MATHEMATICA APPLICANDA
Vol. 48(1) 2020, p. 59–86
doi: 10.14708/ma.v48i1.6465

Daniah Tahir, Ingemar Kaj (Uppsala)
Krzysztof Bartoszek * (Linköping)
Marta Majchrzak, Paweł Parniewski * (Łódź)
Sebastian Sakowski (Łódź)

Using multitype branching models to analyze bacterial pathogenicity

Abstract We apply multitype continuous time Markov branching models to study pathogenicity in *E. coli*, a bacterium belonging to the genus *Escherichia*. First, we examine briefly the properties of multitype branching processes and we also survey some fundamental limit theorems regarding the behavior of such models under various conditions. These theorems are then applied to discrete, state dependent models in order to analyze pathogenicity in a published clinical data set consisting of 251 strains of *E. coli*. We use well established methods, incorporating maximum likelihood techniques, to estimate speciation rates as well as the rates of transition between different states of the models. From the analysis, we not only derive new results, but we also verify some preexisting notions about virulent behavior in bacterial strains.

2010 Mathematics Subject Classification: Primary: 60B12; Secondary: 60J85.

Key words and phrases: Markov models, branching processes, limit theorems, virulence factors, *E. coli* strains.

1. Introduction In this paper, we explore the possibility of utilizing the theory of branching processes into analyzing pathogenicity in bacterial strains. For that purpose, we first review fundamental properties of multitype, continuous time Markov branching processes as well as their behavior in the long time limit. Then, we apply multitype branching models to examine virulence in *E. coli* strains, and perform an in depth analysis on the limits of proportions of *E. coli* in different states of the models. The strains used in this study were isolated from human hosts, and obtained from a previously published data set (by Bartoszek et al. [4]) of pathogenic and nonpathogenic *E. coli* bacteria.

In recent years, considerable research has been conducted regarding the use of branching processes to explore various biological phenomena. A two–type Markov branching model, coined as the 'binary state speciation and extinction' (BiSSE) model, was proposed by Maddison et al. [12] to assess

* KB was supported by Vetenskapsrådets grant no. 2017-04951.
* MM and PP were partially supported by IMB PAS.

the impact of binary characters on rates of diversification—the difference between speciation and extinction rates. The parameters of the BiSSE model describe speciation and extinction in the two types, as well as the transitions that take place from one type to the other. This model was recently used by Bartoszek et al. [4] in order to estimate parameters from genetic data of *E. coli* populations, and predict pathogenicity of various virulence factors (VFs)—agents that enable bacteria to replicate and spread within the host by damaging and eluding its defences [6]. Bartoszek et al. used a modified version of the BiSSE model, in which the extinction rates were assumed to be zero. Another Markov branching model, named as the 'multistate speciation and extinction' (MuSSE) model, was later introduced by FitzJohn [7] as an extension of the BiSSE model to binary traits with more than two states.

In this paper, we apply the MuSSE model with zero extinction rates to estimate state dependent speciation and transition rates for a real, known collection of pathogenic and nonpathogenic strains of *E. coli* bacteria (obtained from [4]) that reside in the human digestive and urinary tracts. The bacterial strains are subdivided into four categories depending on whether or not they carry a VF in the gut and bladder of the human host. Lately, a number of researchers have successfully used the discrete state MuSSE model for estimation of various parameters. For instance, Sachs et al. (2013) used the MuSSE model to estimate trait-dependent diversification and transition rates amongst states consisting of free-living, mutualistic, parasitic, and dual lifestyle bacteria in the Proteobacteria phylum. Using this model, they inferred that proteobacterial mutualist lineages arise from free–living and parasitic ancestors, but rarely transition back to a parasitic or free–living status. Pirie et al. [14] implemented the MuSSE framework to compare diversification rates of 800 species in the plant genus *Erica*, endemic to five geographical regions—Palearctic, Tropical African, Madagascan, Drakensberg and Cape. Arbuckle et al. [1] used the BiSSE and MuSSE models to show that speciation and extinction rates vary across defensive traits in amphibians.

With this study, we ask the following questions:

(a) Is it possible to utilize state dependent branching processes to analyze pathogenic behavior in bacteria?

(b) How do maximum likelihood methods behave when estimating parameters (such as speciation rates and transition rates between states) of multitype models?

(c) Based only on a finite sample, do the estimated parameters provide reasonable information on the almost sure limits of the proportion of bacterial strains, and if so, can they be used further to obtain plausible confidence regions for these limits?

To answer these questions, we proceed by first giving a concise survey of n–type branching processes. More specifically, we recall fundamental theorems

regarding the long time behavior of branching models. These limit theorems are obtained from earlier works of Athreya and Ney [2] and Janson [8]. In order to thoroughly explain the mathematical background, which is pertinent to understanding the forthcoming biological application, we introduce some technical notation as well. For multitype branching models, a matrix known as the mean offspring matrix—whose entries consist of the net growth and transition rates of the process—is of central importance; the limit behavior of the process can be completely characterized by this matrix through its eigenvalues and corresponding eigenvectors. In the coming sections, we evaluate the mean offspring matrix for various sub–models of the multitype branching process, and present interesting results associated with the largest eigenvalue of these matrices. After reviewing the general characteristics of multitype branching processes, we apply 4–type branching sub–models to a known clinical data set of virulent and nonvirulent *E. coli* strains, and, with the help of the MuSSE model, obtain rates of speciation and transition among the 4 states of the models. Using the aforementioned limit theorems, we also perform an in depth analysis on the limits of proportions of *E. coli* strains in different states. Finally, we compare the results obtained from various models and draw some useful inferences regarding pathogenic behavior in virulent bacteria.

2. General multitype branching processes We consider an n–type continuous time Markov branching process $\boldsymbol{X}(t)$, given by the column vector $\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_n(t)\big)'$, $t \geq 0$, where each $X_i(t)$, $i = 1, \ldots, n$, represents the number of type–i particles at time t. The lifetime of each type is assumed to be exponentially distributed with intensity a_i, $i = 1, \ldots, n$. We introduce a vector $\boldsymbol{a}$ whose components comprise of a_i, i.e., $\boldsymbol{a} = (a_1, \ldots, a_n)$. With each type i, we also associate $\boldsymbol{j} = (j_1, \ldots, j_n)$, a vector of nonnegative integer coordinates. Then, the offspring distribution of the n types is specified by the coordinates of $\boldsymbol{p}(\boldsymbol{j})$, where

$$\boldsymbol{p}(\boldsymbol{j}) = \big(p^{(1)}(\boldsymbol{j}), \ldots, p^{(n)}(\boldsymbol{j})\big)$$

and $\sum_{\boldsymbol{j}} p^{(i)}(\boldsymbol{j}) = 1$, for all $i = 1, \ldots, n$. Here, $p^{(i)}(\boldsymbol{j}) = p^{(i)}(j_1, \ldots, j_n)$ gives the probability that a type–i particle creates j_1 type–1 offspring, j_2 type–2 offspring, …, j_n type–n offspring [3]. Letting $\boldsymbol{s} = (s_1, \ldots, s_n)$, the generating function is recognized as $\boldsymbol{f}(\boldsymbol{s}) = \big(f^{(1)}(\boldsymbol{s}), \ldots, f^{(n)}(\boldsymbol{s})\big)$, where, for each $i = 1, \ldots, n$,

$$f^{(i)}(\boldsymbol{s}) = \sum_{\boldsymbol{j}} p^{(i)}(\boldsymbol{j}) \boldsymbol{s}^{\boldsymbol{j}} = \sum_{j_1, \ldots, j_n \geq 0} p^{(i)}(j_1, \ldots, j_n) s_1^{j_1} \cdot \ldots \cdot s_n^{j_n}$$

determines the distribution of the number of various types of offspring produced by a type–i particle. Further, we consider the matrix $\mathcal{A} = \{a_{ik} : i, k = 1, \ldots, n\}$, where

$$a_{ik} = a_i\Big(\frac{\partial f^{(i)}(\boldsymbol{s})}{\partial s_k}\Big|_{\boldsymbol{s}=(1,\ldots,1)} - \delta_{ik}\Big) \quad \text{and} \quad \delta_{ik} = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{otherwise.} \end{cases}$$

The mean matrix of the branching process is given by

$$M(t) = e^{\mathcal{A}t} = \{m_{ik}(t) : i, k = 1, \ldots, n\},$$

with $m_{ik}(t) = \mathbb{E}[X_k(t)|X_i(t) = 1]$ [3]. Following [8], we identify the mean offspring matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \mathcal{A}^T,$$

where T in the superscript denotes the matrix transpose. We let γ be the largest positive eigenvalue of $\boldsymbol{A}$, and, let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the left and right normalized column eigenvectors, respectively, of $\boldsymbol{A}$, corresponding to γ. Thus, $\boldsymbol{u}'\boldsymbol{A} = \gamma\boldsymbol{u}'$ and $\boldsymbol{A}\boldsymbol{v} = \gamma\boldsymbol{v}$.

We now recall some limit results for multitype branching processes that will be used in later sections for the analysis of branching models. These results, numbered 1, 2 and 3 here, are stated formally in Appendix A of the paper as Theorems A.3, A.4 and A.5, respectively.

1. The fundamental limit result for multitype branching processes signifies the existence of a nonnegative random variable W such that $e^{-\gamma t}\boldsymbol{X}(t) \xrightarrow{a.s.} W\boldsymbol{v}$ as $t \to \infty$ [2]. This implies that the number of particles increases to infinity at a speed $e^{\gamma t}$, with the distribution of the types specified by a random multiple of the right eigenvector.
2. For a real nonnegative number N, let T_N be the first time when the total size of the branching process reaches a given level $N > 0$. Then, letting C be the sum of the components of $\boldsymbol{v}$, and under additional assumptions (see Appendix A), the asymptotic distribution of types at T_N is given by the right normalized eigenvector $\boldsymbol{v}$ in the sense that $\boldsymbol{X}(T_N)/N \xrightarrow{a.s.} \boldsymbol{v}/C$ as $N \to \infty$ [8].
3. Under further assumptions on the size of the eigenvalues of the offspring matrix $\boldsymbol{A}$, as $N \to \infty$, the centered and normalized type distribution has a Gaussian limit distribution in the sense that $\sqrt{N}\,(\boldsymbol{X}(T_N)/N - \boldsymbol{v}/C) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma_b})$, where $\mathcal{N}(0, \boldsymbol{\Sigma_b})$ is a multivariate normal distribution and $\boldsymbol{\Sigma_b}$ is the covariance matrix stated explicitly in Appendix A [8].

