2014

Accuracy and efficiency of algorithms for the demarcation of bacterial ecotypes from DNA sequence data

Juan Carlos Francisco, *Wesleyan University*
Frederick M Cohan, *Wesleyan University*
Danny Krizanc, *Wesleyan University*

Accuracy and efficiency of algorithms for the demarcation of bacterial ecotypes from DNA sequence data

Juan Carlos Francisco

Department of Mathematics and Computer Science,
Wesleyan University,
Middletown, CT, USA
Email: jfrancisco@wesleyan.edu

Frederick M. Cohan

Department of Biology,
Wesleyan University,
Middletown, CT, USA
Email: fcohan@wesleyan.edu

Danny Krizanc*

Department of Mathematics and Computer Science,
Wesleyan University,
Middletown, CT, USA
Email: dkrizanc@wesleyan.edu
*Corresponding author

Abstract: Identification of closely related, ecologically distinct populations of bacteria would benefit microbiologists working in many fields including systematics, epidemiology and biotechnology. Several laboratories have recently developed algorithms aimed at demarcating such 'ecotypes'. We examine the ability of four of these algorithms to correctly identify ecotypes from sequence data. We tested the algorithms on synthetic sequences, with known history and habitat associations, generated under the stable ecotype model and on data from *Bacillus* strains isolated from Death Valley where previous work has confirmed the existence of multiple ecotypes. We found that one of the algorithms (ecotype simulation) performs significantly better than the others (AdaptML, GMYC, BAPS) in both instances. Unfortunately, it was also shown to be the least efficient of the four. While ecotype simulation is the most accurate, it is by a large margin the slowest of the algorithms tested. Attempts at improving its efficiency are underway.

Keywords: bacterial ecotypes; demarcation algorithms; stable ecotype model; DNA sequence data; ecotype simulation; AdaptML; GYMC; BAPS.

Reference to this paper should be made as follows: Francisco, J.C., Cohan, F.M. and Krizanc, D. (2014) 'Accuracy and efficiency of algorithms for the demarcation of bacterial ecotypes from DNA sequence data', *Int. J. Bioinformatics Research and Applications*, Vol. 10, Nos. 4/5, pp.409–425.

Biographical notes: Juan Carlos Francisco received his BA in Computer Science with a certificate in Bioinformatics from Wesleyan University in 2011. He is currently working as a software engineer at Groupon in Chicago.

Frederick M. Cohan graduated from Pasadena High School and earned his BS at Stanford in Biological Sciences; he was the first to earn a PhD from Harvard's Organismic and Evolutionary Biology department. He studies the origins of diversity in bacteria. He is intrigued by what is the same and different about species and speciation across all walks of life. He is working to develop a system to identify the most newly divergent products of speciation even when we know little about the ecological and physiological differences between new species. He is a Professor of Biology and Environmental Studies at Wesleyan University.

Danny Krizanc received his BSc from University of Toronto in 1983 and his PhD from Harvard University in 1988, both degrees in Computer Science. He held positions at the Centruum voor Wiskunde en Informatica, Amsterdam, The Netherlands, the University of Rochester, Rochester, New York and Carleton University in Ottawa, Canada before joining the Department of Mathematics and Computer Science at Wesleyan University in 1999. His research focus is the design and analysis of algorithms, especially as applied to distributed computing, networking and bioinformatics.

This paper is a revised and expanded version of a paper entitled 'Demarcation of bacterial ecotypes from DNA sequence data: a comparative analysis of four algorithms, presented at the '2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)', Las Vegas, NV, USA, 23–25 February 2012.

1 Background

The taxonomy of bacteria has provided microbiologists a system for routinely identifying species as closely related groups that differ in their disease-causing properties, ecological roles in biological communities and in their physiological capacities (Rosselló-Mora and Amann, 2001; Kopac and Cohan, 2011). This service has emerged from a pragmatic approach, whereby species are recognised as groups (or clusters) of close relatives separated by large gaps in phenotypic and molecular characters (Rosselló-Mora and Amann, 2001; Vandamme et al., 1996). In the last three decades, bacterial taxonomy has adopted various universal molecular criteria for identifying species as clusters, first based on an indirect measure of shared genome content (i.e. DNA–DNA hybridisation) Wayne et al., 1987; Lan, 1996), and more recently based on measures of sequence identity of shared genes, including 16S rRNA sequence similarity (Stackebrandt and Ebers, 2006), multilocus sequence analysis (Hanage et al., 2006), and most recently, genome-wide average nucleotide identity (Konstantinidis and Tiedje, 2005).

While these universal molecular criteria for species demarcation have successfully focused on identifying species taxa that are divergent in DNA sequence identity, genome content, and physiology (Rosselló-Mora and Amann, 2001); species classification has placed almost no emphasis on ensuring that each individual species is homogeneous in any characteristic (Kopac and Cohan, 2011; Staley, 1909; Sikorski, 2008). It is becoming

clear that a typical species taxon recognised by systematics is highly heterogeneous in its genome content and physiology, as well as its ecology (Touchon et al., 2009; Walk et al., 2009; Kettler et al., 2007; Lefebure and Stanhope, 2007; Marri et al., 2006; Paul et al., 2010; Vernikos et al., 2007). A more fine-grained taxonomy, which recognises the closely related, ecologically distinct populations that are now subsumed by bacterial systematics within a single species taxon, would benefit microbiologists from many subfields. For example, the reification of ecologically diverse clades within *Escherichia coli* (and within other recognised species taxa) has led to error in estimates of population genetic parameters, such as population size and recombination rates (Kopac and Cohan, 2011). Epidemiologists might find it useful to characterise all the ecologically distinct populations within a disease-causing species taxon (Cohan and Perry, 2007). Biotechnologists could also take advantage of a more fine-grained systematics of species: after discovering a bacterium with a valuable enzyme, one could then search for homologs of the enzyme across closely related, ecologically distinct populations if they were recognised by taxonomy (Cohan and Perry, 2007; Jensen, 2010). Vaccine development may be negatively impacted by its focus on the set of 'core genes' shared by an entire species taxon (Kopac and Cohan, 2011; Tettelin et al., 2005).

