Alshawaqfeh *et al. BMC Genomics* 2017, **18**(Suppl 3):228
DOI 10.1186/s12864-017-3605-x

BMC Genomics

RESEARCH Open Access

Inferring microbial interaction networks from metagenomic data using SgLV-EKF algorithm

Mustafa Alshawaqfeh[1*], Erchin Serpedin[1] and Ahmad Bani Younes[2]

From Third International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2016) Seattle, WA, USA. 02-Oct-16

Abstract

Background: Inferring the microbial interaction networks (MINs) and modeling their dynamics are critical in understanding the mechanisms of the bacterial ecosystem and designing antibiotic and/or probiotic therapies. Recently, several approaches were proposed to infer MINs using the generalized Lotka-Volterra (gLV) model. Main drawbacks of these models include the fact that these models only consider the measurement noise without taking into consideration the uncertainties in the underlying dynamics. Furthermore, inferring the MIN is characterized by the limited number of observations and nonlinearity in the regulatory mechanisms. Therefore, novel estimation techniques are needed to address these challenges.

Results: This work proposes SgLV-EKF: a stochastic gLV model that adopts the extended Kalman filter (EKF) algorithm to model the MIN dynamics. In particular, SgLV-EKF employs a stochastic modeling of the MIN by adding a noise term to the dynamical model to compensate for modeling uncertainties. This stochastic modeling is more realistic than the conventional gLV model which assumes that the MIN dynamics are perfectly governed by the gLV equations. After specifying the stochastic model structure, we propose the EKF to estimate the MIN. SgLV-EKF was compared with two similarity-based algorithms, one algorithm from the integral-based family and two regression-based algorithms, in terms of the achieved performance on two synthetic data-sets and two real data-sets. The first data-set models the randomness in measurement data, whereas, the second data-set incorporates uncertainties in the underlying dynamics. The real data-sets are provided by a recent study pertaining to an antibiotic-mediated Clostridium difficile infection. The experimental results demonstrate that SgLV-EKF outperforms the alternative methods in terms of robustness to measurement noise, modeling errors, and tracking the dynamics of the MIN.

Conclusions: Performance analysis demonstrates that the proposed SgLV-EKF algorithm represents a powerful and reliable tool to infer MINs and track their dynamics.

Keywords: Microbial interaction network, Extended Kalman filter, Metagenomics, SgLV-EKF algorithm

Background

The microbiota, a conglomeration of all the bacteria living on/in the human body, is now being extensively studied in order to understand its relevance to the host. Interestingly, it has been suggested in several works that the maintenance of a stable microbial ecosystem is necessary for a healthy life [1]. For instance, a disruption of the stable state of the microbiome, referred to as 'dysbiosis', is directly linked to obesity [2–4], diabetes [5], inflammatory bowel disease (IBD) [6] and cancer [7, 8].

Even though the bacteria have been recognized as playing a key role in defining the health and disease states, their study has represented a challenge in the past due to several reasons. First, the bacteria were mainly studied through cultivation. Many bacterial groups were neither known earlier nor cultivated in a large number

*Correspondence: mustafa.shawaqfeh@tamu.edu
[1]Bioinformatics and Genomic Signal Processing Lab, ECEN Dept., Texas A&M University, College Station, 77843-3128 TX, USA
Full list of author information is available at the end of the article

in a laboratory setting. Second, in vitro measurements do not match real in vivo values because the laboratory conditions do not match the environment of the host [9]. However, recent advances in high-throughput sequencing have overcome these limitations. At present, the sequencing technologies provide the researchers with cross-sectional and longitudinal microbial compositions in different environments.

In particular, longitudinal microbial studies are important because they offer an insight into the dynamics of the bacterial community and its response to external perturbations [10]. In addition to its importance to understand the variations in bacterial populations, such observational studies are promising to discover the regulation mechanisms which are essential to identify bacterial groups that may cause or protect against diseases [11]. Therefore, time series analysis tools are crucial to exploit the temporal information embedded into the time series data.

Bacterial communities comprise a vast number of species with complex relationships including mutualism, competition, parasitism, commensalism, amenalism and neutralism [12]. These interactions can be mediated by natural competition for space and resources or via some symbiotic relationships. For example, substances secreted by one species may be metabolized by another [13, 14]. Additionally, members of bacterial communities can interact indirectly through the immune system [15]. Identifying these interactions is crucial to understand the ecological communities and the underlying regulation activities between microbes. For example, the depletion of a species may affect other species that depend on it for their survival. As an additional example, the oppositional and symbiotic interactions between species contribute to the development and resistance of pathogens [16].

Various methods have been proposed to infer the microbial interaction network (MIN) [12, 17]. These methods can be broadly divided into similarity-based methods and dynamic-based methods. Similarity-based approaches employ a similarity measure to score the pairwise relationship between each pair of microbes. Two microbes are considered to have an interaction if the pairwise similarity score exceeds a predefined threshold. Popular methodologies for constructing similarity-based networks are the correlation coefficient and local similarity analysis (LSA) [18–22]. While these methods are computationally efficient, they present several drawbacks. Firstly, they identify only pairwise relations. Therefore, complex interactions in microbial communities are not captured. Secondly, similarity-based networks are undirected. This means that the inferred interactions are assumed to be bidirectional with equal strengths. However, this represents an invalid biological assumption. Third, a similarity-based approach treats the time series data as a static snapshot, and hence it ignores the temporal dependencies.

On the other hand, dynamic methods overcome these drawbacks and go beyond identifying only the interaction network to build predictive models that enable tracking the bacterial composition over time and their response to external perturbations [12]. Constructing such a model presents two major phases: (*i*) model selection phase, which aims to determine a set of equations to identify the structure of the system; (*ii*) parameter estimation phase, or commonly referred to as system identification, which determines the unknown parameters of the model from the observed data. A common approach in dynamical modeling is to use ordinary differential equations (ODEs). An example of ODE-based dynamical models that have been employed to characterize the microbial interaction network is the generalized Lotka-Volterra (gLV) model [11, 23–25]. gLV has been extensively used due to the following two main features of gLV equations. Firstly, the model parameters directly capture the growth rates and pairwise interactions between all species in the system. Secondly, the gLV model can be extended to account for external stimuli such as the introduction of probiotics, antibiotics or changes in diet [11]. However, ODE-based models consider only the uncertainty caused by the noise in the measurements. Therefore, the randomness in the dynamical model is not considered by such models.

In general, estimating the unknown parameters is embedded within the optimization framework that aims to minimize the error between the model's output and the experimental data. The proposed optimization techniques are broadly divided into integral-based methods and regression-based methods. Integral-based methods are iterative algorithms that search the parameter space for an optimal set. At each iteration, the ODEs are solved via numerical integration to compute the difference between the model output and the available data. The primary drawbacks of integral-based algorithms are the computational burden required to solve the ODEs and the convergence failure due to the integration breakdown [26].

To reduce the computational complexity, regression techniques approximate the derivative terms in the ODE model from the observed data, thereby, converting the ODEs into a regular multivariate regression system. For instance, the parameters of a linearized (via logarithmic transformation) version of the gLV model were estimated by the ridge regression in [11] and the sparse linear regression in [24]. The linearization step restricts the bacterial abundance levels to be strictly positive. This assumption is biologically invalid since it is possible that some bacteria may be totally depleted in some samples. Generally, regression-based methods are computationally efficient and scalable for very large dimensional data [27, 28].

However, their performance relies on the accuracy of the estimated derivatives. Therefore, without a proper denoising preprocessing step, the slope approximation may perform poorly due to the overfitting problem [29]. Additionally, for fast varying observations, an intelligent algorithm is required to track the variation in data and provide an accurate estimate of the derivatives. Estimating the model's parameters is a challenging task due to the following factors: (a) The number of the unknown parameters is much larger than the available observations; (b) The underlying regulation mechanisms that govern the microbial interaction network are nonlinear. The aforementioned literature about inferring the MIN from time series data has not specifically dealt with these two challenges.

