Methodologies for Phylogenetic Inference

Chapter · April 2015

DOI: 10.1002/9780470015902.a0025545

CITATION
1

READS
140

2 authors, including:

Maria Anisimova
Zurich University of Applied Sciences
116 PUBLICATIONS **12,967** CITATIONS

SEE PROFILE

Version 1 a0025545

Methodologies for Phylogenetic Inference

Manuel Gil, *University of Zürich, Zürich, Switzerland*
Maria Anisimova, *Zürich University of Applied Sciences, Wädenswil, Switzerland*

Advanced article

Article Contents

Online posting date: 15th April 2015

Phylogenetic inference from homologous molecular sequences is key to hypothesis testing and problem solving not only in evolutionary biology but also in a wide variety of other fields – from medicine to ecology. Model-based phylogenetic methods rely on Markov substitution models to describe the molecular evolution as a stochastic process of character substitution over time on a phylogenetic tree relating the sequences. Model parameters are estimated with standard statistical inference methods, namely, Bayesian and maximum likelihood approaches. A typical phylogenetic analysis first infers a multiple sequence alignment. Given this alignment, a phylogenetic tree is then estimated together with branch lengths and model parameters. Ideally, the alignment and phylogeny should be estimated simultaneously, amongst others to take alignment uncertainty into account.

Introduction

Molecular data flooding from new generation sequencing (NGS) technologies provide modern biology with an unprecedented amount of data, and with that, the great potential to discover new cellular mechanisms and previously unknown relationships. The most powerful way to mine this data is via evolutionary thinking, rather than focusing on data from one species represented by one individual (Anisimova, 2012). The molecular data we observe at present is due to the events of the past. It is by studying the molecular history of sequences that we can unravel intricate molecular rules. Since Darwin, the molecular history is typically described by a tree structure: it is often referred to as a 'phylogeny' to describe the evolution from a common ancestor by speciation and gene duplication, or a 'genealogy' to describe how genetic material is passed down in populations. Modelling molecular history using trees allows tracing changes through time, bringing a new powerful dimension to any biological study. The time scales may vary from hundreds of millions of years (e.g. for ancient speciation or duplication events) to hours and days, such as in studies of serial viral samples from infected patients. Phylogenetic studies typically rely on analyses of genomic regions in a number of homologous sequences. The homology is often assumed on the basis of sequence similarity stemming from common ancestry, and is established by reconstructing multiple sequence alignments (MSAs).

Phylogenetic analyses are used in a variety of contexts: in forensics [e.g. to provide circumstantial evidence of deliberate HIV transmission (Scaduto *et al*., 2010)], in epidemiology [e.g. to estimate origins of viral epidemics (de Oliveira *et al*., 2006; Worobey *et al*., 2014)], in conservation [e.g. to study effects of climate change (Thuiller *et al*., 2011)] and in medical applications (Nesse *et al*., 2006; Saslis-Lagoudakis *et al*., 2012). New fields such as evolutionary medicine and evolutionary 'omics' are emerging, and starting to frequently rely on our ability to model and reconstruct molecular history from phylogenies. The state-of-the-art phylogenetic inference methods rely on likelihood-based estimation, in a frequentist or Bayesian framework. The computation of phylogenetic likelihood (de Oliveira Martins *et al*., 2013) requires a model of sequence evolution.

eLS subject area: Evolution & Diversity of Life

How to cite:
Gil, Manuel and Anisimova, Maria (April 2015) Methodologies for Phylogenetic Inference. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0025545

Modelling Sequence Evolution

Two processes play a major role in the evolution of molecular sequences: (1) a *process of substitution*, that is, character replacement over time and (2) a *process of insertions and deletions* (*indels*) of sequence fragments. The sequences are thought of evolving on a *phylogeny*, which is represented without the loss of generality by a bifurcating (binary) tree topology with branch lengths. The branch lengths represent the amount of change, which is a product of time, and the rate of molecular evolution. Typically, the leaves of the tree are the observed sequences and the nodes are the ancestral sequences. Characters in the observed sequences, which are related only by substitutions, are termed *homologous*. An MSA estimate for a set of sequences is a data matrix where each row consists of one sequence, often interspersed by gaps reflecting indels, so that potential homologous

characters are placed in the same column [for review on MSA inference, see Loytynoja (2012)].