Identification of closely related, ecologically distinct populations of bacteria is much more difficult than the case for animals and plants. Specialists on an animal taxon usually know ahead of time the phenotypic traits (e.g. in song birds the shape and size of bill) that distinguish most closely related, adaptively divergent groups, and they know the ecological dimensions by which closely related groups frequently diverge (here, insect versus seed predation, and small versus large prey). However, identification of bacterial populations by their phenotypic divergence is difficult because new populations are frequently formed through horizontal genetic transfer events, whereby a biochemical function is introduced into a lineage (Wiedenbeck and Cohan, 2011; Ochman and Davalos, 2006; Gogarten and Townsend, 2005). The nature of these transferred biochemical functions or their sources is impossible to predict, so we need to find universal ways to predict the ecologically distinct populations among close relatives – methods that do not require advance knowledge of the nature of their ecological divergence (Cohan and Perry, 2007).

Several laboratories have recently developed algorithms that aimed to identify bacterial populations with the properties of 'ecotypes' (Koeppel et al., 2008; Hunt et al., 2008; Corander et al., 2008; Barraclough et al., 2009). An ecotype is defined to constitute a paraphyletic or monophyletic group of close relatives that are ecologically interchangeable, in that the members of an ecotype share genetic adaptations to a particular set of habitats, resources, and conditions, and different ecotypes are distinct in their ecological adaptations (Cohan and Perry, 2007; Ward, 1998). In this paper, we initiate the study of such algorithms. Our goal here is to evaluate four algorithms for their ability to discover the recently divergent ecotypes among closely relatives of bacteria. Three of the algorithms (ecotype simulation, GMYC, BAPS) we consider assume a force of cohesion within each ecotype, as seen in the stable ecotype model (Cohan and Perry, 2007), and one (AdaptML) is more agnostic about the dynamic forces controlling diversity within ecotypes (Hunt et al., 2008).

2 Models of bacterial speciation

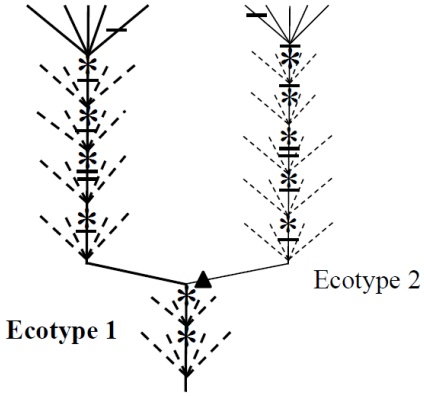
Stable ecotype model: In the stable ecotype model, the homogeneity of ecological adaptations within an ecotype leads to genetic cohesion within the ecotype (Wiedenbeck and Cohan, 2011). That is, sequence diversity within an ecotype is recurrently constrained by one of two cohesive forces, periodic selection and genetic drift. In periodic selection, an adaptive mutant outcompetes all other members of the ecotype, and by virtue of the rare recombination in bacteria (Vos and Didelot, 2009), natural selection favouring the adaptive mutant causes a nearly genome-wide sweep of diversity (Cohan, 2005). Likewise, genetic drift can purge diversity genome-wide among the adaptively homogeneous membership of an ecotype, although genetic drift will be the primary force of cohesion only for ecotypes with extremely small population sizes (Cohan and Perry, 2007; Kuo, et al., 2009). The diversity-purging effects of periodic selection and genetic drift are limited to the membership of a single ecotype. In the case of periodic selection, an adaptive mutant from one ecotype cannot outcompete to extinction the membership of another ecotype, owing to their ecological differences; genetic drift is limited to purging diversity within a single population where any individual can randomly become the ancestor of the entire population in the future.

The dynamics of cohesion within each ecotype, coupled with a lack of cohesion between ecotypes, yields a method for recognising ecotypes through sequence diversity (Figure 1) (Cohan, 2002). While each population is recurrently purged of its diversity, each ecotype accumulates its own unique set of neutral mutations at every gene in the genome (Cohan and Perry, 2007). In principle, this will cause each ecotype to be recognised as a unique sequence cluster. However, such a sequence-based analysis has limited power to discover the most newly divergent ecotypes (i.e. before they have had a chance to accumulate neutral mutations). The stable ecotype model deals with the issue of resolution by assuming that new ecotypes are formed only rarely, such that there is sufficient time for most ecotypes to be recognisable as sequence clusters based on the resolving power of a small number of shared genes. This assumption can be relaxed somewhat as more shared genes are included in an analysis of sequence clusters (up to the resolution of the shared genes of the entire genome).

The ecotype simulation algorithm can analyse either a periodic selection-based or a drift-based cohesion within ecotypes; we will focus on a model of periodic selection dominating the cohesion within ecotypes. This is because recombination rates are always sufficiently low in bacteria (Vos and Didelot, 2009) to foster genome-wide sweeps of diversity (Cohan, 2005) (W. Hanage, personal communication). The GMYC algorithm uses a genetic drift model of cohesion (Barraclough et al., 2009), and BAPS does not explicitly assume any particular model of cohesion, only that the diversity within populations is much less than that between them (Tang et al., 2009).

Habitat preference model: The algorithm AdaptML does not assume cohesion within ecotypes and works on a different principle. It assumes that each ecotype formation event causes at least a quantitative shift in habitat preferences (Hunt et al., 2008). That is, newly divergent ecotypes may overlap in the habitats that they utilise, but they are at least quantitatively different in their habitat preferences. This causes the ecotypes to differ in the frequencies at which they are found in each habitat. In this habitat preference model, ecotypes can be discovered as phylogenetic groups (either monophyletic or paraphyletic) that are significantly different in the habitats from which they are isolated.