To address the challenges mentioned above, we propose a stochastic-based dynamical model that encodes the uncertainties in both the measurements and the dynamics to compensate for modeling errors and capture the complex interactions among the microbiota. Moreover, we propose EKF to jointly estimate the states of the stochastic model and its parameters. EKF is selected because of the following two features. First, EKF can handle the nonlinearities in the dynamic model or the observation model or both via linearization about the current mean and variance. Second, EKF performs the estimation recursively which renders the EKF as a suitable approach for inferring a large number of parameters from a limited number of observations [30]. Although EKF has had success in several biological applications such as gene regulatory networks, signaling pathways and metabolic networks [31–33], it has not been applied to estimate the microbial interaction network from metagenomic time series data. We refer to the combination of the stochastic gLV model with EKF to estimate its parameters as the SgLV-EKF algorithm. The main contributions of this work can be summarized as:

- We improve the conventional modeling of MINs from a nonlinear ODE dynamic model to a more general nonlinear stochastic model to compensate for uncertainties in the model and/or observations.
- We propose the EKF, which has not been proposed in the context of microbial interaction networks, to infer the bacterial interaction network. The EKF is selected due to its inherent ability to estimate the parameters of nonlinear interactions from limited number of observations.
- Comprehensive simulation studies corroborate the fact that the proposed approach outperforms Nelder and Stein's algorithm in terms of robustness to measurement noise, modeling errors, computational efficiency, and tracking the dynamics of the microbial interaction network.

Methods

System model

In this paper, the MIN is modeled as a nonlinear dynamic stochastic system that captures the dynamics of the bacterial abundance level as follows:

$$\begin{aligned} x_i(k+1) &= f_i(\boldsymbol{x}(k)) + w_i(k), \\ y_i(k) &= x_i(k) + v_i(k), \end{aligned} \tag{1}$$

where $i = 1,\ldots,n$ is the state index, $k = 1,\ldots,M$ represents the time-step, M is number of measurement time points, $\boldsymbol{x}(k) \in \Re^n$ denotes the system state vector, and $\boldsymbol{y}(k) \in \Re^n$ stands for the observation vector. In particular, $y_i(k)$ and $x_i(k)$ represent the measured and the actual relative abundance level of the i^{th} bacteria at time k, respectively. The microbial interaction network containing n bacteria is described by the nonlinear function $\boldsymbol{f} = \left[f_1, f_2, \ldots, f_n\right]^T$, where f_i is defined in terms of the discrete-time differential equation (4). Variables $\boldsymbol{w}(k) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(k))$ and $\boldsymbol{v}(k) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}(k))$ represent the zero-mean white Gaussian process noise and measurements noise, respectively, with covariance matrices given by

$$\begin{aligned} E\left\{\boldsymbol{w}(k)\boldsymbol{w}^T(j)\right\} &= \boldsymbol{Q}(k)\delta_{kj}, \\ E\left\{\boldsymbol{v}(k)\boldsymbol{v}^T(j)\right\} &= \boldsymbol{R}(k)\delta_{kj}, \\ E\left\{\boldsymbol{w}(k)\boldsymbol{v}^T(j)\right\} &= \boldsymbol{0} \end{aligned} \tag{2}$$

where $E\{.\}$ denotes the expectation operator and δ_{kj} denotes the Kronecker delta function:

$$\delta_{kj} = \begin{cases} 0 \ \textit{if} \ k \neq j \\ 1 \ \textit{if} \ k = j \end{cases}. \tag{3}$$

Generalized Lotka-Volterra model

The gLV model is a first order nonlinear system of differential equations. In its discrete form, the gLV is represented as a group of first order nonlinear difference equations that relate the dissimilarity between the abundance levels of species at time t with respect to time $t-1$.

Let $\{x_i(t); i = 1,\ldots,n\}$ be the relative abundance level of the i^{th} bacteria at time t whose intrinsic growth rate is g_i. Moreover, let c_{ij} represent the strength of the influence of microbe i onto bacteria j (a.k.a., the 'interaction coefficient'). The gLV model is defined by means of the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = g_i x_i(t) + x_i(t)\sum_{j=1}^{n} c_{ij}x_j(t). \tag{4}$$

The above framework was extended to model the effects of external perturbations (e.g., antibiotics, diets) onto the microbial community structure [11]. This was obtained by adding another term to (4) which modulates the influence of each stimulating source into each member of the ecosystem. Mathematically, let ϵ_{il} represent the 'sensitivity' of the i^{th} microbe in response to the l^{th} stimuli with

signal strength u_l. The resulting gLV model is captured by [11]:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = g_i x_i(t) + x_i(t)\sum_{j=1}^{n} c_{ij}x_j(t) + x_i(t)\sum_{l=1}^{L}\epsilon_{il}u_l(t). \quad (5)$$

We remark in passing that a simplified gLV model was previously employed in [24] to characterize the dynamics of the gut microbiome considering only the interaction between various species. Particularly, the simplistic gLV model is formulated as:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = x_i(t)\sum_{j=1}^{n} c_{ij}x_j(t), \quad (6)$$

where the intrinsic growth rate is ignored compared to (4).

Kalman filter and extended Kalman filter

This section reviews the key features of the Kalman filter and then focuses on formulating the EKF for estimating both the states and parameters of the state space model.

Kalman filter

Under certain conditions, e.g., linearity of model and Gaussian noise, the Kalman filter represents an optimal filter of the system state in the presence of measurement errors. Let assume that the dynamics of a discrete-time system is governed by the following linear model:

$$\boldsymbol{x}(k+1) = \boldsymbol{\Phi}_k\boldsymbol{x}(k) + \boldsymbol{\Gamma}_k\boldsymbol{u}(k) + \boldsymbol{\Lambda}_k\boldsymbol{w}(k), \quad (7)$$

and the observation model is given by

$$\boldsymbol{y}(k) = \boldsymbol{\Psi}_k\boldsymbol{x}(k) + \boldsymbol{\nu}(k), \quad (8)$$

where k is a time-step index, $\boldsymbol{x}(k) \in \Re^n$ represents the system state vector, and $\boldsymbol{y}(k) \in \Re^n$ stands for the observation vector. The variable n denotes the number of states. Variables $\boldsymbol{w}(k) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(k))$ and $\boldsymbol{\nu}(k) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}(k))$ represent the zero-mean multivariate Gaussian noise in the process and measurements, respectively. The initial state, and the noise vectors at each step are all assumed to be mutually independent.

The discrete-time Kalman filter assumes the following steps:

- *Initialization*: at $k = 0$ and for given initial states $\hat{\boldsymbol{x}}^-(0) = \boldsymbol{x}_0$, the initial value of the covariance matrix is given by:

$$\begin{aligned}\boldsymbol{P}^-(0) =& \boldsymbol{P}_{x_0x_0} = \\ & E\left\{(\boldsymbol{x}(0) - \boldsymbol{x}_0)(\boldsymbol{x}(0) - \boldsymbol{x}_0)^T\right\},\end{aligned} \quad (9)$$

  where the superscript $(-)$ denotes a-priori value.

- *Gain*: compute the Kalman gain matrix

$$\boldsymbol{K}(k) = \boldsymbol{P}^-(k)\boldsymbol{\Psi}_k^T\left[\boldsymbol{\Psi}_k\boldsymbol{P}^-(k)\boldsymbol{\Psi}_k^T + \boldsymbol{R}(k)\right]^{-1}. \quad (10)$$

- *Update*: update the state estimate $\hat{\boldsymbol{x}}^+(k)$ and covariance $\boldsymbol{P}^+(k)$ at each measurement

$$\begin{aligned}\hat{\boldsymbol{x}}^+(k) &= \hat{\boldsymbol{x}}^-(k) + \boldsymbol{K}(k)\left[\boldsymbol{y}(k) - \boldsymbol{\Psi}_k\hat{\boldsymbol{x}}^-\right], \\ \boldsymbol{P}^+(k) &= \left[\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{\Psi}_k\right]\boldsymbol{P}^-(k),\end{aligned} \quad (11)$$

  where the superscript $(+)$ denotes the posteriori value.