3. A 4–type branching model Consider a 4–type, continuous time Markov branching process, $\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_4(t)\big)'$, $t \geq 0$, where each $X_i(t)$, $i = 1, \ldots, 4$, gives the number of type–i particles at time t. Let λ_i represent the speciation rate of type–i particles, $i = 1, \ldots, 4$. Further, q_{12} and q_{21} are the rates of transition from type–1 $\to$ 2 and type–2 $\to$ 1, respectively, and similarly, q_{34} and q_{43} the transition rates from type–3 $\to$ 4 and type–4 $\to$ 3, respectively. Finally, m_{32} is assumed to be the transition rate from type–3 to type–2. The branching rates of $\boldsymbol{X}(t)$ are thus given as

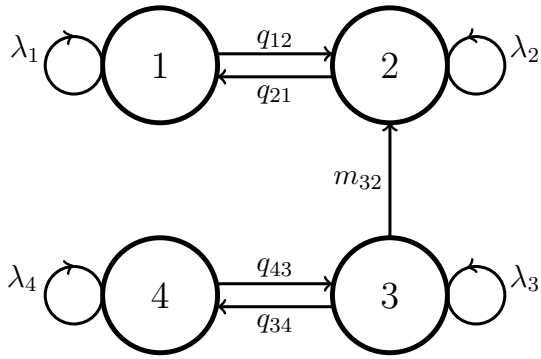

Figure 1: Diagrammatic representation of the speciation and transition parameters in the 4–type branching model $\boldsymbol{X}(t)$.

$$(x_1,x_2,x_3,x_4) \mapsto \begin{cases} (x_1+1,x_2,x_3,x_4) & \lambda_1 x_1 \\ (x_1,x_2+1,x_3,x_4) & \lambda_2 x_2 \\ (x_1,x_2,x_3+1,x_4) & \lambda_3 x_3 \\ (x_1,x_2,x_3,x_4+1) & \lambda_4 x_4 \\ (x_1-1,x_2+1,x_3,x_4) & q_{12}x_1 \\ (x_1+1,x_2-1,x_3,x_4) & q_{21}x_2 \\ (x_1,x_2,x_3-1,x_4+1) & q_{34}x_3 \\ (x_1,x_2,x_3+1,x_4-1) & q_{43}x_4 \\ (x_1,x_2+1,x_3-1,x_4) & m_{32}x_3, \end{cases} \tag{1}$$

where the initial state $\boldsymbol{X}(0)$ can be either $(0,0,1,0)'$ or $(0,0,0,1)'$, since type–3 and type–4 are the dominating types (see Def. A.1). Figure 1 summarizes the parameters used in this branching process. The motivation behind choosing this particular model is that the results derived here will be applied in Section 4 to obtain results and draw inferences from a real data set of *E. coli* bacterial strains. Note that a 4–type branching model would generally consist of 20 parameters: 4 speciation parameters, 4 extinction parameters and 12 parameters representing transition rates between states. Here we have assumed nonextinction, and we have also set 7 (out of 12) transition parameters to zero.

In order to identify the offspring matrix $\boldsymbol{A}$ for this process, we recall notation from Section 2, and let $\boldsymbol{a} = (a_1,\ldots,a_4)$, where

$$a_1 = \lambda_1 + q_{12}, \quad a_2 = \lambda_2 + q_{21}, \quad a_3 = \lambda_3 + q_{34} + m_{32}, \quad a_4 = \lambda_4 + q_{43}.$$

We recognize the offspring distribution of the four states as

$$p^{(1)}(2,0,0,0) = \lambda_1/a_1, \quad p^{(1)}(0,1,0,0) = q_{12}/a_1,$$

$$p^{(2)}(0,2,0,0) = \lambda_2/a_2, \quad p^{(2)}(1,0,0,0) = q_{21}/a_2,$$

$$p^{(3)}(0,0,2,0) = \lambda_3/a_3, \quad p^{(3)}(0,0,0,1) = q_{34}/a_3, \quad p^{(3)}(0,1,0,0) = m_{32}/a_3,$$

$$p^{(4)}(0,0,0,2) = \lambda_4/a_4, \quad p^{(4)}(0,0,1,0) = q_{43}/a_4.$$

Using the above information, we are now able to find the generating functions:

$$f^{(1)}(s_1,s_2,s_3,s_4) = (\lambda_1 s_1^2 + q_{12}s_2)/a_1,$$

$$f^{(2)}(s_1,s_2,s_3,s_4) = (\lambda_2 s_2^2 + q_{21}s_1)/a_2,$$

$$f^{(3)}(s_1,s_2,s_3,s_4) = (\lambda_3 s_3^2 + q_{34}s_4 + m_{32}s_2)/a_3,$$

$$f^{(4)}(s_1,s_2,s_3,s_4) = (\lambda_4 s_4^2 + q_{43}s_3)/a_4.$$

The mean offspring matrix $\boldsymbol{A}$ for this process is defined as

$$\boldsymbol{A} = \begin{pmatrix} \delta_1 & q_{21} & 0 & 0 \\ q_{12} & \delta_2 & m_{32} & 0 \\ 0 & 0 & \delta_3 & q_{43} \\ 0 & 0 & q_{34} & \delta_4 \end{pmatrix},$$

where $\delta_1 = \lambda_1 - q_{12}$, $\delta_2 = \lambda_2 - q_{21}$, $\delta_3 = \lambda_3 - m_{32} - q_{34}$ and $\delta_4 = \lambda_4 - q_{43}$. Letting

$$H_1 = \delta_1 + \delta_2, \quad H_2 = \delta_3 + \delta_4,$$

$$S_1 = \sqrt{(\delta_1 - \delta_2)^2 + 4q_{12}q_{21}}, \quad S_2 = \sqrt{(\delta_3 - \delta_4)^2 + 4q_{34}q_{43}},$$

the eigenvalues of $\boldsymbol{A}$, obtained using `Maple 18.00` [13], are

$$\gamma_1 = \frac{1}{2}(H_1 + S_1), \quad \gamma_2 = \frac{1}{2}(H_1 - S_1), \quad \gamma_3 = \frac{1}{2}(H_2 + S_2), \quad \gamma_4 = \frac{1}{2}(H_2 - S_2).$$

We have

$$\gamma_1 \geq \max\{\delta_1,\delta_2\} \geq \min\{\delta_1,\delta_2\} \geq \gamma_2, \quad \gamma_3 \geq \max\{\delta_3,\delta_4\} \geq \min\{\delta_3,\delta_4\} \geq \gamma_4.$$

It can be seen that $\gamma_1 \geq 0$ and $\gamma_3 \geq 0$ for any set of parameters. Also, $\gamma_1 > 0$ if at least one of δ_1 and δ_2 is strictly positive, or if both are negative, then one is strictly negative, and a similar result holds for γ_3. Further, we also require the eigenvectors of $\boldsymbol{A}$ to be used later in the application of limit theorems A.3–A.5. Thus, using again `Maple 18.00` [13], the left column eigenvectors of $\boldsymbol{A}$, after substantial manual simplification, are found as

$$\boldsymbol{u_1} = \Big(\frac{\frac{1}{2}(G_1+S_1)(\gamma_1 G_3 + G_4)}{q_{21}m_{32}q_{43}}, \frac{\gamma_1 G_3 + G_4}{q_{43}m_{32}}, \frac{\gamma_1 - \lambda_4 + q_{43}}{q_{43}}, 1\Big)',$$

$$\boldsymbol{u_2} = \Big(\frac{-\frac{1}{2}(-G_1+S_1)(\gamma_2 G_3 + G_4)}{q_{21}m_{32}q_{43}}, \frac{\gamma_2 G_3 + G_4}{q_{43}m_{32}}, \frac{\gamma_2 - \lambda_4 + q_{43}}{q_{43}}, 1\Big)',$$

$$\boldsymbol{u_3} = \Big(0, 0, -\frac{1}{2q_{43}}(G_2 - S_2), 1\Big)', \quad \boldsymbol{u_4} = \Big(0, 0, -\frac{1}{2q_{43}}(G_2 + S_2), 1\Big)',$$

and similarly, the right column eigenvectors are computed as

$$\boldsymbol{v_1} = \Big(1, -\frac{1}{2q_{21}}(G_1 - S_1), 0, 0\Big)', \quad \boldsymbol{v_2} = \Big(1, -\frac{1}{2q_{21}}(G_1 + S_1), 0, 0\Big)',$$

$$\boldsymbol{v_3} = \Big(1, \frac{\gamma_3 - \lambda_1 + q_{12}}{q_{21}}, \frac{\gamma_3 G_3 + G_4}{-q_{21}m_{32}}, \frac{\frac{1}{2}(G_2 + S_2)(\gamma_3 G_3 + G_4)}{-q_{21}m_{32}q_{43}}\Big)',$$

$$\boldsymbol{v_4} = \Big(1, \frac{\gamma_4 - \lambda_1 + q_{12}}{q_{21}}, \frac{\gamma_4 G_3 + G_4}{-q_{21}m_{32}}, \frac{\frac{1}{2}(G_2 - S_2)(\gamma_4 G_3 + G_4)}{-q_{21}m_{32}q_{43}}\Big)',$$

where $G_1 = \delta_1 - \delta_2$, $G_2 = \delta_4 - \delta_3$, $G_3 = H_1 - H_2$, and $G_4 = \delta_3\delta_4 - \delta_1\delta_2 - q_{43}(m_{32}+q_{34}) + q_{12}q_{21} + q_{43}m_{32}$. It should be noted that for certain parameter settings, where one of q_{21}, q_{43} or m_{32} vanish, the above expressions do not apply. Instead, alternate formulae have to be found on a case by case basis.

Now let $\gamma = \max\{\gamma_1, \gamma_3\}$ be the largest positive eigenvalue, and let $\boldsymbol{v} = (\nu_1, \nu_2, \nu_3, \nu_4)'$ be the normalized form of the corresponding right eigenvector. Then, from Thm. A.3 (and as described in limit result 1), there exists a nonnegative random variable W, such that as $t \to \infty$,

$$\big(X_1(t)e^{-\gamma t}, \ldots, X_4(t)e^{-\gamma t}\big)' \xrightarrow{a.s.} (\nu_1 W, \ldots, \nu_4 W)'.$$

Moreover, let $\nu_1 + \nu_2 + \nu_3 + \nu_4 = C > 0$ and recall that T_N is the first time when the total number of species reaches a level N. Using limit result 2 (or Thm. A.4), we have that as $N \to \infty$,

$$\frac{\boldsymbol{X}(T_N)}{N} = \frac{\big(X_1(T_N), \ldots, X_4(T_N)\big)'}{N} \xrightarrow{a.s.} \frac{(\nu_1, \ldots, \nu_4)'}{C}. \tag{2}$$

In order to apply the limit result 3 (see also Thm. A.5 in Appendix A), we need more information on the variance characteristics of the branching process. For that purpose, consider first the column vectors, ξ_i , $i = 1, \ldots, 4$, given as

$$\xi_1 = \begin{cases} (1,0,0,0)' & \text{with probability} & \lambda_1/a_1 \\ (-1,1,0,0)' & \text{-"-} & q_{12}/a_1 \end{cases},$$

$$\xi_2 = \begin{cases} (0,1,0,0)' & \text{with probability} & \lambda_2/a_2 \\ (1,-1,0,0)' & \text{-"-} & q_{21}/a_2 \end{cases},$$

$$\xi_3 = \begin{cases} (0,0,1,0)' & \text{with probability} & \lambda_3/a_3 \\ (0,0,-1,1)' & \text{-"-} & q_{34}/a_3 \\ (0,1,-1,0)' & \text{-"-} & m_{32}/a_3 \end{cases},$$

$$\xi_4 = \begin{cases} (0,0,0,1)' & \text{with probability} & \lambda_4/a_4 \\ (0,0,1,-1)' & \text{-"-} & q_{43}/a_4 \end{cases}.$$

Next, define a matrix $\boldsymbol{B}$ as

$$\boldsymbol{B} = \sum_{i=1}^{4} \nu_i a_i \mathbb{E}(\xi_i \xi_i') = \begin{pmatrix} b_1 & -b_2 & 0 & 0 \\ -b_2 & b_3 & -b_4 & 0 \\ 0 & -b_4 & b_5 & -b_6 \\ 0 & 0 & -b_6 & b_7 \end{pmatrix},$$

where $b_1 = a_1\nu_1 + q_{21}\nu_2$, $b_2 = q_{12}\nu_1 + q_{21}\nu_2$, $b_3 = a_2\nu_2 + q_{12}\nu_1 + m_{32}\nu_3$, $b_4 = m_{32}\nu_3$, $b_5 = a_3\nu_3 + q_{43}\nu_4$, $b_6 = q_{34}\nu_3 + q_{43}\nu_4$ and $b_7 = a_4\nu_4 + q_{34}\nu_3$. Now, since $\boldsymbol{A}$ is a 4×4 matrix with 4 distinct eigenvalues, it is diagonalizable. Hence, using Eq. (7) from Appendix A, the matrix $\boldsymbol{\Sigma_I}$ is specified by

$$\boldsymbol{\Sigma_I} = \sum_{j:\gamma_j < \frac{\gamma}{2}} \sum_{k:\gamma_k < \frac{\gamma}{2}} \frac{\hat{u}_j' \boldsymbol{B} \hat{u}_k}{\gamma - \gamma_j - \gamma_k} \hat{v}_j \hat{v}_k',$$

where column vectors $\hat{u}_i$ and $\hat{v}_i$ are determined as

$$\hat{u}_i = \frac{\boldsymbol{u_i}}{\boldsymbol{u_i} \cdot \boldsymbol{v_i}} \quad \text{and} \quad \hat{v}_i = \boldsymbol{v_i}, \quad i = 1, \ldots, 4.$$

