LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 6, 2008, 190–199

A Bayesian compositional estimator for microbial taxonomy based on biomarkers

Karel Van den Meersche[1,2], Karline Soetaert [2], and Jack J. Middelburg[2]*
[1]University of Gent, Department of Biology, Marine Biology Section, Krijgslaan 281-S8, 9000 Gent, Belgium
[2]Netherlands Institute of Ecology (NIOO-KNAW), P.O. Box 140, 4400 AC Yerseke, The Netherlands

Abstract

Determination of microbial taxonomy based on lipid or pigment spectra requires use of a compositional estimator. We present a new approach based on Bayesian inference and an implementation in the open software platform R. The Bayesian Compositional Estimator (BCE) aims not only to obtain a maximum likelihood solution, but also to provide a complete estimate of the taxonomic composition, including probability distributions and dependencies between estimated values. BCE results are compared with those obtained with CHEMTAX. The BCE has not only a similar accuracy, but also extracts more information from the data, the most obvious being standard deviation and covariance estimates.

In aquatic ecology, microorganisms play an important role as primary producers and primary consumers. Determination of their abundance and species composition is traditionally done by microscopy, which has significant drawbacks. Many microbial cells lack clear morphologically distinctive traits. Moreover, cell preparation measures, such as filtering or staining, may damage cells, altering their morphology (Gieskes and Kraay 1983). Smaller phytoplankton, especially picoplankton, are recognized to contribute significantly to marine primary production (e.g., Li et al. 1983; Platt et al. 1983), yet are difficult to determine microscopically and are easily overlooked during cell counts.

If determination up to species level is not necessary, biomarkers provide a powerful alternative to cell counts. In ecology, these are molecules that exclusively or predominantly occur in distinct species, taxonomic, or functional groups. They can be used to identify and quantify broad taxonomic groups when microscopic quantification fails or is considered too time-consuming. A common method is pigment analysis of water samples using high performance liquid chromatography (HPLC), allowing determination of major phytoplankton groups, typically up to class level (e.g., *Cryptophyceae*, *Prasinophyceae*, *Bacillariophyceae*). Lipid analysis using gas chromatography (GC) is another example, which also allows estimating abundances of heterotrophic cells to a certain extent. If a gas chromatographer is coupled to an isotope ratio mass spectrometer, then also stable isotopes can be measured in specific lipids making it possible to link identity with functioning (Boschker and Middelburg 2002; Van Den Meersche et al. 2004).

For biomarkers to be used as a quantification tool, they have to meet several requirements. Within the considered taxon, their concentration must be a reasonably constant fraction of total biomass. Also, biomarkers should be short-lived, or degrade rapidly after cell death, if they are to represent living biomass. A third prerequisite, which is rarely met, is the exclusive occurrence of a biomarker in one or a few taxa. Biomarkers that are unique for one taxon can then be used for quantification of this taxon. Examples are zeaxanthine for cyanobacteria (Mackey et al. 1996), branched fatty acids for bacteria (Gillan et al. 1981; Sargent et al. 1987), and ladderanes for Anammox bacteria (Damsté et al. 2002). However, really unique biomarkers are rare, and more often taxa share a number of biomarkers, or differ merely in the relative importance of various biomarkers. In this case, the whole spectrum of analyzed compounds needs to be taken into account. Many chromatography-based techniques provide such a spectrum of compounds. Estimation of taxonomical composition from these spectra of non-unique biomarkers is possible, although not as straightforward as estimation from unique biomarkers.

The most well-known and widespread technique to estimate microplankton composition from biomarker data in aquatic

*Corresponding author: k.vdmeersche@nioo.knaw.nl

Acknowledgments

We thank Dr. Nicole Dijkman and Dr. Nicolas Van Oostende for providing feedback and testing the method, Prof. Dr. Carlo Heip for guidance and providing research facilities, and two anonymous reviewers for constructive feedback. The research was supported by a grant from the Flemish Fund for Scientific Research (FWO) and extra funding from the University of Gent. This is contribution 4279 from the Netherlands Institute of Ecology.

ecology is the CHEMTAX program (Mackey et al. 1996). It has proven useful as a tool for estimating taxonomic composition of phytoplankton solely from the pigment composition and has been applied in different marine and lacustrine environments (Lionard et al. 2005; Llewellyn et al. 2005; Rodriguez et al. 2002). It has been benchmarked against microscopic counts, and in most cases, leads to acceptable estimates of phytoplankton composition. As the method is not restricted to water samples or pigment data, it has been successfully applied to sediments as well as to phospholipid-derived fatty acids (PLFA) (Dijkman and Kromkamp 2006). The combination of pigment or lipid analysis and CHEMTAX is a suitable, sufficiently accurate, and possibly time-saving alternative to elaborate cell counts, which has been proposed as a monitoring tool for phytoplankton bloom successions in coastal ecosystems (Muylaert et al. 2006).

CHEMTAX not only estimates taxonomic composition, but it also adjusts the taxon-specific biomarker ratios to obtain a better fit to the data. This makes the result less sensitive to the input values for the biomarker ratios and potential errors therein. This was presented as an important feature of the algorithm, and it was shown that even with input values deviating significantly from the true values, the algorithm still converged to the correct result.