- *Propagation*: propagate both the state estimate $\hat{\boldsymbol{x}}(k)$ and covariance $\boldsymbol{P}(k)$ using the posteriori estimate $\hat{\boldsymbol{x}}^+(k)$ and posteriori covariance $\boldsymbol{P}^+(k)$

$$\begin{aligned}\hat{\boldsymbol{x}}^-(k+1) &= \boldsymbol{\Phi}_k\hat{\boldsymbol{x}}^+(k) + \boldsymbol{\Gamma}_k\boldsymbol{u}(k), \\ \boldsymbol{P}^-(k+1) &= \boldsymbol{\Phi}_k\boldsymbol{P}^+(k)\boldsymbol{\Phi}_k^T + \boldsymbol{\Lambda}_k\boldsymbol{Q}(k)\boldsymbol{\Lambda}_k^T.\end{aligned} \quad (12)$$

Extended Kalman filter for parameter estimation

The Kalman filter is the optimum state estimator for a linear state space model observed in Gaussian noise. However, most of the biological systems are nonlinear. This renders the Kalman filter inapplicable in such scenarios. To overcome this challenge, one possible solution is to linearize the nonlinear dynamic system before applying the Kalman filter. This process of approximating the nonlinear system with a linear one while using the Kalman filter results in the EKF. It is worth to mention that although EKF is not necessarily optimal, it was adopted as a standard method to deal with nonlinear systems. The classical extended Kalman filter's domain of convergence depends on the region where the first-order Taylor series linearization adequately approximates the nonlinear dynamics of the system. Therefore, the initializing stage requires the initial state estimate be close enough to the true state.

The general structure of the EKF is to estimate the state vector by minimizing the system variance error. Another useful application of EKF is to estimate the unknown system parameters. Augmenting the state vector to include the unknown parameters as additional states enables an efficient system identification method for nonlinear systems. The same solution is applicable to systems with uncertain parameters but it may lead to poor performance in the estimation process. The augmented system decreases the estimation error caused by imperfect model parameters. Consider the following state space model:

$$\begin{aligned}\boldsymbol{x}(k+1) &= \boldsymbol{f}(\boldsymbol{x}(k);\boldsymbol{\theta}) + \boldsymbol{w}(k), \\ \boldsymbol{y}(k) &= \boldsymbol{h}(k) + \boldsymbol{\nu}(k),\end{aligned} \quad (13)$$

where k is a time index, $\boldsymbol{x} \in \Re^n$ represents the system state vector, $\boldsymbol{y} \in \Re^m$ stands for the observation vector, $\boldsymbol{w} \in \Re^n$ and $\boldsymbol{\nu} \in \Re^m$ denote the system noise

and the measurement noise, respectively. $\boldsymbol{w} \in \Re^n$ and $\boldsymbol{v} \in \Re^m$ are zero-mean white Gaussian stochastic processes with covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively. The dynamic evolution and measurements of the system are governed by the nonlinear functions $\boldsymbol{f} : \Re^n \rightarrow \Re^n$ and $\boldsymbol{h} : \Re^n \rightarrow \Re^m$, respectively, with $\boldsymbol{\theta}$ representing the parameters of the dynamic model. Variables n and m stand for the number of states and number of measurements, respectively.

Let $\boldsymbol{z}$ denote the augmented state vector that includes the parameters of the model as additional states. The vector $\boldsymbol{z}$ is give by:

$$\boldsymbol{z}(k) = \begin{bmatrix} \boldsymbol{x}(k) \\ \boldsymbol{\theta}(k) \end{bmatrix}. \tag{14}$$

The augmented version of the state space model given in Eq. (13) takes the form:

$$\begin{aligned} \boldsymbol{z}(k+1) &= \begin{bmatrix} \boldsymbol{x}(k+1) \\ \boldsymbol{\theta}(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}(k)) \\ \boldsymbol{\theta}(k) \end{bmatrix} + \begin{bmatrix} \boldsymbol{w}(k) \\ \boldsymbol{\eta}(k) \end{bmatrix} \\ &= \boldsymbol{F}(\boldsymbol{z}(k)) + \boldsymbol{\zeta}(k), \\ \boldsymbol{y}(k) &= \boldsymbol{x}(k) + \boldsymbol{\nu}(k), \end{aligned} \tag{15}$$

where $\boldsymbol{\zeta}(k)$ denotes the zero-mean Gaussian white-noise for the augmented dynamic defined by $\boldsymbol{F}$. Constructing the augmented model in (15) assumes that the system parameters are constant (i.e., $\boldsymbol{\theta}(k) = \boldsymbol{\theta}$). Once the augmented state equations are constructed, the standard EKF can be implemented to estimate the states of the augmented system (i.e., $\boldsymbol{z}$), which enables the joint estimation the model states $\boldsymbol{x}$ and its parameters $\boldsymbol{\theta}$. For detailed derivations of the EKF, in both discrete-time and continuous time forms, the authors recommend [34]. The following steps summarize the implementation of EKF:

- *Initialization*: at $k = 0$ and for given initial states $\boldsymbol{z}_0 = [\boldsymbol{x}_0, \boldsymbol{\theta}_0]^T$, the initial value of the covariance matrix is given by:

$$\boldsymbol{P}_0 = \begin{bmatrix} \boldsymbol{P}_{x_0 x_0} & \boldsymbol{P}_{x_0 \theta_0} \\ \boldsymbol{P}_{\theta_0 x_0} & \boldsymbol{P}_{\theta_0 \theta_0} \end{bmatrix}. \tag{16}$$

and

$$\begin{aligned} \hat{\boldsymbol{z}}^-(0) &= E\{\boldsymbol{z}(0)\} = \boldsymbol{z}_0, \\ \boldsymbol{P}^-(0) &= E\left\{(\boldsymbol{z}(0) - \boldsymbol{z}_0)(\boldsymbol{z}(0) - \boldsymbol{z}_0)^T\right\} = \boldsymbol{P}_0. \end{aligned} \tag{17}$$

The initial covariance matrices are given by

$$\begin{aligned} \boldsymbol{P}_{x_0 x_0} &= E\left\{(\boldsymbol{x}(0) - \boldsymbol{x}_0)(\boldsymbol{x}(0) - \boldsymbol{x}_0)^T\right\}, \\ \boldsymbol{P}_{x_0 \theta_0} &= E\left\{(\boldsymbol{x}(0) - \boldsymbol{x}_0)(\boldsymbol{\theta}(0) - \boldsymbol{\theta}_0)^T\right\}, \\ \boldsymbol{P}_{\theta_0 x_0} &= E\left\{(\boldsymbol{\theta}(0) - \boldsymbol{\theta}_0)(\boldsymbol{x}(0) - \boldsymbol{x}_0)^T\right\}, \\ \boldsymbol{P}_{\theta_0 \theta_0} &= E\left\{(\boldsymbol{\theta}(0) - \boldsymbol{\theta}_0)(\boldsymbol{\theta}(0) - \boldsymbol{\theta}_0)^T\right\}. \end{aligned} \tag{18}$$

Assume that $\boldsymbol{z}(0) \sim \mathcal{N}(\boldsymbol{0}, P(0))$.