Finally using Eq. (6), the covariance matrix $\boldsymbol{\Sigma_b}$ is given by

$$\boldsymbol{\Sigma_b} = \frac{1}{C^3} \boldsymbol{M_1} \boldsymbol{\Sigma_I} \boldsymbol{M_2},$$

$$\text{where} \quad \boldsymbol{M_1} = \begin{pmatrix} \nu_1 - C & \nu_1 & \nu_1 & \nu_1 \\ \nu_2 & \nu_2 - C & \nu_2 & \nu_2 \\ \nu_3 & \nu_3 & \nu_3 - C & \nu_3 \\ \nu_4 & \nu_4 & \nu_4 & \nu_4 - C \end{pmatrix} \quad \text{and} \quad \boldsymbol{M_2} = \boldsymbol{M_1}^T.$$

Under the condition that γ is greater than two times the second largest eigenvalue, and using the central limit theorem A.5, we have that

$$\sqrt{N}\left(\frac{\big(X_1(T_N), \ldots, X_4(T_N)\big)'}{N} - \frac{(\nu_1, \ldots, \nu_4)'}{C} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma_b}), \tag{3}$$

as $N \to \infty$.

In the following section, we apply the above 4–type model to a clinical data set of 251 bacterial strains, and obtain insightful results concerning pathogenicity in *E. coli* bacteria. We will take $N = 251$, so that T_N is the first time when the total size of the branching process reaches 251, and hence we can obtain the proportion of strains in various states of the model.

4. Application of the branching model to *E. coli* strains data

From [4], we obtain an *E. coli* data set of $N = 251$ strains, which forms the tips of a given phylogenetic tree (see Fig. 1 and Fig. 3 in [4]). The tree is fixed and describes the genealogical structure of the strains. The tree tips

are grouped into 4 categories: pathogenic and nonpathogenic *E. coli* found in the human gastrointestinal tract, and, pathogenic and nonpathogenic *E. coli* found in the human urinary tract. Whether a bacterial strain is pathogenic or not in any environment, depends on whether or not it is positive for carrying a certain VF (such as toxins, invasins, hemolysins, etc.). Hence, the strains are divided into the following 4 states

1	2	3	4
U_0	U_1	K_1	K_0

where

U_0: negative for VF in the urinary tract,

U_1: positive for VF in the urinary tract,

K_1: positive for VF in the digestive tract,

K_0: negative for VF in the digestive tract.

Virulence Factor (VF)	N_1	N_2	N_3	N_4
astA - heat-stable enterotoxin 1	108	20	7	116
cnf1 - cytotoxic necrotizing factor 1	90	38	3	120
fimG - fimbrial protein	13	115	120	3
fyuA - pesticin receptor protein	50	78	74	49
hly1 - alpha hemolysin	88	40	5	118
iroN - IroN protein	44	84	36	87
iutA - ferric aerobactin receptor	66	62	77	46
papC - fimbrial protein	83	45	33	90
sat - secreted autotransporter toxin	113	15	46	77

Table 1: A list of various VFs and the number of strains, N_i, in each of the four states obtained from [4]. Note that $N_1 + \ldots + N_4 = 251$ for each VF.

From the data given in [4], we choose to analyze 9 VFs. These VFs along with the number of strains, N_i, in each state i, $i = 1, \ldots, 4$, are given in Table 1. We model this data set by applying a 4–type, continuous time Markov branching process $\boldsymbol{X}(t) = (U_0, U_1, K_1, K_0)'$ and branching rates as given in Eq. (1). Figure 1 gives a diagrammatic representation of the parameters used in the model. In the figure, λ_i, $i = 1, \ldots, 4$, represent speciation rates of strains in various states, q_{12}, q_{21}, q_{34}, q_{43} are the transition rates from pathogenic to nonpathogenic states and vice versa, and finally, m_{32} represents migration from state K_1 to U_1. As in [4], extinction rates are set to zero for all 4 states, and, pathogenic and nonpathogenic strains are allowed to transition back and forth in the same environment (urinary tract or intestine). It is well known that *E. coli* travel from the gastrointestinal tract to the bladder in the human host, and cause urinary tract infections ([5], [10]). Hence, we assume that *E. coli* migrate from states K_1 to U_1.

All subsequent analysis of the bacterial data set is carried out in `R` [15]. The discrete MuSSE model [7] is applied to the data set of each VF to

	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
1.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\boldsymbol{q_{12}}$	q_{34}	q_{43}	m_{32}
2.	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
3.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\boldsymbol{q_{12}}$	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
4.	λ_1	λ_2	λ_3	λ_4	$\mathbf{0}$	q_{21}	q_{34}	q_{43}	m_{32}
5.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\mathbf{0}$	q_{34}	q_{43}	m_{32}
6.	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	$\mathbf{0}$	q_{43}	m_{32}
7.	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	$\mathbf{0}$	m_{32}
8.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\boldsymbol{q_{12}}$	$\mathbf{0}$	q_{43}	m_{32}
9.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\boldsymbol{q_{12}}$	q_{34}	$\mathbf{0}$	m_{32}
10.	λ_1	λ_2	λ_3	λ_4	$\mathbf{0}$	q_{21}	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
11.	λ_1	λ_2	λ_3	λ_4	q_{12}	$\mathbf{0}$	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
12.	λ_1	$\boldsymbol{\lambda_1}$	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
13.	λ_1	λ_2	λ_3	$\boldsymbol{\lambda_3}$	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
14.	λ_1	$\boldsymbol{\lambda_1}$	λ_3	λ_4	q_{12}	$\boldsymbol{q_{12}}$	q_{34}	q_{43}	m_{32}
15.	λ_1	λ_2	λ_3	$\boldsymbol{\lambda_3}$	q_{12}	q_{21}	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
16.	λ_1	$\boldsymbol{\lambda_1}$	λ_3	$\boldsymbol{\lambda_3}$	q_{12}	$\boldsymbol{q_{12}}$	q_{34}	$\boldsymbol{q_{34}}$	m_{32}
17.	$\mathbf{0}$	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
18.	λ_1	$\mathbf{0}$	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
19.	λ_1	λ_2	$\mathbf{0}$	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
20.	λ_1	λ_2	λ_3	$\mathbf{0}$	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}

Table 2: A list of 20 parameter constraints used in the analysis. The first row gives parameters for a model in which no constraints are applied, i.e., all parameters are allowed to vary freely. The subsequent rows represent models in which at least one constraint is used: a parameter is either constrained to be zero, or set equal to another parameter. In each row succeeding the first, which is to be used as a reference row in this table, the constrained parameters are highlighted in bold. For example, in the row marked (1), $q_{21} = q_{12}$, while the remaining parameters are free to vary; in row (2), $q_{43} = q_{34}$, etc.

obtain estimates for all parameters. This is achieved by making use of the 'make.musse()' function in the `R` package *diversitree* [7], which allows for maximum likelihood estimation of the models' parameters. During the parameter estimation analysis of each VF, the initial state for the process is determined by the relative probability of observing type–3 and type–4 strains, i.e., $\big(0, 0, N_3/(N_3 + N_4), N_4/(N_3 + N_4)\big)'$. To increase the estimation power of the MuSSE framework, we try out various models in which different constraints are applied on the parameters (the extinction rates already being set to zero), as given in Table 2. From the table, it can be seen that parameters are either set to be zero, or, pairs of parameters are constrained to be equal. This is done by utilizing the 'constrain()' function in the *diversitree* package. The most suitable constraint on the parameters is then chosen using the Bayesian information criterion or BIC [17], that is, for each VF separately, we choose that combination of constrained (or freely varying) parameter estimates which give the lowest BIC.

VF	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}
astA	26.39	0.000	229.6	0	176.0	904.4	407.4	41.68	65.53
cnf1	5.840	108.3	284.8	8.066	1.185	144.6	341.3	14.92	121.5
fimG	4.105	33.75	96.99	3.737	11.74	11.74	3.715	3.715	35.18
fyuA	0.464	51.71	120.7	4.242	0	38.36	47.40	6.445	46.71
hly1	4.630	101.7	265.6	10.58	0	131.1	287.9	17.77	110.4
iroN	1.843	48.50	164.0	0	21.20	60.93	178.0	20.55	51.44
iutA	0.000	64.93	126.6	3.923	49.07	157.0	43.34	5.527	46.21
papC	4.314	100.7	164.9	8.197	1.915	128.5	152.7	16.06	59.41
sat	6.059	165.6	157.4	6.493	0	307.3	102.4	9.691	65.62

Table 3: Parameter values for all VFs in the branching model $\boldsymbol{X}(t)$.

Results The parameter values obtained as a result of the analysis are given in Table 3. We conclude that for 6 out of 9 VFs, whenever the parameters are constrained in some manner, the model is a better fit to the given data set, in contrast to when all the parameters are allowed to vary freely. From the parameter values in Table 3, we have that:

(a) $\lambda_2 > \lambda_1$ for 8 out of 9 VFs, and $\lambda_3 > \lambda_4$ for all VFs. Thus, virulent strains of *E. coli*, in both urinary and digestive environments, speciate at a higher rate than nonvirulent strains.

(b) $q_{21} \geq q_{12}$ and $q_{34} \geq q_{43}$ for all VFs. Thus, *E. coli* bacteria, in both digestive and urinary tracts, lose their pathogenicity at a higher rate as compared to gaining it.

Using the mathematical analyses in previous sections and Appendix A, we also find that in accordance with our assumptions (F1) to (F6), $\gamma_3 > 0$ is the largest eigenvalue for all VFs and $\boldsymbol{v_3} > 0$ the corresponding right eigenvector. Let $\boldsymbol{v} = (\nu_1, \ldots, \nu_4)'$ be the normalized version of $\boldsymbol{v_3}$, and let $C = \nu_1 + \ldots + \nu_4$, as before. Then, from Thm. A.4 and using Eq. (2), we have for $N = 251$,

$$\frac{(N_1, \ldots, N_4)'}{N} \xrightarrow{a.s.} \boldsymbol{p} = (p_1, \ldots, p_4)',$$

where $\boldsymbol{p} = (p_1, \ldots, p_4)' := \boldsymbol{v}/C$ is the limit of the proportions of *E. coli* strains. For all VFs, the values of N_i/N and p_i, $i = 1, \ldots, 4$, are given in Table 4. The sum, $p_2 + p_3$, gives the probability of maintaining each VF in the *E. coli* strains. From Table 4, we infer that the probability of maintaining VFs varies in the strains; it depends on the VF under consideration. Figure 2 compares the value $p_2 + p_3$, with the sum $N_2/N + N_3/N$. It can be seen that *E. coli* strains carrying VFs **fimG**, **fyuA** and **iutA** have a higher probability of being virulent as compared to strains carrying **cnf1**, **hly1** and **sat**.

