Journal of Biomedical Informatics 134 (2022) 104197

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

DEGBOE: Discrete time Evolution modeling of Gene mutation through Bayesian inference using qualitative Observation of mutation Events

Komlan Atitey

Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC 27709, United States

ARTICLE INFO

Keywords:
Discrete-time nonlinear systems
Nonstationary process
Bayesian inference
Gene mutations
Lung cancer progression

ABSTRACT

An important aspect of cancer progression concerns the way in which gene mutations accumulate in cellular lineages. Comprehensive efforts into cataloging cancer genes have revealed that tumors demonstrate variability in genes that accumulate mutations which depend on the presence or absence of other mutations. However, understanding the stochastic process by which mutations arise across the genome is an important open problem of this nature in biology due to modeling discrete variate time-series is the most challenging, and, as yet, least well-developed of all areas of research in time-series. In this paper, a DEGBOE framework is proposed to model the mutation time-series given the sequence data of the gene mutations. The method relates the discrete-time, nonlinear and nonstationary series of gene mutations to the time-varying autoregressive moving average model. It presents the observation as a nonlinear function dependent on two variables: gene mutation, and gene-gene interactions characterizing the effects of the varying presence or absence of other gene mutations on a mutations' occurrence and evolution. DEGBOE is applied to model the dynamics of frequently mutated genes in lung cancer, including EGFR, KRAS, and TP53. The results of the model are analyzed and compared to the original simulated data of the DNA walk, and experimental lung cancer mutations data. It identifies the driver role of TP53 mutations in lung cancer progression.

1. Introduction

Cancer is the second leading cause of death in the United States and across the world [1]. For example, it is estimated that 13 million Americans are living with cancer and 40.8 % of people can expect to be diagnosed with cancer [2]. According to the National Institutes of Health (NIH), cancer cost the United States an estimated $263.8 billion in medical costs and lost productivity in 2010 and the cost of cancer care is expected to escalate more rapidly soon as more expensive targeted treatments are adopted as standards of care [3]. Cancer is known to result primarily from genetic mutations [4,5]. Mutations are heritable changes in an organism's DNA (or RNA, in RNA-based organisms) [6], and spontaneous mutations are those that occur without an exogenous DNA-damaging agent. These include DNA polymerase errors, mutations induced by endogenous agents, and deletions, duplications, and insertions [7]. Several factors other than genetic mutations may also affect carcinogenesis, such as epigenetic modifications [8], the tumor environment [9], and adaptive evolution [10]. However, carcinogenesis is primarily a result of genetic mutations [11,12]. Therefore, the study of tumor evolution at the gene mutation level undoubtedly provides a solid theoretical basis for accurate tumor therapy [13]. This paper focuses on mathematical modeling of the discrete-time occurrence and evolution of somatic mutation events. From the perspective of tumor evolution theory, if we can dynamically model the evolution of somatic mutations, we can improve treatment and prognosis.

Almost every tumor arises from random DNA mutations that corrupt a cell's genes and alter the genetic regulatory circuits that control its functions [6]. Normal cells evolve progressively through increasingly abnormal (dysplastic) states until neoplasia. The formation of a tumor and its development, or cancer progression [14], is a complex process that normally occurs over many years (10 or more). Luckily, not all genetic mutations engender tumor cells: many mutations are irrelevant to normal cell life, while many others make the mutated cells genome so unstable that the cell dies quickly. The requisite number of genetic mutations for tumorigenesis is very high, as shown in Fig. 1, far beyond the number of genetic mutations that occur during normal human life. Sometimes a random DNA mutation provides a normal cell with a sort of genomic instability: even if it does not seem to provide immediate proliferation or survival benefits, genomic instability greatly increases the speed of further genetic mutations and speeds the acquisition of

E-mail address: komlan.atitey@nih.gov.

https://doi.org/10.1016/j.jbi.2022.104197
Received 18 June 2022; Received in revised form 2 August 2022; Accepted 1 September 2022
Available online 6 September 2022
1532-0464/Published by Elsevier Inc.

other characteristics.

Understanding the stochastic process by which mutations arise across the genome is an important open problem of this nature in biology because it is central to identifying mutations that drive cancer [16]. Numerous domains involve modeling highly nonstationary discrete-time and integer-valued stochastic processes where event counts vary dramatically over time or space [17,18]. Also, multiple methods for modeling stationary stochastic processes exist [19,20], but far fewer exist for nonstationary processes because they are difficult to capture with the covariance functions of parametric models. For instances, nonstationary kernels have been introduced for Gaussian processes [21], but these may not be tractable in large datasets due to their computational complexity. More recently, Poisson-gamma models have been considered for dynamics systems [22,23] but these methods have focused on relationships between count variables, not predicting counts based on continuous covariates. Although many methods have been developed for somatic mutations detection [24,25], there is no consensus on the best variant-calling algorithm. However, understanding the statistical properties of genomic sequences is fundamental to recreate the dynamic processes that led to the observed DNA structure and to identify gene-related diseases such as cancer [26,27].

To address the above concerns, this paper proposes the DEGBOE framework to model the temporal occurrence and evolution of frequently mutated lung cancer genes, including the genes EGFR, KRAS, and TP53 [28–30]. The mathematical model of the DEGBOE framework is represented by finite order time-varying autoregressive moving average (ARMA) processes [31] that appropriately describes the discrete-time nonstationary somatic gene mutation event (local hidden variable) considered as a visible representation of system dynamics (time-series) [32,33]. Add to the local hidden variable of gene mutation events, the number of other gene mutations (global hidden variable) interacting with and influencing the dynamic evolution of somatic gene mutations is jointly modeled. The Bayesian rule is used to compute the joint posterior distribution of the local and global hidden variables with the purpose of identifying the influence of the global hidden variable on the local hidden variable in time-series gene mutations. Due to the computation intractability of the joint posterior distribution, a variational Bayesian technique is exploited to convert the joint posterior distribution into a posterior distribution of the local hidden variable that depends on the global hidden variable. The sequential Monte Carlo (SMC) sampling-importance sampling-resampling is applied with Matlab to numerically implement the actual posterior distribution. In application, the proposed method describes a discrete-time evolution of gene mutations in terms of the interaction with a variate number of other gene mutations events.

In particular, the method provides a way of improving inference based on the entire posterior distribution of the time-varying hidden states of a mutation given the sequence data of gene mutations evolution. It adopts multiple computation approaches to address challenges in modeling a discrete-time, nonlinear, and nonstationary process. The proposed framework is applied to model time-series of genomic sequence (DNA walk of human gene 276), and the frequently mutated genes of lung cancer during lung cancer progression. Contributions to this work are resumed into four main parts. In first, a gamma-Gaussian model is proposed to pattern time-discrete nonstationary stochastic processes at any length scale without retraining. Secondly, the modeling of discrete-time nonstationary genomic sequences is performed by using data from the DNA walk of human gene 276. Thirdly, the evolution of lung cancer gene mutations is demonstrated as dependent on the presence or absence of other gene mutations. Finally, the proposed DEGBOE framework identifies the mutations in the gene TP53 as highly dependent on the presence or absence of other gene mutations and drives the

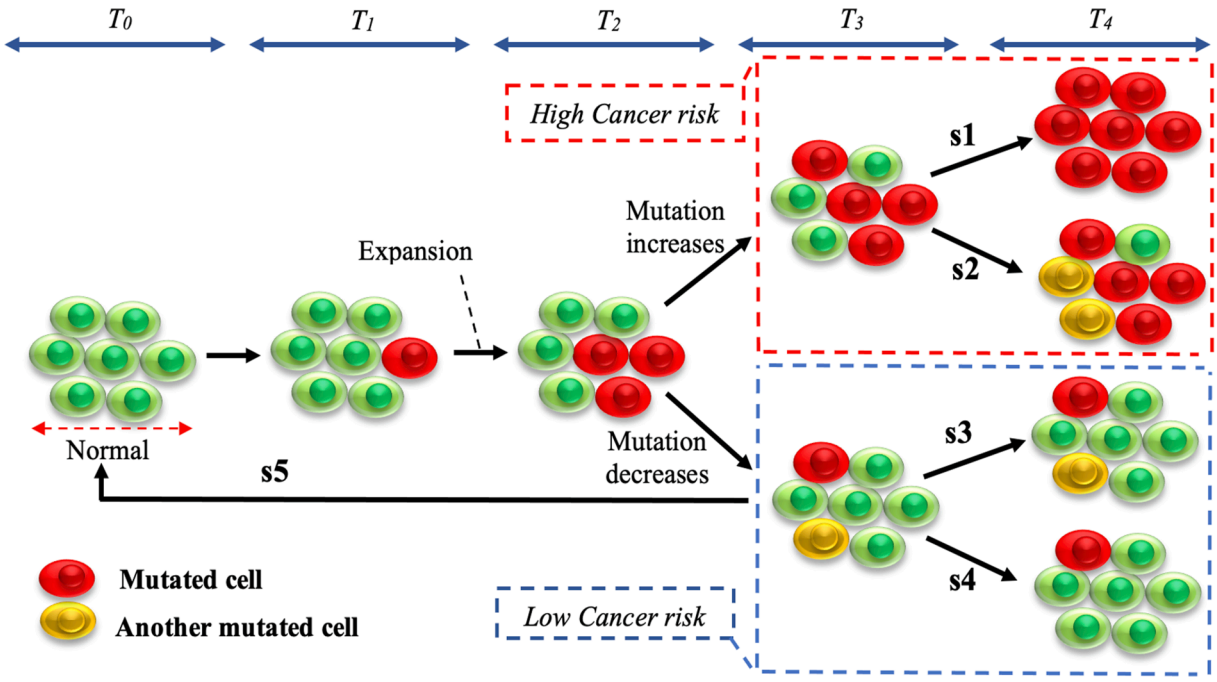

Fig. 1. Dynamics of gene mutation. T0, T1, T2, T3, and T4 represent successive time steps to investigate the population of cellular dynamics for genetic mutation. The mutator dynamics in populations propagated through bottlenecks of different sizes: **s1:** Rapid growth to tumor: The mutation spreads through the population and replaces the resistant (normal) cells. **s2:** An additional driver mutation increases the population of mutated cells, which coexist with normal cells. It leads to intratumor heterogeneity, which refers to biological variations within the same tumor. Commonly, intratumor heterogeneity arises not only from genomic and epigenomic disorders of tumor cells themselves, but also from the influence of the tumor microenvironment [15]. Importantly, intratumor heterogeneity exists among different geographical regions of the same tumor (spatial heterogeneity), as well as between the primary tumor and subsequent local or distant recurrence in the same patient (temporal heterogeneity). **s3:** Equilibrium of cell mutation with no novel cell mutation. At this stage, neither the normal cell nor the mutated cell populations vary. **s4:** Decreasing mutant cell population. **s5:** Mutants go extinct. When the delay before the emergence of the next mutant is long, the population may attain one of the outcomes presented at time T4 (**s1, s2, s3, s4,** and **s5**).

progression of lung cancer.