- *Gain*: compute the Kalman gain matrix

$$\begin{aligned} \boldsymbol{K}(k) &= \boldsymbol{P}^-(k)\boldsymbol{H}^T(\hat{\boldsymbol{z}}^-(k)) \\ &\left[\boldsymbol{H}(\hat{\boldsymbol{z}}^-(k))\boldsymbol{P}^-(k)\boldsymbol{H}^T(\hat{\boldsymbol{z}}^-(k) + \boldsymbol{R}(k)\right]^{-1}, \end{aligned} \tag{19}$$

where $\boldsymbol{H}\left(\hat{\boldsymbol{z}}^-(k)\right) \equiv \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}|_{\hat{\boldsymbol{z}}^-(k)}$.

- *Update*: update the state estimate $\hat{\boldsymbol{z}}^+(k)$ and covariance $\boldsymbol{P}^+(k)$ at each measurement

$$\begin{aligned} \hat{\boldsymbol{z}}^+(k) &= \hat{\boldsymbol{z}}^-(k) + \\ &\quad \boldsymbol{K}(k)\left[\boldsymbol{y}(k) - \boldsymbol{h}(\hat{\boldsymbol{z}}^-(k))\right], \\ \boldsymbol{P}^+(k) &= \left[\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{H}(\hat{\boldsymbol{z}}^-(k))\right]\boldsymbol{P}^-(k). \end{aligned} \tag{20}$$

- *Propagation*: propagate both the state estimate $\hat{\boldsymbol{z}}(k)$ and covariance $\boldsymbol{P}(k)$ using the posteriori estimate $\hat{\boldsymbol{z}}^+(k)$ and posteriori covariance $\boldsymbol{P}^+(k)$

$$\begin{aligned} \hat{\boldsymbol{z}}^-(k+1) &= \boldsymbol{F}\left(\hat{\boldsymbol{z}}^+(k)\right), \\ \boldsymbol{P}^-(k+1) &= \boldsymbol{\Omega}\left(\hat{\boldsymbol{z}}^+(k)\right)\boldsymbol{P}^+(k)\boldsymbol{\Omega}^T\left(\hat{\boldsymbol{z}}^+(k)\right) \\ &\quad + \boldsymbol{Q}(k), \end{aligned} \tag{21}$$

where $\boldsymbol{\Omega}\left(\hat{\boldsymbol{z}}^+(k)\right) \equiv \frac{\partial \boldsymbol{F}(\boldsymbol{z})}{\partial \boldsymbol{z}}|_{\hat{\boldsymbol{z}}^+(k)}$.

For our model of MIN given in Eq. (1), the system dynamics (i.e., $\boldsymbol{f}$) is depicted by the gLV model defined in Eq. (4) and the observation model $\boldsymbol{h}$ is given by the identity function (i.e., $\boldsymbol{h}(\boldsymbol{z}(k)) = \boldsymbol{x}(k)$). The system parameters vector $\boldsymbol{\theta}$ captures the intrinsic growth rates and all the pairwise interaction coefficients between the n bacteria included in the gLV model. In particular, $\boldsymbol{\theta}$ is given by:

$$\boldsymbol{\theta} = \left[g_1, g_2, \ldots g_n, c_{11}, c_{12}, \ldots, c_{nn}\right]^T. \tag{22}$$

Results and discussion

In this section, we compared SgLV-EKF with the current state-of-the-art algorithms proposed for inferring the microbial interaction network using the gLV model. In particular, EKF is compared with two similarity-based algorithms, one algorithm from the integral-based family, and two regression-based algorithms. The first similarity-based algorithm utilizes the Pearson correlation coefficient (PCC) [18], whereas the second algorithm employs the local similarity analysis [22] to quantify the similarity between time series data. For the integral-based algorithm, the gradient free Nelder-Mead algorithm [35] is used to span the parameter space for the optimal solution. For the regression-based techniques, the first regression-based algorithm was developed by Stein et al. in [11] and it employs the regularized linear regression to infer the MIN. We refer to this algorithm as the Stein's algorithm. The second regression-based algorithm is called the learning interactions from microbial time series (LIMITS) algorithm. This algorithm was proposed in [24] and it is based on the sparse linear regression model. It is important to mention that Stein's algorithm involves Tikhonov

regularization parameters. These parameters were set to the same values used in [11]. All the experiments were performed on a Windows 8.1 system with a 3.4 GHz Intel Core i7 processor on a Matlab 8.3.0.

Synthetic data

The MIN inference algorithms are evaluated in their ability to predict: (a) MIN; (b) Variation of the bacterial abundance levels over the time (i.e., states of the dynamic model). An important metric of any interaction network is its ability to recover the topology/structure of the simulated interaction network. Specifically, the accuracy, sensitivity, and specificity of the MIN inference algorithms in predicting the presence and/or absence of interactions. Moreover, to evaluate the ability of the MIN inference algorithms in predicting the dynamic of the bacterial system, we use the relative mean square error (MSE) as a fidelity criterion to measure the error between the observed data and the estimated bacterial abundances. In our evaluation, we define the true positive (TP) as the number of edges that are truly detected, and the false negative (FN) as the number of edges that are not detected. Similarly, if no edges are present, the number of times the algorithm mistakenly predicts the presence of an edge is defined as the false positive (FP). Otherwise, the number of times that the algorithm truly predicts the absence of an edge is defined as the true negative (TN). Sensitivity and specificity are defined as TP/(TP + FN) and TN/(TN + FP), respectively. And accuracy is defined as (TP + TN)/(TP + FN + TN + FP). Ideally, the values of sensitivity, specificity and accuracy are one. An algorithm with low sensitivity value indicates that this algorithm fails in predicting the existing edges (i.e., interactions) in the network. On the other hand, an algorithm with low specificity performance implies that the algorithm suggests the presence of edges that don't exist in reality. We assume the absence of interaction if the absolute value of the interaction strength is less than one tenth the average of the absolute values of the nonzero elements in the simulated network (i.e., $|c_{ij}| < 0.1$).

In order to evaluate the performance of our proposed scheme, a microbial community consisting of 10 bacteria is simulated. A number of 30 time series points (i.e., the microbial abundance levels) are generated using the stochastic gLV model (Eq. 1) with the parameters shown in Fig. 1a. In our simulations, we perform 100 Monte-Carlo simulations, and we present the average of these experiments.

It is pertinent to mention that comparisons with similarity-based methods are limited to the evaluation of the efficiency of the algorithms to identify the presence and/or absence of interactions in the simulated networks for various dynamic/measurement noise levels. This is because similarity methods don't include a mathematical modeling of the microbial community. Hence, similarity methods don't enable predicting the temporal bacterial abundance profiles.

Figure 1b shows the inferred values using EKF for the growth rate and the interaction network of the simulated microbial system. The small differences between the true values of system parameters and the inferred parameters using EKF point out that the proposed EKF-based approach is accurate in terms of estimating the true system parameters.

The performance metrics mentioned above are evaluated under the following simulation set-ups: (a) Measurement noise level (i.e., σ_v^2); (b) System noise level (i.e., σ_w^2).