Figure 1 The stable ecotype model. This model is marked by a much higher rate of periodic selection events (indicated by asterisks) than ecotype formation events (indicated by triangle), such that each ecotype endures many periodic selection events during its lifetime. In the Stable Ecotype model, most ecotypes are discoverable as sequence clusters because longstanding ecotypes have an opportunity to accumulate neutral sequence divergence at every locus (sequence changes indicated by horizontal lines), while diversity within ecotypes is recurrently purged. The wide and thin lines in the figure represent two different ecotypes; each asterisk represents a periodic selection event, which purges diversity within one ecotype only. The dashed lines represent lineages that have gone extinct owing to periodic selection



The two classes of algorithms have their advantages and disadvantages (Cohan and Koeppel, 2008). The three algorithms based on the stable ecotype model have the advantage that they can demarcate ecotypes even when the investigator has no a priori ideas about either the ecological or physiological dimensions by which the ecotypes have diverged; and the disadvantage is that the ecotypes hypothesised by these algorithms must be independently tested for their ecological distinctness. An advantage of AdaptML is that it simultaneously demarcates ecotypes and tests them for differences in their habitat preferences; so no further testing of ecological distinctness is necessary. The disadvantage of AdaptML is that it is limited to identifying ecotypes that are significantly different in preferences to habitat types that are anticipated by the investigators. Interestingly, three recent studies have shown that ecotype simulation and AdaptML reach highly similar ecotype demarcations (Connor et al., 2010; Becraft et al., 2011; Melendrez et al., 2011), indicating that investigators' a priori guesses about the significant habitat types have been fairly accurate. Finally, a disadvantage of each class of algorithms is that their ability to find the adaptively homogeneous ecotypes is limited by the phylogenetic resolution of the sequences analysed, and it is difficult to test whether a putative ecotype is homogeneous in its ecological adaptations (Kopac and Cohan, 2011).

3 The algorithms

Ecotype simulation (ES) (Koeppel et al., 2008): Ecotype simulation takes as input a lineage-through-time plot of sequence diversity (Martin, 2002), where the number of sequence-identity bins is plotted against the sequence identity criteria for binning. This

curve represents the history of splitting of lineages that have survived to the present. The algorithm searches for the combination of parameters that yields the observed curve with maximum likelihood. The parameters include: the rate of formation of new ecotypes (Ω), the rate of periodic selection (σ), and the number of ecotypes (*npop*) in the sample. A stochastic simulation, using a backward coalescence approach, generates the phylogenetic tree for the history of the sample, and then a forward simulation adds mutations to the sequences and generates sequence divergences. The algorithm tries many combinations of parameter values over many orders of magnitude, first through brute force and then by a hill-climbing approach.

Ecotype simulation estimates the parameter values under the assumption that the estimated values of Ω and σ apply to the whole tree. Because these rates are likely to vary over large phylogenetic groups, this approach may be made more accurate by limiting the analysis to closely related organisms.

Next, the algorithm demarcates the individual ecotypes. This is accomplished by recursively analysing smaller subsets of the phylogenetic tree to find the most inclusive clades whose optimal value for the number of ecotypes (*npop*) is 1, while applying the rate values Ω and σ estimated for the whole phylogeny.

A number of sequence-based analyses have used the ecotype simulation algorithm. David Ward and colleagues have used ecotype simulation to identify ecologically distinct populations of *Synechococcus* in hot spring microbial mats (Becraft et al., 2011; Melendrez et al., 2011; Ward et al., 2006), and the hypothesised ecotypes were shown to differ in their associations with temperature and light intensity and quality. The algorithm has also been used for an analysis of *Bacillus* isolates on different slopes of canyons in Death Valley, California (Connor et al., 2010) and in Israel (Koeppel et al., 2008) and on a salinity gradient in Death Valley (S. Kopac, personal communication). The hypothesised ecotypes were found to differ in their associations with solar exposure (Koeppel et al., 2008; Connor et al., 2010), soil texture (Connor et al., 2010), salinity, and rhizospheres (S. Kopac, personal communication). Multiple ecotypes were found within recognised species taxa, and some ecotypes were previously unknown. In addition, ecotype simulation has discovered ecotypes with different virulence properties within the pathogenic species taxa *Legionella pneumophilus* (Cohan et al., 2006) and *Propionibacterium acnes* (A. MacDowell and F.M. Cohan, unpublished research).

Ecotype simulation is currently available for the Windows operating system and has a straightforward graphical user interface.

GMYC (Barraclough et al., 2009): The Generalised Mixed Yule-Coalescent (GMYC) model was developed to demarcate ecologically distinct bacterial populations from sequence data. The software package that utilises GMYC needs only a phylogenetic tree to demarcate strains. However, this tree must be ultrametric – a tree whose root-to-tip paths are equal for all lineages. The authors recommend the use of an algorithm developed by Sanderson (Sanderson, 2003), available in the APE package of the R language (Paradis et al., 2004), to produce an ultrametic tree from sequences.

GMYC assumes a Yule model of speciation followed by a neutral coalescent model within species, where drift is the only process yielding coalescence. The algorithm maximises the likelihood of a transition from the time that new species are being formed to the time when all coalescences are due to drift events within species.

The GMYC algorithm was originally designed to delineate species from sequences for sexually reproducing organisms such as insects (Pons et al., 2006), and it was later shown to be applicable to asexually reproducing organisms such as bdelloid rotifers (Fontaneto et al., 2007). In the later study, GMYC demonstrated that populations of the

asexual *Rotaria* genus had diversified into distinguishable genetic clusters. While GMYC has been applied to bacterial sequence data (Barraclough et al., 2009), users of GMYC have not tested whether the algorithm yields closely related populations that are ecologically distinct.

At the time of writing, the GMYC algorithm is available only upon request. GMYC runs on any computer with a version of the R programming language installed, which includes Mac, Linux, and Windows. The algorithm comes in the form of a bundle of R functions that must be imported into one's local R environment.

*BAPS (*Corander and Tang, 2007*)*: The Bayesian Analysis of Population Structure (BAPS) application refines existing Bayesian approaches to determine the structure of populations from genetic data. It assumes a partition-based mixture model and performs classification using a variant of the Metropolis–Hasting algorithm to identify clusters of sequences, with no explicit model of purging of diversity within clusters; the algorithm can take into account recombination within and between populations (Corander and Tang, 2007).

BAPS was used to perform a large-scale non-phylogenetic analysis of the population structure of bacteria from the genus *Neisseria* and found multiple clusters within recognised species (Tang et al., 2009), in spite of recombination frequencies exceeding the rate of mutation (Vos and Didelot, 2009). However, the authors did not attempt to determine whether the clusters identified by BAPS corresponded to ecologically distinct populations.