In detail, the paper is organized in five more sections which follow this introduction. Section 2 presents the proposed DEGBOE methodology. It first introduces the definition of the time-varying ARMA process and the state-space description of dynamic systems, and then further constructs the nonstationary nature of the model. After that, variational Bayesian inference is presented as a tool to approximate the joint posterior distribution in Section 3. In Section 4, the datasets used to evaluate the DEGBOE framework are presented along with results, and Section 5 illustrates the validation of the DEGBOE framework in modeling gene mutations. Finally, in Section 6, the main conclusions and future work are provided.

2. Materials and methods

In this section, a detailed description of the proposed method is given with a graphical illustration in Fig. 2. It includes the overview of the nonstationary discrete time-series of gene mutations described along with the time-varying ARMA model, and gene-gene interactions.

2.1. Model description

The selection of variables used in the DEGBOE framework relies on the time-series gene mutations in the lung cancer dataset used in this study. The goal is to enable identifying patterns that serve as way to extract actual information from the dataset. Fig. 2(a) presents the lung cancer mutations binary dataset used in this study. The dataset contains 100 mutational samples of lung cancer which displays 12,327 genes (M) on the 100 mutational samples (N) for lung cancer mutations. Each row of the dataset represents a gene, and each column denotes a lung cancer mutational sample collected at different times.

The lung cancer mutations are represented as a discrete timeline of mutational events such that the appearance of a mutation in the timeline corresponds to its estimated place in lung cancer progression, which is proportional to the temporal order of the 100 mutational samples in the dataset. Fig. 2(b) portrays the computational model: the graphical model with observation $z_k = \{z_k\}_{k=1}^{N}$, local hidden variables $\boldsymbol{x_k} = \{x_k\}_{k=1}^{N}$, and global hidden variables $\boldsymbol{\varphi} = \{\varphi_m\}_{m=1}^{M-1}$. The local hidden variable illustrates the temporal evolution of mutation occurrence and evolution through the 100 mutational samples; meanwhile, the global hidden variable characterizes the presence or absence of other mutations influencing new gene mutations.

2.2. Mathematical models

2.2.1. Nonstationary, discrete time-varying gene mutations

This work is interested in the occurrence and evolution of gene mutations over time with respect to their interaction with other mutations. Based on the concept of Bayesian inference, the aim is to estimate the posterior distribution of the time-varying hidden states of a mutation given the sequence data of gene mutations evolution. For simplicity of modeling the time-varying mutation for lung cancer progression, the modeling of gene mutations dynamically enhanced by the interaction with other genes mutations is treated as modeling nonstationary, discrete-time, and stochastic processes where event counts vary dramatically over time. A system of genes is defined as constituted by M populations, whose microscopic state is represented by a scalar latent variable $x \in \mathbb{R}^{+} = [0, \infty]$, and the evolution of each population is determined by interactions between genes mutations as shown in Fig. 2 (b).

Grenier [34] showed that a discrete nonstationary signal of a single population of gene at time k (x_k) can be represented by a finite-order

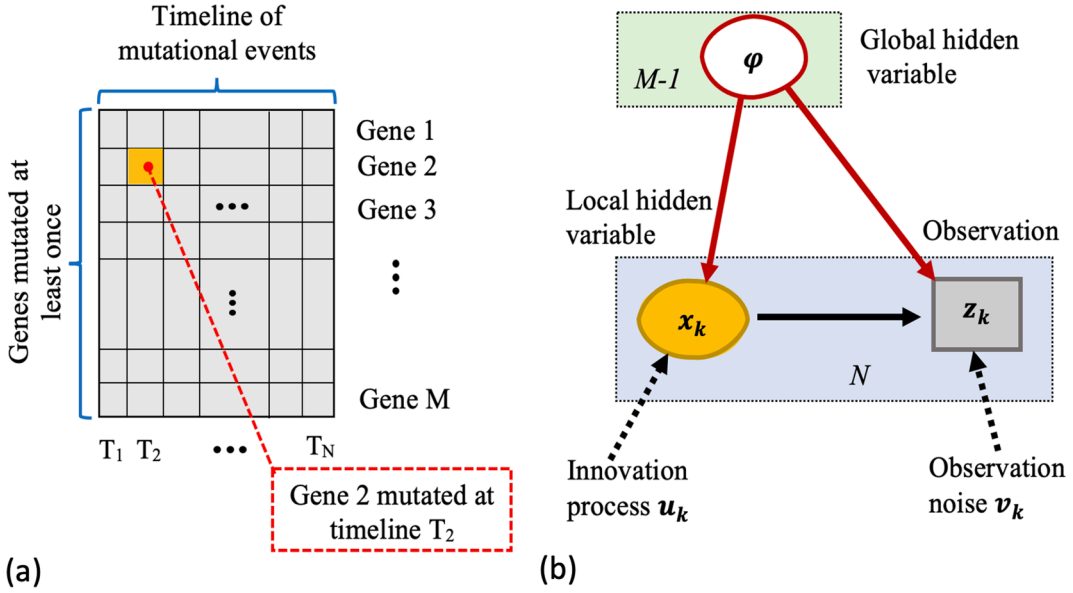

Fig. 2. DEGBOE computational framework for modeling of time-series gene mutations. (a) Lung cancer mutation binary dataset of M number of genes in N mutational samples; (b) Computational model describing the sequence data of genes mutations (observation) in terms of the local hidden variable (gene mutations), and the global hidden variable (effect of the presence or absence of other mutations on gene mutation). The DEGBOE's computational framework considers a discrete Bayesian network consisting of three variables (φ, $\mathbf{x_k}$, and $\mathbf{z_k}$) whose corresponding graphical model is given in Fig. 2(b). The observation $\mathbf{z_k} = \{z_k\}_{k=1}^{N}$ represents the collected sequence data of genes mutations; the local hidden variable $\mathbf{x_k} = \{x_k\}_{k=1}^{N}$ illustrates the occurrence and evolution of gene mutations; and the global hidden variable $\varphi = \{\varphi_m\}_{m=1}^{M-1}$ describes the gene-gene interaction involve in gene mutations for cancer progression. The circle representing the global hidden variable has no incoming arrows, indicating that there will be a known family of distribution for the global hidden variable. The local hidden variable has an incoming arrow from the global hidden variable, denoting that the probability of having a specific gene mutation depends on the presence or absence of other gene mutations, i.e., the probability of the occurrence and evolution of gene mutations depends on gene-gene interactions. Similarly, the sequence data of genes mutations (observation) has an incoming arrow from the global hidden variable. Importantly, the fact that there is an arrow directly connecting the local hidden variable and the observation indicates that, conditioned on the gene-gene interaction, the local hidden variable and the observation are not independent. The structure of the model implies that the joint probability of the set of variables can be written in the form, as: $p(\boldsymbol{\varphi}, \boldsymbol{x}_k, z_k) = p(\boldsymbol{\varphi})p(\boldsymbol{x}_k/\boldsymbol{\varphi})p(z_k/\boldsymbol{\varphi}, \boldsymbol{x}_k)$.

time-varying ARMA(m,n) process of order m and n of the form.

$$x_k = \sum_{i=1}^{m} \phi_{i,(k-i)} x_{k-i} + \sum_{j=1}^{n} \varphi_{j,(k-j)} u_{k-j} + u_k \tag{1}$$

defined by its m autoregressive coefficients $\{\phi_{1:m}\}$, and n moving average coefficients $\{\varphi_{1:n}\}$. These parameters describe the autocorrelation of the output random process x_k generated from the input innovation u_k at the discrete time k. The symbol u_k represents a zero-mean Gaussian innovation process that is correlated in time to allow the time-series to exhibit a wide range of decaying memory properties. The model presented in Equation (1) is a state-space description of dynamic systems [35] with a matrix notation.

$$\boldsymbol{\phi}_k \boldsymbol{x}_k = \boldsymbol{\psi}_k \boldsymbol{u}_k \tag{2}$$

where $\boldsymbol{x}_k = \{x_{1:k}\}$ and $\boldsymbol{u}_k = \{u_{1:k}\}$, and the transition matrices are.

$$\begin{cases} \boldsymbol{\phi}_k = \begin{pmatrix} 1 & -\phi_1 & \cdots & -\phi_m & \cdots & 0 \\ 0 & 1 & -\phi_1 & \cdots & -\phi_m & -\phi_{m+1} \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \vdots & 1 & -\phi_1 & -\phi_2 \\ 0 & \vdots & 0 & 0 & 1 & -\phi_1 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \\ \boldsymbol{\psi}_k = \begin{pmatrix} 1 & \varphi_1 & \cdots & \varphi_n & \cdots & 0 \\ 0 & 1 & \varphi_1 & \cdots & \varphi_n & \varphi_{n+1} \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \vdots & 1 & \varphi_1 & \varphi_2 \\ 0 & \vdots & 0 & 0 & 1 & \varphi_1 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \end{cases} \tag{3}$$

Accordingly, the state vector $\boldsymbol{x}_k$ can be expressed as a linear transformation of innovations [36], i.e.,

$$\boldsymbol{x}_k = \Theta_k \boldsymbol{u}_k \tag{4}$$

where $\Theta_k = \boldsymbol{\phi}_k^{-1} \boldsymbol{\psi}_k$ refers to the transfer function that determines the observability of innovations from states $\boldsymbol{x}_k$ as well as the controllability of the ARMA model [36].