Varying the measurement noise level σ_v^2

First, the six algorithms are tested when varying the Gaussian noise levels in the observed data with variance ranging from 10^{-4} to 10^{-1}. The performance of the six algorithms in terms of their abilities to identify the simulated network is depicted in Fig. 2. As it is clearly depicted by this figure, increasing the noise

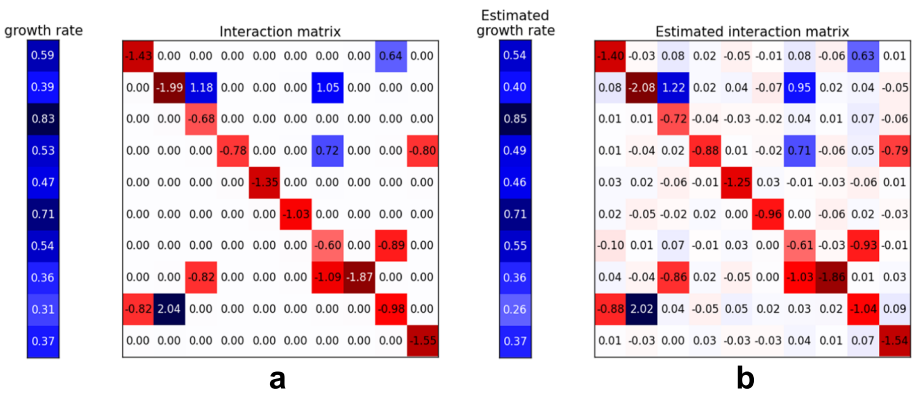

Fig. 1 Simulation of microbial regulatory network. **a** The true values of the model parameters. **b** The inferred parameters

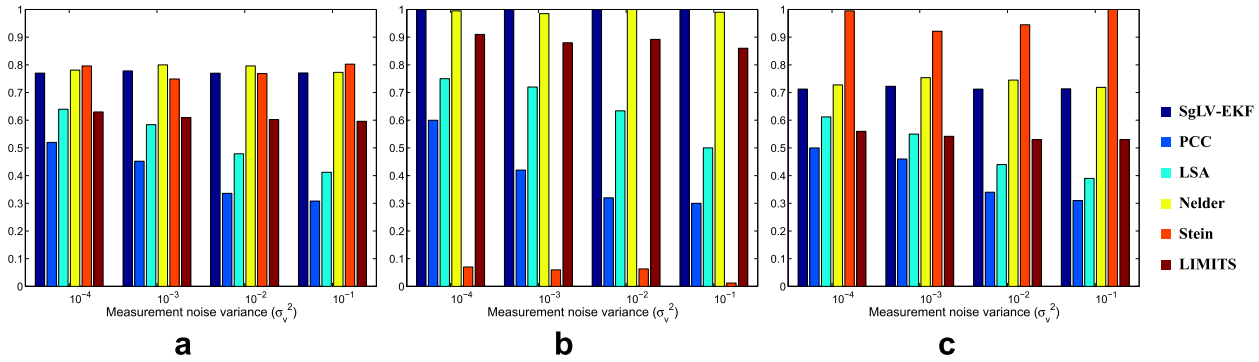

Fig. 2 Evaluation measurements of the efficiency of the six algorithms to identify the presence and absence of interactions in the simulated networks for various measurement noise levels in terms of **a** accuracy, **b** sensitivity and **c** specificity

variance has a slight effect on the performance of the SgLV-EKF, the Nelder and the two regression-based algorithms. On the other hand, the performance of the two similarity-based algorithms (i.e., PCC and LSA) degrades significantly by increasing the noise power. Moreover, similarity-based algorithms show the least accurate performance compared to the other algorithms. The performance of similarity-based techniques may be attributed to two main reasons. The first reason is that the abundance profiles of two microorganisms may be correlated even they don't interact directly. For example, if bacteria A and B do not assume a direct interaction, but both of them rely on the products of bacteria C, then the abundance profiles of A and B are expected to be correlated. The second reason is that the bacterial abundance data provided by the sequencing-based techniques represent the relative fraction of the bacterial abundances rather than their absolute abundances. This compositional nature of the bacterial profiles can lead to unreliable results [36].

For the regression-based algorithms, Stein's algorithm fails to detect the majority of the interactions as depicted from the very low sensitivity values in Fig. 2b, whereas LIMITS algorithm provides more reliable results with consistence accuracy performance around 60%. However, SgLV-EKF outperforms both Stein's and LIMITS algorithms. SgLV-EKF and Nelder's algorithm yield close and stable results over different variance values. However, the execution time of Nelder algorithm is approximately 80 times higher than the SgLV-EKF execution time as it is pointed out by Table 1.

The relative MSE of the predicted bacterial abundance levels is depicted in Fig. 3. The noise level has a negligible effect on the relative MSE of SgLV-EKF and Nelder's algorithm. Both SgLV-EKF and Nelder's algorithm exhibit low MSE errors. The estimated parameters resulted from both Stein's and LIMITS methods lie in the unstable region of the dynamic system. Therefore, they present an infinite MSE error.

Table 1 Average execution time for various methods (seconds)

	SgLV-EKF	Nelder	Stein	LIMITS
Execution Time	2.11	161.62	0.06	1.4

Varying the dynamic noise level σ_w^2

This section evaluates the algorithms performance against uncertainties in the dynamic model. This uncertainty is modeled by a zero mean white Gaussian noise with variance varying from 10^{-7} to 10^{-1}.

Figure 4 presents the accuracy, sensitivity, and specificity of the six MIN inference algorithms. Since our scheme takes into account the randomness in the dynamic model, SgLV-EKF outperforms the other five algorithms in identifying the structure of the interaction network. Moreover, SgLV-EKF provides a robust and reliable performance against the uncertainty in the dynamic model and it exhibits an average accuracy higher than than 75%.

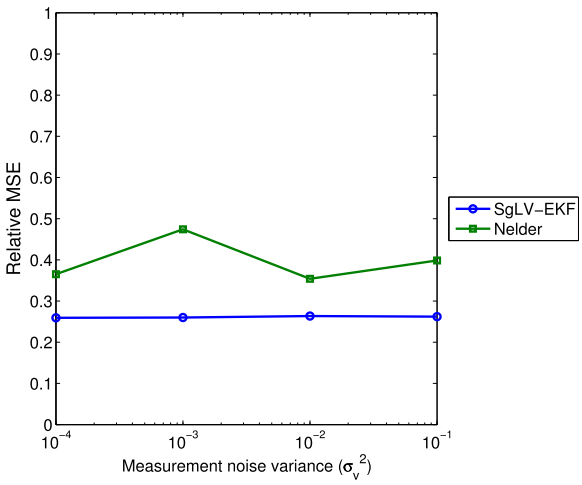

Fig. 3 Relative MSE performance under different measurement noise levels

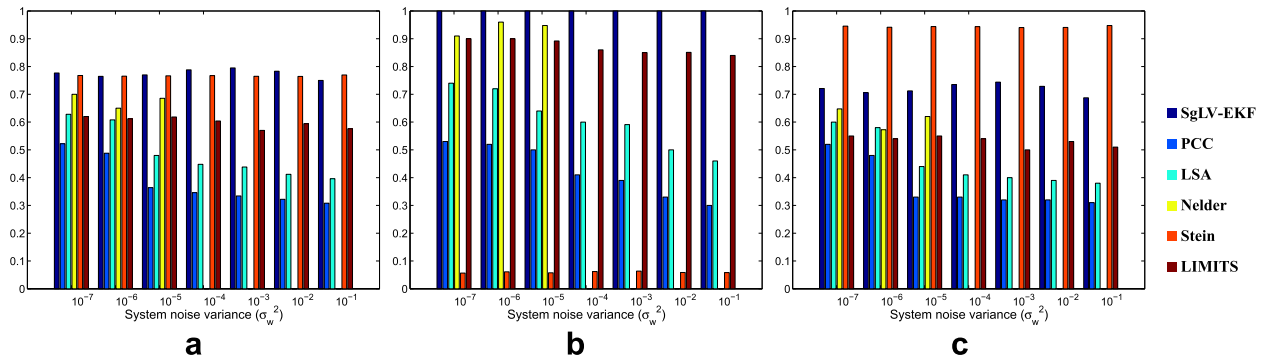

Fig. 4 Evaluation measurements of the efficiency of the six algorithms to identify the presence and absence of interactions in the simulated networks for various dynamic noise levels in terms of **a** accuracy, **b** sensitivity and **c** specificity

On the other hand, due to the presence of a small amount of noise in the dynamic model, the estimation using the other five algorithms is unreliable and inconsistent. In particular, the two similarity-based algorithms show a significant reduction in their accuracy performance due to the increase in the process noise power. For example, for noise level exceeding 10^{-4}, PCC and LSA achieve an average accuracy of only 45% and 35%, respectively. Similar to the results in the previous section, Stein's method failed in inferring the existing interactions as illustrated by the very low sensitivity values in Fig. 4b. For noise power values larger than 10^{-4}, Nelder's method diverged and failed in providing any estimate of the model's parameters. The divergence of Nelder's algorithm combined with the robust performance of the SgLV-EKF algorithm justify our approach of replacing the conventional ODE-based gLV model with a stochastic gLV model that accounts for uncertainties in the system model.