Confidence regions We now apply Thm. A.5 to construct confidence regions for $\boldsymbol{p}$ – the limit of proportions of *E. coli* strains. Using the expression in Eq. (3), let

VF	N_1/N	N_2/N	N_3/N	N_4/N	p_1	p_2	p_3	p_4
astA	0.43	0.08	0.03	0.46	0.83	0.15	0.01	0.01
cnf1	0.36	0.15	0.01	0.48	0.60	0.06	0.02	0.32
fimG	0.05	0.46	0.48	0.01	0.08	0.45	0.44	0.03
fyuA	0.20	0.31	0.29	0.20	0.35	0.32	0.15	0.18
hly1	0.35	0.16	0.02	0.47	0.51	0.08	0.04	0.37
iroN	0.18	0.33	0.14	0.35	0.52	0.35	0.02	0.11
iutA	0.26	0.25	0.31	0.18	0.34	0.20	0.23	0.23
papC	0.33	0.18	0.13	0.36	0.52	0.10	0.07	0.31
sat	0.45	0.06	0.18	0.31	0.55	0.03	0.09	0.33

Table 4: Limit values, p_i, for the the branching model $\boldsymbol{X}(t)$, and the proportion of strains, N_i/N, in each state, $i = 1, \ldots, 4$. Here, $N = 251$, N_i is the number of strains in each state i, and $p_i = \nu_i/C$, where ν_i $(i = 1, .., 4)$ are the components of the normalized eigenvector $\boldsymbol{v}$ and $C = \sum_{i=1}^{4} \nu_i$.

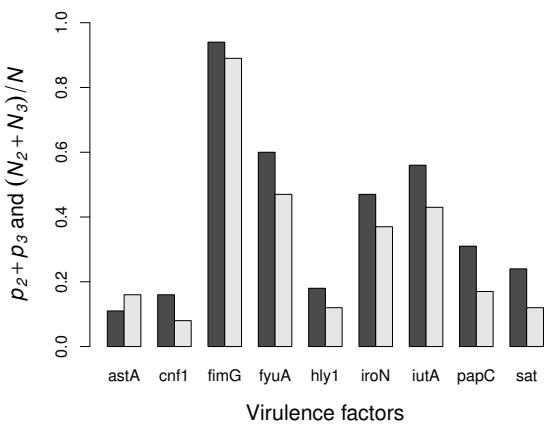

Figure 2: A bar chart comparing the probabilities, $p_2 + p_3$, of maintaining VFs in bacterial strains, with the sum, $(N_2 + N_3)/N$, of strain proportions. The values used in this figure are obtained from Table 4. $p_2 + p_3$ is represented by grey bars and $(N_2 + N_3)/N$ by black bars.

$$D^2 = N \left(\frac{\widetilde{\boldsymbol{X}}(T_N)}{N} - \frac{\widetilde{\boldsymbol{v}}}{C} \right)' (\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{b}})^{-1} \left(\frac{\widetilde{\boldsymbol{X}}(T_N)}{N} - \frac{\widetilde{\boldsymbol{v}}}{C} \right) \tag{4}$$

where vectors and matrices with the first type removed are denoted by a tilde above them, that is, $\widetilde{\boldsymbol{X}}(T_N) = \big(X_2(T_N), \ldots, X_4(T_N)\big)' = (N_2, \ldots, N_4)'$, $\widetilde{\boldsymbol{v}} = (\nu_2, \ldots, \nu_4)'$, and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{b}} = \boldsymbol{\Sigma}_{\boldsymbol{b}}$ with the first row and column removed. Here, it is important to notice that the matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ has rank 3, hence, for the construction of the subsequent confidence regions, we have removed the counts for the first type (alternatively, we can choose to remove any one of the four types) in Eq. (4). From the theory developed in Chapters 4.2 and 5.4

VF	D^2	$(N_2, N_3, N_4)/N$	a	b	c
astA	-	-	-	-	-
cnf1	10.35	(0.15, 0.01, 0.48)	0.651	0.091	0.046
fimG	3.421	(0.46, 0.48, 0.01)	0.352	0.034	0.021
fyuA	5.655	(0.31, 0.29, 0.20)	0.409	0.135	0.101
hly1	8.963	(0.16, 0.02, 0.47)	0.280	0.098	0.060
iroN	-	-	-	-	-
iutA	5.989	(0.25, 0.31, 0.18)	0.488	0.127	0.061
papC	8.499	(0.18, 0.13, 0.36)	0.362	0.097	0.081
sat	6.589	(0.06, 0.18, 0.31)	0.172	0.098	0.045

Table 5: Specification of confidence ellipsoids for limits of proportions of *E. coli* strains. For a significance level $\alpha = 0.01$, the observed generalized squared distance D^2, defined by Eq. (4) and (5), is given as $D^2 \leq \chi^2_3(0.01) = 11.345$. The center of the confidence ellipsoids is represented by $(N_2, N_3, N_4)/N$, while the half-lengths of the axes are denoted by a, b, and c.

of [9], at a given significance level α, a $100(1-\alpha)\%$ joint confidence region for $\boldsymbol{p}$—which can be thought of as the mean of a multidimensional normal distribution—is defined by ellipsoids such that,

$$D^2 \leq \chi^2_3(\alpha), \tag{5}$$

where $\chi^2_3(\alpha)$ is the upper α–level quantile of the χ^2 distribution with 3 degrees of freedom. The quantity D^2 represents the square of the generalized distance from the centre of the confidence ellipsoid to a constant density surface. For various VFs, the observed D^2 value is calculated using Eq. (4), and is given in the second column of Table 5. It can be seen that for **astA** and **iroN**, we are unable to find D^2, since the conditions of Thm. A.5 are not met for these two VFs; the second largest eigenvalue γ_1 is greater than $\gamma_3/2$, i.e., $\gamma_3/2 < \gamma_1 < \gamma_3$. In this case, weak convergence is shown but to a random variable, whose distribution is not characterized, only its existence (Corollary 3.18 in [8]).

The axes of the confidence ellipsoids and their relative lengths are determined using the eigenvalues and corresponding eigenvectors of the positive definite matrix $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{b}}$ [9]. Since $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{b}}$ is a 3×3 matrix with the first type removed, we can find a simultaneous confidence ellipsoid for p_i when $i = 2, 3$ and 4. Let β, ζ, and η be the positive eigenvalues, and $\hat{\beta}$, $\hat{\zeta}$, and $\hat{\eta}$, the corresponding right eigenvectors of $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{b}}$. Then, the axes of the confidence ellipsoids centered at $\widetilde{\boldsymbol{X}}(T_N)/N$, are given as

$$\pm a\,\hat{\beta}\,, \quad \pm b\,\hat{\zeta}\,, \quad \text{and} \quad \pm c\,\hat{\eta},$$

with $a = D\sqrt{\beta/N}$, $b = D\sqrt{\zeta/N}$, and $c = D\sqrt{\eta/N}$ being the half length of the three axes. For various VFs, the axes lengths can be computed, as shown in Table 5.

Since Thm. A.5 is an asymptotic result, and our sample size consists of only 251 data points, we must confirm that Thm. A.5 has been successfully implemented while finding the confidence regions in the aforementioned analysis. To achieve this, we check how the observed proportions (N_i/N) behave for the observed number of strains $(N = 251)$ by making use of simulated trees. For the estimated model parameters, given in Table 3, we obtain the clades evolution using the 'tree.musse()' function of the *diversitree* package in R. For various VFs, excluding **astA** and **iroN**, we simulate 10000 trees, each with 251 tips. To ensure that the root of all simulated trees is a dominating type (assumption (F5) in Appendix A), we take the prior distribution to be concentrated on states 3 and 4, with probabilities equaling the observed relative proportions of types–3 and 4. For this, we use the 'sample()' function of R which randomly selects either of the two states with desired probabilities. Out of the 10000 simulations, we take into consideration in the subsequent analysis, only those trees in which the observed tips have at least one obser-

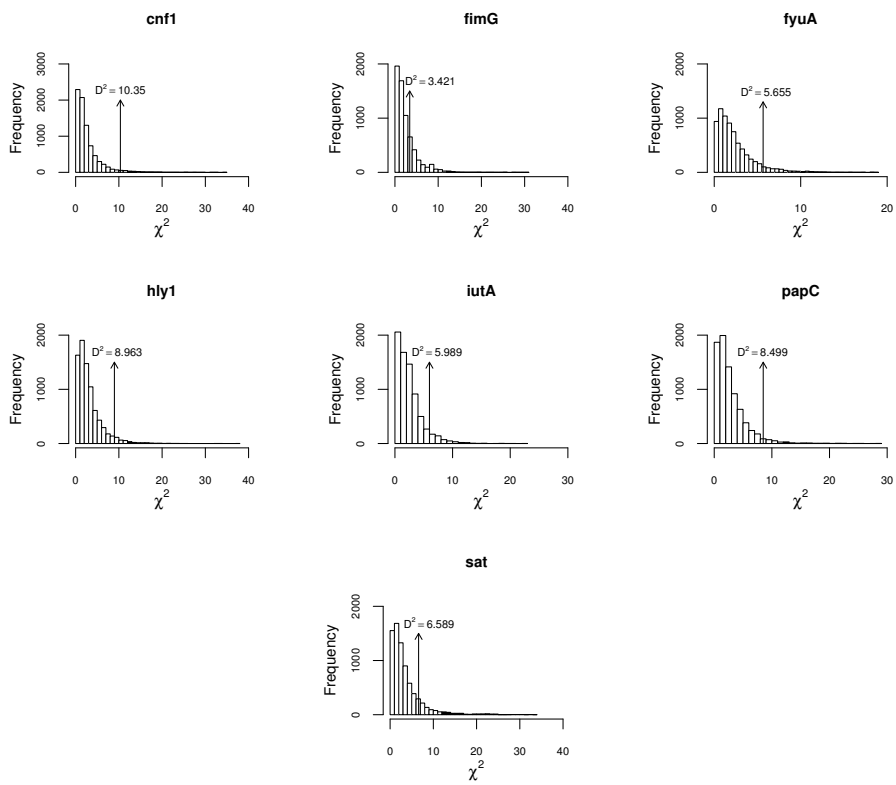

Figure 3: For various VFs, histograms showing the distribution of generalized squared distances for simulated trees. The values less than $\chi^2_3(0.01) = 11.345$ are represented by white bars, and those greater than $\chi^2_3(0.01)$ are given in black. Arrows in each histogram represent the observed D^2 value as obtained in Table 5.

VF	$\mathcal{N}_{sim}$	$\mathcal{F}_{sim}$	f_{sim}
astA	-	-	-
cnf1	7958	0.031	0.02
fimG	6588	0.236	0.01
fyuA	7276	0.074	0.01
hly1	8073	0.046	0.02
iroN	-	-	-
iutA	7405	0.071	0.01
papC	7967	0.096	0.01
sat	7583	0.128	0.03

Table 6: Values obtained from simulating 10000 trees: $\mathcal{N}_{sim}$ denotes the total number of simulations, out of 10000, that are actually used in the analysis, $\mathcal{F}_{sim}$ is that fraction of simulations for which the generalized square distance is greater than the observed D^2 value (given in Table 5), and f_{sim} is that fraction of simulations for which the generalized square distance is greater than $\chi^2_3(0.01) = 11.345$.

vation of type–3 or 4. This is due to the essential nonextinction assumption of Thm. A.5 (cf. Def. A.2). For each VF, the number of simulations that we use out of 10000 is denoted by $\mathcal{N}_{sim}$, and shown in Table 6. From the 251 tip counts of each simulated tree, we calculate the proportions of types, and then obtain the D^2 value given in Eq. (4), i.e., the squared generalized distance from the simulated proportions (with the first coordinate removed) which lie on the surface of an ellipsoid, to the ellipsoid's centre (at $\widetilde{\boldsymbol{v}}/C$). Using the χ^2 distribution with 3 degrees of freedom, we obtain from Eq. (5), how 'far in the tail' the observed D^2 values lie.