Stochastic models of *the substitution process* in molecular sequences over time have been well established (Whelan *et al*., 2001; Anisimova and Kosiol, 2009): a continuous-time Markov substitution model on N characters is defined by an $N \times N$ matrix Q describing the instantaneous substitution patterns ($N = 4$ for nucleotides, 20 for amino acids and 61 for codons using the standard genetic code). For computational convenience the typical assumptions are independence and identical distribution for sites, stationarity, reversibility and homogeneity of the process. The probability transition matrix over time t is computed as $P(t) = \exp(tQ)$, where t is measured in expected substitutions per site. At $t = 0$, this matrix is an identity matrix, that is, the probability of any transition is 0.

Describing *the indel process* in probabilistic terms is more challenging: unlike substitutions, indels often involve several sites, vary in length and may 'overlap', obscuring the underlying mechanisms. Descriptions of empirical indel length distributions (Benner *et al*., 1993; Qian and Goldstein, 2001) are difficult to generalise and have limited value for our ability to describe the evolution of indels. However, these empirical distributions have been suggested for alignment inference, and it has been typical to describe the distribution of indels and their length in a sequence assuming fixed divergence rather than describing the indel process over time. The divergence time and taxa sampling, however, have strong influence on observed indel patterns. Several evolutionary indel models have been proposed.

A simple idea is to include a single-position gap as an additional character in the standard Markov substitution model (McGuire *et al*., 2001), which improves model fit for a given alignment, but distorts the biological reality of indel evolution, mainly because of typical assumptions of site independence (only one-site gaps are described), stationarity (implying constancy of gap frequencies over time) and reversibility (prohibiting independent insertion and deletion rates). Such treatment of gaps leads to the so-called 'linear cost model' (i.e. equal gap opening and extension penalties). Extending this solution to a nonreversible process helps lift the requirement of fixed sequence length in time (Rivas, 2005; Rivas and Eddy, 2008), but poses new challenges, and still (crucially) requires site independence.

The classical TKF91 model (Thorne *et al*., 1991) was the first rigorous statistical description of sequence evolution including indels of arbitrary length using an infinite-state continuous-time birth-death process. Another popular approach to modelling indels uses profile-hidden Markov models or HMMs (Eddy, 1998), defined by a sequence profile. HMMs equivalent to TKF91 may be defined (Durbin *et al*., 1998). To align a pair of sequences, pair-HMMs are used, where emission probabilities of characters reflect substitution patterns, while indels are included as additional hidden states with transition probabilities into those states representing the cost of opening a gap and the transition probabilities within these states representing the cost of a gap extension (Durbin *et al*., 1998). Phylo-HMMs extend pair-HMMs to more than a pair of sequences (Siepel and Haussler, 2004). One notable limitation of HMMs for MSA inference is their reliance on a fixed degree of divergence, yet it can be circumvented by accounting for divergence during the MSA optimisation by adjusting for branch lengths of a guide tree using rough statistics (Löytynoja and Goldman, 2008).

Likelihood-Based Phylogenetic Inference

Because phylogeny inference relies on MSAs, the two entities should ideally be estimated jointly (**Figure 1**). The joint estimation of a tree topology τ with branch lengths $B = \{b_i\}$ and an alignment A from a set of unaligned sequences $S = \{s_i\}$, requires stochastic models of indels and substitutions with parameter sets I (indel rates) and Q (a Markov generator matrix with instantaneous substitution rates). In the following we focus on the structural parameters τ and A and for simplicity omit the numerical parameters $\theta = \{I, Q, B\}$, although they are optimised jointly with τ

