NEUTRAL EVOLUTION AND DISPERAL LIMITATION PRODUCE BIOGEOGRAPHIC PATTERNS IN *MICROCYSTIS AERUGINOSA*

POPULATIONS OF LAKE SYSTEMS

A Thesis Presented

By

Sahar Shirani

to

The Department of Civil & Environmental Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

in the field of

Civil Engineering

Northeastern University
Boston, Massachusetts

May 2017
Molecular observations reveal substantial biogeographic patterns of cyanobacteria within systems of connected lakes. An important question is the relative role of environmental selection and neutral processes in the biogeography of these systems. Here we quantify the effect of genetic drift and dispersal limitation by simulating individual cyanobacteria cells using an agent-based model (ABM). In the model, cells grow (divide), die and migrate between lakes. Each cell has a full genome that is subject to neutral mutation (i.e. the growth rate is independent of the genome). The model is verified by simulating simplified lake systems, for which theoretical solutions are available. Then, it is used to simulate the biogeography of the cyanobacterium *Microcystis aeruginosa* in a number of real systems, including the Great Lakes, Klamath River, Yahara River and Chattahoochee River. Model output is analyzed using standard bioinformatics tools (BLAST, MAFFT). The emergent patterns of nucleotide divergence between lakes are dynamic, including gradual increases due to accumulation of mutations and abrupt changes due to population takeovers by migrant cells (coalescence events). The model predicted nucleotide divergence is heterogeneous within systems and for weakly connected lakes it can be substantial. For example, Lakes Superior and Michigan are predicted to have an average genomic nucleotide divergence of 8,200 bp or 0.14%. The divergence between more strongly connected lakes is much lower. Our results provide a quantitative baseline for future biogeography studies. They show that dispersal limitation can be an important factor in microbe biogeography, which is contrary to the common belief, and could affect how a system responds to environmental change.
ACKNOWLEDGMENTS

We thank Steve Chapra for providing the Great Lakes model and Euan Reavie for cyanobacteria concentration data in the Great Lakes. Peter Furth and Haris Koutsopoulos provided advice on the theoretical aspect. Haiwei Luo helped with the mutation rates. Three anonymous reviewers provided constructive criticism. Financial support was provided by the National Science Foundation (NSFENG/ECCS/1404163) and the MIT Sea Grant College Program, under NOAA Grant Number NA10OAR4170086, MIT SG Project Number 2010-R/RT-2/RC-117.
# TABLE OF CONTENTS

1. Introduction ..............................................................................................................1
2. Model Description ...................................................................................................3
   2.1 Overview ...........................................................................................................4
   2.2 Transport ..........................................................................................................6
   2.3 Growth and Death .............................................................................................7
   2.4 Genomes, Mutation and Quantifying Divergence ..............................................8
   2.5 Downscaling .....................................................................................................9
3. Model Application ..................................................................................................10
   3.1 Comparison to Theory .......................................................................................10
      3.1.1 Basic Concepts ..........................................................................................10
      3.1.2 Stepping Stone Model ..............................................................................11
      3.1.3 Two Boxes with Different Population Sizes .............................................14
   3.2 Lake Systems ..................................................................................................18
      3.2.1 Great Lakes ..............................................................................................18
      3.2.2 Klamath River ..........................................................................................21
      3.2.3 Yahara River ............................................................................................22
      3.2.4 Chattahoochee River ...............................................................................23
   3.3 Model Accuracy ...............................................................................................25
      3.3.1 Parameter Description ..............................................................................25
      3.3.2 Selection of Parameter Values ..................................................................26
4. Results and Discussion ...........................................................................................29
   4.1 Temporal Pattern for Two Segments ...............................................................29
   4.2 Temporal Patterns at the System Scale ............................................................30
   4.3 Dynamic Phylogeny .........................................................................................32
   4.4 Spatial Patterns ...............................................................................................33
   4.5 Comparison among Lake Systems ..................................................................35
   4.6 Relation to other species and systems ..............................................................37
5. Summary and Outlook.................................................................38
6. References....................................................................................40
LIST OF TABLES

Table 3-1 Great Lakes cyanobacteria concentration....................................................19
Table 3-2 Segment Characteristics............................................................................20
Table 3-3 Migration Rates due to advection...............................................................20
Table 3-4 Migration Rates due to diffusion...............................................................21
Table 3-5 Klamath River Characteristics...................................................................22
Table 3-6 Segment characteristics............................................................................22
Table 3-7 Migration Rates..........................................................................................22
Table 3-8 Yahara River Characteristics....................................................................23
Table 3-9 Segment Characteristics............................................................................23
Table 3-10 Migration Rates.......................................................................................23
Table 3-11 Chattahoochee River Characteristics......................................................24
Table 3-12 Segment Characteristics...........................................................................24
Table 3-13 Migration Rates.......................................................................................24
Table 3-14 Number of accepted average nucleotide divergence (\(\delta\)) of each pair of lakes out of total of 36 pairs for the Great Lakes...................................................28
Table 3-15 Number of accepted average nucleotide divergence (\(\delta\)) of each pair of lakes out of total of 3 pairs for the Klamath River.......................................................28
Table 3-16 Number of accepted average nucleotide divergence (\(\delta\)) of each pair of lakes out of total of 10 pairs for the Yahara River.........................................................29
Table 3-17 Number of accepted average nucleotide divergence (\(\delta\)) of each pair of lakes out of total of 10 pairs for the Chattahoochee River...............................................29
LIST OF FIGURES

**Fig. 2.1.** Simulated lake systems. Top row: Geographic representation (not to scale). Downstream ends are marked with an arrow. Bottom row: Flow chart representation. Circles are segments. Area is proportional to *Microcystis* population size (relative within each system). Arrows are advection (solid red) and diffusion (dashed blue) migration rates. Thickness is proportional to square root of migration rate..........................4

**Fig. 2.2.** Linear system of lakes.................................................................6

**Fig. 3.1.** Circular stepping stone system....................................................12

**Fig. 3.2.** Average nucleotide divergence of two subpopulations (δ)................14

**Fig. 3.3.** Two boxes with different population size and one-way migration.........14

**Fig. 3.4.** Average nucleotide divergence of two subpopulations (δ̅) ............18

**Fig. 3.5.** Average nucleotide divergence (δ) of each pair of lakes compared with base case when changing sample interval. Error bars are based on 99% statistical significance......28

**Fig. 4.1.** Sawtooth pattern of nucleotide divergence between Lakes Superior (SUP) and Southern Michigan (SM)..........................30

**Fig. 4.2.** Nucleotide divergence between Western Erie (WE) and other segments of the Great Lakes.................................................................31

**Fig. 4.3** Dynamic phylogeny of *Microcystis aeruginosa* in the Great Lakes........33

**Fig. 4.4.** Heat maps of average nucleotide divergence for each lake system. Average nucleotide divergence (bp / 1 Mbp) in.........................................................34

**Fig. 4.5.** Correlation between model-predicted average nucleotide divergence (δ) and relative import rate (Ri). .................................................................37
1. Introduction

Understanding the mechanisms underlying the production of biogeographic patterns and their dynamics is an important endeavor in ecology and evolution. Two general mechanisms are natural selection to environmental factors and neutral evolution coupled with dispersal limitation. For microbes with large population sizes and high dispersal rates, it is often assumed that natural selection is more important (i.e. “Everything is everywhere but the environment selects” (Baas-Becking, 1934)). However, dispersal limitation has been demonstrated for microbial communities in soil, freshwater, marine water and other habitats (Hanson, Fuhrman, Horner-Devine, & Martiny, 2012). Also, it was recently shown that neutral processes can produce biogeographic patterns at the global ocean scale (Ferdi L. Hellweger, Erik van Sebille, & Neil D. Fredrick, 2014). Here, we explore the role of dispersal limitation in the biogeography of cyanobacteria in lake systems (i.e., lakes connected by streams or canals).