BAPS has Windows, Linux and Mac versions, all with easy to use GUIs. The software package also has a command line equivalent that is designed for batch processing.

*AdaptML (*Hunt et al., 2008*)*: AdaptML places strains into ecologically distinct populations (ecotypes) based on the assumption that the origin of each ecotype is driven by a change in habitat preferences. This algorithm has been used to demarcate ecotypes within the Vibrionaceae in coastal estuaries (Hunt et al., 2008), within *Bacillus* in Death Valley (Connor et al., 2010), and within *Synechococcus* in Yellowstone hot springs (Becraft et al., 2011; Melendrez et al., 2011).

Unlike the cohesion-based algorithms, AdaptML does not accept as input the nucleotide sequences of the desired data set. Rather, AdaptML uses a phylogenetic tree (usually based on sequences) and data specifying the habitat of isolation for each strain. Habitats can have multiple environmental dimensions (e.g. size of marine particle and season of collection in the Vibrionaceae study) (Hunt et al., 2008). The various ecotypes need not be absolutely specialised to different habitats; the algorithm can detect ecotypes that are quantitatively different in their habitat associations. The algorithm assumes a hidden Markov model for the evolution of habitat associations and maximises the likelihood of associations of the strains observed on the tree.

AdaptML runs on any operating system with a version of Python installed, which includes the Mac, Windows, and Linux operating systems. While it does not have a GUI, instructions for running the script are detailed on the authors' website. There is also a web app version of the algorithm where one can upload the data set and have AdaptML demarcations emailed.

In the present study, we perform a comparative evaluation of each of these four algorithms using algorithmically generated bacterial data with *a priori* knowledge of the ecotypes. This data set is generated under the assumptions of the stable ecotype model, with periodic selection acting as the force of cohesion within ecotypes. We also present the algorithms' demarcations of *Bacillus* strains in the Radio Facility Wash canyon in Death Valley (Connor et al., 2010). Finally, we compare the running times of the algorithms.

An extended abstract of this paper appeared in the proceedings of the Second IEEE International Conference on Computational Advances in Bio and Medical Sciences (Francisco et al., 2012). This version greatly expands both the background and discussion sections from that version and includes additional results displayed in two new figures and one new table.

4 Results and discussion

4.1 Analysis of in silico generated sequences

Over all, ecotype simulation produced the closest match to the known ecotype demarcations (mean VI over all parameter values = 1.37), followed by GMYC (2.25), BAPS (2.38), and AdaptML (2.77). Even with absolute specialisation of ecotypes to different habitats ($\gamma = 0$), AdaptML was the least accurate (Table 1).

Table 1 Comparison of ecotype demarcations for in silico-generated ecotypes. Results are reported as Variation of Information, for $v = 50$, for all values of Ω, σ, and *npop*. AdaptML scores are taken from data sets where ecotypes are absolutely specialised to habitats ($\gamma = 0$), representing the most accurate demarcations made by AdaptML

Rate of ecotype formation	Rate of periodic selection	Variation of Information Score					
		Number of ecotypes					
		npop = 10		*npop* = 20		*npop* = 30	
Ω = .019	σ = .11	0.69	ES	0.70	ES	0.63	ES
		1.71	GMYC	2.21	GMYC	3.16	GMYC
		1.34	BAPS	1.93	BAPS	2.51	BAPS
		1.17	AdaptML	2.64	AdaptML	4.09	AdaptML
	σ = 1.1	0.51	ES	0.39	ES	0.33	ES
		1.58	GMYC	1.99	GMYC	2.91	GMYC
		1.15	BAPS	1.83	BAPS	2.44	BAPS
		1.26	AdaptML	2.80	AdaptML	4.11	AdaptML
	σ = 11	0.57	ES	0.33	ES	0.25	ES
		1.70	GMYC	1.86	GMYC	2.73	GMYC
		1.09	BAPS	1.78	BAPS	2.48	BAPS
		1.22	AdaptML	2.60	AdaptML	4.05	AdaptML
Ω = .19	σ = .11	1.63	ES	1.47	ES	1.40	ES
		1.83	GMYC	2.48	GMYC	2.88	GMYC
		2.09	BAPS	2.99	BAPS	3.56	BAPS
		1.65	AdaptML	2.88	AdaptML	4.05	AdaptML
	σ = 1.1	1.23	ES	1.09	ES	1.03	ES
		1.66	GMYC	2.15	GMYC	2.87	GMYC
		1.56	BAPS	2.46	BAPS	3.26	BAPS
		1.46	AdaptML	2.90	AdaptML	4.04	AdaptML
	σ = 11	1.03	ES	0.88	ES	0.89	ES
		1.83	GMYC	1.97	GMYC	2.47	GMYC
		1.28	BAPS	2.27	BAPS	2.99	BAPS
		1.51	AdaptML	2.86	AdaptML	4.05	AdaptML
Ω = 1.9	σ = .11	1.91	ES	2.15	ES	2.63	ES
		2.24	GMYC	2.20	GMYC	3.00	GMYC
		2.22	BAPS	2.87	BAPS	3.59	BAPS
		1.94	AdaptML	2.71	AdaptML	3.95	AdaptML
	σ = 1.1	2.11	ES	2.39	ES	2.68	ES
		2.14	GMYC	2.33	GMYC	2.69	GMYC
		2.23	BAPS	3.00	BAPS	3.63	BAPS
		1.91	AdaptML	2.86	AdaptML	3.81	AdaptML
	σ = 11	2.22	ES	2.72	ES	3.09	ES
		2.11	GMYC	2.05	GMYC	2.07	GMYC
		1.85	BAPS	2.55	BAPS	3.29	BAPS
		1.85	AdaptML	2.75	AdaptML	3.79	AdaptML

Ecotype simulation showed the greatest fidelity to the known ecotype demarcations in every parameter combination (Table 1). Setting *npop* to 10 narrowed the gap in accuracy between ecotype simulation and the other algorithms, and also resulted in AdaptML producing more accurate demarcations than GMYC and BAPS.