To capture the nonstationary nature of the model, the innovation process $\boldsymbol{u}_k$ is modeled as fractional Gaussian noise with mean $E[\boldsymbol{u}_k] = 0$ under stationary increments [37]. The increment process $\Delta u(k) = u(k) - u(k-1)$ has Hurst exponent H, which determines the correlation at all time scales k. It also represents a measure of long-range dependence in time-series as well as a quantitative measure of the fractal nature of a DNA sequence [38]. The autocovariance function $\gamma_u(k)$ of the increments that follows from Chiang *et al.* [39] is given by.

$$\gamma_u(k) = E[u_{k+\tau} u_k] = \sigma_u^2 / 2 \left[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H} \right] \tag{5}$$

with the variance σ_u [40], and the Hurst exponent fixed in the range $0 < H < 1$ for the nonstationary nature of the model. Hence, for $k = 1, 2, \cdots, N$ with N illustrating the length of the sequence $\boldsymbol{x}_k$ in Fig. 2(a), and by taking $\rho_u(k) = 1/2 \left[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H} \right]$ such that $\gamma_u(k) = \sigma_u^2 . \rho_u(k)$, the covariance matrix $\boldsymbol{C}_{u_k}$ of the zero-mean Gaussian vector $\boldsymbol{u}_k$ can be expressed for any integer values as.

$$\boldsymbol{C}_{u_k} = \sigma_u^2 \boldsymbol{R}_{u_k} \tag{6}$$

where:

$$\boldsymbol{R}_{u_k} = \begin{pmatrix} \rho_u(1) & \rho_u(2) & \cdots & \rho_u(N) \\ \rho_u(2) & \rho_u(1) & \cdots & \rho_u(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_u(N-1) & \rho_u(N-2) & \cdots & \rho_u(2) \\ \rho_u(N) & \rho_u(N-1) & \cdots & \rho_u(1) \end{pmatrix}. \tag{7}$$

Consequently, the covariance matrix for the zero-mean Gaussian distributed $\boldsymbol{x}_k$ could be derived as.

$$\boldsymbol{C}_{x_k} = \Theta_k \boldsymbol{C}_{u_k} \Theta_k^T \tag{8}$$

Thus, the evolution of the i^{th} population, for $i \in \{1, 2, \cdots, N\}$, is described by the distribution function $f_i = f_i(k, x)$ over the state x at time k such that f_i is Gaussian: $\boldsymbol{x}_k \; \mathcal{N}\left(\boldsymbol{x}_k; \mu_{x_k}, \boldsymbol{C}_{x_k}\right)$ with zero mean ($\mu_{x_k} = 0$) and covariance $\boldsymbol{C}_{x_k}$.

2.2.2. Gene-gene interactions

The complexity of cellular systems does not arise from the independent action of many different genes, but is rather the result of extensive genetic interactions among them. Therefore, in addition to the latent hidden variable x_k (local hidden variable) describing the occurrence and evolution of a gene mutation, the number of other gene mutations interacting with and influencing the dynamic growth of a mutation is modeled as a global hidden variable (φ). Furthermore, the observation is introduced with a nonlinearity function g and noise v_k, i.e., the noise observation z_k as.

$$z_k = g(x_k, \varphi, v_k) \tag{9}$$

where the nonlinearity g is constrained by the requirement that the likelihood $p(z_k / x_k, \varphi)$ of x_k and φ is computable up to a proportionality constant. Overall, the class of the model involved observations, global hidden variables, and local hidden variables illustrated in Fig. 2(b). The N observations are $z_k = z_{1:N}$; the vector of global hidden variables is $\boldsymbol{\varphi} = \varphi_{1:M-1}$; and the N local hidden variables are $\boldsymbol{x}_k = \boldsymbol{x}_{1:N}$.

The joint posterior distribution for the global hidden variables and local hidden variables is computed, following Bayes' rule, proportional to the likelihood times the prior distribution, such as:

$$p(\boldsymbol{x}_k, \boldsymbol{\varphi} / z_k) \propto p(z_k / \boldsymbol{x}_k, \boldsymbol{\varphi}) p(x_k, \varphi / z_k) \tag{10}$$

with likelihood $p(z_k / \boldsymbol{x}_k, \boldsymbol{\varphi})$ and prior distribution $p(x_k, \varphi / z_k)$. Due to the computational intractability of the posterior distribution [41], it is necessary to approximate the posterior distribution of the hidden variables (local and global) given the observations z_k for computational tractability. Accordingly, the result of Equation (10) is used to develop a stochastic variational inference and convert the posterior distribution of both local and global hidden variables $p(\boldsymbol{x}_k, \boldsymbol{\varphi} / z_k)$ into a posterior distribution of local hidden variables.

3. Bayesian inference in time-varying progression of mutation

This section presents the variational Bayesian inference as a tool to approximate the joint posterior distribution of the local and global hidden variables for computational tractability.

3.1. Variational Bayesian inference

The general starting point of the variational inference framework is a joint probability density function over observed variables z_k and hidden (local, global) variables $\boldsymbol{h}_k = \{\boldsymbol{x}_k, \boldsymbol{\varphi}\}$,

$$p(z_k, \boldsymbol{h}_k) = p(\boldsymbol{h}_k) p(z_k / \boldsymbol{h}_k) \tag{11}$$

where $p(\boldsymbol{h}_k)$ is referred to as the prior density and $p(z_k / \boldsymbol{h}_k)$ as the likelihood. Therefore, given an observed value z, the aims of the variational inference are to: (1) determine the conditional density of h given z as the posterior density $p(h/z)$, (2) evaluate the marginal density of the

observed data as log model evidence following:

$$lnp(z_k) = \ln \int_{h} p(z_k, \boldsymbol{h}_k) d\boldsymbol{h}. \tag{12}$$

Hence, the log model evidence allows for comparing different models in their plausibility to explain observed data z_k. Particularly, due to the fact that a model comprising multiple hidden random variables (e.g., local and global variables) leads to a computational intractability, as the case in this work, the lower bound approximation of the log model evidence is used to achieve the two aims of the variational inference. As a result, variational inference in effect replaces an integration problem with an optimization problem by decomposing the log model evidence into a core of the variational inference approach as:

$$lnp(z_k) = F[q(\boldsymbol{h}_k)] + KL[q(\boldsymbol{h}_k) \| p(\boldsymbol{h}_k / z_k)] \tag{13}$$

where $q(\boldsymbol{h}_k)$ is the variational density denoting an arbitrary probability density over the local hidden variable used to approximate the (in practice, typically intractable) joint posterior distribution $p(\boldsymbol{h}_k/z_k)$ since $\boldsymbol{h}_k$ includes local and global hidden variables. Equation (13) presents the log model evidence comprising the sum of two quantities of theoretic information. Firstly, the variational free energy, defined as.

$$F[q(\boldsymbol{h}_k)] = \int_{h} q(\boldsymbol{h}_k) \ln \left(\frac{p(z_k, \boldsymbol{h}_k)}{q(\boldsymbol{h}_k)} \right) d\boldsymbol{h} \tag{14}$$

and, secondly, the Kullback–Leibler (KL) divergence, which measures the distance between the true posterior distribution $p(\boldsymbol{h}_k/z_k)$ and the variational density $q(\boldsymbol{h}_k)$,

$$KL[q(\boldsymbol{h}_k) \| p(\boldsymbol{h}_k / z_k)] = \int_{h} q(\boldsymbol{h}_k) \ln \left(\frac{q(\boldsymbol{h}_k)}{p(\boldsymbol{h}_k / z_k)} \right) d\boldsymbol{h} \tag{15}$$

Therefore, maximizing the variational free energy hence minimizes the KL divergence between the variational density $q(\boldsymbol{h}_k)$ and the true posterior density $p(\boldsymbol{h}_k/z_k)$ and renders the variational free energy a better approximation of the log model evidence. As the local and global hidden variables $(\boldsymbol{x}_k, \boldsymbol{\varphi})$ may be used interchangeably, Equation (14) is rewritten as.

$$F[q(\boldsymbol{x}_k, \boldsymbol{\varphi})] = \iint q(\boldsymbol{x}_k, \boldsymbol{\varphi}) \ln \left(\frac{p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi})}{q(\boldsymbol{x}_k, \boldsymbol{\varphi})} \right) d\boldsymbol{x} d\boldsymbol{\varphi} \tag{16}$$

There is a tradeoff in choosing $q(\boldsymbol{x}_k, \boldsymbol{\varphi})$ expressive enough to approximate the posterior well, and simple enough to lead to a tractable approximation. A common choice is a fully factorized distribution, also called mean field distribution.

3.2. Mean field approximation of the variational free energy

Here, a detail of a mean field approximation of the variational free energy is provided as a used tool to construct a simple bound for the KL divergence [42,43]. A mean field approximation assumes that all latent (hidden) variables are independent as $q(\boldsymbol{x}_k, \boldsymbol{\varphi}) = q(\boldsymbol{x}_k) q(\boldsymbol{\varphi})$. It simplifies Equation (16) in a form.

$$F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = \iint q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) ln \left[\frac{p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi})}{q(\boldsymbol{x}_k) q(\boldsymbol{\varphi})} \right] d\boldsymbol{x} d\boldsymbol{\varphi}. \tag{17}$$

Based on the quotient rule for logarithms, the Equation (17) is converted to:

$$\begin{aligned} F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = & \iint q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) \ln [p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) - \ln q(\boldsymbol{x}_k)] d\boldsymbol{x} d\boldsymbol{\varphi} \\ & - \iint q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) \ln q(\boldsymbol{\varphi}) d\boldsymbol{x} d\boldsymbol{\varphi}. \end{aligned} \tag{18}$$

It can be reformulated in respect of the local hidden variable $(\boldsymbol{x}_k)$ into the form.

$$\begin{aligned} F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = & \iint q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) \ln [p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) - \ln q(\boldsymbol{x}_k)] d\boldsymbol{x} d\boldsymbol{\varphi} \\ & - \int q(\boldsymbol{\varphi}) \ln q(\boldsymbol{\varphi}) \left[\int q(\boldsymbol{x}_k) d\boldsymbol{x} \right] d\boldsymbol{\varphi} \end{aligned} \tag{19}$$

The fact that the integral of the probability density function $(q(\boldsymbol{x}_k))$ is 1 shows that the second term in Equation (19) turns to $\int q(\boldsymbol{\varphi}) \ln q(\boldsymbol{\varphi}) d\boldsymbol{\varphi}$ which can be considered as a constant $(\boldsymbol{c}^{te})$ as.

$$\begin{aligned} F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = & \int q(\boldsymbol{x}_k) \left[\int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) d\boldsymbol{\varphi} \right] d\boldsymbol{x} \\ & - \int q(\boldsymbol{x}_k) \ln q(\boldsymbol{x}_k) \left[\int q(\boldsymbol{\varphi}) d\boldsymbol{\varphi} \right] d\boldsymbol{x} - \boldsymbol{c}^{te} \end{aligned} \tag{20}$$