Several studies have explored the biogeography of microbes in systems of connected lakes. Van der Gucht et al. (2007) explored dispersal limitation of bacteria within the De Maten system, a collection of small connected lakes, using 16S rRNA fingerprinting. Seven of the 42 OTUs (operational taxonomic units) found in the system were present in all eleven lakes examined. After subtracting out the effect of environmental variables, space did not significantly explain any of the variability, suggesting the system is well connected. Miller and McMahon (2011) investigated the cyanobacteria ecology in several connected lakes in the Yahara Watershed using molecular methods. They found that variability of community composition within lakes is greater than between lakes. Miller, Beversdorf, Chaston, and McMahon (2013) and Beversdorf, Miller, and McMahon (2013) also studied this system and found spatial variability of community composition within and between lakes. Davis et al. (2014) explored genetic connectivity of Microcystis aeruginosa in Lakes St. Clair, Erie and Ontario using the mcyA toxin gene. They found six groups at 99% identity that were present throughout the system, suggesting high connectivity. Dyble, Fahnenstiel, Litaker, Millie, and Tester (2008) also found similar results when comparing mycB genes from Saginaw Bay (Lake Huron) and Lake Erie. Their
study suggests that the difference in genetic composition could affect the produced microcystin toxin concentration values in each location. Zwirglmaier, Keiz, Engel, Geist, and Raeder (2015) investigated the connectivity of bacteria in the Oster Lakes, a collection of small connected lakes, using 16S rRNA sequencing. At the 97% identity level, some OTUs were present throughout the system and others were confined to individual lakes. The distribution of OTUs was related to environmental parameters suggesting a role for environmental selection. The similarity of lake communities decreased with distance and neighboring lakes shared more OTUs than non-neighboring lakes, suggesting dispersal limitation may also be a factor in this system. Otten, Crosswell, Mackey, and Dreher (2015) investigated the connectivity of Microcystis aeruginosa within the Klamath River, which includes a number of impoundments. Based on nucleotide identity of a highly-conserved gene, they found a single population (OTU) at the 97% level, but multiple subpopulations when considering single nucleotide polymorphism (SNP). However, the temporal succession of these subpopulations is consistent across most of the system, which suggests they are well connected in space. Hayden and Beman (2015) studied the biogeography of the microbial community in a number of lakes in Yosemite National Park using 16S rRNA sequencing. They found that some of the variability can be explained by environmental variables as well as geographic distance, suggesting that both environmental selection and dispersal limitation may be important in this system. In summary, these empirical studies show that environmental selection as well as neutral processes may play a role in the biogeography of microbes in lake systems.

We propose that mechanistic modeling can also contribute to understanding the relative role of these processes in producing biogeographic patterns in these systems. Modeling natural selection is challenging, because the linkage between the genome and fitness is difficult to quantify. However, the mechanisms underlying neutral evolution and dispersal limitation, including growth, death, mutation and transport are relatively well understood for connected lakes (i.e., vs. soil or unconnected lake systems), making it possible to make quantitative predictions without prior calibration (e.g., dispersal parameter, Condit et al. (2002)).
In the context of population genetics, lakes can be considered islands within a “sea” of land (Reche, Pulido-Villena, Morales-Baquero, & Casamayor, 2005). The biogeography of island systems has been studied extensively from a theoretical perspective (Hamilton, 2011). The basic “island model” is based on assumptions of equal population size and migration rate (Wright, 1943). The theory is further developed by introducing the “stepping stone model” which assumes the migration rate between subpopulations is a function of their distance rather than a constant value (Strobeck, 1987).

Although these analytical models can produce estimates of nucleotide divergence for numerous scenarios, their simplifying assumptions, such as symmetrical migration patterns and equal subpopulation sizes, are not applicable to natural lake systems. Typical phytoplankton ecology models can be used to predict biogeographic patterns produced by environmental selection (Barton, Dutkiewicz, Flierl, Bragg, & Follows, 2010), but they generally use a population-level approach and implicitly assume “everything is everywhere” so they cannot be used to simulate dispersal limitation. Numerical simulations, including Monte Carlo and agent-based models (ABM), are not constrained by these assumptions (F. L. Hellweger, E. van Sebille, & N. D. Fredrick, 2014).

Here we explore the role of neutral processes in producing biogeographic patterns in cyanobacteria populations in systems of connected lakes. How much nucleotide divergence can be produced by neutral mutation and maintained by dispersal limitation? Are there any temporal or spatial patterns that arise from the properties of the system? We present an ABM that tracks individual cyanobacteria cells including their growth, death, mutation and migration in natural lake systems with non-uniform properties (e.g. population size, migration rate). Biogeography is quantified as genomic nucleotide divergence and computed from the DNA sequence of the cells in the model. We apply the model to a number of lake systems, which shows that neutral processes can produce substantial biogeographic patterns in those systems.

2. Model Description
The model simulates a number of lake systems, including the Great Lakes, Klamath River, Yahara River and Chattahoochee River (Fig. 2.1). These systems were selected because they cover a range of characteristics (e.g., size, connectivity) and have been the subject of past microbe biogeography studies (i.e., they are test beds for studying microbial biogeography, see literature review above). The model methodology and input are described in this section (see Model Application for details of input).

Fig. 2.1. Simulated lake systems. Top row: Geographic representation (not to scale). Downstream ends are marked with an arrow. Bottom row: Flow chart representation. Circles are segments. Area is proportional to 
*Microcystis* population size (relative within each system). Arrows are advection (solid red) and diffusion (dashed blue) migration rates. Thickness is proportional to square root of migration rate. (A) Great Lakes, including Lakes Superior (SUP), Michigan (Northern (NM) and Southern (SM)), Huron (Northern (NH) and Southern (SH)), Erie (Western (WE), Central (CE) and Eastern (EE)) and Ontario (ONT). (B) Klamath River, including Upper Klamath Lake (UKL), Copco Reservoir (COP) and Iron Gate Reservoir (IG). (C) Chattahoochee River including Lake Lanier (LAN), West Point Lake (WPL), Lake Harding (HAR), Walter F. George Reservoir (WFG) and Lake Seminole (SEM). (D) Yahara River, including Lakes Mendota (MEN), Wingra (WIN), Monona (MON), Waubesa (WAU) and Kegonsa (KEG).
2.1 Overview

The systems are generally sub-divided into individual lakes, although some lakes were omitted because they are small (e.g., Lakes Ewauna and Keno and Boyle Reservoir on the Klamath River) and some were divided into segments because they are large and/or have different characteristics (e.g., Western, Central and Eastern Lake Erie, Fig. 2.1). Each segment is represented as a well-mixed volume that supports a dynamic cyanobacteria population and is connected to others through advection and diffusion. Individual cells are simulated using an agent-based modeling (ABM) approach (Kreft et al., 2013). For model parameterization, *Microcystis aeruginosa* is used as a representative species of cyanobacteria.

Microbe population models are typically based on a mass balance in an Eulerian framework and that constitutes a good starting point here. Each segment is represented as a well-mixed volume. Microbes are transported between segments by advection and diffusion. For advection, a number of methods (e.g., centered or backward difference) are available (Chapra, 2008). Here, a backward difference is used because it is most compatible with the agent-based implementation (see Downscaling). The diffusion flux is based on the concentration gradient between the segments. Within each lake the microbes grow and die. For a simple linear system of lakes (Fig. 2.2), the mass balance equation is:

\[
\frac{d c_i}{dt} = \frac{Q_{i-1,i} c_{i-1}}{V_i} - \frac{Q_{i,i+1} c_i}{V_i} + E'_{i-1,i} (c_{i-1} - c_i) + E'_{i,i+1} (c_{i+1} - c_i) + V_i k_g c_i - V_i k_d c_i
\]  

(2-1)

Where \( V \) = volume (m\(^3\)), \( c \) = microbe concentration (cells/m\(^3\)), \( t \) = time (d), \( Q \) = advective flow (m\(^3\)/d), \( E' \) = bulk diffusion coefficient (m\(^3\)/d), \( k_g \) = growth rate (1/d) and \( k_d \) = death rate (1/d).

For the agent-based model, transport is based on migration rates and it is useful to put the mass balance in those terms. For this, \( k_{ma} \) (1/d) and \( k_{md} \) (1/d) are defined as migration rate due to advection and diffusion, respectively:

\[
(2-2.a)
\]
Using these migration rates, the mass balance becomes:

\[
\frac{d}{dt} V_i = k_{ma,i-1,i} V_{i-1} c_{i-1} - k_{ma,i,i+1} V_i c_i + k_{md,i-1,i} V_{i-1} c_{i-1} - V_i c_i + k_{md,i+1,i} V_{i+1} c_{i+1} - k_{md,i,i} V_i c_i + V_i k_{gl} c_i - V_i k_{dl} c_i
\]

(2-3)

**Fig. 2.2.** Linear system of lakes.

### 2.2 Transport

Cells are transported based on advection and diffusion migration rates in a stochastic manner. For example, the advection migration rate for cells from Northern Michigan (NM) to Northern Huron (NH) \((k_{ma, NM, NH}, l/d)\) is calculated as the corresponding flow rate divided by the volume of Northern Lake Michigan. The probability of a cell advecting during a time step \((\Delta t)\) is \(k_{ma, NM, NH} \Delta t\). For each time step and each cell, if a random number drawn from a standard uniform distribution is less than this probability the cell is moved. Migration by diffusion is calculated in a similar manner based on a diffusion migration rate. Our simulations explore divergence over long time scales (e.g. up to 20k years) and seasonal dynamics in transport are not resolved. For the Great Lakes, advection and diffusion migration rates are based on a previous water quality model (Chapra & Dolan,
2012) and for other systems advection migration rates are based on measured flow rates (see Model Application).

