-dependent ARMA coefficients $a_i\,[n]$ and $b_i\,[n]$. Consider the system function, $H\,(z)$, of a stationary ARMA model (whose coefficients a_i and b_i are constant, i.e., independent of time). We know that

$$H(z)=\frac{\Sigma_{k=0}^{q} b_k z^{-k}}{\Sigma_{k=0}^{p} a_k z^{-k}}=\frac{\Pi_{k=1}^{q}(1-r_k z^{-1})}{\Pi_{k=1}^{p}(1-p_k z^{-1})}, \qquad (16)$$

where $\{r_k\}_{k=1}^{q}$ (resp. $\{p_k\}_{k=1}^{p}$) are the zeros (resp. poles) of the system function. From the fact that a stationary white noise process has a at spectrum, we observe that the closer (in L_2 distance) the zeros and poles are, the flatter is the spectrum of the process. Following the same reasoning locally for non-stationary processes, we define the curve of randomness, $CR\,[n]$, of the non-stationary process $x\,[n]$ by

$$\begin{cases} CR[n]=\min_{(r_k[n],p_k[n])}\left(\frac{1}{q}\sum_{k=1}^{q}|\,r_k[n]-p_k[n]\,|+\frac{1}{p-q}\sum_{k=q+1}^{p}|\,p_k[n]\,|\,\right), & \text{if } p>q; \\ CR[n]=\min_{(r_k[n],p_k[n])}\left(\frac{1}{p}\sum_{k=1}^{p}|\,r_k[n]-p_k[n]\,|+\frac{1}{p-q}\sum_{k=p+1}^{q}|\,r_k[n]\,|\,\right), & \text{if } q>p; \\ CR[n]=\min_{(r_k[n],p_k[n])}\left(\frac{1}{p}\sum_{k=1}^{p}|\,r_k[n]-p_k[n]\,|\,\right), & \text{if } p=q. \end{cases} \qquad (17)$$

where the minimum is taken over all pairs ($r_k\,[n]$, $p_k\,[n]$). Observe that the case $p < q$ is obtained from the $p > q$ case by interchanging the roles of r_k and p_k, and the indices p and q. The curve of randomness defined in Eq. (17) is a measure of how close the zeros and the poles of the system function are, and therefore, is a measure of how flat the system function is, and how close is the process from a white noise. The index of randomness, $IR(p, q)$, of a TD-ARMA(p, q), is then defined as the average of the curve of randomness, i.e.,

$$IR(p,q)=\frac{1}{N}\sum_{n=0}^{N-1}CR[n]. \qquad (18)$$

In particular, the index of randomness of a TD-ARMA(1,1) ($x\,[n] + a[n - 1]x[n - 1] = w[n] + b[n]w[n - 1]$) is given by

$$IR(1,1)=\frac{1}{N}\sum_{n=0}^{N-1}|\,a[n]-b[n]\,|. \qquad (19)$$

Observe that the index of randomness of a white noise process is equal to zero. We say that the sequence $x_1\,[n]$ with index of randomness IR_1 is more random than the sequence $x_2\,[n]$ with index of randomness IR_2 if the index of randomness of the former is lower than the index of randomness of the latter, i.e., $IR_1 < IR_2$.