The results are shown in the form of histograms in Figure 3. We conclude that the observed proportions are close to the ellipsoid's centre, i.e., the quantile corresponds to a level greater than a cutoff value of $\alpha = 0.01$, and thus for the given number of strains, $N = 251$, we are indeed close to the asymptotic regime. From the histograms, we obtain the fraction of simulations, $\mathcal{F}_{sim}$, which give a value of the generalized square distance greater than the observed value of D^2, as well as the fraction of simulations, f_{sim}, for which the generalized squared distance is greater than the critical value, $\chi^2_3(0.01) = 11.345$. For each VF, these values are shown in Table 6.

5. Two variations of the branching model We now model the same *E. coli* data set [4] by applying another 4–type, Markov branching process $\boldsymbol{Y}(t) = (U_0, U_1, K_1, K_0)'$ with branching rates as shown in Figure 4. All parameters of speciation and transition are the same as in the original process $\boldsymbol{X}(t)$, except for the migration rate, m_{32}, which is replaced by m_{42} in this case. Since *E. coli* are assumed to travel from the intestinal to the urinary tract and cause infections, here we assume that bacteria migrate from the nonvirulent state K_0 to the virulent state U_1. The mean offspring matrix, $\hat{\boldsymbol{A}}$,

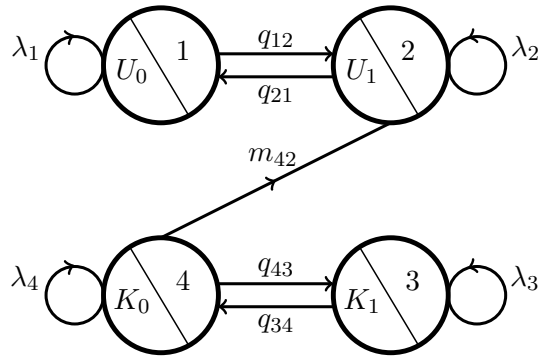

Figure 4: Diagrammatic representation of the branching process $\boldsymbol{Y}(t)$. The four states are similar to the ones in the branching model $\boldsymbol{X}(t)$. The parameters λ_i, $i = 1, \ldots, 4$, represent speciation rates in each state, q_{12}, q_{21}, q_{34} and q_{43} are the transition rates between states, m_{42} represents migration from state K_0 to U_1.

for this process is given as

$$\hat{\boldsymbol{A}} = \begin{pmatrix} \lambda_1 - q_{12} & q_{21} & 0 & 0 \\ q_{12} & \lambda_2 - q_{21} & 0 & m_{42} \\ 0 & 0 & \lambda_3 - q_{34} & q_{43} \\ 0 & 0 & q_{34} & \lambda_4 - q_{43} - m_{42} \end{pmatrix},$$

with eigenvalues

$$\hat{\gamma}_1 = \frac{1}{2}\Big(\lambda_1 - q_{12} + \lambda_2 - q_{21} + \sqrt{(\lambda_1 - q_{12} - \lambda_2 + q_{21})^2 + 4q_{12}q_{21}}\Big),$$

$$\hat{\gamma}_2 = \frac{1}{2}\Big(\lambda_1 - q_{12} + \lambda_2 - q_{21} - \sqrt{(\lambda_1 - q_{12} - \lambda_2 + q_{21})^2 + 4q_{12}q_{21}}\Big),$$

$$\hat{\gamma}_3 = \frac{1}{2}\Big(\lambda_3 - m_{42} - q_{34} + \lambda_4 - q_{43} + \sqrt{(\lambda_3 + m_{42} - q_{34} - \lambda_4 + q_{43})^2 + 4q_{34}q_{43}}\Big),$$

$$\hat{\gamma}_4 = \frac{1}{2}\Big(\lambda_3 - m_{42} - q_{34} + \lambda_4 - q_{43} - \sqrt{(\lambda_3 + m_{42} - q_{34} - \lambda_4 + q_{43})^2 + 4q_{34}q_{43}}\Big).$$

Using `Maple 18.00` [13], the left and right column eigenvectors of $\hat{\boldsymbol{A}}$ are obtained as

$$\hat{\boldsymbol{u}}_{\boldsymbol{1}} = \Big(\frac{\frac{1}{2}(G_1 + S_1)(\hat{\gamma}_1\hat{G}_3 + \hat{G}_4)}{q_{21}m_{42}q_{34}}, \frac{(\hat{\gamma}_1\hat{G}_3 + \hat{G}_4)}{q_{34}m_{42}}, 1, \frac{\hat{\gamma}_1 - \lambda_3 + q_{34}}{q_{34}}\Big)',$$

$$\hat{\boldsymbol{u}}_{\boldsymbol{2}} = \Big(\frac{\frac{1}{2}(G_1 - S_1)(\hat{\gamma}_2\hat{G}_3 + \hat{G}_4)}{q_{21}m_{42}q_{34}}, \frac{(\hat{\gamma}_2\hat{G}_3 + \hat{G}_4)}{q_{34}m_{42}}, 1, \frac{\hat{\gamma}_2 - \lambda_3 + q_{34}}{q_{34}}\Big)',$$

$$\hat{\boldsymbol{u}}_{\boldsymbol{3}} = \Big(0, 0, 1, \frac{1}{2q_{34}}(\hat{G}_2 + \hat{S}_2)\Big)', \quad \hat{\boldsymbol{u}}_{\boldsymbol{4}} = \Big(0, 0, 1, \frac{1}{2q_{34}}(\hat{G}_2 - \hat{S}_2)\Big)',$$

and

$$\hat{\boldsymbol{v}}_{\mathbf{1}} = \Big(1, -\frac{1}{2q_{21}}(G_1 - S_1), 0, 0\Big)', \quad \hat{\boldsymbol{v}}_{\mathbf{2}} = \Big(1, -\frac{1}{2q_{21}}(G_1 + S_1), 0, 0\Big)',$$

$$\hat{\boldsymbol{v}}_{\mathbf{3}} = \Big(1, \frac{\hat{\gamma}_3 - \lambda_1 + q_{12}}{q_{21}}, \frac{\frac{1}{2}(\hat{G}_2 - \hat{S}_2)(\hat{\gamma}_3\hat{G}_3 + \hat{G}_4)}{q_{21}m_{42}q_{34}}, \frac{\hat{\gamma}_3\hat{G}_3 + \hat{G}_4}{-q_{21}m_{42}}\Big)',$$

$$\hat{\boldsymbol{v}}_{\mathbf{4}} = \Big(1, \frac{\hat{\gamma}_4 - \lambda_1 + q_{12}}{q_{21}}, \frac{\frac{1}{2}(\hat{G}_2 + \hat{S}_2)(\hat{\gamma}_4\hat{G}_3 + \hat{G}_4)}{q_{21}m_{42}q_{34}}, \frac{\hat{\gamma}_4\hat{G}_3 + \hat{G}_4}{-q_{21}m_{42}}\Big)',$$

respectively, where

$$\hat{G}_2 = \delta_4 - m_{42} - \lambda_3 + q_{34}, \quad \hat{G}_3 = \delta_1 + \delta_2 - \delta_4 - \lambda_3 + q_{34} + m_{42},$$

$$\hat{G}_4 = \lambda_1 q_{21} - \lambda_1\lambda_2 + \lambda_2 q_{12} + \lambda_3\lambda_4 - \lambda_3 m_{42} - \lambda_3 q_{43} - \lambda_4 q_{34} + q_{34}m_{42},$$

$$\hat{S}_2 = \sqrt{(\hat{G}_2)^2 + 4q_{34}q_{43}},$$

while δ_1, δ_2, δ_4, G_1, and S_1 are defined previously in Section 3. Similar to the first model $\boldsymbol{X}(t)$, the package *diversitree* [7] in `R` is applied to each VF and estimates of all parameters are obtained, as shown in Table 7. We apply constraints given in Table 2, except that m_{32} is now replaced by m_{42} in the table. The best constraint on the parameters is again chosen using the BIC.

VF	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{42}
astA	3.031	138.9	0	86.05	27.85	448.9	8.210	18.47	27.88
cnf1	7.054	114.8	3.608	96.15	0	159.4	4.108	2.972	34.91
fimG	0.000	8.344	33.47	260.3	87.44	21.58	11.07	0	553.5
fyuA	3.192	68.53	117.9	0	8.990	67.68	124.2	24.75	14.25
hly1	4.768	109.4	0	86.64	0.823	145.7	8.042	17.18	27.30
iroN	0.875	48.39	4.562	107.5	0	31.93	6.113	31.19	42.69
iutA	0	75.26	157.2	7.472	36.64	158.8	104.9	3.066	12.50
papC	4.046	113.5	147.7	0	2.051	140.7	201.5	26.92	13.11
sat	5.793	173.1	0	107.1	0.000	318.2	14.71	65.11	34.35

Table 7: Parameter values for the branching process $\boldsymbol{Y}(t)$.

VF	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
astA	0.42	0.06	0.14	0.38
cnf1	0.43	0.14	0.02	0.41
fimG	0.09	0.46	0.44	0.01
fyuA	0.34	0.21	0.17	0.28
hly1	0.43	0.12	0.11	0.34
iroN	0.31	0.36	0.15	0.18
iutA	0.13	0.08	0.30	0.49
papC	0.51	0.09	0.10	0.30
sat	0.50	0.04	0.28	0.18

Table 8: Limit values $\hat{p}_i$ $(i = 1, \ldots, 4)$ for the the branching model $\boldsymbol{Y}(t)$.

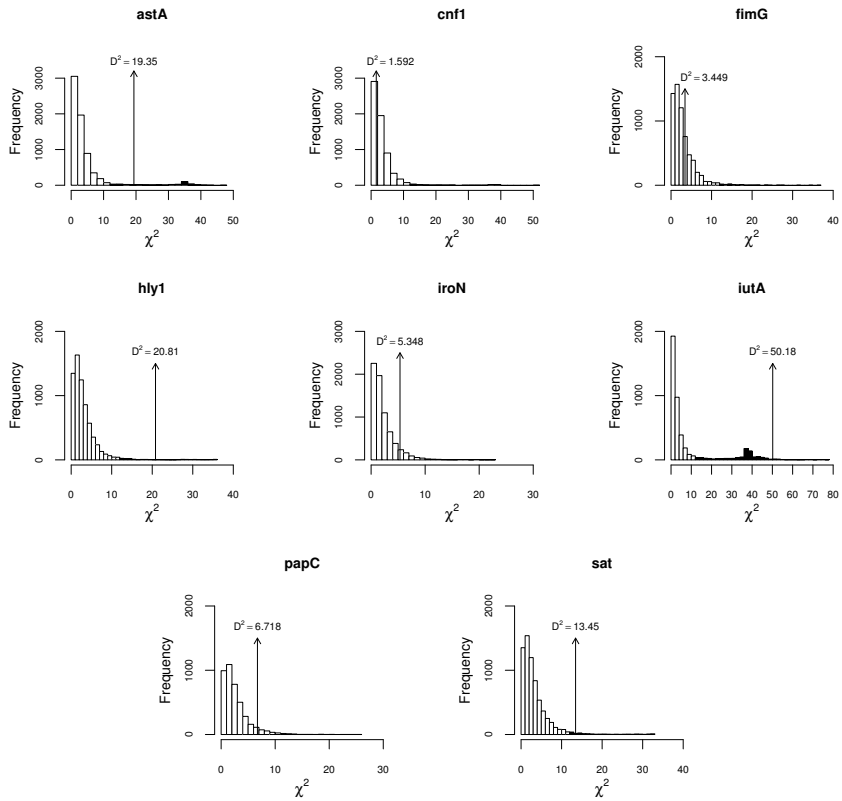

Figure 5: Distribution of generalized squared distances for those VFs for which assumptions of Thm. A.5 hold under the model $\boldsymbol{Y}(t)$. The values less than $\chi^2_3(0.01) = 11.345$ are represented by white bars, and those greater than 11.345 are given in black. Arrows in each histogram represent the observed D^2 value.