We assume that transport via the major natural connections of the lakes dominates and that other mechanisms, including minor waterways (locks and canals), air, birds or boats (via roads or against currents) (Kristiansen, 1996) are not important. Those mechanisms do transport microbes and theoretically even a single microbe transported by them can take over the population of the destination lake and disrupt the population divergence. However, for the neutral case, this probability is proportional to the number of microbes transported (see Model Application). A conservative back-of-the-envelope calculation illustrates that this is unlikely to be important for the bird scenario. 17 million waterfowl migrate annually through the Great Lakes (waterfowl days, (Moore, Badzinski, Cuthbert, & Wires, 2016)). If it is assumed that each of them transports 1 L of water (or an equivalent amount of microbes) per day from Lake Ontario to Lake Erie, that would amount to 2.00E7 L. This is over seven orders of magnitude lower than the amount of water going the other way over Niagara Falls (2.00E14 L). By the same argument, import from sources outside of the lake system is unlikely to affect the neutral nucleotide divergence.

2.3 Growth and death

Growth (division) and death events are stochastic. The probability of a cell dividing within a given time step is calculated based on a growth rate \( k_g \) in the same manner as advection is calculated above. The model uses a constant growth rate of 0.53 \( /d \) based on estimates for phytoplankton in freshwater (Reynolds & Irish, 1997). The death rate is assigned to simulate a constant population size in each segment. Theoretically, this could be achieved by using a death rate equal to the growth rate and correcting for inflow and outflow. However, since the growth, death and transport processes are stochastic, there will always be a small net increase or decrease, resulting in a drift in the population size. To avoid this and stabilize the population, a feedback control is added to the death rate formulation. Specifically, the death rate varies from a base value (equal to the growth rate) with the population size in each segment. When the population size drops below a specified
carrying capacity the death rate decreases leading to a rise in population and vice versa. This approach allows us to simulate dynamic, but relatively constant population sizes in each segment. Population sizes are based on site specific information (see Model Application). For example, for the Great Lakes, they are based on observed cyanobacteria concentration in each segment (Reavie, Barbiero, Allinger, & Warren, 2014).

The model does not resolve seasonal population dynamics. It assumes cells are present in the water column and grow continuously. However, phytoplankton populations can exhibit strong temporal variability (succession, blooms) and some species or strains may only represent a substantial portion of the total population over a short time period (e.g., overwintering in the sediment bed, (Kutovaya et al., 2012)). Ignoring this variability does not introduce a significant error into our analysis for the following reason. When cells are absent from the water column (i.e., in the sediment bed), they do not grow or mutate and are not subject to transport (i.e., via outflow in the water column). Consequently all processes in the model are “on hold”. Explicitly considering this variability would effectively break up the time series of divergence (e.g., Fig. 4.1), but the long-term average nucleotide divergence would not be affected. Seasonal variability in the population size will also not affect the nucleotide divergence as long as the relative population sizes do not change (i.e., all segments exhibit the same temporal pattern) (see Downscaling). We confirm this with time-variable simulations (see Results and Discussion).

2.4 Genomes, Mutation and Quantifying Divergence

Each cell has a genome consisting of an array of ATGC letters (nucleotides). The model can be run with an actual DNA sequence, and we present results from a simulation with the 5.84E6 bp Microcystis aeruginosa genome (Fig. 4.3). However, for computational convenience, we use a smaller synthetic 1.00E5 bp sequence of random nucleotides for most simulations (see Model Description).

At division, at a specified mutation rate (m, mutations/bp/generation), a nucleotide may change. The mutation rate is based on a literature value for E. coli (m = 5.4E-10
mutations/bp/generation), which should be a good estimate for Microcystis aeruginosa considering the similar genome sizes (E. coli: 4.64E6 bp, Microcystis aeruginosa: 5.84E6 bp) (Drake, Charlesworth, Charlesworth, & Crow, 1998; Lynch et al., 2016). Several estimates of point mutation rates for stress resistance in Microcystis aeruginosa are available ((García-Villada et al., 2004), 1.8E-6; (López-Rodas et al., 2007), 3.1E-7 – 3.6E-7; (Costas, Flores-Moya, & López-Rodas, 2008), 1.1E-6 – 1.1E-5; (del Mar Fernández-Arjona et al., 2013), 7.1E-7). However, those rates were obtained using fluctuation analyses, which is subject to bias (Long et al., 2016). Also, the values are much higher than that for E. coli, which is inconsistent with the observation that mutation rates are similar for microbes with similar genome sizes (Drake et al., 1998; Lynch et al., 2016). We suspect that the mutation rates may be higher at these resistance loci and do not reflect a genome-wide average, which is what is required by the model. Our results can readily be converted for different mutation rates (see Results and Discussion). The mutation rate was increased to account for base pair changes by recombination. Estimates of the relative rate of recombination to mutation (r / m) were developed for a number of prokaryotes, including Microcystis aeruginosa (r / m = 18.3), using a coalescent-based method implemented in the ClonalFrame program (Tanabe, Kasai, & Watanabe, 2007; Vos & Didelot, 2009). Microcystis aeruginosa has a relatively high r / m ratio, ranking 7th out of 48 prokaryotes summarized by Vos and Didelot (2009), but it also is the only free-living freshwater cyanobacteria on that list. The corresponding value was added to the mutation rate (m + r = 1.04E-8 changes/bp/division).

The model writes the genomes of a sample of cells from each lake at a specified time interval to a file. They are then analyzed using BLAST to yield their average nucleotide identity (1-\(\bar{\pi}\)). We quantify biogeography as the average nucleotide divergence (\(\bar{\delta}\)), which is the amount of nucleotide difference between two populations that can be attributed to dispersal limitation among them. It is calculated as the difference between the total nucleotide difference (\(\bar{\pi}_{xy}\), based on a sample of cells from both segments x and y) and the local nucleotide diversities (\(\bar{\pi}_{x}, \bar{\pi}_{y}\), based on a sample of cells from segment x and y only) (F. L. Hellweger et al., 2014).
2.5 Downscaling

The model simulates individual cells using an agent-based approach. In each time step, each individual is subject to growth, death and migration. For the agent based model, transport, growth (division) and death are discrete stochastic events. That is, the probability of a cell dying during a time step ($\Delta t$) is $k_{dl}\Delta t$. Division and migration events are calculated in a similar manner.

Using the actual cell population as the agent population in the model would mean simulating a very large number of cells (e.g. $\sim 10^{22}$ for the Great Lakes, see Model Application). We therefore divide the number of real cells in each subpopulation by a constant value which results in a smaller number of agents, called “super-individuals”, that is computationally tractable and can still capture the effect of proportional population sizes (Ferdi L. Hellweger & Bucci, 2009; F. L. Hellweger et al., 2014).

This downscaling does not affect the results, because the nucleotide divergence between two lakes is a function of the relative and not the absolute population size ((F. L. Hellweger et al., 2014), see Model Application). However, the total nucleotide difference ($\overline{\pi}_{xy}$) and local nucleotide diversities ($\overline{\pi}_x, \overline{\pi}_y$) depend on absolute population size, so the model does not make realistic predictions about these parameters. To verify our model, it is applied to a number of simplified scenarios and the results are compared to those of theory (see Model Application).

3. Model Application

3.1 Comparison to Theory

3.1.1 Basic Concepts

The nucleotide difference ($\pi$, mutations/genome) between two independently evolving cells is a function of the mutation rate ($k_u$, mutations/genome/generation) and time to their common ancestor, the coalescence time ($T_c$, generations):
\[ \pi = 2 k_u T_c \] (3-1.a)

Equation (3-1.a) shows that \( \pi \) increases linearly with \( T_c \). The time-average \( \pi \) between two coalescence events spaced \( T_c \) apart is \( 0.5 \times 2 \times k_u \times T_c \). For multiple coalescence events we have to consider that the coalescence time is a random variable. This means that the average value of \( \pi \) cannot simply be calculated directly from the average coalescence time using Equation (3-1.a). Rather, the distribution of \( T_c \) has to be considered.

Consider for example, two scenarios with the same average \( T_c \), but different distributions. One scenario has a coalescence event in a short time followed by a coalescence event after a longer time (\( T_c = 1, 49 \)). Another scenario has two coalescence events evenly spaced (\( T_c = 25, 25 \)). According to (S3-1.a), the time-average \( \pi \) is \( ((k_u \times 1 \times 1 + k_u \times 49 \times 49)/50) = 48.04 k_u \) for the first scenario and \( ((k_u \times 25 \times 25 + k_u \times 25 \times 25)/50) = 25 k_u \) for the second scenario.