Results

All genome sequences considered in this paper have been extracted from the NIH website http://www.ncbi.nlm.nih.gov. The algorithms were implemented in MATLAB. The DNA sequences were mapped to numerical sequences using the purine-pyrimidine DNA walk proposed in [1]. In our simulations, the recursive least squares algorithm was found to best estimate the time-dependent coefficients of the TD-ARMA model. We used the MATLAB function *rarmax*, which implements the recursive least-squares algorithm for TD-ARMA models. The choice of the orders p and q of the model were determined experimentally as follows: For each genomic sequence, we computed 100 TD-ARMA models corresponding to the orders (1, 1) up to (10, 10). The best model was chosen to be the one that minimizes the average squared error between the actual and the fitted sequences. Our extensive simulations on various DNA sequences from different organisms show that most of the sequences are best fitted with low-order TD-ARMA models, e.g., TD-ARMA(1,1), TD-ARMA(1,2) and TD-ARMA(2,1). Figure 3 shows the DNA walk of the Human gene 276 and its TD-ARMA(1,1) fitted sequence. Observe that the TD-ARMA(1,1) model accurately fits this gene sequence. The estimated time-varying coefficients $a\,[n]$ and $b\,[n]$ are displayed in Fig. 4 for both the coding and non-coding regions of this gene. Their statistical differences are not clear from the plot of the time-series coefficients. The curves of randomness of the coding and non-coding regions are displayed in Fig. 5. Table 1 shows the index of randomness of various gene sequences. For concise representation, the column titles have been abbreviated as follows: "C. length" (resp."N.C. length") denotes the length (in base pairs) of the coding (resp. non-coding) segment of the gene. The total length of the gene is the sum of the lengths of its coding and non-coding regions. "C. (p, q)" (resp. "N.C. (p, q)") denotes the optimal TD-ARMA parameters (p, q) for the coding (resp. non-coding) region of the gene. Finally, "C. IR" (resp. "N.C. IR") is the index of randomness of the coding (resp. non-coding) segment of the gene. Observe that, in all considered genes, the index of randomness of both the coding and non-coding segments are strictly positive, and the index of randomness of the coding region is consistently lower than the index of randomness of the non-coding region (recall that the index of randomness of a white noise is zero). These observations bring to bear two important conclusion: (i) Both the coding and non-coding sequences are not random, as attested by an index of randomness greater than zero. (ii) The coding sequences are "whiter" than the non-coding sequences. This conclusion revokes previous work on statistical correlation of DNA sequences, which, based on stationary time-series analysis, presumed that coding DNA behaves like random sequences [1-3,5,6,10]; and supports the conflicting view that both coding and

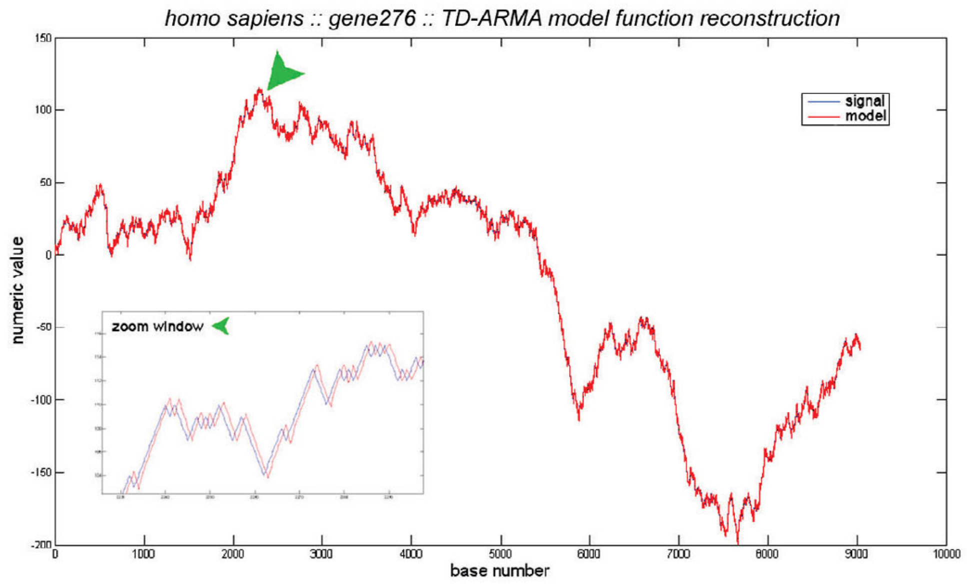

Figure 3
TD-ARMA modeling. TD-ARMA modeling of the Human gene 276: The blue signal is the DNA walk, and the red signal is the TD-ARMA(1,1) fitted signal. The TD-ARMA(1,1) model accurately fits the genomic signal.