Figure 5 presents the relative MSE for the SgLV-EKF and Nelder's algorithms. The results of Stein's and LIMITS algorithms are omitted here for the same reason mentioned in the previous section. SgLV-EKF shows a consistent performance against system noise.

Real data

To further demonstrate the capability of SgLV-EKF algorithm in inferring the microbial interaction networks, we considered two realistic time series datasets. Recently, an investigation to assess the effect of antibiotics on the intestinal microbial community infected with C. difficile was carried out in [37]. In this study, DNA sequences were taken from the cecum and the ileum of 9 mice models. The sequences generated from this study were analyzed in [11] to obtain the OTUs profiles of each sample. The OTU assignment retains the ten most abundant genera (listed in Table 2) in addition to C. difficile which together account for approximately 90% of the total 16S rRNA sequences. In this paper, the time series data belonging to two mice under different conditions were considered.

It is pertinent to remember that, for the moment, no complete microbial interactions database reference is available to objectively evaluate the results obtained based on real data sets. However, the results can be assessed by evaluating their consistency with biological assumptions and their agreement with previous studies. For more accurate evaluation, the identified interactions need further analysis via high-throughput experiments. Also, since the constructed MINs include only a subset (i.e., 11 OTUs) of the total OTUs presented in the samples, an edge between two microbes may not necessarily indicate a direct interaction. For example, if two microbes are co-regulated by another microbe which is not included in the 11 OTUs, these two microbes may exhibit an interaction between them.

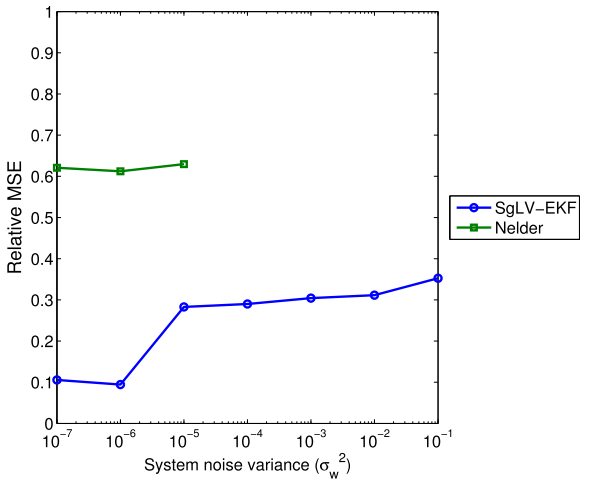

Fig. 5 Relative MSE performance under different dynamic noise levels. For σ_w^2 is larger than 10^{-4}, Nelder's algorithm diverges

Table 2 OTUs that are considered in the construction of the MINs of the two realistic datasets

OUT 1:	Barnesiella
OUT 2:	Undefined genus of Lachnospiraceae
OUT 3:	Unclassified Lachnospiraceae
OUT 4:	Other
OUT 5:	Blautia
OUT 6:	Undefined genus of unclassified Mollicutes
OUT 7:	Akkermansia
OUT 8:	Coprobacillus
OUT 9:	Clostridium difficile
OUT 10:	Enterococcus
OUT 11:	Undefined genus of Enterobacteriaceae

Dataset-1: Gut microbiota of mouse model infected by C. difficile

This dataset consists of 4 time points taken over two weeks and it belongs to the mouse with ID 8. This mouse received spores of C. difficile and was used to determine the impact of the pathogen (i.e., C. difficile) on the native gut microbiota. Figure 6 depicts the measured bacterial abundance level time series data $\mathbf{x}_i$ and its predicted values $\hat{\mathbf{x}}_i$. The results show that the SgLV-EKF was successful in tracking the bacterial abundance level.

The predicted values for the growth rates and the MIN are depicted in Fig. 7. The inferred growth rates are consistent with the biological assumptions in the sense that they are all positive. Moreover, their range $[0.2 - 0.83]$ agrees with typical growth rate ranges $[0.43 - 1.46]$ [25] and $[0.2 - 0.9]$ [11]. The comparable growth rates for the bacterial populations may indicate the existence of a balance state in the bacterial ecosystem. In other words, the environment is not dominated by one or few species with significantly higher growth rates.

The negative values of the diagonal elements in the inferred interaction matrix Fig. 7 are consistent with the underlying biology. This is because the negative values indicate that each species would reach the carrying capacity even in the absence of the other species [11]. Intriguingly, even Coprobacillus exhibits low abundance levels as shown in Fig. 6, the inferred MIN suggests Coprobacillus as the bacteria with the strongest interactions (i.e., larger interaction coefficients values) with other members in the microbial community. In particular, Coprobacillus inhibits all other microbes except Akkermansia and Blautia. Interstingly, all the bacteria exhibit inhibitory activity against C. difficile except Enterococcus, Undefined genus of Lachnospiraceae, and Undefined genus of unclassified Mollicutes which positively interact with the pathogen. This positive interaction agrees with previous results in [38]. The predicted MIN suggests C. difficile to negatively impact Blautia and Coprobacillus. This complies with the findings in [39] that show that both Blautia and Coprobacillus are among the top genera that are depleted in patients infected by C. difficile. Moreover, the inferred MIN shows that Barnesiella is negatively interacting with Enterococcus. This agrees with the results found in [40]. The constructed microbial interaction network is displayed in Fig. 8.

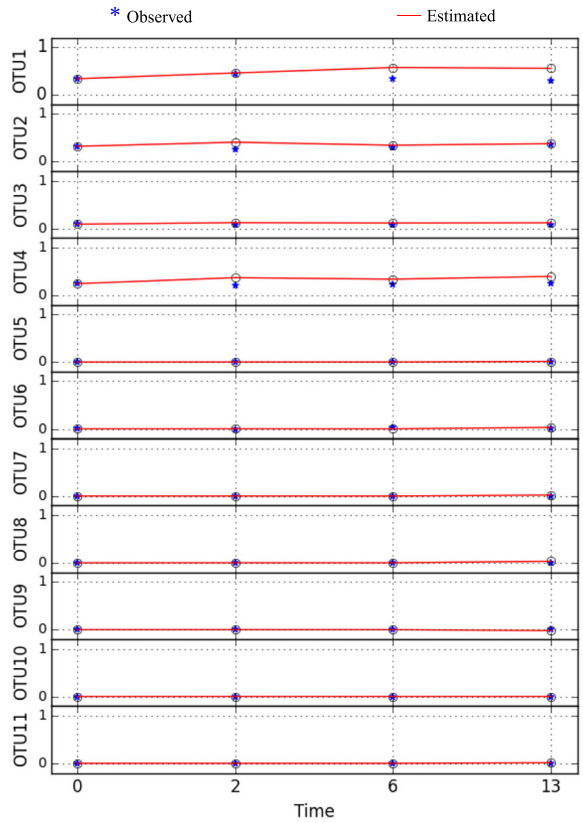

Fig. 6 Time series of observed and predicted bacterial abundance levels in relation to Dataset-1