In a similar way, the integral of the probability density function $q(\boldsymbol{\varphi})$ in Equation (20) is 1. This makes variational free energy in the form:

$$F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = \int q(\boldsymbol{x}_k) \left[\int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) d\boldsymbol{\varphi} \right] d\boldsymbol{x} - \int q(\boldsymbol{x}_k) \ln q(\boldsymbol{x}_k) d\boldsymbol{x} - \boldsymbol{c}^{te} \tag{21}$$

Then, the variational free energy can be reformulated as.

$$\begin{aligned} F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = & \int q(\boldsymbol{x}_k) \left[\ln \left(\exp \int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) d\boldsymbol{\varphi} \right) \right] d\boldsymbol{x} \\ & - \int q(\boldsymbol{x}_k) \ln q(\boldsymbol{x}_k) d\boldsymbol{x} - \boldsymbol{c}^{te} \end{aligned} \tag{22}$$

$$F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = \int q(\boldsymbol{x}_k) \left[\ln \left(\frac{\exp \int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) d\boldsymbol{\varphi}}{q(\boldsymbol{x}_k)} \right) \right] d\boldsymbol{x} - \boldsymbol{c}^{te} \tag{23}$$

$$F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = - \int q(\boldsymbol{x}_k) \left[\ln \left(\frac{q(\boldsymbol{x}_k)}{\exp \int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) d\boldsymbol{\varphi}} \right) \right] d\boldsymbol{x} - \boldsymbol{c}^{te} \tag{24}$$

which can be re-expressed in terms of the Kullback-Leibler divergence as.

$$F[q(\boldsymbol{x}_k), q(\boldsymbol{\varphi})] = - KL \left[q(\boldsymbol{x}_k) \| exp \left(\int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) \right) d\boldsymbol{\varphi} \right] - \boldsymbol{c}^{te} \tag{25}$$

As a result, maximizing the negative KL divergence by setting $q(\boldsymbol{x}_k) = exp\left(\int q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) \right) d\boldsymbol{\varphi}$ thus maximizes the variational free energy. It leads to.

$$\ln q(\boldsymbol{x}_k) = \int d\boldsymbol{\varphi} q(\boldsymbol{\varphi}) \ln p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) \tag{26}$$

The joint probability $p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi})$ in Equation (26) is a product of the likelihood $p(z_k/\boldsymbol{x}_k, \boldsymbol{\varphi})$ and the joint posterior distribution for the global hidden variables and latent hidden variables of Equation (10) following.

$$p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) = p(\boldsymbol{x}_k, \boldsymbol{\varphi} / z_k) \prod_{k=1}^{N} p(z_k / \boldsymbol{x}, \boldsymbol{\varphi}) \tag{27}$$

For the posterior density, the mean-field approximation $p(\boldsymbol{x}_k, \boldsymbol{\varphi}/z_k) \approx q(\boldsymbol{x}_k) q(\boldsymbol{\varphi})$ [44] is considered. It converts Equation (27) into:

$$p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) = q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) \prod_{k=1}^{N} p(z_k / \boldsymbol{x}, \boldsymbol{\varphi}) \tag{28}$$

Equation (26) is written finally as:

$$\ln q(\boldsymbol{x}_k) = \int d\boldsymbol{\varphi} q(\boldsymbol{\varphi}) \ln \left[q(\boldsymbol{x}_k) q(\boldsymbol{\varphi}) \prod_{k=1}^{N} p(z_k / \boldsymbol{x}, \boldsymbol{\varphi}) \right] \tag{29}$$

Equation (29) gives a tractable form of the density of the dynamics of the local hidden variable (gene mutations) in terms of the global hidden variable (gene-gene interactions) and the observation (data of mutation sequences).

3.3. Probability density function of the dynamic of gene mutation

Thurley et al.'s [45] studies regarding the response time distributions of cell-to-cell communication networks proved that the interaction between cells or genes can be described by gamma distributions, as the gamma distribution is an asymmetric (right-skewed) distribution with a single peak at $k > 0$. Relying on this result, the global hidden variable φ is defined as gamma-distributed random variables such that: $q(\boldsymbol{\varphi}) = G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi})$ with mean $\langle\varphi\rangle = a_\varphi/b_\varphi$, dependent on the shape and scale parameters a_φ and b_φ respectively.

The N observations ($z = z_{1:N}$) are normally distributed with the mean and variance both being random variables of local and global hidden variables, respectively, such that.

$$p(z_k/\boldsymbol{x}_k, \boldsymbol{\varphi}) = \prod_k \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi}) \tag{30}$$

Accordingly, the joint probability density function of the hidden variables and observations is given finally by the product of the Gaussian density $\boldsymbol{x}_k$ and a Gamma density $\boldsymbol{\varphi}$, in order that Equation (27) satisfies.

$$p(z_k, \boldsymbol{x}_k, \boldsymbol{\varphi}) = \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \prod_k \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi}) \tag{31}$$

The non-independent Gaussian-Gamma prior density in Equation (31) has the advantage of belonging to the conjugate-exponential class and allows for the derivation of an exact analytical solution for the form of the posterior distribution. Hence, reporting Equation (31) into 29 gives.

$$\ln q(\boldsymbol{x}_k) = \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln[G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) \prod_k \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi})] \tag{32}$$

which is converted into a simple form as.

$$\ln q(\boldsymbol{x}_k) = \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) + \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) + \sum_k \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi}) \tag{33}$$

The expression of $\ln q(\boldsymbol{x}_k)$ in Equation (33) is broken down into 3 different terms for which analyses will make it computable in terms of the variable of interest $\boldsymbol{x}_k$. Consequently, as the first term, $\int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi})$, does not depend on the local random variables or the latent state of interest, this integrand could be discarded and replaced by a normalization term $\mho$. The second term, $\int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln \mathcal{N}\left(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}\right)$, depends on $\boldsymbol{x}_k$ but not on φ, so it can be simplified into $\ln \mathcal{N}\left(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}\right)$. Accordingly, Equation (33) can be simplified to.

$$\ln q(\boldsymbol{x}_k) = \ln \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) + \sum_k \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi}) + \ln \frac{1}{\mho} \tag{34}$$

Rewriting $\ln \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi})$ into a simple form as $(z_k - \boldsymbol{x}_k)\varphi(z_k - \boldsymbol{x}_k) + c$ where c stands for the normalization constant, the required integral becomes the expectation of a Gamma distribution such as.

$$\sum_k \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \ln \mathcal{N}(z_k/\boldsymbol{x}, \boldsymbol{\varphi}) = \sum_k (z_k - x_k) \left[\int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \varphi \right] (z_k - x_k). \tag{35}$$

In addition, by defining $\langle\varphi\rangle = \int d\varphi G(\boldsymbol{\varphi}; \boldsymbol{a_\varphi}, \boldsymbol{b_\varphi}) \varphi$ as the expected value of the global random variables, Equation (35) is converted into:

$$\ln q(\boldsymbol{x}_k) = \ln \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) + \sum_k (z_k - x_k) \langle\varphi\rangle (z_k - x_k) + \ln \frac{1}{\mho}. \tag{36}$$

From Equation (36), $q(\boldsymbol{x}_k)$ is deduced in the form:

$$q(\boldsymbol{x}_k) = \frac{1}{\mho} \mathcal{N}(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}) \exp\left[\sum_k (z_k - x_k) \langle\varphi\rangle (z_k - x_k) \right] \tag{37}$$

Here, the expected value of the global hidden variable φ can be characterized in terms of the shape and scale parameters such that $\langle\varphi\rangle = \boldsymbol{a_\varphi} \boldsymbol{b_\varphi}^{-1}$. The covariance $\boldsymbol{C}_{\boldsymbol{x}_k}$ included in the approximated density $q(\boldsymbol{x}_k)$ provides the Hurst exponent (Equations 5–8) that is used as a measure of long-term memory during the time series.

4. Results

This section presents the results of the proposed DEGBOE framework over simulated genomic sequence data of DNA walk and experimental time-varying lung cancer mutation data.

4.1. Genomic sequence data

The DEGBOE framework is first applied to model genomic sequences that exhibit nonstationary statistical behavior. Early, Zielinski et al. modeled successfully the nonstationary genomic sequences with a time-dependent ARMA model [26] through standard statistical tests to verify that genomic sequences are not stationary and their nonstationary nature varies and is often more complex than a simple trend [46,47]. Hence, Zielinski et al. efficiently estimated the time-varying ARMA coefficients and showed that ARMA (1,1) fits the DNA walk of the human gene 276 sequence. It is necessary to recall that, genome walking is a method for determining the DNA sequences of unknown genomic regions flanking a region of known DNA sequence. The genomic working is significantly useful for capturing homologous genes in new species when areas in the target gene exhibit strong sequence conservation to characterized species [48]. Also, it is valuable to capture the dynamics of DNA through one-dimensional random walks [49]. Accordingly, following Zielinski et al.'s [26] results, this work generates a simulated DNA walk of human gene 276 in terms of the genome size (Dataset 1) with appropriate coefficients ($\phi_1 = 0.35, \varphi_1 = 0.4$), of the ARMA (1,1) model. The simulated data of DNA walk is used to assess the performance of the DEGBOE's framework in modeling a nonstationary genomic sequence. It should be noted that genome size refers to the amount of DNA contained in a haploid genome expressed in terms of the number of base pairs [50,51].

4.2. Time-varying lung cancer mutation data

To investigate the performance of the DEGBOE framework in modeling the occurrence and evolution of gene mutations in lung cancer progression, the DEGBOE framework is applied to modeling of frequently mutated genes, including EGFR, KRAS, and TP53 [28–30], using a lung cancer mutations dataset (Dataset 2) provided by Auslander et al. [52]. The mutation events are described throughout the dataset (Fig. 2(a)) as dummy variables (independent variables) that equal 0 when gene mutation is absent or 1 when mutation occurs.

In addition to modeling somatic mutations, this dataset allows to examine the global variable describing the number of gene mutations required to activate any new gene mutation for lung cancer progression. With the purpose to make the computation more flexible, the binary discrete feature variables of the mutations data are preprocessed by converting the binary sequence of mutations into a magnitude of mutation by evaluating the following function. Let a be the value of a dummy variable: $a = \{0, 1\}$; M and N represent, respectively, the number of rows (number of genes) and columns (number of temporal mutational samples) of the lung cancer mutational dataset. The magnitude of a gene mutation in a temporal mutational sample denoted as λ_j is determined as $\lambda_j = a_{1j} \times 100/\sum_{i=1}^{N} a_{ij}$, for $j = \{1, 2, \cdots, M\}$, so that the greater the number of mutations in the temporal gene mutation

sample, the lower the magnitude of a single gene mutation. Thus, the $M \times N$ lung cancer mutational dataset presents discrete-time gene mutations with different time-varying magnitudes.