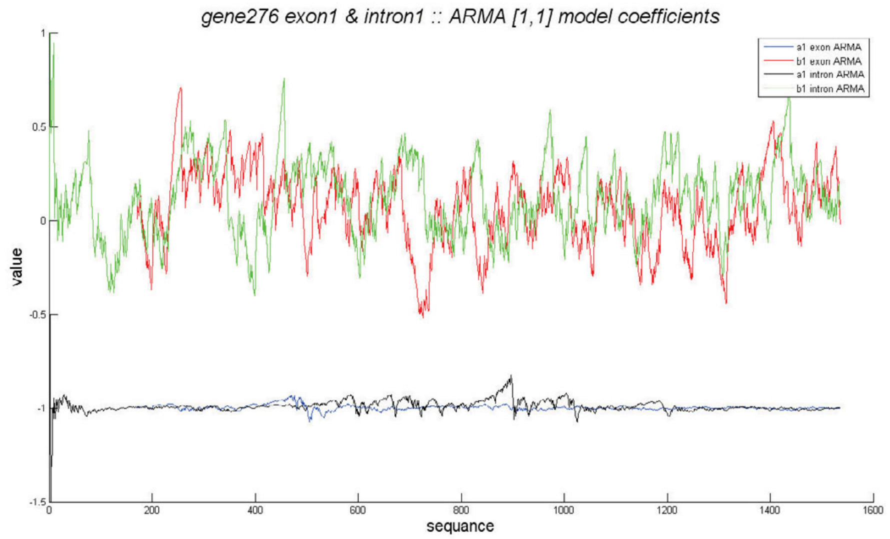

Figure 4
TD-ARMA coefficients estimation. Estimation of the TD-ARMA(1,1) coefficients of the Human gene 276. The TD-ARMA(1,1) model is given by $x[n] + a[n - 1]\, x[n - 1] = w[n] + b[n - 1]\, w[n - 1]$. The blue and black (resp. red and green) curves depict the time series $a[n]$ (resp. $b[n]$) for the coding and non-coding regions of the gene, respectively.