VF	$\hat{\mathcal{N}}_{sim}$	$\hat{\mathcal{F}}_{sim}$	$\hat{f}_{sim}$
astA	6885	0.040	0.05
cnf1	6473	0.647	0.02
fimG	6476	0.291	0.02
fyuA	-	-	-
hly1	6777	0.010	0.03
iroN	7012	0.080	0.00
iutA	4473	0.010	0.19
papC	4157	0.064	0.01
sat	6776	0.02	0.03

Table 9: Results of the confidence regions' analysis for $\hat{\boldsymbol{p}}$ corresponding to the process $\boldsymbol{Y}(t)$. $\hat{\mathcal{N}}_{sim}$ represents that number of simulated trees (out of 10000) in which the observed tips have at least a single observation of the dominating type–3 or 4. $\hat{\mathcal{F}}_{sim}$ is the fraction of simulations for which the generalized squared distance is greater than the observed D^2 value. $\hat{f}_{sim}$ is that fraction of simulations for which the generalized squared distance is greater than $\chi^2_3(0.01) = 11.345$.

For each VF, $\hat{\gamma}_3$ is found to be the largest eigenvalue, with corresponding normalized eigenvector, say $\hat{\boldsymbol{v}}$. We again apply Thm. A.4 and using the estimated parameters, obtain the limiting values $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_4)' = \hat{\boldsymbol{v}}/C$, where C is the sum of coordinates of $\hat{\boldsymbol{v}}$. The values of $\hat{\boldsymbol{p}}$ are given in Table 8. From Tables 7 and 8, we infer that: (a) $\lambda_2 > \lambda_1$ for all VFs and $\lambda_3 < \lambda_4$ for 6 out of 9 VFs, (b) $q_{21} > q_{12}$ for 8 out of 9 VFs and $q_{34} > q_{43}$ for 5 out of 9 VFs, (c) the probability, $\hat{p}_2 + \hat{p}_3$, of prevalence of VFs in *E. coli* strains varies with each VF—for instance, the VF **fimG** has the maximum probability of being maintained. An analysis of the confidence regions for $\hat{\boldsymbol{p}}$ is also carried out, similar to the one for $\boldsymbol{p}$ in Section 4. Figure 5 and Table 9 give the results for this analysis.

We apply yet another 4–type branching process $\boldsymbol{Z}(t) = (U_0, U_1, K_1, K_0)'$ to the same data set. The branching rates are as shown in Figure 6. This time we include migration from both virulent and nonvirulent states in the gastrointestinal tract to the virulent state of the urinary tract, using parameters m_{32} and m_{42}. The mean offspring matrix $\bar{\boldsymbol{A}}$ is given as

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \lambda_1 - q_{12} & q_{21} & 0 & 0 \\ q_{12} & \lambda_2 - q_{21} & m_{32} & m_{42} \\ 0 & 0 & \lambda_3 - q_{34} - m_{32} & q_{43} \\ 0 & 0 & q_{34} & \lambda_4 - q_{43} - m_{42} \end{pmatrix}.$$

The eigenvalues of $\bar{\boldsymbol{A}}$ are

$$\bar{\gamma}_1 = \frac{1}{2}\Big(\lambda_1 - q_{12} + \lambda_2 - q_{21} + \sqrt{(\lambda_1 - q_{12} - \lambda_2 + q_{21})^2 + 4q_{12}q_{21}}\Big),$$

$$\bar{\gamma}_2 = \frac{1}{2}\Big(\lambda_1 - q_{12} + \lambda_2 - q_{21} - \sqrt{(\lambda_1 - q_{12} - \lambda_2 + q_{21})^2 + 4q_{12}q_{21}}\Big),$$

$$\bar{\gamma}_3 = \frac{1}{2}\Big(\lambda_3 - m_{32} - q_{34} + \lambda_4 - m_{42} - q_{43} + \sqrt{(\lambda_3 - m_{32} - q_{34} - \lambda_4 + m_{42} + q_{43})^2 + 4q_{34}q_{43}}\Big),$$

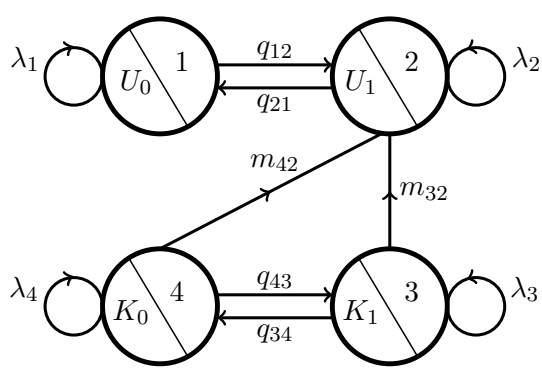

Figure 6: Branching rates for the model $\boldsymbol{Z}(t)$. λ_i, $i = 1, \ldots, 4$, represent speciation rates, q_{12}, q_{21}, q_{34} and q_{43} are the rates of transition between states, m_{32} and m_{42} are migration rates from state K_1 to U_1 and K_0 to U_1, respectively.

$$\bar{\gamma}_4 = \frac{1}{2}\Big(\lambda_3 - m_{32} - q_{34} + \lambda_4 - m_{42} - q_{43} - \sqrt{(\lambda_3 - m_{32} - q_{34} - \lambda_4 + m_{42} + q_{43})^2 + 4q_{34}q_{43}}\Big).$$

The left and right column eigenvectors of $\bar{\boldsymbol{A}}$ are obtained as

$$\bar{\boldsymbol{u}}_{\boldsymbol{1}} = \Big(\frac{\frac{1}{2}(G_1 + S_1)(\bar{\gamma}_1\bar{G}_3 + \bar{G}_4)}{q_{21}G_5}, \frac{\bar{\gamma}_1\bar{G}_3 + \bar{G}_4}{G_5}, 1, \frac{G_6}{G_5}\Big)',$$

$$\bar{\boldsymbol{u}}_{\boldsymbol{2}} = \Big(\frac{-\frac{1}{2}(-G_1 + S_1)(\bar{\gamma}_2\bar{G}_3 + \bar{G}_4)}{q_{21}G_{5_2}}, \frac{\bar{\gamma}_2\bar{G}_3 + \bar{G}_4}{G_{5_2}}, 1, \frac{G_{6_2}}{G_{5_2}}\Big)',$$

$$\bar{\boldsymbol{u}}_{\boldsymbol{3}} = \Big(0, 0, 1, \frac{1}{2q_{34}}(\bar{G}_2 + \bar{S}_2)\Big)', \quad \bar{\boldsymbol{u}}_{\boldsymbol{4}} = \Big(0, 0, 1, \frac{1}{2q_{34}}(\bar{G}_2 - \bar{S}_2)\Big)'.$$

and

$$\bar{\boldsymbol{v}}_{\boldsymbol{1}} = \Big(1, -\frac{1}{2q_{21}}(G_1 - S_1), 0, 0\Big)', \quad \bar{\boldsymbol{v}}_{\boldsymbol{2}} = \Big(1, -\frac{1}{2q_{21}}(G_1 + S_1), 0, 0\Big)',$$

$$\bar{\boldsymbol{v}}_{\boldsymbol{3}} = \Big(1, \frac{\bar{\gamma}_3 - \lambda_1 + q_{12}}{q_{21}}, \frac{(\bar{\gamma}_3\bar{G}_3 + \bar{G}_4)G_8}{q_{21}G_7}, \frac{-(\bar{\gamma}_3\bar{G}_3 + \bar{G}_4)G_9}{q_{21}G_7}\Big)',$$

$$\bar{\boldsymbol{v}}_{\boldsymbol{4}} = \Big(1, \frac{\bar{\gamma}_4 - \lambda_1 + q_{12}}{q_{21}}, \frac{(\bar{\gamma}_4\bar{G}_3 + \bar{G}_4)G_{8_2}}{q_{21}G_7}, \frac{-(\bar{\gamma}_4\bar{G}_3 + \bar{G}_4)G_{9_2}}{q_{21}G_7}\Big)'.$$

respectively, where

$$\bar{G}_2 = d_4 - \delta_3, \qquad \bar{G}_3 = \delta_1 + \delta_2 - \delta_3 - d_4,$$
$$\bar{G}_4 = \hat{G}_4 - m_{32}d_4, \qquad \bar{S}_2 = \sqrt{(\bar{G}_2)^2 + 4q_{34}q_{43}},$$
$$G_5 = m_{32}(\bar{\gamma}_1 - d_4) + m_{42}q_{34}, \qquad G_{5_2} = m_{32}(\bar{\gamma}_2 - d_4) + m_{42}q_{34},$$

$$G_6 = m_{42}\bar{\gamma}_1 - m_{42}\lambda_3 + m_{32}m_{42} + m_{32}q_{43} + m_{42}q_{34},$$
$$G_{6_2} = m_{42}\bar{\gamma}_2 - m_{42}\lambda_3 + m_{32}m_{42} + m_{32}q_{43} + m_{42}q_{34},$$
$$G_7 = m_{42}^2q_{34} - m_{32}^2q_{43} - m_{32}m_{42}g_2,$$

$$G_8 = m_{42}(-\bar{\gamma}_3 + d_4) + m_{32}q_{43}, \qquad G_{8_2} = m_{42}(-\bar{\gamma}_4 + d_4) + m_{32}q_{43}$$
$$G_9 = m_{32}(-\bar{\gamma}_3 + \delta_3) + m_{42}q_{34}, \qquad G_{9_2} = m_{32}(-\bar{\gamma}_4 + \delta_3) + m_{42}q_{34},$$

with $d_4 = \lambda_4 - q_{43} - m_{42}$ and δ_1, δ_2, δ_3, G_1, S_1, $\hat{G}_4$ as defined in earlier sections. Model analysis is carried out in `R`, and as before, the initial state is determined by the relative probability of observing type–3 and type–4 strains. We use the same constraints given in Table 2, except that both m_{32} and m_{42} are allowed to vary. In addition, we use 21 more constraints in which the constraints from Table 2 are repeated using $m_{32} = m_{42}$. The best constraints are chosen according to BIC, and the parameter values obtained are given in Table 10. From the parameters, we compute the eigenvalues and corresponding eigenvectors of $\bar{\boldsymbol{A}}$. $\bar{\gamma}_3$ is found to be the largest eigenvalue for all VFs. Moreover, $\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_4)'$, denoting the limits of proportions of *E. coli*

VF	λ_1	λ_2	λ_3	λ_4	q_{12}	q_{21}	q_{34}	q_{43}	m_{32}	m_{42}
astA	3.031	138.8	0	86.05	27.86	448.8	8.210	18.47	0.000	27.88
cnf1	0	24.84	3.768	107.5	2.343	136.6	2.027	2.706	2.053	33.95
fimG	4.487	35.10	103.2	4.034	0	11.17	1.832	2.014	30.41	2.125
fyuA	3.705	57.57	116.2	0	8.254	56.65	69.75	12.72	36.23	0.000
hly1	4.768	109.4	0	86.64	0.823	145.7	8.046	17.18	0.000	27.30
iroN	1.902	48.60	164.7	0	20.77	60.34	181.2	20.79	51.59	0.000
iutA	0	71.13	155.1	7.056	38.11	148.3	47.56	1.058	9.561	9.561
papC	4.493	99.45	166.2	0	2.081	126.5	203.0	25.97	49.39	0.027
sat	6.003	164.8	141.0	0	0.000	299.9	124.1	17.00	44.77	0.000

Table 10: Parameter estimates for the branching model $\boldsymbol{Z}(t)$.