4.3. Performance evaluation

The performance of the DEGBOE framework is demonstrated with different values of Hurst exponent (H1 = 0.5, H2 = 0.7, H3 = 0.9) which allow for the determination of the relative trend of time-series of genomic sequences and gene mutations. Numerical studies were conducted using Matlab; the codes used in this study were written in Matlab version Matlab_R2021a. In addition, R packages were also used with R code for data preprocessing and visualization. All the materials can be found online through the GitHub link provided in Appendix A.

4.3.1. Genomic sequences modeling

Fig. 3 shows the DNA walk of human gene 276 (Signal) and its corresponding DEGBOE fitted sequences (Model). The mean square error (MSE) is used to confirm the model's fit to the true data in relation to Hurst exponent. Let recall that the MSE measures the amount of error in statistics models by assessing the average squared difference between the signal and the model.

The outcomes from applying the DEGBOE computational framework relate its performance to Hurst exponent (H) values used. As can be observed in Fig. 3, the Fig. 3(a) displays comparison between the true DNA walk (Signal) and the fitted DNA walk (Model) with the DEGBOE framework on the basis of Hurst exponent equivalent to H3 = 0.9. Fig. 3 (b) matches the true DNA walk and the fitted DNA walk with the DEGBOE framework in terms of the Hurst exponent corresponding to H2 = 0.7. Fig. 3C compares the true DNA walk and the fitted DNA walk modeled with the DEGBOE framework with Hurst exponent in proportion to H1 = 0.5. The shaded area of the DNA walks in Fig. 3(a)-3 (c) represents the 95 % confidence interval for the DNA walks standard deviation. The mean square error (MSE) is used to confirm the model's fit to the true data (Fig. 3(d)). It is worth noting that the lower the MSE, the closer the model's values are to the true data. It shows that when the Hurst exponent (H) is large, the model's fit to the DNA walk of human gene 276 (Signal) is better than when H is small. Thus, the good performance of the DEGBOE computational framework depends on high Hurst exponent values.

4.3.2. Modeling of discrete time somatic mutation

Each studied genes discrete-time mutation is modeled by applying the DEGBOE framework described in this study. Fig. 4 demonstrates the performance of the DEGBOE framework with different values of Hurst exponents (H) in modeling the temporal mutations of the genes EGFR, KRAS, and TP53 in lung cancer progression.

Fig. 4(a), 4(d) and 4(g) compare the true magnitude of mutations (Signal) and the fitted magnitude of mutations (Model) of gene EGFR for different values of Hurst exponent equivalent to H1 = 0.5, H2 = 0.7, and H3 = 0.9 respectively. By way of illustration, Fig. 4(b), 4(e), and 4(h) juxtapose the true magnitude of mutations (Signal) and the fitted magnitudes (Model) of gene KRAS for different values of Hurst exponent used as H1 = 0.5, H2 = 0.7, and H3 = 0.9 respectively. Finally, Fig. 4(c), 4(f), and 4(i) relate the true magnitudes of mutations (Signal) to the fitted magnitudes (Model) of gene TP53 mutations for different values of Hurst exponent utilized comparable to H1 = 0.5, H2 = 0.7, and H3 = 0.9 respectively. The shaded area of magnitudes of mutations represents the 95 % confidence interval for the magnitude's standard deviation. Comparing the true magnitude to the fitted magnitude for the three gene mutations shows the dependency of the performance of the DEGBOE framework to the Hurst exponent: a high value of Hurst exponent lead to

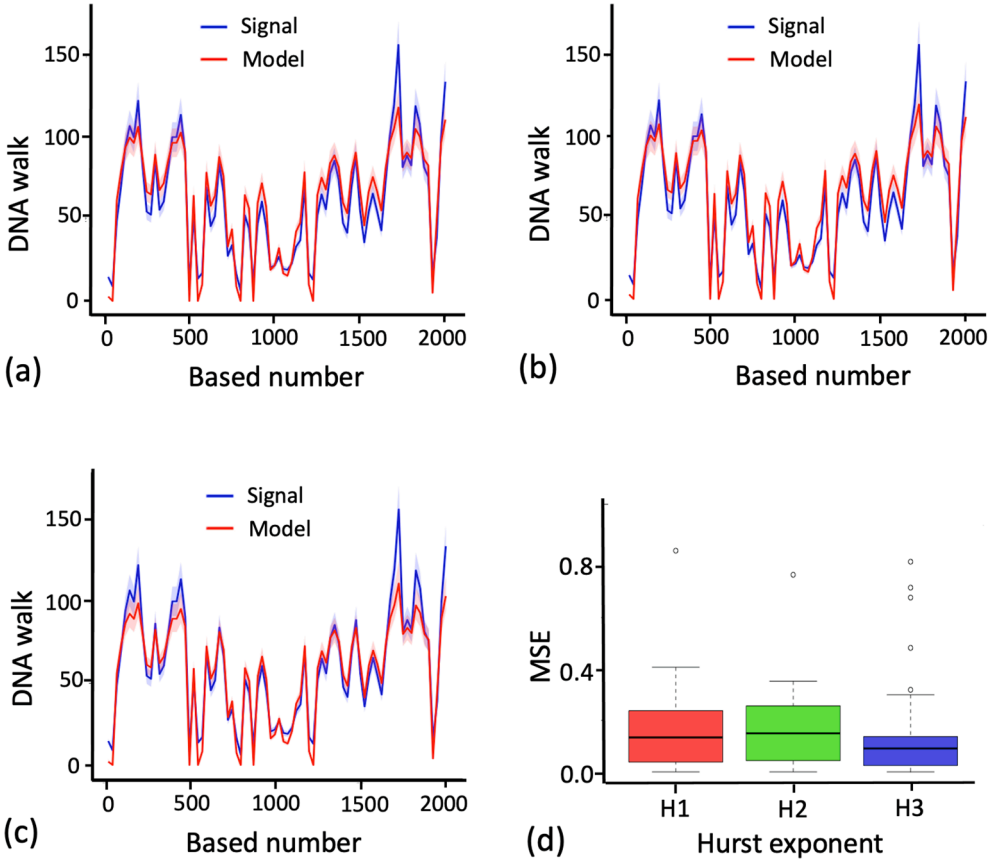

Fig. 3. The performance of the DEGBOE framework in modeling of the DNA walk of human gene 276. The Signal represents the true data, and the Model portrays the implementation of the DEGBOE framework. (a) H3 = 0.9; (b) H2 = 0.7; (c) H1 = 0.5; D) MSE.

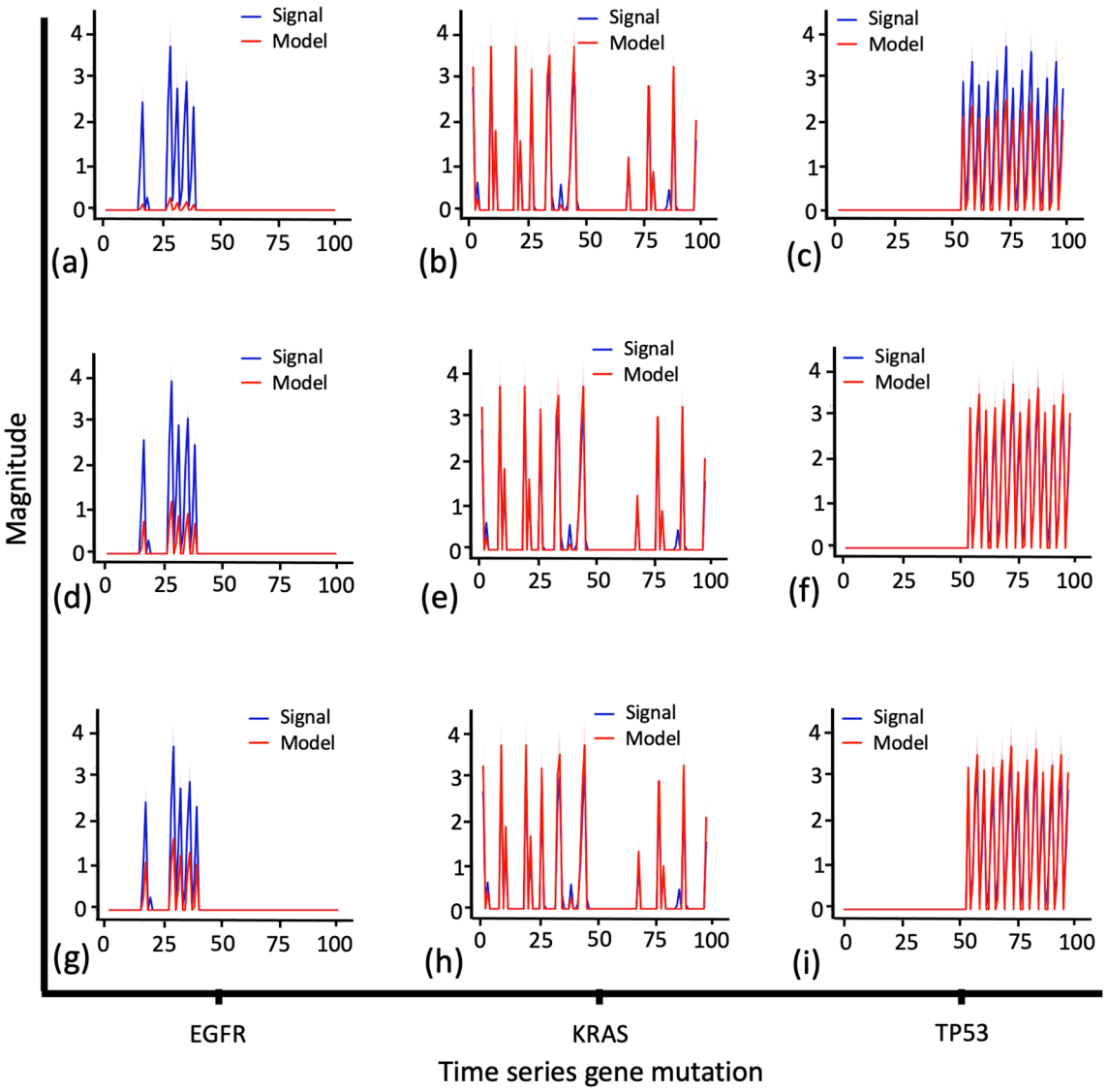

Fig. 4. The DEGBOE modeling of time-varying somatic mutations magnitudes. The graph displays the time series gene mutations of EGFR, KRAS, and TP53 from the left to the right column, respectively. (a) EGFR for H = 0.5; (b) KRAS for H = 0.5; (c) TP53 for H = 0.5; (d) EGFR for H = 0.7; (e) KRAS for H = 0.7; (f) TP53 for H = 0.7; (g) EGFR for H = 0.9; (h) KRAS for H = 0.9; (i) TP53 for H = 0.9.

a high performance of the DEGBOE framework in modeling of lung cancer gene mutation for lung cancer progression.