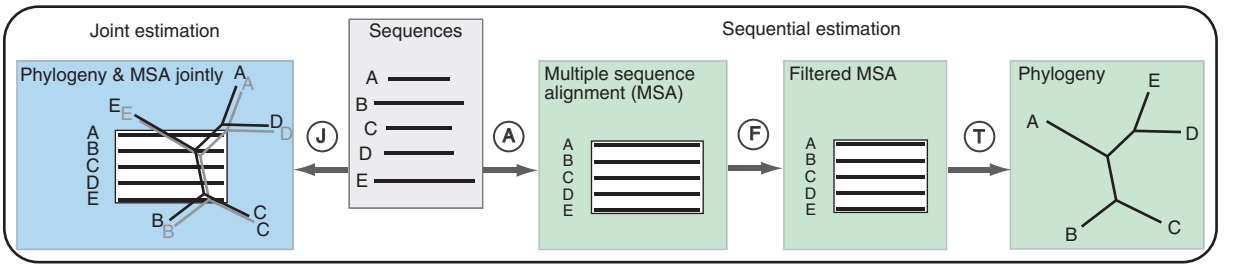

Figure 1 **Overview of a standard work-flow for phylogenetic inference**. Molecular sequences are modelled to be evolving on a phylogenetic tree (phylogeny) according to a character substitution and insertion–deletion (indel) process. The tree topology describes common ancestry by speciation and gene duplication. The branch lengths represent the amount of change. Characters in the observed sequences related by substitutions only are termed homologous. A multiple sequence alignment (MSA) is a matrix where each row consists of one sequence, enriched by gaps to reflect indels, so that homologous characters are matched in the same column.
Ideally, owing to their interrelation, homology and phylogeny (blue box) should be estimated jointly using a model of substitution and indel (J). Amongst other advantages, the joint approach allows taking uncertainty in the MSA into account (for instance, by marginalising it out).
However, a sequential approach (green boxes) is prevalent. Here, first an MSA is reconstructed (A), often followed by filtering (F) – that is, removing unreliable columns – promoted as a way to increase the signal-to-noise ratio of the MSA. The actual tree estimation step (T) typically assumes a substitution model and treats the gaps in the MSA as missing data.

and A. At the core of the joint probabilistic inference is the likelihood function $L(\tau, A|S) = \Pr(S|\tau, A)$, which is used to obtain parameter estimates using either a Bayesian or a maximum likelihood (ML) approach.

The first joint methods have adopted a *Bayesian* approach (Fleissner *et al*., 2005; Lunter *et al*., 2005; Redelings and Suchard, 2005), where the objective of inference is the posterior probability of parameters in focus:

$$\Pr(A, \tau|S) = \frac{L(A, \tau|S)\Pr(A|\tau)\Pr(\tau)}{\sum_{\tau,A} L(A, \tau|S)\Pr(A|\tau)\Pr(\tau)} = \frac{L(A, \tau|S)\Pr(A|\tau)\Pr(\tau)}{\Pr(S)}$$

where we write for brevity $\Pr(A)$ instead of $\Pr(A = a)$ for any random variable A on a realised constant a. The uncertainty is taken into account by specifying prior distributions $\Pr(\tau)$ and $\Pr(A|\tau)$ (as well as for all numerical parameters θ omitted here). The posterior distribution $\Pr(\tau|S)$ obtained from Bayesian estimation packages is often used only to obtain a single estimate of a tree, for example, the maximum *a posteriori* (MAP) tree:

$$\Pr(\tau|S) = \sum_{A} \Pr(A, \tau|S)$$

In contrast, the *frequentist* paradigm enables the method of the ML, which aims to obtain point estimates of parameters of interest. In the context of joint inference, an optimal alignment–tree pair $(\widehat{A}, \widehat{\tau})$ can be obtained by maximising the log-likelihood function:

$$(\widehat{A}, \widehat{\tau}) = \arg\max_{\tau,A} \log L(\tau, A|S)$$

The point estimate of a topology may be obtained by marginalising over the nuisance parameter:

$$\tau = \arg\max_{\tau} \log L(\tau|S) = \arg\max_{\tau} \log \sum_{A} L(\tau, A|S)$$

The most widespread practice, however, adopts a two-step approach (**Figure 1**). First, an MSA is reconstructed and subsequently a phylogeny and substitution model parameters are estimated from the MSA using a Bayesian or ML approach (Yang and Rannala, 2012). Yet, because the alignment and tree estimation problems are interdependent, separating the two comes with a number of problems. First, MSA methods either align sequences progressively using a rough guide-tree, or ignore their evolutionary history altogether (Gotoh, 1999). Conditioning MSAs on inaccurate trees, however, introduces biases which are propagated to the subsequent tree-building step (Redelings and Suchard, 2005). Second, the models used in the two steps are often different, thereby introducing inconsistencies into the estimation procedure. Third, inferred MSAs are treated as fixed and correct entities in the tree-building step without properly taking into account alignment uncertainty.

Alignment Uncertainty

In the context of the two-step approach, several MSA packages provide uncertainty estimates for inferred alignment columns, but they are rarely used in the subsequent tree building. Instead, 'unreliable' columns are often removed before the phylogenetic analysis with ad hoc filtering methods or even by eye (Anisimova *et al*., 2010). The hope is to increase the signal-to-noise ratio of the alignment, however, at the cost of throwing away valuable phylogenetic information from variable and indel-rich regions (Dessinomz and Gil, *et al*., 2010). Yet, basing the analysis on a single MSA, in particular a filtered one, may lead to an underestimate of the variability of parameter estimates, such as substitution rates and the frequency of positively selected site (Wong *et al*., 2008). In the context of joint alignment-tree inference, the Bayesian approach grants access to a posterior distribution of the parameters of interest (i.e. alignment and phylogeny) rather than obtaining just one point estimate of an alignment-tree pair. Such an approach allows estimating posterior probabilities for alignment columns.