VF	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$
astA	0.42	0.06	0.14	0.38
cnf1	0.22	0.12	0.02	0.64
fimG	0.06	0.36	0.56	0.02
fyuA	0.42	0.26	0.12	0.20
hly1	0.42	0.12	0.11	0.35
iroN	0.55	0.37	0.01	0.06
iutA	0.07	0.06	0.59	0.28
papC	0.56	0.09	0.07	0.28
sat	0.52	0.03	0.11	0.34

Table 11: Limits of proportions of *E. coli* strains for the branching process $\boldsymbol{Z}(t)$.

VF	$\bar{\mathcal{N}}_{sim}$	$\bar{\mathcal{F}}_{sim}$	$\bar{f}_{sim}$
astA	6757	0.042	0.06
cnf1	6810	0.004	0.01
fimG	6996	0.182	0.03
fyuA	-	-	-
hly1	6845	0.012	0.02
iroN	-	-	-
iutA	5421	0.129	0.15
papC	7033	0.023	0.01
sat	6733	0.110	0.04

Table 12: For the model $\boldsymbol{Z}(t)$, $\bar{\mathcal{N}}_{sim}$ represents the number of simulations (out of 10000) in which at least one observation of the dominating type–3 or 4 was obtained, $\bar{\mathcal{F}}_{sim}$ gives that fraction of simulations for which the generalized squared distance is greater than the observed value of D^2, and $\bar{f}_{sim}$ is that fraction of simulations for which the generalized squared distance is greater than $\chi^2_3(0.01) = 11.345$.

strains and calculated using Thm. A.4, is given in Table 11. From the tables, we conclude that: (a) $\lambda_2 > \lambda_1$ for all VFs and $\lambda_3 > \lambda_4$ for 6 out of 9 VFs. (b) $q_{21} > q_{12}$ and $q_{34} \geq q_{43}$ for 6 VFs. (c) $m_{32} \geq m_{42}$ for 6 out of 9 VFs. The sum $\bar{p}_2 + \bar{p}_3$ gives the probability of prevalence of various VFs in *E. coli* strains. An analysis regarding the confidence regions for the limits of strain proportions was also performed, similar to the one for the previous models.

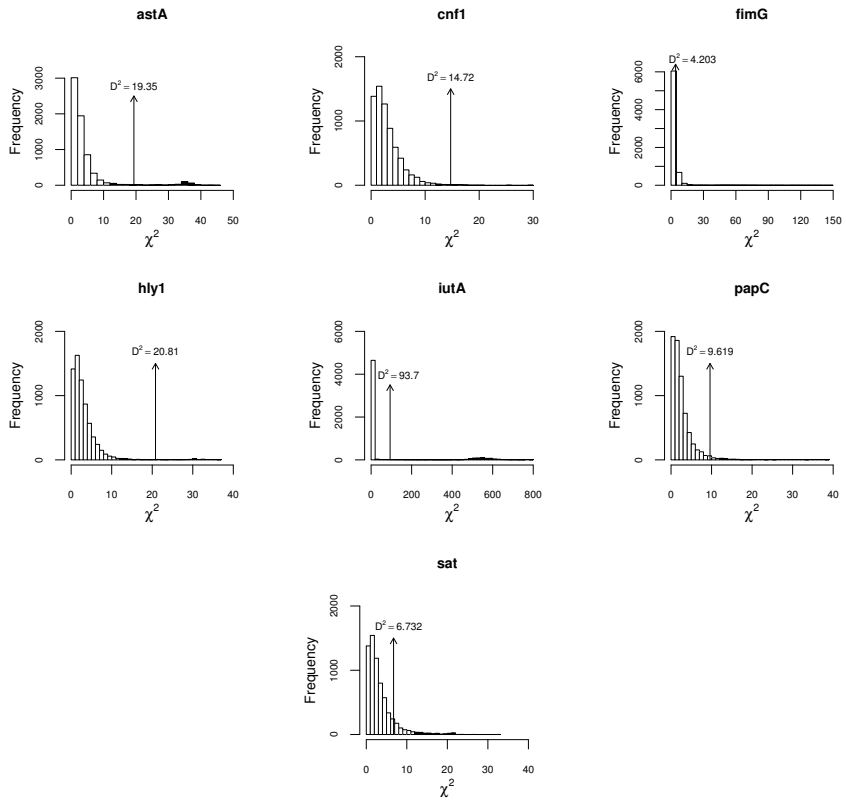

Figure 7: For the process $\boldsymbol{Z}(t)$, these histograms give the distribution of the generalized squared distances (obtained from Eq. (4)) of 10000 simulated trees for those VFs for which Thm. A.5 is satisfied. The values less than $\chi^2_3(\alpha) = 11.345$, for $\alpha = 0.01$, are represented by white bars, while the remaining are given in black. Arrowed lines on each histogram represent the observed D^2 value for each VF.

The results are displayed in Figure 7 and Table 12.

6. Conclusions We now list some concise but useful inferences drawn from the mathematical analysis of the three models. Here, we would also like to state that in a future work, we plan to carry out a rigorous analysis of a larger data set of *E. coli* strains, which would not only add to the current results, but also lead to more comprehensive and solid biological conclusions.

1. *E. coli* strains carrying a VF speciate at faster rates as compared to strains which do not carry a VF, in urinary tracts of the human host (Tables 3, 7 and 10). This result remains fairly constant in all three models. A similar result was also obtained in [4], where it was shown that speciation rates were higher for virulent *E. coli* strains as compared to nonvirulent strains that were isolated from human urine samples. The same result, however, does not hold for strains in the digestive tract; according to the process $\boldsymbol{X}(t)$ and

$\boldsymbol{Z}(t)$, most virulent bacterial strains speciate faster than nonvirulent strains, but the opposite is true under the model $\boldsymbol{Y}(t)$.

2. *E. coli* bacteria lose their virulence at a higher rate as compared to gaining it, in both urinary and digestive tracts of the human hosts. Under all three models, this behavior is exhibited for majority of the VFs (Tables 3, 7 and 10). This is consistent with the fact that bacteria maintain their virulence only if conditions are favorable for host invasion or colonization, otherwise, they lose their pathogenicity, since the expression of VFs is costly to maintain and tends to decrease bacterial fitness [4, 11].
3. Pathogenic and nonpathogenic bacteria in the gut migrate to the urinary tract at different rates. This result is inferred from the analysis of the third branching model $\boldsymbol{Z}(t)$ with two migration rates (Table 10). For 5 out of 9 VFs, pathogenic bacteria are found to migrate faster than the nonpathogenic ones. However, compared separately, from Tables 3 and 7, $m_{32} > m_{42}$ for most VFs (8 out of 9).
4. The probability of maintaining virulence in bacterial strains (p_2+p_3, $\hat{p}_2+\hat{p}_3$ and $\bar{p}_2+\bar{p}_3$ in the models $\boldsymbol{X}(t)$, $\boldsymbol{Y}(t)$ and $\boldsymbol{Z}(t)$, respectively) varies with the VF under consideration—see Tables 4, 8 and 11. However, under all three branching models, it is consistently seen that the VF **fimG** has the highest chance of prevailing in virulent bacterial strains.
5. The confidence region analysis of the limits of proportions of *E. coli* strains shows that Thm. A.5 is more successfully implemented to the data set under the first two models, $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$, as compared to the third model, $\boldsymbol{Z}(t)$. From figures 3, 5, and 7 we conclude that for Thm. A.5, the branching process $\boldsymbol{X}(t)$ is a better fit for analyzing strains carrying **astA**, **fimG**, **fyuA**, **hly1**, **iutA**, and **sat**, while $\boldsymbol{Y}(t)$ is suitable for **cnf1**, **iroN** and **papC**.

A. Appendix In this section, we state the limit theorems that are applied in this paper for the analysis of bacterial strains data. We first list some important definitions and assumptions that relate to n–type branching processes.

DEFINITION A.1 A type–p is said to be *dominating* if it is possible to obtain every other type, say type–q ($q = 1, \ldots, n$, and $q \neq p$), in a branching process that starts with a single type–p particle. The set of all dominating types is called the *dominating class* [8].

DEFINITION A.2 A branching process becomes *essentially extinct* if there are no particles from the dominating class at some time instance [8].

Assumptions:

(F1) Following a branching event involving a type–p particle, the number of type–p particles either increases, or decreases by at most one, while the number of type–q particles, $q = 1, \ldots, n$, $q \neq p$, always increases.

(F2) Following a branching event involving a type–p particle, all changes in the number of type–q particles, $q = 1, \ldots, n$, have finite means and variances, that is, the process never explodes.

(F3) The largest eigenvalue, γ, of the mean offspring matrix $\boldsymbol{A}$ is positive, i.e., the multitype branching process is supercritical, hence, there exists a positive probability of nonextinction.

(F4) The largest eigenvalue of $\boldsymbol{A}$ is simple, i.e., the algebraic multiplicity of γ is one.

(F5) There exists a dominating type–p, such that the multitype branching process starts with at least one particle of type–p.

(F6) The largest eigenvalue, γ, belongs to the dominating class.

THEOREM A.3 *(Theorem 1 in [2]) For the branching process* $\boldsymbol{X}(t)$,

$$\lim_{t\to\infty} e^{-\gamma t}\boldsymbol{X}(t) = W\boldsymbol{v}$$

exists with probability 1, where W *is a nonnegative random variable,* γ *is the largest eigenvalue of the mean offspring matrix* $\boldsymbol{A}$*, and* $\boldsymbol{v}$ *is the normalized right eigenvector of* $\boldsymbol{A}$ *corresponding to* γ*.* ■

Let $\boldsymbol{b} \in \mathbb{R}^n$ be a fixed column vector and let $N \geq 0$. Define

$$T_{\boldsymbol{b}}(N) = \min\{t \geq 0 : \boldsymbol{b} \cdot \boldsymbol{X}(t) \geq N\}$$

to be the first time when $\boldsymbol{b} \cdot \boldsymbol{X}(t)$ exceeds N. Also let $\boldsymbol{b} \cdot \boldsymbol{v} > 0$. Then, conditioned on essential nonextinction, $T_{\boldsymbol{b}}(N) < \infty$ for all $N \geq 0$ and $\boldsymbol{b} \cdot \boldsymbol{X}(t) \xrightarrow{a.s.} \infty$ as $t \to \infty$ (Lemma 3.14 in [8]).

We now state two more limit theorems, which are applied in the main sections of the paper with the special case $\boldsymbol{b} = (1, 1, \ldots, 1) \in \mathbb{R}^n$.

THEOREM A.4 *(Theorem 3.15 in [8]) Assume (F1) to (F6) and let* $\boldsymbol{b}\cdot\boldsymbol{v} > 0$*. Then, conditioned on essential nonextinction, as* $N \to \infty$*,*

$$\frac{\boldsymbol{X}\big(T_{\boldsymbol{b}}(N)\big)}{N} \xrightarrow{a.s.} \frac{\boldsymbol{v}}{\boldsymbol{b}\cdot\boldsymbol{v}}.$$

■

Let Δ be the set of all eigenvalues, γ_i, of $\boldsymbol{A}$. Of course, the largest eigenvalue γ also belongs to Δ. There exist projection matrices $\boldsymbol{P}_{\gamma_i}$, such that $\sum_{\gamma_i\in\Delta} \boldsymbol{P}_{\gamma_i} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Let $\boldsymbol{P}$ be the projection matrix onto the sum of the eigenspaces corresponding to $\gamma_i < \gamma/2$, that is, $\boldsymbol{P} = \sum_{\gamma_i\in\Delta_1} \boldsymbol{P}_{\gamma_i}$, where $\Delta_1 = \{\gamma_i \in \Delta : \gamma_i < \gamma/2\}$. Also, let $\xi_i = (\xi_{i1}, \ldots, \xi_{in})'$ be a random column vector with integer coordinates, denoting the change in population if a branching event occurs at a particle

of type–i, and define the matrix $\boldsymbol{B}$ as $\boldsymbol{B} = \sum_{i=1}^{n} \nu_i a_i \boldsymbol{B}_i$, where ν_i represents the coordinates of the eigenvector $\boldsymbol{v}$ and $\boldsymbol{B}_i = \mathbb{E}(\xi_i \xi_i')$. Furthermore, define another matrix $\boldsymbol{\Sigma_I}$ as

$$\boldsymbol{\Sigma_I} = \int_0^\infty \boldsymbol{P} e^{s\boldsymbol{A}} \boldsymbol{B} (\boldsymbol{P} e^{s\boldsymbol{A}})' e^{-s\gamma} ds,$$

where $s \in \mathbb{R}$ [8] .