The MSE metric is also used to measure how close the magnitude of gene mutations provided by the DEGBOE framework are to the measured magnitudes (Fig. 5).

Fig. 5 demonstrates that a higher Hurst exponent value generates a more accurate DEGBOE's model of gene mutations. In Fig. 5A, the MSE provides information about the performance of the DEGBOE framework

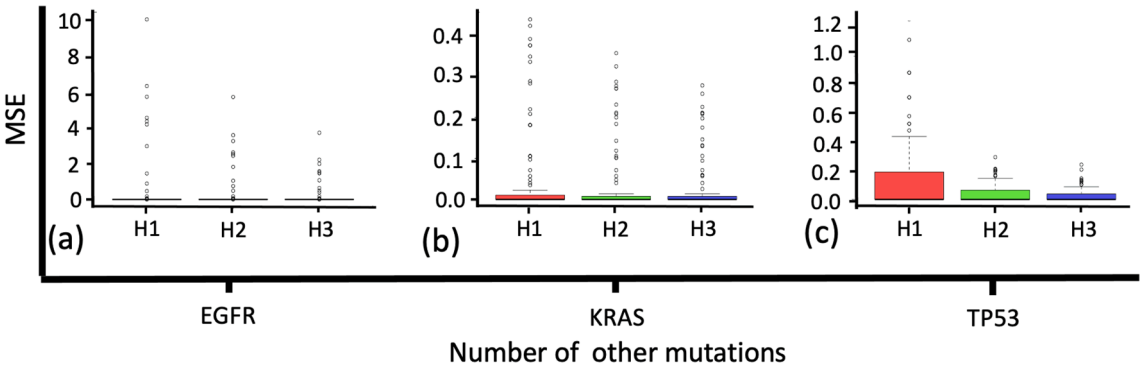

Fig. 5. Evaluate the performance of the DEGBOE framework with the MSE metric with varying Hurst exponents: H1 = 0.5, H2 = 0.7, and H3 = 0.9 for the three different genes. (a) EGFR; (b) KRAS; (c) TP53.

by allowing a comparison of the actual difference between the true (Signal) and the fitted (Model) magnitude of gene EGFR mutation for different Hurst exponent values H1, H2, and H3. With regard to Fig. 5B, the MSE shows the actual difference between the true and the fitted magnitude of gene KRAS mutations for different Hurst exponent values H1, H2, and H3. Finally, in Fig. 5C, the MSE enables a comparison of the actual difference between the true and the fitted magnitude of the gene TP53 mutation for different Hurst exponent values H1, H2, and H3. For all the three genes, the DEGBOE framework presents a good performance in modeling of gene mutations at high values of Hurst exponent.

Outcomes of the DEGBOE framework fit to the time-series gene mutations with large values of Hurst exponent. It demonstrates a long trend of mutations of the genes EGFR, KRAS, and TP53 in lung cancer progression.

4.3.3. *Evaluation of gene mutation rate*

Results of this study show how a high Hurst exponent leads to the ideal performance of the DEGBOE framework in modeling of time-series gene mutations. Next, it is interesting to deal with summarizing the density of gene mutations in terms of the number of interactions of other mutated genes during lung cancer progression. Let recall that the density of the discrete-time gene mutations (local hidden variable) in the DEGBOE framework describes the progression of a gene mutation as dependent on the presence or absence of mutations in one or more other genes. In practice, the density of gene mutations characterized the presence or absence of other gene mutations as the mean of the global hidden variable presented in Equation (37).

According to a probability of a mutational event represents a rate of gene mutations [53], a sample of gene mutations density in this work is comparable to a gene mutation rate. As a result, the sequential Monte Carlo (SMC) sampling-importance sampling-resampling (SISR) framework is used to draw 2,000 samples from gene mutations densities of genes EGFR, KRAS, and TP53. The number of SMC-SISR iterations contributes then to effectively describes a sequence of gene mutations in correlation with other genes mutations. In this regard, among the 12,327 genes identified in lung cancer dataset (Fig. 2(a)), 10 different samples of number of gene are generated as: $N1, N2, \cdots, N10$ corresponding to 1; 1001; 2001; 3001; 4001; 5001; 6001; 7001; 8001; 9001 and 10,001 respectively. It relies on the statement that a gene mutation in each of the 10 samples affect the occurrence and progression of new mutations. Therefore, the mean of the global hidden variables in the density of gene mutation (Equation (37)) is then replaced iteratively by each of the 10 different possible interactions a gene mutation may have with other genes . Analyses of the corresponding 10 different densities of discrete-time gene mutation occurrence and progression is processed. As seen in Fig. 6, mutations of genes EGFR, KRAS, and TP53 are found to respond differently to the 10 possible different interactions with other genes mutations.

The upper and lower box hinges of the boxplots correspond to the first and third quartiles; the horizontal line within the box matches to the median; and the vertical lines span the minimum and maximum expression values. Analysis of mutation rates shows that EGFR (Fig. 6 (a)) displays a relatively consistent mutation rate over the 10 different samples; hence, the interquartile range (the count of extreme values) of its mutation rate does not present notable variability. However, KRAS and TP53 manifest different medians as well as interquartile ranges in their interactions with other gene mutations (Fig. 6(b) and 6(c)). For instance, the time-varying mutation rate of KRAS shows its lowest and highest medians at N4 and N1 respectively, and its highest interquartile range at N8 and N10. The mutation rate of TP53 exhibits a much larger interquartile range for N4, and N6. Table 1 shows the variability of rates of mutations of the genes EGFR, KRAS, and TP53 for different

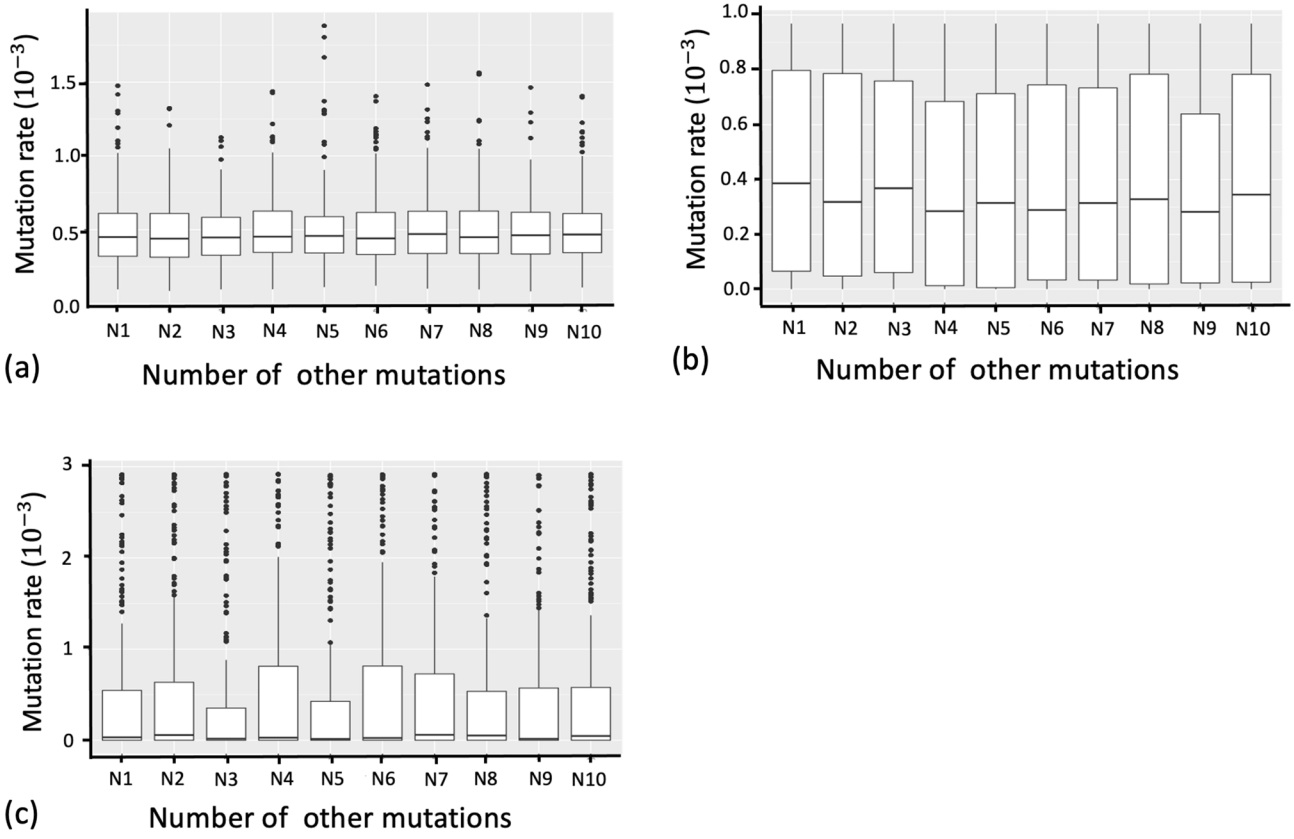

Fig. 6. Boxplot representation of 10 different connections with gene mutations. The sequential Monte Carlo (SMC) sampling-importance sampling-resampling (SISR) is used with Matlab implementation to draw 2,000 samples from the density of mutation variables allocated to each of the 3 genes studied. Each sample of the density represents the probability of a mutational event or the mutation rate. (a) EGFR; (b) KRAS; (c) TP53.

Table 1
Interquartile range of mutation rates for genes EGFR, KRAS, and TP53 in lung cancer progression.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
EGFR(10^{-5})	2	2.1	1.9	2	1.9	2	2	2	2	1.9
KRAS(10^{-5})	7.3	7.4	7.2	6.9	7.4	7.3	7.3	9.5	7.1	9.7
TP53(10^{-4})	5	6	3.5	8	3.8	7.8	7.4	4.5	4.6	4.7

interactions with the presence or absence of other genes mutations.

