Journal of Theoretical Biology 398 (2016) 85–95

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Bayesian model selection framework for identifying growth patterns in filamentous fungi

Xiao Lin [a], Gabriel Terejanu [a,*], Sajan Shrestha [b], Sourav Banerjee [b], Anindya Chanda [c]

[a] Department of Computer Science and Engineering, University of South Carolina, 315 Main St, Swearingen Bldg. 3A01L, Columbia, SC 29208, USA
[b] Department of Mechanical Engineering, University of South Carolina, United States
[c] Department of Environmental Health Sciences, University of South Carolina, United States

HIGHLIGHTS

- Model selection for identifying growth patterns in the presence of model error.
- Methodology for modeling the structural uncertainties in mathematical models.
- Bayesian model comparison is a mathematically formalized Occam's razor.
- Numerical results using three fungal growth models in the context of simulated data.
- Model complexity plays an important role in identifying growth patterns in fungi.

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form
22 February 2016
Accepted 11 March 2016
Available online 19 March 2016

Keywords:
Fungal growth
Model discrepancy
Statistical modeling
Model calibration

ABSTRACT

This paper describes a rigorous methodology for quantification of model errors in fungal growth models. This is essential to choose the model that best describes the data and guide modeling efforts. Mathematical modeling of growth of filamentous fungi is necessary in fungal biology for gaining systems level understanding on hyphal and colony behaviors in different environments. A critical challenge in the development of these mathematical models arises from the indeterminate nature of their colony architecture, which is a result of processing diverse intracellular signals induced in response to a heterogeneous set of physical and nutritional factors. There exists a practical gap in connecting fungal growth models with measurement data. Here, we address this gap by introducing the first unified computational framework based on Bayesian inference that can quantify individual model errors and rank the statistical models based on their descriptive power against data. We show that this Bayesian model comparison is just a natural formalization of Occam's razor. The application of this framework is discussed in comparing three models in the context of synthetic data generated from a known true fungal growth model. This framework of model comparison achieves a trade-off between data fitness and model complexity and the quantified model error not only helps in calibrating and comparing the models, but also in making better predictions and guiding model refinements.

1. Introduction

Mycelial expansion of filamentous fungi has an enormous impact on agriculture, animal and human health and environmental sustainability (Redecker et al., 2000; Hawksworth, 2001; Gow and Gadd, 1995; Blain, 1975). The expansion of mycelium in different environments is orchestrated by several physical and biochemical mechanisms (Leeder et al., 2011). A number of models have been developed in the past decades to describe mycelial growth. Generally, these models can be categorized into two groups: continuous and discrete. A continuous model, appearing in the form of differential equations, usually describes mycelial network by its average properties, such as biomass density and hyphal density. The model proposed by Edelstein (1982) for example, falls into this category and is the basis of many models. The model consists of various branching and anastomosis and could be calibrated for different species. However, the model is based on the assumption that unlimited nutrient is supplied and does not include the interaction between fungi and its environment. Later work (Edelstein and Segel., 1983; Davidson and Park, 1998) improved this model by introducing processes such as

* Corresponding author.
E-mail addresses: lin65@email.sc.edu (X. Lin), terejanu@cec.sc.edu (G. Terejanu), sajan@email.sc.edu (S. Shrestha), banerjes@cec.sc.edu (S. Banerjee), achanda@mailbox.sc.edu (A. Chanda).

http://dx.doi.org/10.1016/j.jtbi.2016.03.021
0022-5193/

uptake and translocation. Boswell et al. (2003) further incorporated the hyphal division into active and inactive according to whether they are involved in the translocation of internal metabolites. This model was shown to be applicable to the nutritionally heterogeneous environments. Unlike the continuous models that are based on averaging the fungal properties, discrete models consider the growth of each hypha and can produce a realistic visualization of fungal mycelium (Boswell and Davidson, 2012). However, the construction of such a complex mycelial network is computationally expensive and it is much more difficult to include various physical and biochemical processes into a discrete model than into a continuous model (Boswell and Davidson, 2007).

Mathematical structures that describe the complexity of these fungal growth patterns vary greatly. Thus, it is important to have access to a computational tool capable to rank alternative models in the light of data. This raises the issue of model selection, which is significant in several aspects. First, mycologists can identify the growth model that best describes the experimental data. This can determine the dominant growth patterns for various fungi species grown in diverse environments; these include the study of fungal growth patterns in presence of antifungal drugs. Second, developing a new model is a nontrivial task. It consists of identifying dominant growth patterns, describing them mathematically, and validating whether the new model has improved descriptive power over its predecessors. Choosing from existing models on the other hand, can save time and energy in quickly identifying dominant growth patterns and guide the development of the new model. In addition, for a series of models derived from the same predecessor, the complex ones may not always be the better choice than the simpler ones. According to Occam's razor, a complex model should only be chosen if it offers a significant improvement (MacKay, 2003). This principle should direct the development of a new model, when additional processes are incorporated. Thus rigorous model selection can be an important tool in studying fungal growth.