THEOREM A.5 *(Corollary 3.16 in[8]) Assume (F1) to (F6) and $\boldsymbol{b} \cdot \boldsymbol{v} > 0$. Suppose further that $\gamma/2 > \gamma_{(n-1)}$, where $\gamma_{(n-1)}$ is the second largest eigenvalue, i.e., the second largest eigenvalue is less than half the largest eigenvalue. Then, conditioned on essential nonextinction, as $N \to \infty$,*

$$\sqrt{N}\left(\frac{\boldsymbol{X}\left(T_{\boldsymbol{b}}(N)\right)}{N} - \frac{\boldsymbol{v}}{\boldsymbol{b}\cdot\boldsymbol{v}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma_b}),$$

where

$$\boldsymbol{\Sigma_b} = (\boldsymbol{b}\cdot\boldsymbol{v})^{-1}\left(\boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{b}'}{\boldsymbol{b}\cdot\boldsymbol{v}}\right)\boldsymbol{\Sigma_I}\left(\boldsymbol{I} - \frac{\boldsymbol{b}\boldsymbol{v}'}{\boldsymbol{b}\cdot\boldsymbol{v}}\right) \tag{6}$$

is the covariance matrix with rank $n-1$. ■

For the application in Sections 3–5, we consider the special case in which $\boldsymbol{A}$ is a diagonalizable matrix. Then, there exist column vectors $\hat{u}_i$ and $\hat{v}_i$, such that, $\hat{u}_i'\boldsymbol{A} = \gamma_i \hat{u}_i'$, $\boldsymbol{A}\hat{v}_i = \gamma_i \hat{v}_i$, and $\hat{u}_i \cdot \hat{v}_j = \delta_{ij}$ for $i, j = 1, \ldots, n$. Using Lemma 5.3 in [8], the matrices $\boldsymbol{P}$ and $\boldsymbol{P}e^{s\boldsymbol{A}}$ can now be expressed as

$$\boldsymbol{P} = \sum_{j:\gamma_j \in \Delta_1} \hat{v}_j \hat{u}_j' \quad \text{and} \quad \boldsymbol{P}e^{s\boldsymbol{A}} = \sum_{j:\gamma_j \in \Delta_1} e^{s\gamma_j} \hat{v}_j \hat{u}_j',$$

respectively, and hence $\boldsymbol{\Sigma_I}$ becomes

$$\boldsymbol{\Sigma_I} = \sum_{j:\gamma_j \in \Delta_1} \sum_{k:\gamma_k \in \Delta_1} \frac{\hat{u}_j' \boldsymbol{B} \hat{u}_k}{\gamma - \gamma_j - \gamma_k} \hat{v}_j \hat{v}_k'. \tag{7}$$

Author Contributions: Development of mathematical models and analytical calculations, IK, DT and KB; provision of biological data, MM, PP and SS; implementation of statistical software, KB and DT; data analysis, manuscript writing and editing, DT.

Funding: KB was supported by Vetenskapsrådets grant no. 2017-04951. MM and PP were partially supported by IMB PAS.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this article:

BiSSE	binary state speciation and extinction	VF	virulence factor
MuSSE	multistate speciation and extinction	**astA**	heat-stable enterotoxin
cnf1	cytotoxic necrotizing factor	**fimG**	fimbrial protein
fyuA	pesticin receptor protein	**hly1**	alpha hemolysin
iroN	IroN protein	**iutA**	ferric aerobactin receptor
papC	fimbrial protein	**sat**	secreted autotransporter toxin

References

[1] K. Arbuckle and M. P. Speed, *Antipredator defenses predict diversification rates*, PNAS, 112, 13597–13602 (2015). doi: 10.1073/pnas.1509811112. Cited on p. 60.

[2] K. B. Athreya, *Some results on multitype continuous time Markov branching processes*, Ann. Math. Stat. 39, 347–357 (1968). doi: 10.1214/aoms/1177698395; MR 0221600 Cited on pp. 61, 62, and 82.

[3] K. B. Athreya and P. E. Ney, *Branching Processes*, Dover Publications Inc. New York, 1972. Cited on pp. 61 and 62.

[4] K. Bartoszek, M. Majchrzak, S. Sakowski, A. B. Kubiak–Szeligowska, I. Kaj, P. Parniewski, *Predicting pathogenicity behavior in Escherichia coli population through a state dependent model and TRS profiling*, PLOS Comput. Biol. 14, e1005931 (2018). doi: 10.1371/journal.pcbi.1005931; PubMed Central PMCID: PMC5809097 . Cited on pp. 59, 60, 66, 67, 73, 80, and 81.

[5] S. L. Chen, M. Wu, J. P. Henderson, T. M. Hooton, M. E. Hibbing, S. J. Hultgren, J. I. Gordon, *Genomic diversity and fitness of E. coli strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection*, Sci. Transl. Med. 5, 184ra60 (2013). doi: 10.1126/scitranslmed.3005497. Cited on p. 67.

[6] A. S. Cross, *What is a virulence factor?* Crit. Care. 12, 196 (2008). doi: 10.1186/cc7127. Cited on p. 60.

[7] R. G. FitzJohn, *Diversitree: comparative phylogenetic analyses of diversification in* `R`, Methods Ecol. Evol. 3, 1084–1092 (2012). doi: 10.1111/j.2041-210X.2012.00234.x. Cited on pp. 60, 67, 68, and 75.

[8] S. Janson, *Functional limit theorems for multitype branching processes and generalized Pólya urns*, Stoch. Proc. Appl. 110, 177–245 (2004). doi: 10.1016/j.spa.2003.12.002. Cited on pp. 61, 62, 71, 81, 82, and 83.

[9] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, Pearson Education Inc. New Jersey, 2007. Cited on p. 71.

[10] I. Jorgensen, P. C. Seed, *How to Make It in the Urinary Tract: A Tutorial by Escherichia coli*, PLoS Pathog. 8, e1002907 (2012). doi: 10.1371/journal.ppat.1002907. Cited on p. 67.

[11] S. Kitamoto, H. Nagao-Kitamoto, P. Kuffa, N. Kamada, *Regulation of*

virulence: the rise and fall of gastrointestinal pathogens, J. Gastroenterol. 51, 195–205 (2016). doi: 10.1007/s00535-015-1141-5. Cited on p. 81.

[12] W. P. Maddison, P. E. Midford, S. P. Otto, *Estimating a Binary Character's Effect on Speciation and Extinction*, Syst. Biol. 56, 701–710 (2007). doi: 10.1080/10635150701607033. Cited on p. 59.

[13] Maple 18.00 (2014). Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario. Cited on pp. 64 and 74.

[14] M. D. Pirie, E. G. H. Oliver, A. Mugrabi de Kuppler, B. Gehrke, N. C. Le Maitre, M. Kandziora, D. U. Bellstedt, *The biodiversity hotspot as evolutionary hot-bed: spectacular radiation of Erica in the Cape Floristic Region*, BMC Evol. Biol. 16, 190 (2016). doi: 10.1186/s12862-016-0764-3. Cited on p. 60.

[15] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/, 2016. Cited on p. 67.

[16] J. L. Sachs, R. G. Skophammer, N. Bansal, J. E. Stajich, *Evolutionary origins and diversification of proteobacterial mutualists*, Proc. R. Soc. B 281: 20132146 (2013). doi: 10.1098/rspb.2013.2146. Cited on p. 60.

[17] G. E. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6, 461–454 (1978). doi: 10.1214/aos/1176344136. Cited on p. 68.

Zastosowanie wielorodzajowego procesu gałązkowego do analizy patogenności bakterii

Daniah Tahir, Ingemar Kaj, Krzysztof Bartoszek,
Marta Majchrzak, Paweł Parniewski, Sebastian Sakowski

Streszczenie W celu zbadania patogenności szczepów *E. coli*, bakterii z rodzaju *Escherichia*, użyto wielorodzajowego markowskiego procesu gałązkowego z czasem ciągłym. W pierwszej kolejności zrobiono przegląd własności wykorzystywanego procesu wraz z najważniejszymi wynikami granicznymi opisującymi zachowanie procesu przy różnych założeniach. Następnie przyjęto konkretny model, w którym rozgałęzenia są zależne od stanu oraz zastosowano go do badania patogenności w zestawie 251 szczepów *E. coli* pochodzących z II Centralnego Szpitala Klinicznego Wojskowej Akademii Medycznej. Parametry modelu, tempa narodzin oraz mutacji, zostały uzyskane metodą największej wiarygodności. Przedstawiona w artykule analiza potwierdza znane własności patogenności bakterii oraz sugeruje nowe ścieżki pracy badawczej.

2010 Klasyfikacja tematyczna AMS (2010): 60B12; 60J85.

Słowa kluczowe: czynniki patogenności, model markowski, proces gałązkowy, szczepy *E. coli*, twierdzenia graniczne.

Daniah Tahir obtained her PhD in Applied Mathematics and Statistics from Uppsala University, Sweden, in 2019 and MPhil in Mathematics from National University of Sciences and Technology, Pakistan, in 2013. Her research interests include stochastic processes in evolutionary biology, phylogenetics, and epidemiological modeling.

Ingemar Kaj is a Professor in Mathematical Statistics at Uppsala University since 1992 with research interests in stochastic processes, random fields, long range dependence, and stochastic models in evolutionary biology and other application areas.

Krzysztof Bartoszek is currently a lecturer of Statistics at the University of Linköping. He is a Computational Biology graduate from the University of Cambridge (MPhil), and has a doctorate in Mathematical Statistics, obtained in 2013 at the University of Gothenburg. His main interests are associated with stochastic processes in phylogenetics.

Sebastian Sakowski received his PhD degree in Computer Science from Silesian University of Technology in 2011, and MSc in Computer Science from University of Lodz in 2004. He is a member of Polish Bioinformatics Society. His current interests include automata theory, DNA computing, bioinformatics and computational biology.

Marta Majchrzak, Ph.D., is a molecular biologist at the Laboratory of Molecular Genetics, Institute of Medical Biology, Polish Academy of Sciences, Poland. Her research interest are in biochemistry and microbiology.

Pawel Parniewski, Ph.D., is the Head of the Laboratory of Molecular Genetics at the Institute of Medical Biology, Polish Academy of Sciences, Poland. His research interests are in microbiology and molecular biology.

Daniah Tahir, Ingemar Kaj
Uppsala University, Department of Mathematics
Uppsala SE-751 06, Sweden
E-mail: daniahtahir@gmail.com; ingemar.kaj@math.uu.se

Krzysztof Bartoszek
Linköping University, Dept. of Comp. & Inf. Sci.
Linköping SE-581 83, Sweden
E-mail: krzysztof.bartoszek@liu.se, krzbar@protonmail.ch

Marta Majchrzak, Paweł Parniewski
Polish Academy of Sciences, Institute of Medical Biology, Łódź, Poland
E-mail: aktram@poczta.onet.pl; pparniewski@cbm.pan.pl

Sebastian Sakowski
University of Łódź, Fac. of Math. & Comp. Sci.
E-mail: sakowski@math.uni.lodz.pl

Communicated by: Mirosław Lachowicz

(Received: 25th of February 2019 ; revised: 22nd of January 2020)