Evaluating Statistical Branch Supports

Likewise, the Bayesian approach provides means to estimate posterior probabilities of inferred clades (branch partitions) by summarising a sample of trees from the posterior distribution on the MAP tree. This psychology is dictated by the fact that many downstream analyses rely on a single tree estimate and typically do not allow for uncertainty in either tree or alignment. However, using 'the single best estimate' defies the whole purpose of the Bayesian inference. Still, even this approach comes at a computational cost, especially when sampling the joint alignment-tree space, bringing additional issues of defining sensible priors for A and τ (including branch lengths), monitoring convergence and the decreasing feasibility for realistically large datasets. Consequently, even in the Bayesian framework, it is a common practice to infer an MSA and a tree in two steps.

In the frequentist framework, the joint alignment-tree inference is not fully developed. Rather, the estimation of branch supports relies on the bootstrap. The nonparametric variant requires the data to be a random sample from some conceptual probability distribution. In the two-step approach discussed earlier, during tree estimation the data is an alignment whose m columns are assumed to be realisations of i.i.d. random variables so that bootstrap replicate alignments can be formed by randomly selecting m columns with replacement. To evaluate the variability of a tree estimate, a set of 100–1000 replicates are produced and the tree ML estimation function is applied to each to obtain a collection of trees. The bootstrap trees can be used as an estimate of the distribution of the ML tree estimator. Most commonly, a set of bootstrap trees is summarised by labelling each internal branch B of the original tree estimate with the percentage of bootstrap trees containing B. The interpretation of the bootstrap branch support is, however, disputed. Most commonly it is thought to measure reproducibility or variability, as it does not capture systematic errors. For example, a biased method that always (wrongly) infers the same branch during a bootstrap run, will obtain the maximal (100%) support for that branch. Further, the repeated estimation during bootstrapping is time-consuming.

As a result, faster branch support methods have been developed. These faster heuristic approaches appear to produce bootstrap supports that are generally less conservative compared to the standard bootstrap (Stamatakis *et al*., 2008; Anisimova *et al*., 2011; Minh *et al*., 2013).

Approximate branch support methods (e.g. aLRT, aBayes) were also developed in the frequentist setting. Their computation relies on approximate likelihoods of alternative sub-tree configurations around the branch of interest. Such methods are fast and were shown to be accurate and powerful if the model reflects the data well (Anisimova and Gascuel, 2006; Anisimova *et al*., 2011).

However, because none of the methods are capable of correctly disentangling biases from nonsystematic error, and each method has its weaknesses and advantages, it is a good practice to use several different approaches to evaluate branches. This allows finding the robust ones or, in contrast, those with conflicting signals. In fact, branch support values have become a standard requirement for publication of studies using inferred phylogenies.

Additional Considerations

Assuming a substitution model that accurately describes the observed sequences is important not only for the estimation of branch supports but for the accuracy of phylogenetic tree inference in general. Model fit is usually evaluated using likelihood-based information criteria, such as AIC and BIC, or likelihood ratio tests for nested models (Posada, 2012). Under the true model, likelihood-based methods are known to be asymptotically unbiased; that is, with growing data size, their estimates converge to true values. Both ML and Bayesian appear even to be robust to moderate deviations from model assumptions. The extent of deviation that can be tolerated can be evaluated by simulation specific to the expected scenario.

This article focused on phylogenetic inference for single loci, which is at the cornerstone of evolutionary genomics. Often, however, the aim is to study the phylogenetic signal from multiple loci in order to reconstruct a species tree. This can be accomplished using concatenated gene sequences (the super-tree approach) or by putting together multiple trees reconstructed from single genes (the super-matrix approach). Better still, the ML or Bayesian approaches for multiple loci offer the best statistical properties of phylogenetic inference (Ren *et al*., 2009). Importantly, any such method has to account for the variability of evolutionary patterns (substitution rates, character composition or other biases) across genes. In recognition that species and gene trees may be distinct, sophisticated probabilistic approaches to gene-species tree reconciliation have been developed (Szöllősi *et al*., 2015).

While ML and Bayesian approaches should be preferred for their statistical properties, in some cases, alternative, faster, methods can be employed successfully (Yang and Rannala, 2012). Most prominent are parsimony (inferring trees explaining the data with minimal number of mutational events) and pairwise distance methods [e.g. neighbour-joining (Saitou and Nei, 1987), BioNJ (Gascuel, 1997)]. Parsimony can be used in the absence of a sensible substitution model. Distance methods rely on pairwise evolutionary distances, which are a model-based summary statistic for molecular sequences, representing the leaf-to-leaf path lengths of the underlying phylogenetic tree. Both approaches are faster than the likelihood methods discussed here, work well for lower levels of divergence and can be very useful in the era of large phylogenies. Further, they provide reasonably good starting trees for likelihood-based heuristic searches in ML and Bayesian settings.