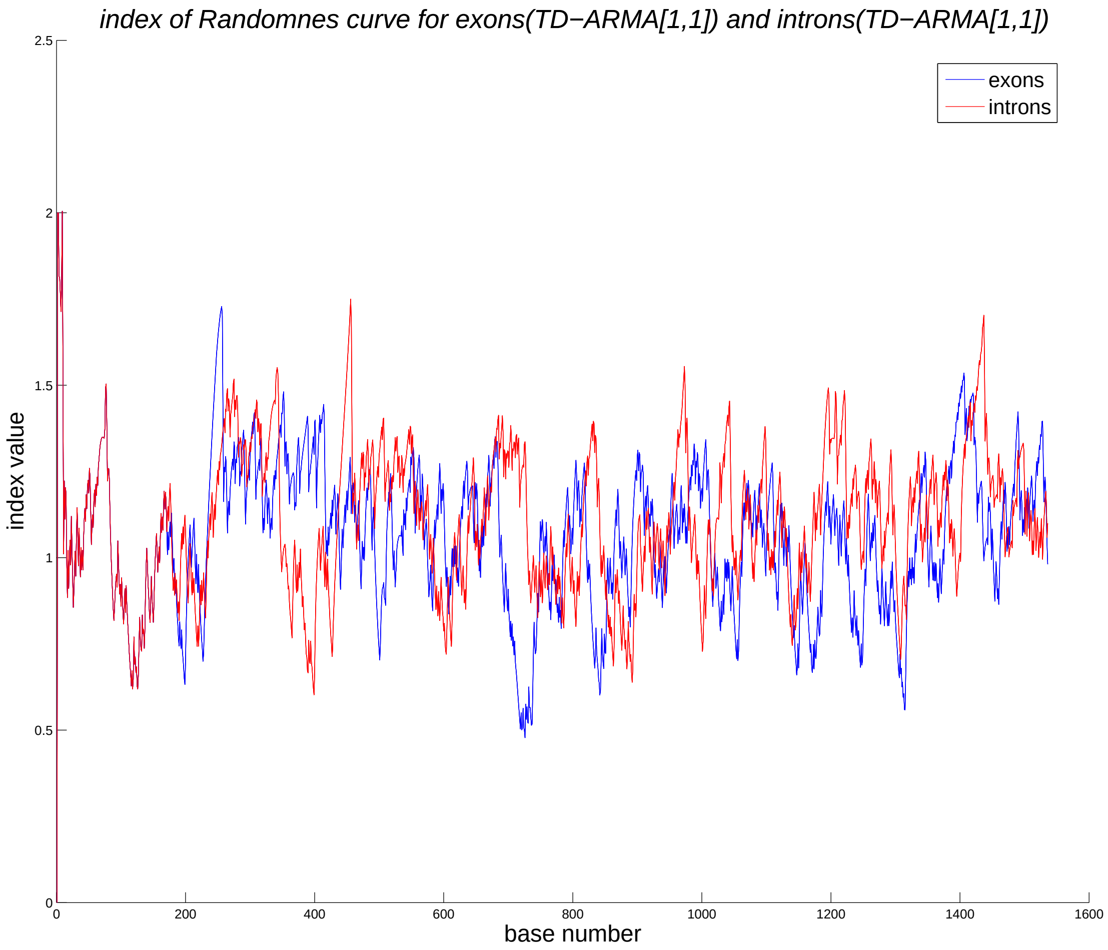

Figure 5
Curve of randomness. The curves of randomness of the coding and non-coding regions of the Human gene 276 are shown in blue and red, respectively. The index of randomness of the coding sequence is equal to 1.0603, whereas its corresponding value for the non-coding sequence is equal to 1.0627.

Table 1: Index of Randomness of the Coding and Non-Coding segments of Various Gene Sequences

Gene NIH accession number	C. length	C. (*p*, *q*)	C. IR	N.C. length	N.C. (*p*, *q*)	N.C. IR
Ashbya gossypii (fungus) AE016815	180953	(1,1)	0.9466	674919	(1,1)	0.9860
Aspergillus fumigatus (form of fungus) CM000169	1227993	(2,1)	0.9870	1835394	(1,1)	1.0683
Candida albicans (form of yeast) AP006852	373390	(1,1)	1.0282	570789	(1,1)	1.0429
Candida albicans AP006852	373390	(1,1)	1.0282	570789	(3,1)	1.0429
fission yeast GI:157310483	753661	(1,1)	1.0402	1654671	(1,1)	1.0642
fruit fly AE002620	21399	(1,1)	1.0084	1222832	(1,2)	1.1075
fruit fly AE002725	11316	(1,1)	1.0145	659655	(1,1)	1.0320
Homo sapiens hs-gene277 NG-004750	1639	(1,1)	1.0688	6573	(1,1)	1.0808
Homo sapiens hs-gene276 NG-004750	1536	(1,1)	1.0603	6672	(1,1)	1.0627

non-coding sequences are not random [11,12,9,17-20]. In particular, our conclusion is in accordance with the evolutionary periodogram analysis conducted in [11,12].

Conclusion

In this paper, we modelled the non-stationary genomic sequences by a time-dependent autoregressive moving average (TD-ARMA) model. By expressing the time-dependent coefficients as linear combinations of parametric basis functions, we were able to transform a linear non-stationary problem into a linear time-invariant problem. Subsequently, we proposed three methods to estimate the time-dependent coefficients: Mean -square, least-squares, and recursive least-squares algorithms. Based on the estimated TD-ARMA coefficients, we defined an index of randomness to quantify the degree of randomness of both coding and non-coding sequences. We found that both coding and non-coding sequences are not random. However, a higher index of randomness attests that coding sequences are "whiter" than non-coding sequences. These results corroborate the evolutionary periodogram analysis of genomic sequences performed in [11] and [12], and revoke the stationary analysis' conclusion that coding DNA behaves like random sequences.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JSZ derived the different estimation algorithms of the TD-ARMA parameters and performed the simulations. NB proposed the use of the non-stationary analysis and the index of randomness as a basis for statistical inference and biophysical interpretation of genomic data, derived the different estimation algorithms of the TD-ARMA parameters, and drafted the manuscript. DS proposed the use of the non-stationary analysis and the index of randomness as a basis for statistical inference and biophysical interpretation of genomic data and derived the different estimation algorithms of the TD-ARMA parameters. WO proposed the use of TD-ARMA modeling as a non-stationary model of genomic sequences. All authors read and approved the final manuscript.