The most straightforward way to evaluate the performance of a model is to determine how well it fits the observational data using classical measures such as mean square error. However, in the context of comparing different models, these type of measures are generally inappropriate. They favor heavily parameterized models able to “fit anything”, leading to over-fitting rather than an improved description of the true biology. This is in contrast with the aim of mathematical modeling, which is to simplify the complex process in the real world so as to study the key properties of the phenomenon of interest (Boswell and Davidson, 2007). Complicated models capable to mirror the reality are meaningless to researchers as they describe spurious physical processes. Thus, when comparing different models there is more to consider. Namely, a good model selection approach should be a trade-off between data fitness and model complexity.

Two of the most commonly used approaches for model selection are Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both are based on the likelihood function which reflects how well the model fits the data and both have an additional penalty term for the number of parameters in the model. Penalty term in BIC is positively correlated with the number of observations and generally larger than the penalty in AIC (Kirk et al., 2013; Kadane and Lazar., 2004). However, the number of parameters may not reflect the real complexity since different models could have different mathematical structures. Bayesian inference uses relative probabilities to compare the models under consideration. This general measure, which naturally embodies Occam's razor, favors models that fit the data well while penalizing models that rely heavily on the data to adjust parameters. Smith and Spiegelhalter (1980) pointed that under specific conditions, AIC and BIC are two particular cases of Bayesian model selection. In this work, we adopt a full Bayesian approach to the model selection problem.

Bayesian model selection in biological applications has been recently reported in Refs. Kirk et al. (2013) and Vanlier et al. (2014). However, an important issue which is less addressed is the existence of model uncertainty. It should be noted that the model selected among a set of candidates may not be the best in describing the real process. It is just better than the other candidates and it is quite possible that there is still a big discrepancy between its output and the real process. Unfortunately, this model discrepancy seems to be neglected in biological models, which causes the difficulty of describing real world data and restricts the use of advanced model-based approaches. In fact, no model could fully describe the real process. The discrepancy which is due to missing physics and numerical approximations is inevitable for mathematical models. Brynjarsdòttir and O'Hagan (2014) illustrate the importance of recognizing model discrepancy by giving a simple example where the model used for simulation is slightly different from the true model that generated observational data. It is shown that the prediction as well as the estimation of the parameters can be biased and over-confident.

The inclusion of model discrepancy in the mathematical model is important not only for improving model predictions but also for understanding the deficiencies of the model, which can lead to further improvements. Thus, it is necessary to include model discrepancy in model-based studies. In this paper, the discrepancy is modeled using a parametric structure and calibrated with observational data as proposed by Kennedy and O'Hagan (2001). We will show that the introduced model discrepancy offers a way to evaluate the predictive capability of the best model in the context of model selection. If the model discrepancy is very large, the model cannot be used even though it outperforms other candidates.

A variety of computational approaches have been developed to solve Bayesian inverse problems. Among them, Markov Chain Monte Carlo (MCMC) algorithms have became the main computational workhorse in scientific computing for solving statistical inverse problems. However, most MCMC approaches will fail, when model error is introduced. Model error causes the likelihood function and the prior probability density function to overlap in the regions of small value, which in turn decreases the acceptance rate of MCMC algorithms. This yields highly correlated samples far away from the region of high density. Furthermore, due to a big difference in the support of the prior probability density function and the likelihood function the evidence estimator used in Bayesian model selection has a large variance. Thus the computational approaches for solving Bayesian inverse problems need to be selected carefully in the presence of model error.

In this paper, an adaptive multilevel-sampling algorithm (Prudencio and Schulz, 2012) is used to obtain samples from the posterior distribution. Instead of sampling directly from the posterior, the algorithm samples a set of intermediate distributions corresponding to likelihood functions that are progressively flatten to increase the overlap with the prior. Besides sampling challenging posterior distributions, the multilevel sampling algorithm can be used to build an evidence estimator that has a significantly lower variance by aggregating the partial evidence calculated at each intermediate step.

All these developments are assembled into a comprehensive framework to perform Bayesian model selection as well as quantify the model discrepancy to study fungal growth. The models proposed by Edelstein (1982) are used to illustrate our approach, where different growth patterns are generated and compared, and the discrepancy of each model is quantified. These models have been selected thanks to their simple structure to provide the reader with an overview on how to apply this framework and interpret the results of the proposed analysis, which is the goal of