Additionally, a visual summary of density for time-varying rates of gene mutation is presented on the basis of the number of SMC-SISR samples (2000 samples) used. It also considers the gene-gene interaction as the number of different gene mutations whose presence or absence influences the progression of new gene mutations. Fig. 7(a) provides the marginal density of the gene EGFR's mutation rate. For all the 10 extends of gene interaction, the relative densities of the gene EGFR's mutations present a single maximum mutation rate varying in the range $[2.10^{-4}, 5.10^{-4}]$. It shows how the density of mutation rate of EGFR does not present a notable difference in terms of the interactions with other genes' mutations. For all the 10 extends of gene mutation interactions, the relative mutation densities of gene KRAS present two maximums of the mutation rate corresponding to the lowest and highest values of the mutation rate in (Fig. 7(b)). However, the densities of KRAS mutations rates do not manifest sensible variability in respect of the number of other mutations that influence gene KRAS's mutations. Overall, EGFR, and KRAS display both an overlap of densities. It demonstrates how EGFR and KRAS mutations do not significantly depend on the presence or absence of other gene mutations. With regard to mutations of gene TP53 in respect to the 10 extends of interactions of other genes mutations, the relative densities of gene TP53's mutations present a single maximum of mutation rate corresponding to the lowest values of its mutation rate (Fig. 7(c)). Also, it is very noticeable that the densities of TP53 display significant variability in respect of the maximum values referring to the 10 extends of interactions with other genes mutations in lung cancer progression.

Unlike the results from EGFR and KRAS, the gene TP53 is identified as exhibiting a high rate of mutations that vary mostly in terms of the presence or absence of other gene mutations (Fig. 7). This outcome is less pronounced with EGFR and KRAS, which display a low variability of mutation rates in terms of the presence or absence of other gene mutations, especially the gene EGFR. It leads to classifying the influence of the three genes' mutations in driving lung cancer mutations in ascending order as EGFR, KRAS, and TP53.

5. Discussion

This work addresses the problem of modeling the stochastic process by which gene mutations arise across the genome for cancer progression. Particularly, on the grounds of Bayesian networks, the DEGBOE computational framework adopts different computation approaches to address challenges in modeling a discrete-time, nonlinear, and nonstationary process. In the construction of the method, the model supports the central knowledge that the variability in gene that accumulates mutations for tumors progression depend on the presence or absence of other mutations highlighting gene-gene interaction. The output of the model is a posterior distribution which relates the evolution of gene mutations to the gene-gene interaction allowing one to evaluate the amplitude of a gene mutation given the set of genes in the tumor microenvironment. With reference to modeling the most frequently mutated lung cancer genes EGFR, KRAS, and TP53, the DEGBOE framework gauges the significance of the evolution of gene mutations in the context of a large background of genetic variants, and it presents a

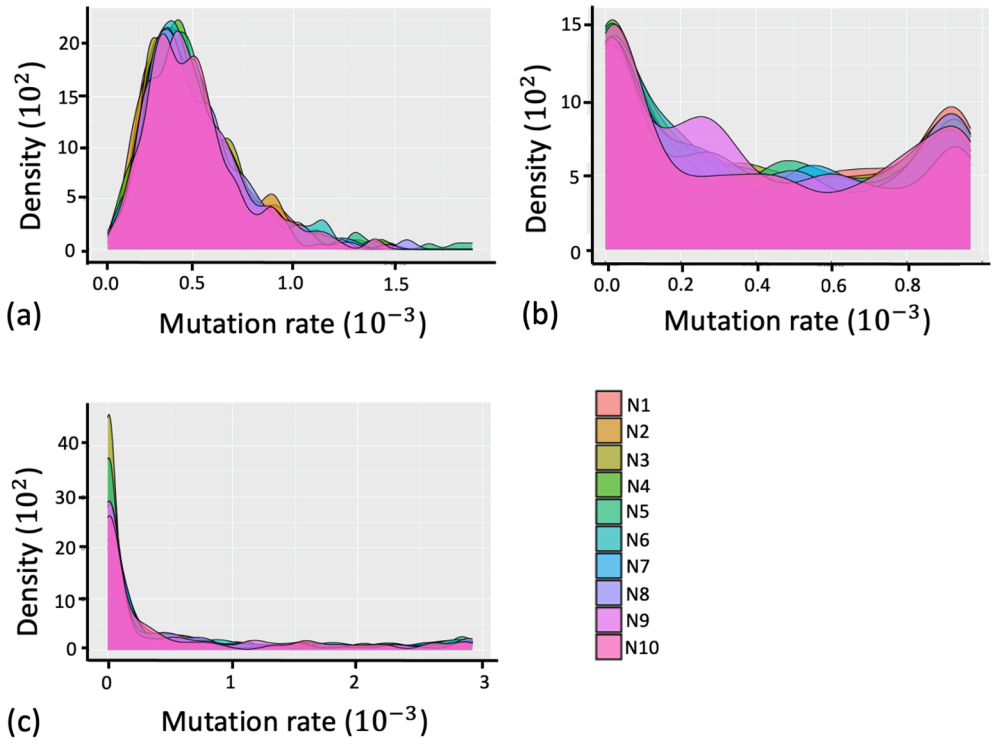

Fig. 7. Density of mutation rate. It is drawn on the basis of the number of SMC-SISR samples used corresponding to the mutation rates, which depends on the presence or absence of influences of other mutations on the evolution of new gene mutations. (a) EGFR; (b) KRAS; (c) TP53.

good performance in terms of correctly identifying the driver role of TP53 mutations in lung cancer progression. The proposed computational framework enables the evaluation of variability in mutation rates modeled as a nonstationary, discrete-time, and stochastic process that remains the source of genetic variation [54].

Evidence has shown that the DEGBOE framework is suitable for analyzing genomic sequences that exhibit discrete, nonstationary statistical behavior [55,56] by considering the evolution of genomic sequences that do not depend on the environment. Additionally, based on discrete mutational data that exhibit binary sequences of lung cancer mutations, the DEGBOE serves to model the temporal magnitude of gene mutations during a series of 100 mutational samples of lung cancer evolution. Furthermore, because the framework performs well in modeling the evolution of discrete nonstationary gene mutations, DEGBOE manages to evaluate the varying rates of gene mutations with respect to the presence or absence of other gene mutations responsible for the occurrence and evolution of new mutations.

First, the findings show that DEGBOE accurately modeled a DNA walk of human gene 276. In addition, it shows that the good performance of the framework depends on the value of the Hurst exponent (H), which is a measure of long-range dependence in time-series and represents a quantitative measure of the fractal nature of a DNA sequence [38]. Second, the DEGBOE framework preserves the time scale between time-series mutation events in modeling the magnitude of gene mutations. Furthermore, results of the study showed a high influence and a role of driver mutation of gene TP53 in the progression of lung cancer for two reasons. Primarily, TP53 presents a high maximum density of mutation rate, comparing it to the maximum densities of the other two more frequently lung cancer mutated genes (EGFR, and KRAS). Secondary, the maximum density of mutation rate of TP53 is obtained at a smaller value of its mutation rate. It shows how low mutation rates of TP53 is optimal enough to influence mutations of other genes for lung cancer progression.

These results fit to early research works which have identified TP53 as a frequently mutated gene in human cancers. For example, the abnormality of the TP53 gene is one of the most significant events in lung cancers and plays an important role in the tumorigenesis of lung epithelial cells [55,56]. Additionally, experimental studies have recognized TP53 mutations as a mutation which drive lung cancer progression [57,58]. The similarity of those results to the outcomes of this study justifies the good performance of the DEGBOE framework in modeling the time varying of gene mutation in connection with the presence or absence of other genes mutations.

The DEGBOE computational approach in modeling the occurrence and evolution of gene mutations for cancer progression has its limitations. It was most appropriate when the time-series gene mutation depends only on the presence or absence of other gene mutations which highlight gene-gene interactions. Commonly, some acquired mutations occur spontaneously and randomly in genes. Other mutations are caused by environmental factors that induce genomic instability or disrupt cellular metabolism, arises from many different pathways, such as telomere damage, centrosome amplification, epigenetic modifications, and DNA damage from endogenous and exogenous sources [61]. It is likely that by integrating into the DEGBOE' s framework all the environmental factors that could potentially influence gene mutation for cancer progression, the computational framework may have the advantage of identifying different evolutionary patterns contributing to cancer progression with high accuracy. In the future, for the sake of computational efficiency, the DEGBOE computational framework would be properly extended at the level of the number of environmental factors that highly control gene mutations for cancer progression.

6. Conclusion

In this work, the DEGBOE framework is proposed for time-series modeling, including a DNA walk and the evolution of frequently mutated genes (EGFR, KRAS, and TP53) for lung cancer progression. The results using two time-series data of DNA walk of human gene 276, and lung cancer mutations have been reported in several cases the accuracy of the DEGBOE framework, which depends on the Hurst exponent (H) describing a long-term dependent process. Firstly, a high H resulted in good performance of the framework in modeling a nonstationary genomic sequence of DNA walk. Secondly, the DEGBOE framework preserves a time scale between time-series mutations' events in modeling the magnitude of gene mutations. Furthermore, analyses of densities of mutation rates of the three genes studied in this work led to identifying gene TP53 mutations as the mutation, which is highly dependent on the presence or absence of other genes mutations as well as driving the progression of lung cancer. Finally, this study highlights the importance of using the variational Bayesian framework to convert the intractable joint posterior distribution into a simple form of density for analysis. The variational Bayesian framework is, thus, particularly useful for researchers applying the DEGBOE framework to open biology questions presenting nonlinear, discrete, and nonstationary properties.

In future work, in addition to gene-gene interactions, new factors that allow to capture dependencies in genes with mutations will be included as input variables. For instance, the environment, the size as well as the structure of the population represent some of the factors on which gene mutations depend. Therefore, including those factors in inputs variables of the DEGBOE framework will extend the performance of the framework to describe more complex mechanisms and dynamics behind genes mutations in cancer progression.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Komlan Atitey: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104197.