The coalescence time follows an exponential distribution. The problem is analogous to calculating the average waiting time for buses, which has been solved in transportation engineering (Larson 1981). Following this approach, the time-average nucleotide divergence is:

\[ \bar{\pi} = 2 k_u [ (\bar{T}_c^2 + \sigma_{T_c}^2) / (2 \bar{T}_c^2) ] \] (3-1.b)

In which \( \sigma_{T_c}^2 \) is the variance of \( T_c \) that equals to \( \bar{T}_c \). Therefore,

\[ \bar{\pi} = 2 k_u \bar{T}_c \] (3-1.c)

Thus, although Equation (3-1.a) cannot generally be applied to yield the time-average \( \pi \) for time-average \( T_c \), the fact that \( T_c \) is exponentially distributed means that it can be used in this case.
3.1.2 Stepping Stone Model

The circular stepping stone model assumes the migration rate \( m \) (transport/generation) (note that here we use nomenclature consistent with (Slatkin 1993)), whereas in the main paper \( m \) refers to mutation rate) between demes (i.e. subpopulations, lakes) is a function of their distance. In this context, it is assumed that the model consists of 4 subpopulations, all with the same number of individuals \( N \) and symmetric migration rates, as shown in Fig. 3.1.

![Circular stepping stone system](image)

**Fig. 3.1** Circular stepping stone system.

Slatkin (Slatkin, 1993) calculated the average coalescence time for a diploid population with \( d \) subpopulations as:

\[
T_{c,i} = 2Nd \\
\bar{T}_{c,ij} = 2Nd + \frac{(d - k)k}{2m}
\]  

(3-2.a)  
(3-2.b)

Where \( T_{c,i} \) (generations) is the average coalescence time for two individuals from the same subpopulation and \( \bar{T}_{c,ij} \) is the average coalescence time for two individuals from different subpopulations that are \( k \) steps apart.

Strobeck (Strobeck, 1987) derived the equations for nucleotide difference in a stepping stone model for a diploid population as:
\[
\bar{\pi}_i = d4Nk_u \quad (3-3.a)
\]
\[
\bar{\pi}_{ij} = d4Nk_u + \frac{k_u}{m} k(d - k) \quad (3-3.b)
\]

Where \(\bar{\pi}_i\) (mutations/genome) is the average nucleotide difference between two cells from one subpopulation and \(\bar{\pi}_{ij}\) is the corresponding value for two subpopulations. These equations can also be obtained by applying Equation (3-1.c) to Equations (3-2.a) and (3-2.b). For a haploid population all \(2N\)s in above formulas are substituted by \(N\). Subtraction of these two values provides the average nucleotide divergence of two subpopulations that are \(k\) steps apart (\(\bar{\delta}\), mutations/genome):

\[
\bar{\delta} = \frac{k_u}{m} k(d - k) \quad (3-4)
\]

The migration rate on a per generation basis \((m)\) can be converted to a per time basis \((k_m)\) using the generation time \((t_g = \ln(2) / k_g, d)\):

\[
k_m = \frac{m}{t_g} \quad (3-5)
\]

To test the full model, it is used to simulate this scenario. Specifically, the simulation is done for four lakes with equal population size of 3000 agents and equal migration rates of 5E-5 (1/d) between adjacent lakes. The growth and mutation rates are as for the Great Lakes simulation (see Lake Systems). Average nucleotide divergence (\(\bar{\delta}\)) and local diversity (\(\bar{\pi}_i\)) and total diversity (\(\bar{\pi}_{ij}\)) are calculated based on BLAST comparison of the genomes of a sample of cells.

The results (Fig. 3-2) show that after some time, the model values are close to those predicted by theory with slight differences that can be attributed to stochastic and sampling variability.
Fig. 3.2. Average nucleotide divergence of two subpopulations ($\delta$). (A) one step apart. (B) two steps apart. Red lines represent the running average of the model values and green lines represent the theoretical value.

3.1.3 Two Boxes with Different Population Sizes

We derive the theory for two subpopulations, each with a different number of individuals and migration from one box to the other by advection. These assumptions are applicable to many natural two-lake systems.

Fig. 3.3. Two boxes with different population size and one-way migration.
The number of cells \((N)\) transported with a specific migration rate \((m, 1/\text{gen})\) (note that here we use nomenclature consistent with Slatkin (1993), whereas in the main paper \(m\) refers to mutation rate) from lake 1 to lake 2 \((N_{TR})\) is:

\[
N_{TR} = m_{1,2} N_1 t 
\]

(3-6.a)

The probability of any one cell taking over the population in lake 2 \((P_{TO})\) is:

\[
P_{TO} = \frac{1}{N_2} 
\]

(3-6.b)

Therefore, the number of takeovers \((N_{TR+TO})\) of the population in lake 2 by migrants from lake 1 can be calculated as:

\[
N_{TR+TO} = N_{TR} P_{TO} = m_{1,2} t \frac{N_1}{N_2} 
\]

(3-6.c)

The average time between takeovers, i.e. the coalescence time \((\bar{T}_c, \text{generations})\), is:

\[
\bar{T}_c = \frac{t}{N_{TR+TO}} = \frac{t}{m_{1,2} t \frac{N_1}{N_2}} = \frac{1}{m_{1,2} \frac{N_1}{N_2}} 
\]

(3-6.d)

The equation is consistent with that of Slatkin (Slatkin, 1993) (Equation 3-2) for \(N_1 = N_2\) and 2 \(m=m_{1,2}\) (to account for two-way migration).

Assuming the difference between two independently evolving cells that descend from the same ancestor increases in a linear, clockwise manner based on mutation rate \((k_u, \text{mutation}/\text{genome}/\text{generation})\) and time, the nucleotide divergence can be estimated as:

\[
\delta_{12} = \bar{T}_c k_{u1} + \bar{T}_c k_{u2} 
\]

(3-7.a)
Since the coalescence time $T_c$ is an exponential random variable, when computing average nucleotide divergence, the average coalescence time ($\bar{T}_c$) should be accounted for:

$$\bar{\delta}_{12} = [(\bar{T}_c^2 + \sigma_{T_c}^2) / (2\bar{T}_c)] k_{u1} + [(\bar{T}_c^2 + \sigma_{T_c}^2) / (2\bar{T}_c)] k_{u2}$$

(3-7.b)

In which $\sigma_{T_c}^2$ is the variance of $T_c$ that equals to $\bar{T}_c$. Therefore,

$$\bar{\delta}_{12} = \bar{T}_c k_{u1} + \bar{T}_c k_{u2}$$

(3-7.c)

$$\tilde{\delta}_{12} = \frac{1}{m_{1,2}} \frac{N_2}{N_1} (k_{u1} + k_{u2})$$

(3-7.d)

This equation can be converted to one using migration rate in units of per time using Equation 3-5:

$$\bar{\delta}_{12} = \frac{k_g}{k_{m,1,2} \ln(2)} \frac{N_2}{N_1} (k_{u1} + k_{u2})$$

(3-7.e)

This equation can be compared to the relative import rate:

$$R_I = \frac{k_{m,1,2} N_1}{k_g N_2}$$

(3-8)

The average nucleotide divergence is proportional to $N_2 / (k_{m,1,2} N_1)$, where the relative import rate is proportional to $(k_{m,1,2} N_1) / N_2$. Therefore, an increase in the relative import rate results in a corresponding decrease in the nucleotide divergence. Equations 3-7.e and 3-8 can be combined (with $k_{u1} = k_{u2}$):

$$\tilde{\delta}_{12} = \frac{2 k_u}{\ln(2) R_I}$$

(3-9)

To test the full model, it is used to simulate this scenario. Specifically, there are two lakes with different population sizes of 2000 and 3000 agents. The migration rate is 5E-
Simulations with smaller upstream population and larger downstream population and vice versa are presented. The growth rate and the mutation rate are as for the Great Lakes simulation (see Lake Systems). Average nucleotide divergence ($\bar{\delta}$) and local diversity ($\bar{\pi}_i$) and total diversity ($\bar{\pi}_{ij}$) are calculated based on BLAST comparison of the sample genomes.

The results (Fig. 3.4) show that after some time, the model values are close to those predicted by theory with slight differences which can be attributed to stochastic and sampling variability.
Fig. 3.4. Average nucleotide divergence of two subpopulations ($\bar{\delta}$) A) migration from smaller subpopulation to bigger subpopulation B) migration from bigger subpopulation to smaller subpopulation. Red lines represent the running average of the model values and green lines represent the theoretical value.

3.2 Lake Systems

3.2.1 Great Lakes

The Great Lakes consists of five major waterbodies, but often models use a finer segmentation to capture gradients or patterns within the lakes. Here, the starting point was the model of Chapra and Dolan which has 21 segments (Chapra & Dolan, 2012). However, those segments differed significantly in size. For example, Green Bay is resolved using 7 segments with areas as small as 52 km², whereas Lake Superior is modeled as one segment with an area of 79800 km². The resulting large difference between the populations is problematic when using an agent-based approach. Specifically, to allow for sufficient agents in the smaller segments and maintaining relative population sizes, the total number
of agents would need to be very large. This is computationally prohibitive. Further, it is not necessary to use that many agents for the larger segments to resolve the population heterogeneity. To address this issue, several segments were combined. The resulting model has 9 segments, as shown in Fig. 2.1A.