References

Anisimova M (ed.) (2012) Evolutionary genomics: statistical and computational methods. Vols 1 and 2. *Methods in Molecular Biology*. New York: Humana Press, Springer.

Anisimova M, Cannarozzi G and Liberles D (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends in Evolutionary Biology* **2** (1): e7.

Anisimova M and Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* **55** (4): 539–552.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C and Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology* **60** (5): 685–699.

Anisimova M and Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution* **26** (2): 255–271.

Benner SA, Cohen MA and Gonnet GH (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology* **229** (4): 1065–1082.

de Oliveira Martins L, Mallo D and Posada D (2013) Phylogenetic likelihood. *eLS*. Chichester: John Wiley & Sons, Ltd.

de Oliveira T, Pybus OG, Rambaut A, *et al.* (2006) Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* **444** (7121): 836–837.

Dessimoz C and Gil M (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology* **11:** R37.

Durbin R, Eddy SR, Krogh A and Mitchison G (1998) *Biological Sequence Analysis*. Cambridge, MA: Cambridge University press.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14** (9): 755–763.

Fleissner R, Metzler M and von Haeseler A (2005) Simulteneous statistical alignment and phylogeny reconstruction. *Systematic Biology* **54** (4): 548–561.

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14** (7): 685–695.

Gotoh O (1999) Multiple sequence alignment: algorithms and applications. *Advances in Biophysics* **36**: 159–206.

Loytynoja A (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. *Methods in Molecular Biology* **855**: 203–235.

Löytynoja A and Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320** (5883): 1632–1635.

Lunter G, Miklos I, Drummond A, Jensen JL and Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**: 83.

McGuire G, Denham MC and Balding DJ (2001) Models of sequence evolution for DNA sequences containing gaps. *Molecular Biology and Evolution* **18** (4): 481–490.

Minh BQ, Nguyen MA and von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* **30** (5): 1188–1195.

Nesse RM, Stearns SC and Omenn GS (2006) Medicine needs evolution. *Science* **311** (5764): 1071.

Posada D (2012) *Selection of phylogenetic models of molecular evolution*. eLS. Chichester: John Wiley & Sons, Ltd.

Qian B and Goldstein RA (2001) Distribution of Indel lengths. *Proteins* **45** (1): 102–104.

Redelings BD and Suchard MA (2005) Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology* **54** (3): 401–418.

Ren F, Tanaka H and Yang Z (2009) A likelihood look at the supermatrix-supertree controversy. *Gene* **441** (1–2): 119–125.

Rivas E (2005) Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics* **6**: 63.

Rivas E and Eddy SR (2008) Probabilistic phylogenetic inference with insertions and deletions. *PLoS Computational Biology* **4** (9): e1000172.

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4** (4): 406–425.

Saslis-Lagoudakis CH, Savolainen V, Williamson EM, *et al.* (2012) Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences of the United States of America* **109** (39): 15835–15840.

Scaduto DI, Brown JM, Haaland WC, *et al.* (2010) Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* **107** (50): 21242–21247.

Siepel A and Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology* **11** (2–3): 413–428.

Stamatakis A, Hoover P and Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology* **57** (5): 758–771.

Szöllősi GJ, Tannier E, Daubin V and Boussau B (2015) The inference of gene trees with species trees. *Systematic Biology* **64** (1): e42–e62.

Thorne JL, Kishino H and Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33** (2): 114–124.

Thuiller W, Lavergne S, Roquet C, *et al.* (2011) Consequences of climate change on the tree of life in Europe. *Nature* **470** (7335): 531–534.

Whelan S, Lio P and Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17** (5): 262–272.

Wong KM, Suchard MA and Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* **319** (5862): 473–476.

Worobey M, Han GZ and Rambaut A (2014) A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* **508**: 254–257.

Yang Z and Rannala B (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13** (5): 303–314.

Further Reading

Felsenstein J (2003) *Inferring Phylogenies*, 2nd (664 pages) ISBN-10: 0-87-893177-5 edn. Sinauer Associates.

Semple C and Steel M (2003) *Phylogenetics*. Oxford Lecture Series in Mathematics and Its Applications 24. New York: Oxford University Press (256 pages) ISBN-13: 978-0-19-850942-4.

Yang Z (2006) *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press (376 pages) ISBN-10: 0-19-856702-2.