Changing the values of Ω (rate of ecotype formation) had little effect on the accuracy of AdaptML and BAPS. GMYC and ecotype simulation performed worse at the highest level of Ω. GMYC improved when Ω was decreased by a factor of 10, while ecotype simulation stayed at mostly the same level of accuracy at low Ω.

Increasing or decreasing σ (rate of periodic selection) did not affect the accuracy of each algorithm as much as modifying Ω, but BAPS and GMYC had notable decreases in accuracy with a higher rate of periodic selection.

We found that the value of *npop* had a significant effect on the accuracy of three of the four algorithms: ecotype simulation was largely unaffected when *npop* was raised or lowered, while AdaptML, BAPS and GMYC suffered a loss of accuracy with a higher number of ecotypes.

Why does ecotype simulation outperform the other algorithms? Perhaps it is because the data were generated under a periodic selection model, and ES is the only algorithm that explicitly assumes periodic selection to be the force of cohesion within ecotypes (Cohan and Koeppel, 2008). GMYC does not take into account periodic selection (Barraclough et al., 2009) and so might perform better when drift is the primary force of cohesion within bacterial ecotypes. While the BAPS algorithm does not explicitly include a model for the purging of diversity within ecotypes, it may also perform better under a drift model.

With regard to the lower performance of AdaptML, we note that our testing gave the algorithm the best possible chance to identify ecotypes. In the synthesis of sequence data, every ecotype formation event was coupled with a change in habitat; moreover, there were no ecotype formation events involving habitat changes in environmental dimensions not being analysed. Also, we allowed complete habitat specialisation ($\gamma = 0$), which should give the algorithm its maximum resolution (Hunt et al., 2008). We therefore conclude that even when the investigators know the habitats over which ecotypes are specialised, AdaptML offers lower resolution to identify the ecotypes than the other algorithms, especially ecotype simulation. AdaptML will have even lower resolution when ecotypes diverge in habitats that are not known by the investigators (Cohan and Koeppel, 2008). Nevertheless, we note that AdaptML is extremely valuable in that it simultaneously identifies ecotypes and finds the environmental dimensions by which they have diverged (Cohan and Koeppel, 2008).

Finally, we compared the VI scores of the different values of γ among all AdaptML runs (Figure 2). As expected, accuracy was negatively correlated with γ, with AdaptML being the most accurate when γ was 0 (with absolute specialisation of each ecotype to a different habitat).

4.2 Analysis of Bacillus sequences

In the analysis of *Bacillus* sequences, the demarcation results of ecotype simulation, AdaptML, and BAPS were broadly similar, and AdaptML and BAPS yielded identical ecotypes. At the top of the phylogeny (Figure 3), Putative Ecotypes (P.E.) 1–5 identified by ecotype simulation were also identified by AdaptML and BAPS, except that P.E. 2

and 3 of ES were identified as a single ecotype by the other two algorithms. At the bottom of Figure 3, the clade identified as five ecotypes by ES (P.E. 7–11) was lumped into a single ecotype by AdaptML and BAPS.

Figure 2 AdaptML VI scores for different values of γ. AdaptML demarcates ecotypes most accurately (with low VI) when ecotypes are highly specialised to different habitats (i.e., when γ is low). These values were obtained with the *Bacillus*-based middle values of Ω and σ, with $npop = 30$

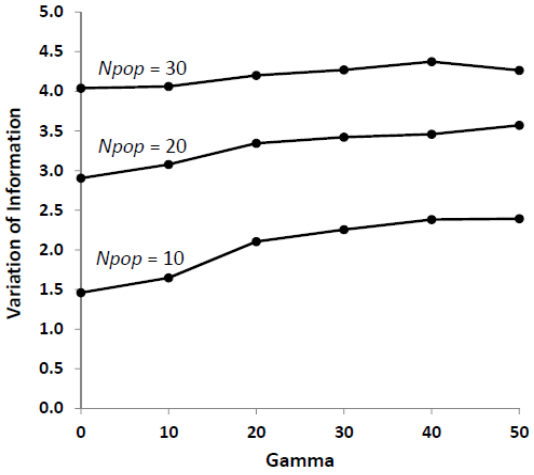
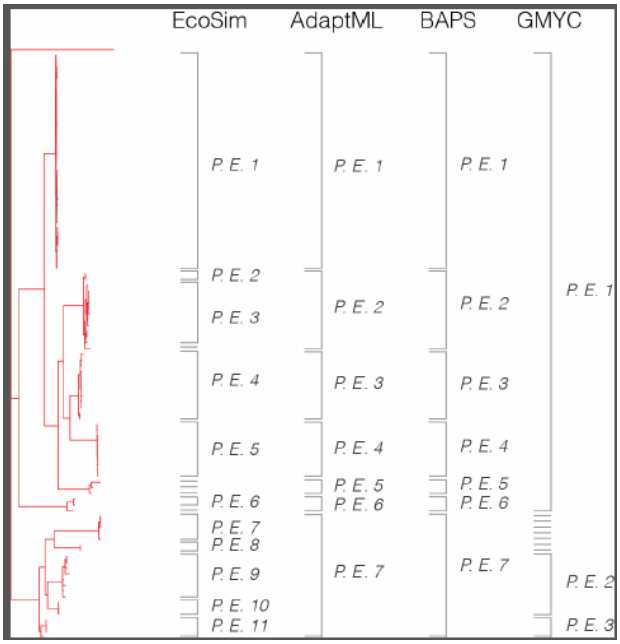


Figure 3 Maximum likelihood phylogeny of *Bacillus* strains with demarcations. Demarcations into putative ecotypes found by Ecotype Simulation (ES), AdaptML, BAPS, and GMYC, for strains isolated from Radio Facility Wash, Death Valley (Connor et al., 2010) (see online version for colours)



GMYC identified many fewer ecotypes than any of the other algorithms. What was identified as six ecotypes, or more by the other three algorithms at the top of Figure 3 (P.E. 1–6 plus singleton ecotypes identified by ES) was demarcated as a single ecotype by GMYC.