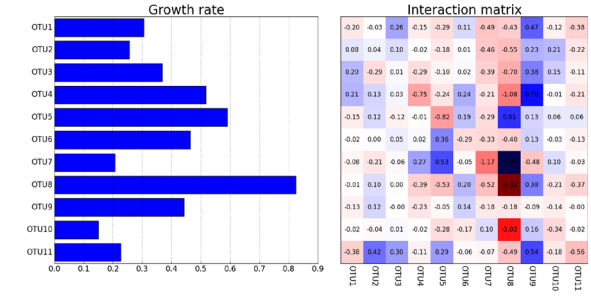

Fig. 7 Inferred growth rates and interaction matrix in relation to Dataset-1

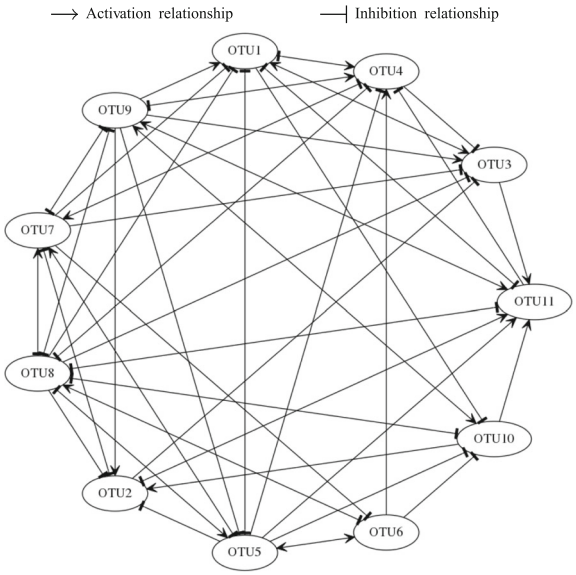

Fig. 8 Inferred microbial interaction network in relation to Dataset-1

Dataset 2: Gut microbiota of mouse model infected by C. difficile and treated with clindamycin

This dataset consists of 11 time points taken over 23 days and it belongs to the mouse with ID 9. At the first day of experiment, this mouse was injected with a single dose of clindamycin, and on the following day received spores of C. difficile. This experiment aimed to investigate the impact of the antibiotic (i.e., clindamycin) on the intestinal bacterial structure. The bacterial abundance levels and their estimated values are presented in Fig. 9. It is clear that the inferred model provides a fairly good prediction of the bacterial abundance data. By comparing the abundance levels in the two datasets, it is clear that the clindamycin antibiotic alters the structure of the microbial community. In particular, Barnesiella and the undefined genus of Lachnospiraceae are severely depleted in response to clindamycin. On the other hand, Enterococcus, the undefined genus of Enterobacteriaceae and more importantly C. difficile exhibit an increase in their abundance levels. This suggests that the induced dysbiosis in the bacterial community from its normal state due to the clindamycin antibiotic facilitates the colonization of C. difficile.

The inferred interaction matrix shown in Fig. 10 supports these findings. For example, the simultaneous increase in the abundance levels of Enterococcus, the undefined genus of Enterobacteriaceae and C. difficile can be explained by the mutualistic (i.e.,+|+) relationships between them. The inferred growth rates shown in Fig. 10 are all positive and ranging between 0.2 and 0.89. This complies with the biological assumption as discussed earlier. The obtained microbial interaction network is shown in Fig. 11.

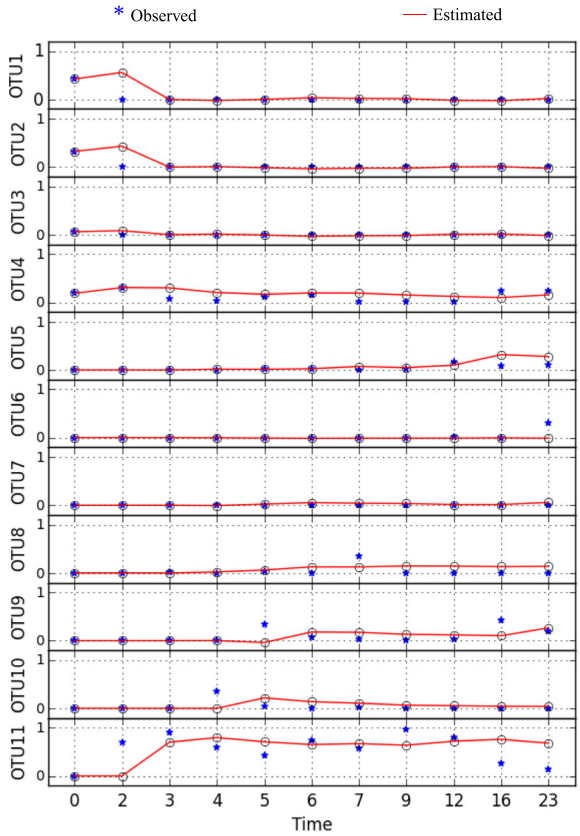

Fig. 9 Time series of observed and predicted bacterial abundance levels in relation to Dataset-2

Conclusions

In this work, we propose the SgLV-EKF algorithm to model the microbial dynamic and infer their interactions. In particular, we replace the conventional model of MIN formulated as a gLV dynamic model with a with a stochastic gLV model. The introduced stochastic model accounts for the uncertainties in the model and/or measurements. The proposed stochastic model accounts for the uncertainty in the model by adding a noise term in the dynamic

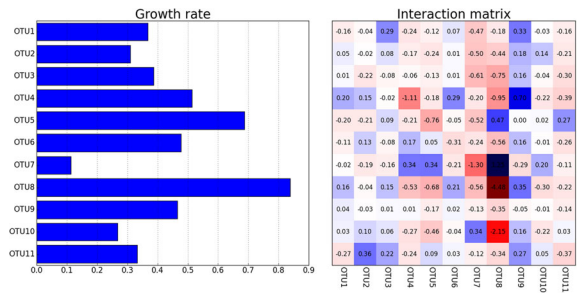

Fig. 10 Inferred growth rates and interaction matrix in relation to Dataset-2

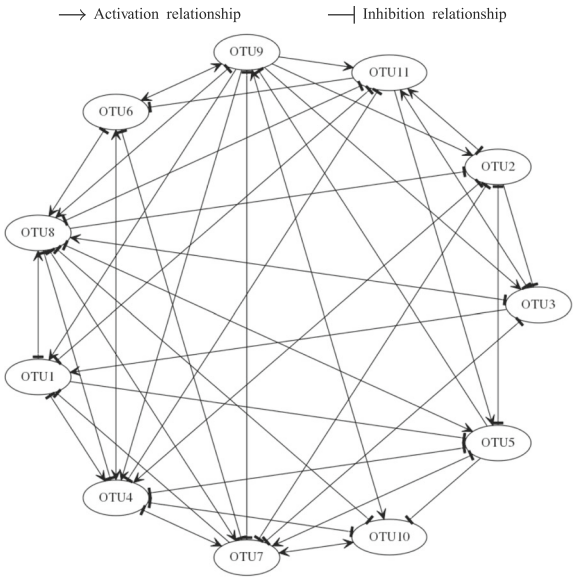

Fig. 11 Inferred microbial interaction network in relation to Dataset-2

equation. Moreover, to deal with the challenges of inferring MIN (i.e., nonlinear dynamics and limited number of observations), we propose EKF to jointly estimate the bacterial abundance levels and their interactions. The online and recursive nature of EKF enables fast and reliable estimation of the model's parameters from short time series data.

The performance of the proposed SgLV-EKF algorithm is compared with two similarity-based algorithms (i.e., PCC and LSA), one integral-based algorithm (i.e., Nelder's algorithm) and two regression-based algorithms (i.e., Stein's and LIMITS algorithms) in the presence of synthetic as well as realistic data sets by varying the noise levels in both the measurements and dynamic model.