Population sizes for each segment are based on observed cyanobacteria concentrations (cells/ml) (Table 3-1) (Reavie et al., 2014). The data corresponds to the 9 segments of the model, which has some lakes resolved using two or more segments. This allows us to capture the observed differences of cell population within one lake, particularly Lake Erie, which has significantly different cyanobacteria concentration between western and eastern parts.

The concentration is multiplied by the volume of the photic zone (Table 3-2) to yield the number of cells in each segment. The calculations are done on a finer segmentation (21 segments of Chapra’s model) and then lumped together. Therefore the volumes used are slightly higher than the photic depth multiplied by area in Table 3-2. Because of computational constraints, the model simulates a fraction of the actual number of cells as agents (Table 3-2).

Segments are connected by advection and diffusion, meaning there is an outflow from upstream to downstream as well as some diffusive exchange across open boundaries. Flow and diffusion rates are from Chapra’s model (Chapra & Dolan, 2012), and they are used to compute the migration rates (see Model Description) (Table 3-3).

A growth rate of 0.53/d was assigned based on estimates of in situ growth rates of phytoplankton in freshwater (Reynolds & Irish, 1997). A constant mutation rate is used throughout the model, as described in the main paper.
Table 3-1 Great Lakes cyanobacteria concentration (Reavie et al., 2014).

<table>
<thead>
<tr>
<th>Segment</th>
<th>Superior</th>
<th>Southern Michigan</th>
<th>Northern Michigan</th>
<th>Northern Huron</th>
<th>Southern Huron</th>
<th>Western Erie</th>
<th>Central Erie</th>
<th>Eastern Erie</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average CYA Concentration</td>
<td>1920.16</td>
<td>2240.83</td>
<td>2677.04</td>
<td>1274.5</td>
<td>27075.92</td>
<td>10172.11</td>
<td>9480.23</td>
<td>6242.69</td>
<td></td>
</tr>
</tbody>
</table>

Table 3-2 Segment Characteristics. Average Depth and Area are from Chapra’s model (Chapra & Dolan, 2012). Photic Depth (http://www.lakeaccess.org/ecology) and Average Depth are compared with each other and the minimum of the two is multiplied by area to result in photic volume. Relative import rate is the rate of cells migrating from the upstream lakes ($N_{up}k_{ma}$) divided by the rate of cell growth in the destination lake ($N_{on}k_{g}$).

<table>
<thead>
<tr>
<th>Segment</th>
<th>Average Depth (km)</th>
<th>Photic Depth (km)</th>
<th>Area (km²)</th>
<th>Photic Volume (km³)</th>
<th>Population (cells)</th>
<th>Population (agents)</th>
<th>Relative Import Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUP</td>
<td>0.148</td>
<td>0.040</td>
<td>79800</td>
<td>3264.57</td>
<td>6.27E+21</td>
<td>3.13E+03</td>
<td>-</td>
</tr>
<tr>
<td>SM</td>
<td>0.081</td>
<td>0.025</td>
<td>28663</td>
<td>716.58</td>
<td>1.61E+21</td>
<td>8.02E+02</td>
<td>-</td>
</tr>
<tr>
<td>NM</td>
<td>0.102</td>
<td>0.025</td>
<td>24935</td>
<td>697.79</td>
<td>1.87E+21</td>
<td>9.34E+02</td>
<td>9.72E-05</td>
</tr>
<tr>
<td>NH</td>
<td>0.081</td>
<td>0.028</td>
<td>16900</td>
<td>984.22</td>
<td>1.25E+21</td>
<td>6.27E+02</td>
<td>9.78E-04</td>
</tr>
<tr>
<td>SH</td>
<td>0.068</td>
<td>0.028</td>
<td>20900</td>
<td>609.73</td>
<td>9.67E+20</td>
<td>4.83E+02</td>
<td>6.72E-04</td>
</tr>
<tr>
<td>WE</td>
<td>0.007</td>
<td>0.019</td>
<td>3680</td>
<td>28</td>
<td>7.58E+20</td>
<td>3.79E+02</td>
<td>1.85E-03</td>
</tr>
<tr>
<td>CE</td>
<td>0.018</td>
<td>0.019</td>
<td>15390</td>
<td>274</td>
<td>2.79E+21</td>
<td>1.39E+03</td>
<td>8.67E-03</td>
</tr>
<tr>
<td>EE</td>
<td>0.026</td>
<td>0.019</td>
<td>6150</td>
<td>116.85</td>
<td>1.11E+21</td>
<td>5.53E+02</td>
<td>8.82E-03</td>
</tr>
<tr>
<td>ONT</td>
<td>0.086</td>
<td>0.020</td>
<td>18960</td>
<td>388.68</td>
<td>2.43E+21</td>
<td>1.21E+03</td>
<td>3.80E-03</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-</td>
<td>-</td>
<td>7080.42</td>
<td>1.90E+22</td>
<td>9.52E+03</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 3-3 Migration Rates due to advection (1/d).

<table>
<thead>
<tr>
<th></th>
<th>SUP</th>
<th>SM</th>
<th>NM</th>
<th>NH</th>
<th>SH</th>
<th>WE</th>
<th>CE</th>
<th>EE</th>
<th>ONT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.6E-05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.9E-05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.6E-04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.7E-04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7.6E-04</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.7E-02</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3-4 Migration Rates due to diffusion (1/d).

<table>
<thead>
<tr>
<th></th>
<th>SUP</th>
<th>SM</th>
<th>NM</th>
<th>NH</th>
<th>SH</th>
<th>WE</th>
<th>CE</th>
<th>EE</th>
<th>ONT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SM</td>
<td>0</td>
<td>0</td>
<td>3.8E-03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NM</td>
<td>0</td>
<td>3.9E-03</td>
<td>0</td>
<td>1.7E-04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NH</td>
<td>0</td>
<td>0</td>
<td>1.2E-04</td>
<td>0</td>
<td>3.3E-03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.4E-03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.2E-03</td>
<td>0</td>
</tr>
<tr>
<td>CE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.2E-04</td>
<td>0</td>
<td>2.0E-03</td>
<td>0</td>
</tr>
<tr>
<td>EE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.6E-03</td>
<td>0</td>
</tr>
<tr>
<td>ONT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.2.2 Klamath River

The Klamath River system consists of three major reservoirs connected via the Klamath River (Fig. 2.1B) located in the northwestern United States (i.e. Oregon and California). In order to maintain the observed relative population sizes and keep the model computationally feasible (see discussion in Great Lakes Section), other lakes and reservoirs along the river with low population sizes had to be omitted. Each reservoir is modeled as a segment. The arrangement represents a linear system of lakes starting at Upper Klamath Lake (UKL) flowing downstream to the Pacific Ocean. Population sizes are based on observed cyanobacteria concentration (Pacificorp.com, 2016) (cells/ml) (Table 3-5). The migration rate is based on advection only since the lakes do not share open boundaries (Table 3-7). Flow rates are based on USGS gauges discharge data. Growth and mutation rates are the same as in the Great Lakes model.
Table 3-5 Klamath River Characteristics.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Upper Klamath Lake</th>
<th>Copco</th>
<th>Iron Gate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average CYA Concentration (cells/mL)</td>
<td>71.90</td>
<td>506.65</td>
<td>278.20</td>
</tr>
</tbody>
</table>

Table 3-6 Segment characteristics (Pacificorp.com, 2016).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>UKL</td>
<td>4.30E-03</td>
<td>2.49E+02</td>
<td>1.07E+00</td>
<td>7.70E+16</td>
<td>1.38E+04</td>
<td>-</td>
</tr>
<tr>
<td>COP</td>
<td>1.00E-02</td>
<td>4.00E+00</td>
<td>4.00E-02</td>
<td>2.03E+16</td>
<td>3.62E+03</td>
<td>7.24E-04</td>
</tr>
<tr>
<td>IG</td>
<td>1.20E-02</td>
<td>3.80E+00</td>
<td>4.56E-02</td>
<td>1.27E+16</td>
<td>2.27E+03</td>
<td>1.06E-02</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-</td>
<td>-</td>
<td>1.16E+00</td>
<td>1.11E+17</td>
<td>1.96E+04</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3-7 Migration Rates (1/d).