Overall, ecotype simulation identified more ecotypes than the other algorithms and no algorithm split a single ecotype hypothesised by ecotype simulation (with one exception – that P.E. 7 and 8 of ES were split into a number of single-strain ecotypes by GMYC). The apparent sensitivity of ES raises the question of whether this algorithm is seeing ecotypes that are not really there. Previous data show that the various putative ecotypes hypothesised by ES, but not identified by the other algorithms, are ecologically distinct and merit recognition as ecotypes. For example, at the top of the phylogeny, P.E. 1–6 of ES has been shown to be significantly heterogeneous in their associations with variations in solar exposure and soil texture (Becraft et al., 2011). While AdaptML and BAPS largely identified these putative ecotypes, GMYC missed them entirely (they are all within GMYC's P.E. 1). At the bottom of the phylogeny, P.E. 7–11 of ES, not discerned by either AdaptML or BAPS, were previously shown also to be significantly heterogeneous in their environmental associations (Connor et al., 2010). We conclude ecotype simulation can resolve real, closely related ecotypes that the other algorithms cannot.

4.3 Running time

We compared the running times of the four algorithms on synthetic data sets of different sizes, and found that AdaptML, GMYC, and BAPS performed demarcations much more quickly than ES (Table 2). AdaptML, GMYC, and BAPS all completed demarcations on the order of a few seconds, even for data sets of 50 sequences whereas ES required on the order of tens of minutes for the same inputs. GMYC was the fastest algorithm, taking no more than one second on average, AdaptML followed, and BAPS was the slowest of the three faster algorithms.

Table 2 Run times. Run times (in seconds) of each algorithm when analysing synthetic sets of sequences produced under the parameter values estimated for *Bacillus*

Algorithm	*20 sequences*	*30 sequences*	*50 sequences*
ES	69.8	384	2390
AdaptML	1.54	1.57	1.64
GMYC	0.201	0.292	0.549
BAPS	4.80	5.15	6.12

5 Conclusions

The algorithm ecotype simulation proved the most accurate of the algorithms studied, in analyses of synthetic sequence data and sequences obtained from closely related *Bacillus* strains. AdaptML and BAPS yielded overall similar demarcations as ES, with somewhat less accuracy, but the GMYC algorithm was unable to identify any of the *Bacillus* ecotypes that had previously been shown to be ecologically distinct. While ES is the most

accurate, it is by far the slowest of the algorithms tested. If this algorithm is to be adapted to analysing the huge data sets that are routinely sampled from environmental DNA, it will have to be made much faster. Improvements to the algorithm are currently under investigation.

6 Methods

6.1 Generation of sequences for analysis

Comparing the accuracy of the four algorithms required us to generate data sets where the strain membership of each ecotype was known. This involved generating a history of a clade in silico for a sample of ν organisms, stemming from a single ancestral sequence. The clade history was based on the rate of ecotype formation (Ω), the rate of periodic selection (σ), and the number of ecotypes (*npop*) within the sample. We used the ecotype simulation algorithm to generate for each organism its sequence and its ecotype affiliation. Also, each ecotype was assigned a habitat preference. The ancestral organism was labelled as having habitat preference 'A' and each ecotype formation event was associated with a change in habitat preference, either from 'A' to 'B' or from 'B' to 'A'.

For each combination of parameter values, we simulated multiple replications of histories of the ν organisms.

We generated data sets with $\nu = 20, 30, 50$ and based the simulation on three values each of Ω and σ. The middle values of Ω and σ were 0.19 and 1.1, respectively, based on prior ecotype simulation analysis of *Bacillus* in a Death Valley canyon (Connor et al., 2010). Lower and upper values of Ω and σ represented 1/10x and 10x of the *Bacillus* values. *Npop* values of 10, 20 and 30 were used, except when they were greater than or equal to the number of sequences. We ran 100 replications for each combination of input parameters and report mean values.

6.2 Preparing the input

For each run, the maximum likelihood tree construction algorithm PhyML was used to build a phylogenetic tree from the generated sequences, since maximum likelihood trees worked best with the AdaptML algorithm. This tree was converted into an ultrametric chronogram using Sanderson's non-parametric rate smoothing algorithm (Sanderson, 2003), which is included in the APE package. This ultrametric tree was used as input for GMYC.

AdaptML requires the habitat from which each strain (or sequence) was isolated. While this information was readily available for our real *Bacillus* sequences, the habitat of isolation of each of the simulated sequences needed to be derived from the habitat preference of the strain's ecotype ('A' or 'B'). Ecotypes of habitat preference 'A' referred habitat '1' and ecotypes of habitat preference 'B' preferred habitat '2'. We took into account different levels of specialisation to habitat using the parameter γ, ranging from 0 (absolute specialisation) to 50 (no specialisation). Members of ecotypes with habitat preference 'A' were isolated from habitat '1' with probability $1–(\gamma/100)$ and from habitat '2' with probability $\gamma/100$; symmetrically, members of ecotypes with preference 'B' were isolated from habitat '2' with probability $1–(\gamma/100)$ and from habitat '1' with

probability $\gamma/100$. The habitat of isolation of an individual organism was determined by a random number based on the probability of isolation from each habitat.

Ecological dimensions were marked in both real and generated sequences as required by AdaptML. The generated sequences were placed into two environmental categories in one dimension. We tested AdaptML with specialisation values (γ) of 0, 10, 20, 30, 40 and 50, using the middle (*Bacillus*-based) values of Ω and σ, with $npop = 30$.

FASTA sequences were converted into XLS format for compatibility with the BAPS package.

For all of the algorithms, we stored the ecotypes demarcated and the strains in these ecotypes in a MySQL database for easy and quick extraction and analysis.

6.3 Bacillus sequences

Bacillus strains were isolated from Radio Facility Wash, a west-running canyon in Death Valley, consisting of habitats with three levels of solar exposure, including the canyon's sunny south-facing slope, the shadier and cooler north-facing slope, and the arroyo at the bottom (Connor et al., 2010). These solar exposure habitats served as the single dimension of ecology we used as environmental input into AdaptML. DNA extraction, PCR amplification, partial sequencing of three genes, and concatenation of the genes were performed as previously described (Connor et al., 2010), and we used PhyML to produce a maximum likelihood tree.