It is observed that Stein's algorithm, an example of regression-based algorithms, is computationally efficient. However, it consistently exhibits a very low sensitivity indicating its failure to detect the majority of the interactions. This renders Stein's algorithm unreliable and inaccurate for estimating the MIN. This inaccuracy is because its sensitivity to the selection of the regularization parameters and the approximation of the derivatives in the ODE model. Particularly, the authors in [11] applied the forward difference to estimate the derivatives, which represents a coarse approximation of the slope of the bacterial abundance profiles. The LIMITS algorithm, a second example of regression-based algorithms, achieves more reliable and consistent performance compared to Stein's algorithm. However, this improvement comes at the cost of increased computational time. The reason behind increasing the execution time of LIMITS algorithm is the bagging procedure implemented in LIMITS to reduce the bias caused by the 'errors-in-variables' problem [41]. Similar to Stein's algorithm, the performance of LIMITS algorithm is sensitive to the accuracy of the approximation used to evaluate the derivatives in the ODE model.

Nelder's algorithm, an implementation of the integral-based approaches, offers close results to the SgLV-EKF when varying the measurements noise. However, Nelder's algorithm failed to compensate for randomness in the dynamic system as the algorithm diverges in the presence of noise with power exceeding 10^{-4}. Moreover, SgLV-EKF is more computationally efficient due to its sequential structure. The main virtue of similarity-based algorithms is that they are computationally efficient. However, similarity-based methods can capture only pairwise relationships between species. This renders these methods incapable of handling the existing complex interactions in microbial communities.

Overall, the robustness against uncertainty in measurements and/or model, the enhanced accuracy relative to the state-of-the-art algorithms, and the reduced computational time make SgLV-EKF a promising approach to model the microbial dynamics and infer the interactions among microbes.

Abbreviations

EKF: Extended Kalman Filter; gLV: Generalized Lotka-Volterra; LIMITS: Learning interactions from MIcrobial time series; LSA: Local similarity analysis; MIN: Microbial interaction network; ODE: Ordinary differential equation OTU: Operational taxonomic unit; PCC : Pearson correlation coefficient; SgLV-EKF: Stochastic gLV model with extended Kalman filter (EKF)

Funding

The publication costs of this article was funded by Texas A&M University.

Availability of data and materials

The two datasets can be found in the supplementary material of [11].

Authors' contributions

MA conceived of the study, developed the framework, conducted the analysis and simulations, provided the results interpretation and wrote the manuscript. ABY contributed to the framework development and analysis, participated in the analysis and simulations, and helped in reviewing the manuscript. ES provided an overall guidance, participated in the mathematical and biological interpretation of the results, and was involved in drafting the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of BMC Genomics Volume 18 Supplement 3, 2017: Selected original research articles from the Third International Workshop on Computational Network Biology: Modeling,

Analysis, and Control (CNB-MAC 2016): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-3.

Publisher's Note

Author details
[1]Bioinformatics and Genomic Signal Processing Lab, ECEN Dept., Texas A&M University, College Station, 77843-3128 TX, USA. [2]AE Dept., Khalifa University, Abu Dhabi, UAE.

Published: 27 March 2017

References

1. Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. Expert Rev Anti-Infect Ther. 2010;8(4):435–54.
2. Flint HJ. Obesity and the gut microbiota. J Clin Gastroenterol. 2011;45: S128–32.
3. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.
4. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science. 2013;341(6150):1241214.
5. Larsen N, Vogensen FK, Van Den Berg F, Nielsen DS, Andreasen AS, Pedersen BK, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PloS ONE. 2010;5(2):e9085.
6. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79.
7. Moore W, Moore LH. Intestinal floras of populations that have a high risk of colon cancer. Appl Environ Microbiol. 1995;61(9):3202–7.
8. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, et al. Human gut microbiome and risk of colorectal cancer. J Natl Cancer Inst. 2013. doi:10.1093/jnci/djt300.
9. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
10. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biology. 2008;6(11):e280.
11. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013;9(12):1–11.
12. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
13. Bucci V, Nadell CD, Xavier JB. The evolution of bacteriocin production in bacterial biofilms. Am Nat. 2011;178(6):E162–73.
14. Klitgord N, Segre D. Environments that induce synthetic microbial ecosystems. PLoS Comput Biol. 2010;6(11):e1001002.
15. Khosravi A, Mazmanian SK. Disruption of the gut microbiome as a risk factor for microbial infections. Curr Opin Microbiol. 2013;16(2):221–7.
16. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol. 2012;8(7):e1002606.
17. Song HS, Cannon WR, Beliaev AS, Konopka A. Mathematical modeling of microbial community dynamics: a methodological review. Processes. 2014;2(4):711–52.
18. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15(7):1.
19. Eiler A, Heinrich F, Bertilsson S. Coherent dynamics and association networks among lake bacterioplankton taxa. ISME J. 2012;6(2): 330–42.
20. Fuhrman JA, Steele JA. Community structure of marine bacterioplankton: patterns, networks, and relationships to function. Aquat Microb Ecol. 2008;53(1):69.
21. Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics. 2006;22(20):2532–8.
22. Xia LC, Steele JA, Cram JA, Cardon ZG, Simmons SL, Vallino JJ, et al. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. BMC Syst Biol. 2011;5(2):1.
23. Mounier J, Monnet C, Vallaeys T, Arditi R, Sarthou AS, Hélias A, et al. Microbial interactions within a cheese microbial community. Appl Environ Microbiol. 2008;74(1):172–81.
24. Fisher CK, Mehta P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. PLoS ONE. 2014;9(7):1–10.
25. Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD. Mathematical modeling of primary succession of murine intestinal microbiota. Proc Natl Acad Sci. 2014;111(1):439–44.
26. Tsai KY, Wang FS. Evolutionary optimization with data collocation for reverse engineering of biological networks. Bioinformatics. 2005;21(7): 1180–8.
27. Voit E, Chou IC. Parameter estimation in canonical biological systems models. Int J Syst Synthetic Biol. 2010;1:1–19.
28. Chou IC, Martens H, Voit EO. Parameter estimation in biochemical systems models with alternating regression. Theor Biol Med Model. 2006;3(1):25.
29. Zhan C, Yeung LF. Parameter estimation in systems biology models using spline approximation. BMC Syst Biol. 2011;5(1):14.
30. Corigliano A, Mariani S. Parameter identification in explicit structural dynamics: performance of the extended Kalman filter. Comput Methods Appl Mech Eng. 2004;193(36):3807–35.
31. Lillacci G, Khammash M. Parameter estimation and model selection in computational biology. PLoS Comput Biol. 2010;6(3):e1000696.
32. Wang Z, Liu X, Liu Y, Liang J, Vinciotti V. An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series. IEEE/ACM Trans Comput Biol Bioinformatics (TCBB). 2009;6(3):410–9.
33. Albiol J, Robuste J, Casas C, Poch M. Biomass estimation in plant cell cultures using an extended Kalman filter. Biotechnol Prog. 1993;9(2): 174–8.
34. Crassidis JL, Junkins JL. Optimal Estimation of Dynamic Systems. In: Chapman & Hall/CRC Applied Mathematics & Nonlinear Science. London: CRC Press; 2011.
35. Nelder JA, Mead R. A simplex method for function minimization. ComputerJournal. 1965;7(4):308–13.
36. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.
37. Buffie CG, Jarchum I, Equinda M, Lipuma L, Gobourne A, Viale A, et al. Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. Infect Immun. 2012;80(1):62–73.
38. Donskey CJ, Ray AJ, Hoyen CK, Fuldauer PD, Aron DC, Salvator A, et al. Colonization and infection with multiple nosocomial pathogens among patients colonized with vancomycin-resistant enterococcus. Infect Control Hosp Epidemiol. 2003;24(4):242–5.
39. Antharam VC, Li EC, Ishmael A, Sharma A, Mai V, Rand KH, et al. Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. J Clin Microbiol. 2013;51(9):2884–92.
40. Ubeda C, Bucci V, Caballero S, Djukovic A, Toussaint NC, Equinda M, et al. Intestinal microbiota containing Barnesiella species cures vancomycin-resistant Enterococcus faecium colonization. Infect Immun. 2013;81(3):965–73.
41. Fuller WA. Properties of some estimators for the errors-in-variables model. Ann Stat. 1980;8:407–22.