References

[1] S. Abedon, E. Bartom, Multiplicity of Infection, in: Brenner's Encyclopedia of Genetics, second ed., Elsevier Inc, 2013, pp. 509–510.
[2] D.J. Weatherspoon, A. Chattopadhyay, S. Boroumand, I. Garcia, Oral cavity and oropharyngeal cancer incidence trends and disparities in the United States: 2000–2010, CancerEpidemiol. 39 (2015) 497–504.
[3] A.B. Mariotto, K. Robin Yabroff, Y. Shao, E.J. Feuer, M.L. Brown, Projections of the cost of cancer care in the United States: 2010–2020, J. Natl Cancer Inst. 103 (2011) 117–128.
[4] G. Ciriello, M.L. Miller, B.A. Aksoy, Y. Senbabaoglu, N. Schultz, C. Sander, Emerging landscape of oncogenic signatures across human cancers, Nat. Genet. 45 (2013) 1127–1133.
[5] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (2011) 646–674.
[6] E. Pienaar, J.J. Linderman, D.E. Kirschner, Emergence and selection of isoniazid and rifampin resistance in tuberculosis granulomas, PLoS One. 13 (2018) e0196322.
[7] W.A. Rosche, P.L. Foster, Determining mutation rates in bacterial populations, Methods 20 (2000) 4–17.

[8] M. Stahl, N. Kohrman, S.D. Gore, T.K. Kim, A.M. Zeidan, T. Prebet, Epigenetics in cancer: a hematological perspective, PLoS genetics. 12 (2016) e1006193.
[9] G. Schneider, M. Schmidt-Supprian, R. Rad, D. Saur, Tissue-specific tumorigenesis: context matters, Nat. Rev. Cancer 17 (2017) 239–253.
[10] L.M. Almassalha, G.M. Bauer, J.E. Chandler, S. Gladstein, I. Szleifer, H.K. Roy, et al., The greater genomic landscape: the heterogeneous evolution of cancer, Cancer Res. 76 (2016) 5605–5609.
[11] B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz Jr, K. W. Kinzler, Cancer genome landscapes, Science 339 (2013) 1546–1558.
[12] R.A. Weinberg, Coming full circle—from endless complexity to simplicity and back again, Cell 157 (2014) 267–271.
[13] K. Rezvani, R. Rouce, E. Liu, E. Shpall, Engineering natural killer cells for cancer immunotherapy, Mol. Ther. 25 (2017) 1769–1781.
[14] C. Bianca, M. Delitala, On the modelling of genetic mutations and immune system competition, Comput. Math. Appl. 61 (2011) 2362–2375.
[15] T. Gerashchenko, E. Denisov, N. Litviakov, M. Zavyalova, S. Vtorushin, M. Tsyganov, et al., Intratumor heterogeneity: nature and biological significance, Biochemistry (Moscow). 78 (2013) 1201–1215.
[16] M.S. Lawrence, P. Stojanov, P. Polak, G.V. Kryukov, K. Cibulskis, A. Sivachenko, et al., Mutational heterogeneity in cancer and the search for new cancer-associated genes, Nature 499 (2013) 214–218.
[17] L. Mihaylova, A. Carmi, Particle algorithms for filtering in high dimensional state spaces: A case study in group object tracking, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE, 2011, pp. 5932–5935.
[18] K. Atitey, P. Loskot, L. Mihaylova, Variational Bayesian inference of hidden stochastic processes with unknown parameters, arXiv preprint arXiv:191100757, 2019.
[19] P. Loskot, K. Atitey, L. Mihaylova, Comprehensive review of models and methods for inferences in bio-chemical reaction networks, Front. Genet. 549 (2019).
[20] C.J. Paciorek, M.J. Schervish, Spatial modelling using a new class of nonstationary covariance functions, Environmet.: Official J. Int. Environmet. Soc. 17 (2006) 483–506.
[21] C. Paciorek, M. Schervish, Nonstationary covariance functions for Gaussian process regression, Adv. Neural Informat. Process. Syst. 16 (2003).
[22] A. Schein, H. Wallach, M. Zhou, Poisson-gamma dynamical systems, Adv. Neural Informat. Process. Syst. 29 (2016).
[23] D. Guo, B. Chen, H. Zhang, M. Zhou, Deep Poisson gamma dynamical systems, Adv. Neural Informat. Process. Syst. 31 (2018).
[24] K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat. Biotechnol. 31 (2013) 213–219.
[25] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, et al., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. 22 (2012) 568–576.
[26] J.S. Zielinski, N. Bouaynaya, D. Schonfeld, W. O'Neill, Time-dependent ARMA modeling of genomic sequences, BioMed Central, BMC bioinformatics, 2008, pp. 1–9.
[27] T.E. Ouldridge, DNA nanotechnology: understanding and optimisation through simulation, Mol. Phys. 113 (2015) 1–15.
[28] Y. Chen, J.-X. Shi, X.-F. Pan, J. Feng, H. Zhao, Identification of candidate genes for lung cancer somatic mutation test kits, Gene. Mol. Biol. 36 (2013) 455–464.
[29] L.H. Araujo, P.E. Lammers, V. Matthews-Smith, R. Eisenberg, A. Gonzalez, A. G. Schwartz, et al., Somatic mutation spectrum of non–small-cell lung cancer in African Americans: a pooled analysis, J. Thoracic Oncol. 10 (2015) 1430–1436.
[30] J.D. Campbell, A. Alexandrov, J. Kim, J. Wala, A.H. Berger, C.S. Pedamallu, et al., Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas, Nat. Genet. 48 (2016) 607–616.
[31] N.-C. Huang, J. Aggarwal, On linear shift-variant digital filters, IEEE Trans. Circ. Syst. 27 (1980) 672–679.
[32] D. Möst, D. Keles, A survey of stochastic modelling approaches for liberalised electricity markets, Eur. J. Oper. Res. 207 (2010) 543–556.
[33] A.-H. ZeS, O.M. Alsmadi, A.M. Al-Smadi, M.I. Zaqout, M.S. Saraireh, ARMA model order and parameter estimation using genetic algorithms, Math. Comput. Modell. Dyn. Syst. 18 (2012) 201–221.
[34] Y. Grenier, Time-dependent ARMA modeling of nonstationary signals, IEEE Trans. Acoust. Speech Signal Process. 31 (1983) 899–911.
[35] P.C. Phillips, Trending time series and macroeconomic activity: Some present and future challenges, J.Economet. 100 (2001) 21–27.
[36] F.J. Nogales, A.J. Conejo, Electricity price forecasting through transfer function models, J. Oper. Res. Soc. 57 (2006) 350–356.
[37] O. Løvsletten, Consistency of detrended fluctuation analysis, Phys. Rev. E 96 (2017), 012141.
[38] M. Corona-Ruiz, F. Hernandez-Cabrera, J.R. Cantú-González, O. González-Amezcua, A.F. Javier, A stochastic phylogenetic algorithm for mitochondrial DNA analysis, Front. Genet. 10 (2019) 66.
[39] J.-Y. Chiang, J.-W. Huang, L.-Y. Lin, C.-H. Chang, F.-Y. Chu, Y.-H. Lin, et al., Detrended fluctuation analysis of heart rate dynamics is an important prognostic factor in patients with end-stage renal disease receiving peritoneal dialysis, PloS one. 11 (2016) e0147282.
[40] L. Zunino, D.G. Pérez, A. Kowalski, M. Martín, M. Garavaglia, A. Plastino, et al., Fractional Brownian motion, fractional Gaussian noise, and Tsallis permutation entropy, Physica A: Statistical Mechanics and its Applications. 387 (2008) 6057–6068.
[41] M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley, Stochastic variational inference, J. Mach. Learn. Res. (2013).
[42] V. Jain, F. Koehler, E. Mossel, The mean-field approximation: Information inequalities, algorithms, and complexity, Conf. Learn. Theory: PMLR (2018) 1326–1347.
[43] H. Zhu, H. Leung, Z. He, A variational Bayesian approach to robust sensor fusion based on Student-t distribution, Inf. Sci. 221 (2013) 201–214.
[44] Y. Chen, G. Rangarajan, J. Feng, M. Ding, Analyzing multiple nonlinear time series with extended Granger causality, Phys. Lett. A 324 (2004) 26–35.
[45] K. Thurley, L.F. Wu, S.J. Altschuler, Modeling cell-to-cell communication networks using response-time distributions, Cell Syst. 6 (355–67) (2018) e5.
[46] N. Farsad, N. Shlezinger, A.J. Goldsmith, Y.C. Eldar, Data-driven symbol detection via model-based machine learning. 2021 IEEE Statistical Signal Processing Workshop (SSP), 2021.
[47] N. Bouaynaya, D. Schonfeld, Nonstationary analysis of coding and noncoding regions in nucleotide sequences, IEEE J. Sel. Top. Signal Process. 2 (2008) 357–364.
[48] F.M. Shapter, D.L. Waters, Genome walking, Springer, Cereal Genomics, 2014, pp. 133–146.
[49] A.P. Singh, S. Mishra, S. Jabin, Sequence based prediction of enhancer regions from DNA random walk, Sci. Rep. 8 (2018) 1–12.
[50] T. Hai, M.G. Hartman, The molecular biology and nomenclature of the activating transcription factor/cAMP responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis, Gene 273 (2001) 1–11.
[51] N.A. Skrypina, A.V. Timofeeva, G.L. Khaspekov, L.P. Savochkina, R. S. Beabealashvilli, Total RNA suitable for molecular biology analysis, J. Biotechnol. 105 (2003) 1–9.
[52] N. Auslander, Y.I. Wolf, E.V. Koonin, In silico learning of tumor evolution through mutational time series, Proc. Natl. Acad. Sci. 116 (2019) 9501–9510.
[53] L. Natarajan, C.C. Berry, C. Gasche, Estimation of spontaneous mutation rates, Biometrics. 59 (2003) 555–561.
[54] H.R. Johnston, B.J. Keats, S.L. Sherman, Population genetics. Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics, Elsevier (2019) 359–373.
[55] A. Mogi, H. Kuwano, TP53 mutations in nonsmall cell lung cancer, J. Biomed. Biotechnol. 2011 (2011).
[56] S. Toyooka, T. Tsuda, A.F. Gazdar, The TP53 gene, tobacco exposure, and lung cancer, Hum. Mutat. 21 (2003) 229–239.
[57] C. Wang, S. Zhang, B. Ma, Y. Fu, Y. Luo, TP53 mutations upregulate RCP expression via Sp1/3 to drive lung cancer progression, Oncogene 41 (2022) 2357–2371.
[58] J. Zhu, M.A. Sammons, G. Donahue, Z. Dou, M. Vedadi, M. Getlik, et al., Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth, Nature 525 (2015) 206–211.