6.4 Variation of information

We used the metric Variation of Information (VI) (Meila, 2003), a criterion for comparing two partitions of the same data set, to determine the closeness of each algorithm's ecotype demarcations to the canonical demarcations generated in silico. In brief, the metric works as follows:

A clustering C is a partition of a data set D into mutually disjoint subsets $C_1, C_2, \ldots, C_k$ called clusters. Let the number of data points in D and in cluster C_k be n and n_k, respectively. The probability of a random point of D being in cluster C_k can be specified by the following random variable:

$$P(k) = \frac{n_k}{n}$$

Using this random variable, we then specify the entropy associated with a particular clustering C:

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k)$$

Let $P(k)$ and $P'(k')$ denote the random variables associated with clustering C and C', respectively. We now define $P(k, k')$, the probability that a point belongs to cluster C_k in clustering C and to cluster $C'_{k'}$ in clustering C':

$$P(k,k') = \frac{|C_k \cap C'_{k'}|}{n}$$

We can then define the mutual information between the two clusterings (how much information each clustering has about the other) as equal to the mutual information between the associated random variables:

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')}$$

Finally, VI is the sum of two terms. The first term measures the amount of information about C that we lose when going from clustering C to C'. The second term measures the amount of information about C' that we gain.

$$VI(C,C') = \left[H(C) - I(C,C')\right] + \left[H(C') - I(C,C')\right]$$

This expression can be simplified as follows:

$$VI(C,C') = H(C) + H(C') - 2I(C,C')$$

VI is always non-negative, symmetric and observes the triangle inequality, i.e. VI is a metric on clustering. By comparing clustering of ecotype demarcations to the 'true' demarcation, each run of each algorithm can be scored with a more accurate demarcation receiving a score closer to 0.

6.5 Running-time tests

We tested the running time of each algorithm on a workstation with a dual-core Intel Core i7 processor running at 2.66 GHz and with 8 GB of RAM, running the Windows 7 operating system. We present the mean run times that each algorithm required to analyse synthetic data sets constructed using the parameter values estimated previously for *Bacillus*, with $\Omega = 0.19$, $\sigma = 1.1$, $npop = 2$ for $\gamma = 20$, $npop = 5$ for $\gamma = 30$, and $npop = 10$ for $\gamma = 50$.

Acknowledgements

This work was funded by NSF FIBR grant EF-0328698 and by research funds from Wesleyan University. All authors worked on the design of the experiments. JCF performed the experiments. All authors participated in the writing of manuscript and have read and approved the final version. Publication of this work was supported by Wesleyan University.

References

Barraclough, T.G., Hughes, M., Ashford-Hodges, N. and Fujisawa, T. (2009) 'Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data', *Biology Letters*, Vol. 5, No. 3, pp.425–428.

Becraft, E., Cohan, F.M., Kühl, M., Jensen, S. and Ward, D.M. (2011) 'Fine-scale distribution patterns of *Synechococcus* ecological diversity in the microbial mat of Mushroom Spring, Yellowstone National Park', *Applied and Environmental Microbiology*, Vol. 77, pp.7689–7697.

Cohan, F.M. (2002) 'What are bacterial species?', *Annual Review of Microbiology*, Vol. 56, pp.457–487.

Cohan, F.M. (2005) 'Periodic selection and ecological diversity in bacteria', in Nurminsky, D. (Ed.): *Selective Sweep*, Landes Bioscience, Georgetown, Texas, pp.78–93.

Cohan, F.M. and Koeppel, A.F. (2008) 'The origins of ecological diversity in prokaryotes', *Current Biology*, Vol. 18, pp.R1024–R1034.

Cohan, F.M. and Perry, E.B. (2007) 'A systematics for discovering the fundamental units of bacterial diversity', *Current Biology*, Vol. 17, pp.R373–R386.

Cohan, F.M., Koeppel, A. and Krizanc, D. (2006) 'Sequence-based discovery of ecological diversity within *Legionella*', in Cianciotto, N.P., Abu Kwaik, Y., Edelstein, P.H., Fields, B.S., Geary, D.F., Harrison, T.G., Joseph, C., Ratcliff, R.M., Stout, J.E. and Swanson, M.S. (Eds): *Legionella: State of the Art 30 Years after Its Recognition*, ASM Press, Washington, DC, pp.367–376.

Connor, N., Sikorski, J., Rooney, A.P., Kopac, S., Koeppel, A.F., Burger, A., Cole, S.G., Perry, E.B., Krizanc, D., Field, N.C., Slaton, M. and Cohan, F.M. (2010) 'The ecology of speciation in *Bacillus*', *Applied and Environmental Microbiology*, Vol. 76, pp.1349–1358.

Corander, J. and Tang, J. (2007) 'Bayesian analysis of population structure based on linked molecular information', *Mathematical Biosciences*, Vol. 205, pp.19–31.

Corander, J., Marttinen, P., Siren, J. and Tang, J. (2008) 'Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations', *BMC Bioinformatics*, Vol. 9, p.539.

Fontaneto, D., Herniou, E.A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C. and Barraclough, T.G. (2007) 'Independently evolving species in asexual bdelloid rotifers', *PLoS Biology*, Vol. 5, No. 4, p.e87.

Francisco, J.C., Cohan, F.M. amd Krizanc, D. (2012) 'Demarcation of bacterial ecotypes from DNA sequence data: a comparative analysis of four algorithms', *IEEE 2nd International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, Las Vegas, pp.1–6.

Gogarten, J.P. and Townsend, J.P. (2005) 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews Microbiology*, Vol. 3, No. 9, pp.679–687.

Hanage, W.P., Fraser, C. and Spratt, B.G. (2006) 'Sequences, sequence clusters and bacterial species', *Philosophical Transactions of the Royal SocietyB*, Vol. 361, No. 1475, pp.1917–1927.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J. and Polz, M.F. (2008) 'Resource partitioning and sympatric differentiation among closely related bacterioplankton', *Science*, Vol. 320, No. 5879, pp.1081–1085.

Jensen, P.R. (2010) 'Linking species concepts to natural product discovery in the post-genomic era', *Journal of Industrial Microbiology and Biotechnology*, Vol. 37, No. 3, pp.219–224.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G.M., Richardson, P. and Chisholm, S.W. (2007) 'Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*', *PLoS Genetics*, Vol. 3, No. 12, p.e231.