<table>
<thead>
<tr>
<th>Segment</th>
<th>UKL</th>
<th>COP</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKL</td>
<td>0.00E+00</td>
<td>1.00E-04</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>COP</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>3.49E-03</td>
</tr>
<tr>
<td>IG</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
</tbody>
</table>

3.2.3 Yahara River

This system includes four major lakes that are connected via the Yahara River and a smaller lake which connects to Lake Monona from outside of the chain (Fig. 2.1D). All the lakes are located near Madison, WI metropolitan area and have experienced significant cyanobacteria blooms in recent years (Beversdorf et al., 2013; Miller et al., 2013). Cyanobacteria populations and segment characteristics were based on the values reported by Beversdorf et al. (Beversdorf, Chaston, Miller, & McMahon, 2015). Flow rates are
based on USGS gauges discharge data. Growth and mutation rates are the same as in the Great Lakes model.

**Table 3-8** Yahara River Characteristics (Beversdorf et al., 2015; PDX, 2016).

<table>
<thead>
<tr>
<th>Segment</th>
<th>Mendota</th>
<th>Wingra</th>
<th>Monona</th>
<th>Waubesa</th>
<th>Kegonsa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average CYA Concentration (cells/mL)</td>
<td>100000</td>
<td>20000</td>
<td>100000</td>
<td>100000</td>
<td>330000</td>
</tr>
</tbody>
</table>

**Table 3-9** Segment Characteristics (Beversdorf et al., 2015; Miller et al., 2013).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MEN</td>
<td>4.00E-03</td>
<td>3.94E+01</td>
<td>1.58E-01</td>
<td>1.58E+19</td>
<td>1.58E+04</td>
<td>-</td>
</tr>
<tr>
<td>WIN</td>
<td>1.70E-03</td>
<td>1.30E+00</td>
<td>2.21E-03</td>
<td>4.42E+16</td>
<td>4.40E+01</td>
<td>-</td>
</tr>
<tr>
<td>MON</td>
<td>2.70E-03</td>
<td>1.33E+01</td>
<td>3.58E-02</td>
<td>3.58E+18</td>
<td>3.58E+03</td>
<td>2.03E-02</td>
</tr>
<tr>
<td>WAU</td>
<td>1.70E-03</td>
<td>8.43E+00</td>
<td>1.43E-02</td>
<td>1.43E+18</td>
<td>1.43E+03</td>
<td>5.35E-02</td>
</tr>
<tr>
<td>KEG</td>
<td>1.70E-03</td>
<td>1.30E+01</td>
<td>2.21E-02</td>
<td>7.29E+18</td>
<td>7.29E+03</td>
<td>1.11E-02</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-</td>
<td>-</td>
<td>2.32E-01</td>
<td>2.81E+19</td>
<td>2.81E+04</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 3-10** Migration Rates (1/d).

<table>
<thead>
<tr>
<th>Segment</th>
<th>MEN</th>
<th>WIN</th>
<th>MON</th>
<th>WAU</th>
<th>KEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEN</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>2.41E-03</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>WIN</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>4.49E-03</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>MON</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>1.12E-02</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>WAU</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>2.97E-02</td>
</tr>
<tr>
<td>KEG</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
</tbody>
</table>

3.2.4 Chattahoochee River

The Chattahoochee River is located in southeastern United States (i.e. Georgia, Alabama and Florida) and includes five main reservoirs (i.e. Lake Lanier, West Point Lake, Lake
Harding, Walter F. George Reservoir and Lake Seminole). Cyanobacteria populations for this system was based on different reports from different organizations (Pacific Northwest Environmental Research Lab., 1975) (Morris, 1978) (Lium, Stamer, Ehlke, Faye, & Cherry, 1979) (Levy & Flock). Flow rates are based on USGS gauges discharge data. Growth and mutation rates are the same as in the Great Lakes model.

**Table 3-11** Chattahoochee River Characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Lanier</th>
<th>West Point</th>
<th>Harding</th>
<th>Walter F. George</th>
<th>Seminole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average CYA Concentration (cells/mL)</td>
<td>14170</td>
<td>31250</td>
<td>6000</td>
<td>4651</td>
<td>100000</td>
</tr>
</tbody>
</table>

**Table 3-12** Segment Characteristics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LAN</td>
<td>5.10E-03</td>
<td>1.50E+02</td>
<td>7.65E-01</td>
<td>1.08E+19</td>
<td>7.23E+03</td>
<td>-</td>
</tr>
<tr>
<td>WPL</td>
<td>3.75E-03</td>
<td>1.05E+02</td>
<td>3.93E-01</td>
<td>1.23E+19</td>
<td>8.18E+03</td>
<td>9.89E-03</td>
</tr>
<tr>
<td>HAR</td>
<td>2.55E-03</td>
<td>2.37E+01</td>
<td>6.04E-02</td>
<td>3.62E+17</td>
<td>2.40E+02</td>
<td>2.07E+00</td>
</tr>
<tr>
<td>WFG</td>
<td>1.11E-02</td>
<td>1.83E+02</td>
<td>2.03E+00</td>
<td>9.44E+18</td>
<td>6.29E+03</td>
<td>1.75E-02</td>
</tr>
<tr>
<td>SEM</td>
<td>9.00E-04</td>
<td>1.52E+02</td>
<td>1.37E-01</td>
<td>1.37E+19</td>
<td>9.12E+03</td>
<td>1.43E-02</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-</td>
<td>-</td>
<td>3.38E+00</td>
<td>4.66E+19</td>
<td>3.11E+04</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 3-13** Migration Rates (1/d).

<table>
<thead>
<tr>
<th></th>
<th>LAN</th>
<th>WPL</th>
<th>HAR</th>
<th>WFG</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAN</td>
<td>0.00E+00</td>
<td>5.87E-03</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>WPL</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>3.20E-02</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>HAR</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>2.40E-01</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>WFG</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>1.09E-02</td>
</tr>
<tr>
<td>SEM</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
<td>0.00E+00</td>
</tr>
</tbody>
</table>
3.3 Model Accuracy

The model is stochastic and repeat simulations with different seed values for the random number generator will produce different results. We quantify the stochastic variability by running the model repeatedly and then estimating the standard error of the mean of the results (see Fig. 4.4).

The accuracy depends on a number of parameters. For example, increasing the simulation time will increase the accuracy. However, the parameters also affect the simulation and output processing time, which is quite demanding for this model. The aim is to adjust the parameters to reduce the computing time as much as possible without compromising the accuracy of results. We first describe each of the parameters, followed by the methodology used to select the parameter values and results.

3.3.1 Parameter Description

Sample Interval

The sample interval refers to how often the genomes of a sample of cells are saved.

Sample Size

The sample size refers to the number of cells sampled from each segment at each sample interval for BLAST analysis.

Genome Size

The genome size is the length of the genome in nucleotides.
Population Size (N)

The population size is the total number of agents simulated. It is approximately equal to the sum of \( K \) for each segment. The model simulates a small fraction of the real cells in the system. We previously showed that the total population size does not affect the nucleotide divergence in a two-box system (F. L. Hellweger et al., 2014). The theoretical models (Stepping-stone and Two-box) in Section S3 also show that the nucleotide divergence is independent of the total population size, although the relative population size does matter. However, there is a lower limit, where the model results will be affected (e.g. consider one or two cells in a segment).

3.3.2-Selection of Parameter Values

The strategy used to select the parameter values is to vary a parameter and compare the results to a base case. If the results are insensitive to the change, the parameter value is acceptable. We start with the base case and estimate the variance due to stochastic variability. This is done by running the model multiple times with different random number seed values. We calculate the average nucleotide divergence (\( \bar{\delta} \)) of each segment pair for each run. Averaging \( \bar{\delta} \) of each pair over multiple runs provides a good estimate to use as the base case. In addition, the standard deviation for each pair can be used when calculating the statistical significance (Equation 3-10).

Then we change other parameters, one at a time, to half and twice their value in the base case and calculate their difference with the base case based on 99% statistical significance:

\[
|\bar{\delta} - \bar{\delta}_{\text{base case}}| < \alpha \sigma_{\text{base case}} \tag{3-10}
\]

In which \( \alpha = 2.58 \) for 99% statistical significance. If there are multiple significantly different pairs in one scenario, its results are not acceptable.

As an example, Fig. 3.5 shows the comparison of average nucleotide divergence (\( \bar{\delta} \)) for each pair of segments of the Great Lakes system when changing the sample interval
parameter. Decreasing the sample interval to a tenth its base case value results in significant changes and is therefore not acceptable. However, increasing the value by a factor of two results in acceptable changes.

**Fig. 3.5.** Average nucleotide divergence ($\bar{\delta}$) of each pair of lakes compared with base case when changing sample interval. Error bars are based on 99% statistical significance.
The results for all parameters is summarized in Table 3-14 to 3-17. Numbers show the number of accepted average nucleotide divergence ($\tilde{\delta}$) of each pair of lakes out of total number of pairs for each system (see Fig. 4.3).