Acknowledgements

The publication of this paper was partially supported by the Arkansas IDeA Networks of Biomedical Research Excellence (INBRE).

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 9, 2008: Proceedings of the Fifth Annual MCBIOS Conference. Systems Biology: Bridging the Omics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/9?issue=S9

References

1. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE: **Long-range correlations in nucleotide sequences.** *Nature* 1992, **356(6365):**168-170.
2. Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE: **Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis.** *Physical Review E* 1995, **51:**5084-5091.
3. Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M: **Scaling features of noncoding DNA.** *Physica A* 1999, **273:**1-18.
4. Li W, Holste D: **Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome.** *Physical Review E* 2005, **71:**041910.
5. Podobnik B, Shao J, Dokholyan NV, Zlatic V, Stanley HE, Grosse I: **Similarity and dissimilarity in correlations of genomic DNA.** *Physica A* 2006, **373:**497-502.
6. Carpena P, Bernaola-Galvan P, Coronado AV, Hackenberg M, Oliver JL: **Identifying chracteristic scales in the human genome.** *Physical Review E* 2007, **75:**032903.
7. Li W: **Expansion-modification systems: a model for spatial 1/f spectra.** *Physical Review A* 1991, **43(10):**5240-5260.
8. Dodin G, Levoir P, Cordier C: **Triplet Correlation in DNA Sequences and Stability of Heteroduplexes.** *Journal of Theoretical Biology* 1996, **183:**341-343.
9. Voss RF: **Evolution of long-range fractal correlations and 1/f noise in DNA base sequences.** *Physical Review Letters* 1992, **68:**3805-3808.
10. Li W, Kaneko K: **Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence.** *Europhysics Letters* 1992, **17:**655.
11. Bouaynaya N, Schonfeld D: **Non-Stationary Analysis of Genomic Sequences.** In *IEEE Statistical Signal Processing Workshop Madison, WI*; 2007:200-204.
12. Bouaynaya N, Schonfeld D: **Non-stationary Analysis of Coding and Non-coding Regions in Nucleotide Sequences.** *IEEE Journal of Selected Topics in Signal Processing* 2008.
13. ADAK S: **Time-dependent spectral analysis of nonstationary time series.** *Journal of the American Statistical Association* 1998, **93(444):**1488-1501.
14. Cramer H: **On some classes of nonstationary stochastic processes.** In *Proceedings of the Berkeley Symppsium on Math, Statistics, and Probability Los Angeles, CA*; 1961.
15. Grenier Y: **Rational nonstationary spectra and their estimation.** *ASSP Workshop on Spectral Estimation* 1981.
16. Huang NC, Aggarwal JK: **On linear Shift-variant digital filters.** *IEEE Transactions on Circuits and Systems* 1980, **27(8):**672-679.
17. Prabhu VV, Claverie JM: **Correlations in intronless DNA.** *Nature* 1992:359-782.
18. Chatzidimitriou-Dreismann CA, Larhammar D: **Long-range correlations in DNA.** *Nature* 1993, **361:**212.
19. Pande VS, Grosberg AY, Tanaka T: **Nonrandomness in protein sequences – evidence for a physically driven stage of evolution.** *Proceedings of the National Academy of Sciences* 1994, **91(26):**12972-12975.
20. Guharay S, Hunt BR, York JA, White OR: **Correlations in DNA sequences across the three domains of life.** *Physica D* 2000, **146(1–4):**.
21. Berthelsen CL, Glazier JA, Skolnick MH: **Global fractal dimension of human DNA sequences treated as pseudorandom walks.** *Physical Review A* 1992, **45(12):**8902-8913.
22. Grenier Y: **Time-Dependent ARMA Modeling of Nonstationary Signals.** *IEEE Transactions on Acoustics, Speech, and Signal Processing* 1983, **31(4):**899-911.
23. Hayes MH: *Statistical digital signal processing and modeling. Wiley* 1996.
24. Ljung L: *System Identification – Theory for the User* second edition. *Prentice Hall*; 2006.