Koeppel, A., Perry, E.B., Sikorski, J., Krizanc, D., Warner, W.A., Ward, D.M., Rooney, A.P., Brambilla, E., Connor, N., Ratcliff, R.M., Nevo, E. and Cohan, F.M. (2008) 'Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics', *Proceedings of the National Academy of Sciences*, Vol. 105, pp.2504–2509.

Konstantinidis, K.T. and Tiedje, J.M. (2005) 'Genomic insights that advance the species definition for prokaryotes', *Proceedings of the National Academy of Sciences of the USA*, Vol. 102, No. 7, pp.2567–2572.

Kopac, S. and Cohan, F.M. (2011) 'A theory-based pragmatism for discovering and classifying newly divergent bacterial species', in Tibayrenc, M. (Ed.): *Genetics and Evolution of Infectious Diseases*, Elsevier, London, pp.21–41.

Kuo, C.H., Moran, N.A. and Ochman, H. (2009) 'The consequences of genetic drift for bacterial genome complexity', *Genome Research*, Vol. 19, No. 8, pp.1450–1454.

Lan, R. (1996) 'Reeves PR: gene transfer is a major factor in bacterial evolution', *Molecular Biology and Evolution*, Vol. 13, No. 1, pp.47–55.

Lefebure, T. and Stanhope, M.J. (2007) 'Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition', *Genome Biology*, Vol. 8, No. 5, p.R71.

Marri, P.R., Hao, W. and Golding, G.B. (2006) 'Gene gain and gene loss in *Streptococcus*: is it driven by habitat?', *Molecular Biology and Evolution*, Vol. 23, No. 12, pp.2379–2391.

Martin, A.P. (2002) 'Phylogenetic approaches for describing and comparing the diversity of microbial communities', *Applied Environmental Microbiology*, Vol. 68, No. 8, pp.3673–3682.

Meila, M. (2003) 'Comparing clusterings by the variation of information', in Schölkopf, B. and Warmuth, M.K. (Eds): *Learning Theory and Kernel Machines*, Springer, Berlin, pp.173–187.

Melendrez, M.C., Lange, R.K., Cohan, F.M. and Ward, D.M. (2011) 'Influence of molecular resolution on sequence-based discovery of ecological diversity among *Synechococcus* populations in an alkaline siliceous hot spring microbial mat', *Applied and Environmental Microbiology*, Vol. 77, pp.1359–1367.

Ochman, H. and Davalos, L.M. (2006) 'The nature and dynamics of bacterial genomes', *Science*, Vol. 311, No. 5768, pp.1730–1733.

Paradis, E., Claude, J. and Strimmer, K. (2004) 'APE: analyses of phylogenetics and evolution in R language', *Bioinformatics*, Vol. 20, No. 2, pp.289–290.

Paul, S., Dutta, A., Bag, S.K., Das, S. and Dutta, C. (2010) 'Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*', *BMC Genomics*, Vol. 11, 103p.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. and Vogler, A.P. (2006) 'Sequence-based species delimitation for the DNA taxonomy of undescribed insects', *Systematics Biology*, Vol. 55, No. 4, pp.595–609.

Rosselló-Mora, R. and Amann, R. (2001) 'The species concept for prokaryotes', *FEMS Microbiol Review*, Vol. 25, No. 1, pp.39–67.

Sanderson, M.J. (2003) 'r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock', *Bioinformatics*, Vol. 19, No. 2, pp.301–302.

Sikorski, J. (2008) 'Populations under microevolutionary scrutiny: what will we gain?', *Archives of Microbiology*, Vol. 189, pp.1–5.

Stackebrandt, E. and Ebers, J. (2006) 'Taxonomic parameters revisited: tarnished gold standards', *Microbiology Today*, Vol. 33, pp.152–155.

Staley, J.T. (1909) 'The bacterial species dilemma and the genomic-phylogenetic species concept', *Philosophical Transactions of the Royal Society of London B*, Vol. 361, No. 1475, pp.1899–1909.

Tang, J., Hanage, W.P., Fraser, C. and Corander, J. (2009) 'Identifying currents in the gene pool for bacterial populations using an integrative approach', *PLoS Computational Biology*, Vol. 5, p.e1000455.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al (2005) 'Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"', *Proceedings of the National Academy of Sciences of the USA*, Vol. 102, No. 39, pp.13950–13955.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O. et al: (2009) 'Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths', *PLoS Genetics*, Vol. 5, No. 1, p.e1000344.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K. and Swings, J. (1996) 'Polyphasic taxonomy, a consensus approach to bacterial systematics', *Microbiol Review*, Vol. 60, No. 2, pp.407–438.

Vernikos, G.S., Thomson, N.R. and Parkhill, J. (2007) 'Genetic flux over time in the Salmonella lineage', *Genome Biology*, Vol. 8, No. 6, p.R100.

Vos, M. and Didelot, X. (2009) 'A comparison of homologous recombination rates in bacteria and archaea', *ISME Journal*, Vol. 3, No. 2, pp.199–208.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M. and Whittam, T.S. (2009) 'Cryptic lineages of the genus *Escherichia*', *Applied and Environmental Microbiology*, Vol. 75, No. 20, pp.6534–6544.

Ward, D.M. (1998) 'A natural species concept for prokaryotes', *Current Opinion in Microbiology*, Vol. 1, No. 3, pp.271–277.

Ward, D.M., Bateson, M.M., Ferris, M.J., Kühl, M., Wieland, A., Koeppel, A., Cohan, F.M. (2006) 'Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units linking microbial community composition, structure and function', *Philosophical Transactions of the Royal Society Series B*, Vol. 361, pp.1997–2008.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P. and Truper, H.G. (1987) 'Report of the ad hoc committee on reconciliation of approaches to bacterial systematics', *International Journal of Systemtic Bacteriology*, Vol. 37, pp.463–464.

Wiedenbeck, J. and Cohan, F.M. (2011) 'Origins of bacterial diversity through horizontal gene transfer and adaptation to new ecological niches', *FEMS Microbiology Reviews*, Vol. 35, pp.957–976.