**Table 3-14**  Number of accepted average nucleotide divergence ($\tilde{\delta}$) of each pair of lakes out of total of 36 pairs for the Great Lakes.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value Used</th>
<th>Acceptable Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1.00E+01</td>
<td>35/36</td>
</tr>
<tr>
<td>Time (y)</td>
<td>1.00E+04</td>
<td>36/36</td>
</tr>
<tr>
<td>Genome Size</td>
<td>1.00E+05</td>
<td>36/36</td>
</tr>
<tr>
<td>Sample Intervals</td>
<td>2.00E+02</td>
<td>36/36</td>
</tr>
<tr>
<td>Agents</td>
<td>9.52E+03</td>
<td>36/36</td>
</tr>
</tbody>
</table>

**Table 3-15**  Number of accepted average nucleotide divergence ($\tilde{\delta}$) of each pair of lakes out of total of 3 pairs for the Klamath River.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value Used</th>
<th>Acceptable Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1.00E+02</td>
<td>0/3</td>
</tr>
<tr>
<td>Time (y)</td>
<td>2.00E+03</td>
<td>1/3</td>
</tr>
<tr>
<td>Genome Size</td>
<td>1.00E+05</td>
<td>2/3</td>
</tr>
<tr>
<td>Sample Intervals</td>
<td>5.00E+02</td>
<td>3/3</td>
</tr>
<tr>
<td>Agents</td>
<td>1.96e+04</td>
<td>3/3</td>
</tr>
</tbody>
</table>

* For X4.
Table 3-16- Number of accepted average nucleotide divergence ($\bar{\delta}$) of each pair of lakes out of total of 10 pairs for the Yahara River.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
<th>Used</th>
<th>Acceptable Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1.00E+02</td>
<td>4/10</td>
<td>9/10</td>
</tr>
<tr>
<td>Time (y)</td>
<td>2.00E+03</td>
<td>6/10</td>
<td>7/10</td>
</tr>
<tr>
<td>Genome Size</td>
<td>1.00E+05</td>
<td>6/10</td>
<td>7/10</td>
</tr>
<tr>
<td>Sample Intervals</td>
<td>5.00E+02</td>
<td>0/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Agents</td>
<td>2.81e+04</td>
<td>10/10</td>
<td>7/10</td>
</tr>
</tbody>
</table>

Table 3-17- Number of accepted average nucleotide divergence ($\bar{\delta}$) of each pair of lakes out of total of 10 pairs for the Chattahoochee River.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
<th>Used</th>
<th>Acceptable Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>2.00E+02</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Time (y)</td>
<td>3.00E+03</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Genome Size</td>
<td>1.00E+05</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Sample Intervals</td>
<td>5.00E+02</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Agents</td>
<td>3.11E+04</td>
<td>10/10</td>
<td>10/10</td>
</tr>
</tbody>
</table>

4. Results and Discussion

4.1 Temporal Pattern for Two Segments

The nucleotide divergence between two segments exhibits some interesting population genetic patterns as illustrated by Lakes Superior (SUP) and Southern Michigan (SM) (Fig. 4.1). The simulation starts with all cells having the same genome. Due to independent evolution of the populations in the lakes, their nucleotide differences accumulate continuously in a clock-wise manner. Cells constantly migrate between the segments. The
model is neutral, so a migrant cell has the same chance of survival as any other cell in the segment. However, since it is vastly outnumbered by the resident cells its chance of establishment and taking over the population is low. Most migrants and their offspring die and the divergence continues. A migrant population may grow and temporarily become a significant fraction of the total population, which can result in a transient decrease in divergence. Occasionally, a migrant may take over the entire population of the destination lake, causing an abrupt complete drop in the nucleotide divergence between the lakes. These are referred to as coalescence events. The divergence then increases gradually again and the pattern repeats itself. The takeovers are random events that occur at a frequency depending on the system topology, relative population sizes and migration rates (see Model Application for equations corresponding to simplified systems).

Fig. 4.1. Sawtooth pattern of nucleotide divergence between Lakes Superior (SUP) and Southern Michigan (SM).

4.2 Temporal Patterns at the System Scale

The dynamics of nucleotide divergence at the scale of the entire system reveals the connectivity among the lakes. This is illustrated in Fig. 4.2 using the nucleotide divergence between Western Erie (WE) and other segments of the Great Lakes. When comparing all segments, a sawtooth pattern is evident, reflecting system-wide divergence and coalescence events (Fig. 4.2A). The divergence in the system is dominated by the divergence between SUP and SM. Downstream lakes generally have low divergence to one of these lakes and higher divergence (corresponding to the divergence between SUP and SM) to the other.
Fig. 4.2. Nucleotide divergence between Western Erie (WE) and other segments of the Great Lakes. (A) between WE and all other lakes of the system (B) between WE and the two source lakes, Superior and Michigan (C) between WE and better connected lakes, Huron and Ontario.

Looking more closely at the divergence between Western Erie and the two upstream lakes of this branched system, Lakes Superior and Michigan, again shows a sawtooth pattern for the maximum divergence (Fig. 4.2B, note different scale). This reflects the divergence of SUP and SM/NM populations (see previous paragraph). SM-WE and NM-WE pairs follow a fairly similar pattern of nucleotide divergence since SM and NM are relatively well connected. Interestingly, WE switches between the upstream lakes. This reflects alternating takeovers of the WE population by cells from SUP and SM/NM. For example, around 100 y, the divergence with SM/NM is low and that with SUP is high. Then the pattern flips around 110 y because of a takeover by a cell from SUP which reduces
divergence with SUP and increases divergence from SM/NM. This pattern emerges, because the frequency of coalescence events between WE and SM/NM or SUP is higher than that between SM/NM and SUP. A coalescence event between SUP and SM therefore results in a drop in divergence across the system.

For lakes that are closer together, Lakes Huron, Erie and Ontario, the divergence is lower and more variable, reflective of their higher connectivity (Fig. 4.2C). An increase in divergence within Erie (CE-WE, EE-WE) is often preceded by increased divergence between Huron (NH-WE, SH-WE). This is because takeovers follow the flow direction and occur first in Huron and then Erie.

4.3 Dynamic phylogeny

The dynamics of divergence between the segments is also evident in phylogenetic trees generated from the genomes of the model cells (Fig. 4.3). The model writes the genomes of a randomly selected cell from each segment to a file. These files are then analyzed using MAFFT and BIO::Phylo bioinformatics tools to generate a phylogenetic tree. The *Microcystis aeruginosa* genome, which was used to initiate this simulation, is also included (MC in Fig. 4.3A&B). The divergence vs. time plot (Fig. 4.3C) shows that, at year 305, the divergence between SUP and NM is relatively high and that between SUP and WE is low (indicating that the WE population is presently of SUP origin). This is also reflected in the phylogenetic tree (Fig. 4.3A). At year 365, SUP and NM have recently experienced a coalescence event. The divergence between SUP and NM is relatively low and that between SUP and WE is similar (indicating that the WE population is presently of NM origin), which is also reflected in the corresponding phylogenetic tree (Fig. 4.3B). Overall, the phylogenetic tree at the later time, which follows more closely a coalescence event, has shorter branches, consistent with the idea of a system-wide takeover (see previous section). Despite the periodic collapse of the tree for the model segments, the divergence with the original *Microcystis aeruginosa* genome continues to grow as the system moves forward in time. See Movie S1 for an animation covering the whole 500-year simulation period.
4.4 Spatial Patterns

The spatial pattern of average nucleotide divergence reflects the system topology. Fig. 4.4A presents the average nucleotid divergence ($\delta$, bp / 1 Mbp) between each of the segments for the Great Lakes system. SM and NM share a large open boundary and their populations are relatively well-mixed and consequently, their nucleotide divergence is relatively low. SUP and NM are both upstream of NH and downstream from there, the system follows a linear flowpath. That means NM and SUP are not connected by advection but only by diffusion (via NH) with relatively low rates. Consequently, the average nucleotide divergence of SM/NM and SUP segments is higher than the rest of the system. The spatial patterns predicted by the model are intuitively correct. For example, the nucleotide divergence increases with distance upstream of ONT – a typical distance-decay pattern.
(Fig. 4.4A, right column). Segments within the same lake (e.g. SM-NM or WE-CE-EE) have relatively low divergence. The connectivity among segments is a function of the migration rates and relative population sizes. This can also be expressed as a relative import rate (Lindström, Forslund, Algesten, & Bergström, 2006) which is the rate of cell import divided by the rate of cell growth. Due to the large population size in WE, the relative import rate is only 0.19% and the divergence with SH is relatively high (235 bp/1 Mbp). The relative import rate of ONT is 0.38% and the divergence with EE is lower (89 bp/1 Mbp).

**Fig. 4.4.** Heat maps of average nucleotide divergence for each lake system. Average nucleotide divergence (bp / 1 Mbp) in (A) Great Lakes (B) Klamath River (C) Chattahoochee River (D) Yahara River. SEM is the standard error of the mean estimated from repeat simulations with different seed values for the random number generator (see Model Application). *Segments are not connected or **No coalescence observed during simulation (value represents lower bound, value not included in calculation of standard deviation).
The predicted average nucleotide divergences are quite heterogeneous within the Great Lakes system, ranging from 10 (SM-NM) to 1378 (SUP-SM) bp / 1 Mbp. This highlights the effect of the topology and individual properties of lakes (or segments) and their connectivity, and the need for a model that accounts for these factors.

Overall the magnitude of nucleotide divergence is relatively low. The largest divergence is predicted between Lakes Superior and Michigan at about 1400 bp / 1 Mbp (0.14%), which equates to 8,200 bp for *Microcystis aeruginosa*. The magnitude of the divergence corresponds 99.86% average nucleotide identity which is substantially above what is often considered to be a species (>95% identity, (Konstantinidis, Ramette, & Tiedje, 2006)).

These results are not directly comparable to observations, because real genomes are subject to environmental selection, which could maintain differences (if local environmental factors are different) or reduce them (if a strongly beneficial mutation sweeps through the entire system). However, our results provide a baseline for the effect of neutral processes in the divergence between these lakes. When two *Microcystis aeruginosa* cells are sequenced from Lakes Superior and Michigan and their genomes are compared, 8,200 of the bp differences can be attributed to neutral evolution. Anything further would have to be due to environmental selection.

4.5 Comparison among Lake Systems

The lake systems simulated here vary in topology and properties (i.e., population size, migration rate, Fig. 2.1), which affects the predicted biogeographic patterns. One important factor is diffusion in a branched system. In the Great Lakes, SM/NM and SUP are connected via NH by diffusion, which allows for coalescence events. In the Yahara River, WIN and MEN are not connected by diffusion. Therefore, the model will not predict coalescence for these segments and their nucleotide divergence will increase indefinitely (Fig. 4.4D).
Connectivity is also a function of flow rate and relative population sizes. For example, in the Yahara River system WIN and MEN are both upstream of MON and they have comparable flow rates. However, the model predicts no coalescence between WIN and MON in the 20,000 year simulation period. Interestingly, this is not due to weak connectivity between these lakes (the theoretical coalescence time is about 50 years, see Model Application), but a result of overwhelming input of cells from MEN (due to larger population size). The relative import rate with respect to MEN is 9%, which means 9% of the cells in MON were born in MEN. Thus, the population of MON is dominated by input from MEN and frequent takeovers from that population do not allow for or interrupt any takeovers from the WIN population. The large influx of cells from MEN “protects” MON from takeovers by WIN. This highlights complexity of real systems and limitations of existing theoretical approaches. Theoretical takeover rates and relative import rates suggest these two segments are relatively well connected, but the full model, which accounts for the complexity shows that they are not connected.

The segments of the Chattahoochee River have similar population sizes and high flow rates (see Model Application). This high connectivity results in relatively low and uniform average nucleotide divergence (Fig. 4.4C).

Overall, the system-wide average nucleotide divergence varies over three orders of magnitude in the order Great Lakes > Klamath River > Yahara River > Chattahoochee River. This is consistent with the average relative import rate of these systems. For segments that are directly connected the average nucleotide divergence correlates with the relative import rate (Fig. 4.5). The values differ from the theoretical relationship for a simple two-box model (see Model Application). Differences can be explained by several factors, including diffusion, branching and number of segments. When there are more than two segments in series, the divergence between segments 2 and 3 can be affected by takeovers of segment 2 by cells from segment 1, which is not considered in the simplified theory.
**Fig. 4.5.** Correlation between model-predicted average nucleotide divergence ($\delta$) and relative import rate ($R_i$). Only directly-connected pairs are included. WIN-MON is excluded because no coalescence event was observed during the simulation (see text). 2-Box theory is $\delta = 2 k_a / (\ln 2 R_i)$, where $k_a$ is mutation rate per genome per generation, see Model Application.

### 4.6 Relation to other species and systems

These results are specific to *Microcystis aeruginosa* and the four lake systems studied here. They can be related to other species in these four systems, because the divergence is proportional to the growth and mutation rates (see above). This would require estimates of growth and mutation rates. The growth rate we use is a general estimate for freshwater phytoplankton (Reynolds & Irish, 1997), but the mutation rate is specific to *Microcystis aeruginosa* (using $m$ from *E. coli* and $r/m$ for *Microcystis aeruginosa*). Estimates of $r/m$ for 48 bacteria and archaea are available from Vos and Didelot (2009), but *Microcystis aeruginosa* is the only free living, freshwater cyanobacteria in this database.

Relating these results to other systems will be more difficult, because the divergence is a function of the system topology, population sizes and transport rates. For a simple two-lake system with one way upstream to downstream connection, the nucleotide divergence is proportional to the population size of the downstream lake and inversely proportional to the population size of the upstream lake (see Model Application).

These results can be compared to those obtained previously for the heterotroph *Pelagibacter ubique* in the global ocean (Ferdi L. Hellweger et al., 2014). For example,
there we found an average divergence of 0.2% between Hawaii and the Gulf of Alaska. Some of the difference can be attributed to a lower growth rate (0.14 / d, 4× lower, corresponds to lower divergence) and higher mutation rate (3.5E-8 changes/bp/division, 3× higher, corresponds to higher divergence) in the ocean simulation. However, the effect of these difference approximately cancels out and the comparison suggests that the connectivity of Lakes Superior and Michigan and that of Hawaii and the Gulf of Alaska are of similar magnitude.

5. Summary and Outlook

For microbes, biogeographic patterns are often assumed to be due to environmental selection and neutral processes are considered less important. We explored the role of neutral evolution and dispersal limitation in producing biogeographic patterns for cyanobacteria in systems of connected lakes using an agent-based model. Application to the Great Lakes and other lake systems showed that neutral processes can produce substantial biogeographic patterns. Dispersal limitation can have important consequences for how a lake may respond to change, like nutrient input or climate. In the absence of dispersal limitation, it can be assumed that a lake may instantaneously adopt a new population that is best suited for the new conditions. However, when there is dispersal limitation, there may be a time lag until the new fittest species can establish itself. For example, a warmer climate may favor a different cyanobacteria population in Lake Huron. Good candidates may be the population from more southern lakes of the system, like Lake Erie. Dispersal limitation suggests that this migration may not happen instantaneously. The level of divergence between populations predicted here can be important. For example, the presence/absence of a toxin producing gene may have important consequences for environmental and public health (Dyble et al., 2008). Our results provide a quantitative baseline of neutral nucleotide divergence that can complement past and future observations in these test bed systems.

Our model constitutes a significant step in realism compared to existing theoretical approaches. Specifically, by allowing for any topology and non-uniform population sizes
and migration rates, the model is able to make predictions for real systems, which is not possible using existing models (e.g., stepping stone model). A distinguishing feature of the model is that it simulates individual cells with whole genomes, which opens it up to analysis using standard bioinformatics tools, like BLAST and MAFFT. The model system or microbes are thus analyzed in the same manner as their real-life counterparts. We expect that as our models become more realistic, the workflow around them will become more similar to that around observations. However, the present model is still relatively simple and there is room for improvement. First, our simulations use constant population sizes and migration rates. We showed that simple time-variable patterns, like overwintering or seasonal dynamics, do not affect the nucleotide divergence at the time scale considered here. However, it is possible that the interaction of migration rates and population sizes can affect dispersal. An example is the tendency of cyanobacteria to bloom during dry, low flow periods. At longer time scales, the gradual natural or anthropogenic eutrophication, climate change, invasive species (e.g. zebra mussel invasion) or management actions (e.g. wastewater treatment, phosphate detergent ban) may play a role. Further, our model does not include environmental selection. Selection can be simulated at the gene level (Ferdi L. Hellweger, 2013), but at the genome scale this would require relating the DNA sequence to the growth or death rates. Although there are examples where individual basepair changes have been related to function of proteins (e.g. Proteorhodopsins, (Fuhrman, Schwalbach, & Stingl, 2008)), this is not yet possible at the genome scale. However, genome’s wide properties, like the GC content, can be related to growth rate (Raghavan, Kelkar, & Ochman, 2012) and the present model could be modified to simulate this.

These results and those for the open ocean reported previously (F. L. Hellweger et al., 2014) suggest that neutral processes are substantial, but relatively weak drivers of microbe biogeography in these well-connected systems. Predicted average nucleotide divergence is generally <0.5%, although for some segments that are not or very weakly connected, the model does not predict a coalescence event. This level of divergence is below typical species delineation of >5% (Konstantinidis et al., 2006). However, observations increasingly resolve small differences in genomic content (Kashtan et al., 2014) and neutral processes need to be considered when explaining biogeographic patterns.
REFERENCES