However, in many cases, CHEMTAX has problems with identifiability of biomarkers (Latasa 2007). On one hand, during calculations, several local optima may be present, implying several mathematically viable solutions, and the algorithm may select the wrong one. For this reason, it is recommended to verify the results with microscopic observations of the same sample. The user can then narrow the range in which ratio values are allowed to vary and thus force the algorithm into the direction of one single solution. On the other hand, if the problem is underdetermined, several runs may end up in different results. The composition matrix is then unidentifiable. Especially when direct observations of occurring taxa are not available, it becomes hard to draw sensible conclusions from the data. Latasa (2007) proposed iterative CHEMTAX runs to improve convergence to one solution. He could show that in many cases this is the correct solution, but exceptions remain. Finally, CHEMTAX uses an optimization routine; it estimates just one solution and does not allow estimating uncertainties on the estimated composition.

In this paper, we propose a technique that tackles these shortcomings. Instead of searching one optimal result, we use prior knowledge on biomarker ratios and sampled data in terms of probability distributions to assess the probability distribution of the sample compositions. This expected probability distribution, also called posterior probability distribution, is obtained with Bayesian inference, which involves fitting a probability model to data (Gelman et al. 2004). Although the fundamentals of this type of statistics were established in the 18th century, mathematical difficulties in calculating the posterior probability distribution postponed its wide-range application until the second half of the 20th century, when numerical solutions became possible thanks to computers. Today, Markov Chain Monte Carlo (MCMC) algorithms are widely available and can be used to sample the posterior probability distribution numerically (Gilks et al. 1996). The result is a full, multidimensional posterior probability distribution of the parameters, that also describe relationships between parameters. The full posterior probability distribution can be used to extract means, confidence intervals, covariances, and maximum likelihood, as well as to recognize unidentifiable parameters.

In this paper, we first introduce the mathematical foundations of the Bayesian Compositional Estimator (BCE) for taxonomical determination based on biomarker profiles. This BCE will be assessed and benchmarked against CHEMTAX. This Bayesian method not only allows investigators to fully use available data (prior information), but also provides uncertainty estimates to the final outcome. In the discussion, we then deal with identifiability, uncertainty, and some fundamental choices concerning the use of prior distributions in Bayesian inference.

Materials and procedures

To formalize the problem of estimating compositions from biomarker ratios and biomarker data, we define three matrices: an input data matrix B containing biomarker ratios in (field) samples, an input ratio matrix A containing the biomarker ratios for several taxonomic groups, and an unknown compositional matrix X. Each row of A contains the biomarker composition of one taxon, while each row of B contains the biomarker composition of one sample. After solving the model, each row of X will contain the taxonomic composition (the relative proportion of a taxon) of each sample. All elements of X are positive, and the row sum equals 1. The product of A and X approximates or equals the data matrix B:

$$\mathbf{X}_{s\times t}\mathbf{A}_{t\times b} \cong \mathbf{B}_{s\times b}$$

s is the number of samples, t the number of taxa, and b the number of biomarkers. The mathematical core problem is the estimation of X, given A and B. This is a multidimensional linear inverse problem. Three cases can be considered. If the number of biomarkers is smaller than the number of taxa, the problem has more unknowns ($s * t$) than equations ($s * b$), and there exist an indefinite number of solutions (the problem is underdetermined). If there are as many biomarkers as there are taxa, and all matrix rows are linearly independent, a unique solution $X = BA^{-1}$ can be found, but this solution is not necessarily positive-definite (all elements > 0). In general, one will have more biomarkers than taxonomical groups. Then an over-determined system is obtained, which can be solved with a (constrained) least squares regression. The constraints are that the elements of the compositional matrix X have to be positive (inequalities), and they should sum to 1 for each sample (equalities). Different algorithms are available for solving

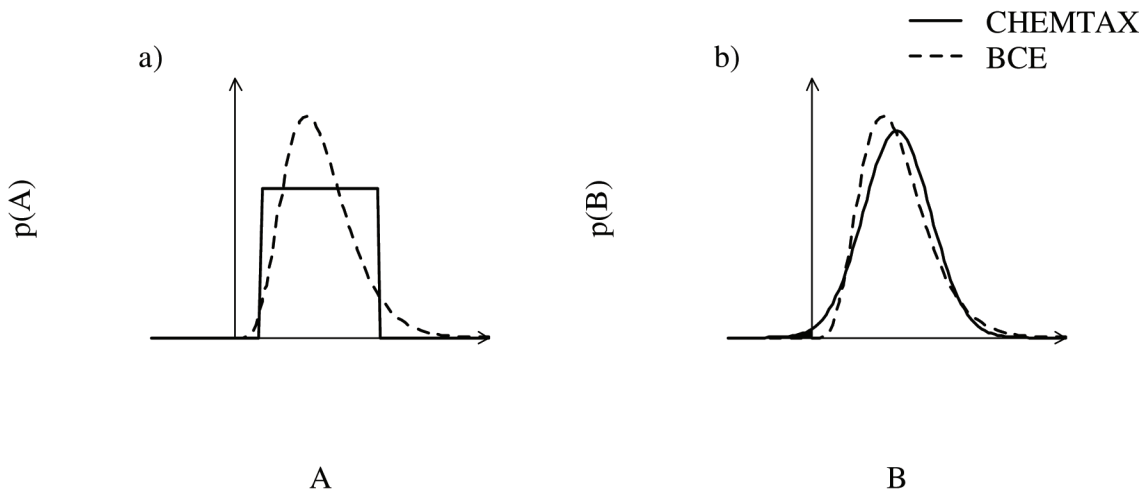

Fig. 1. Prior probability distributions of the ratio matrix A (a) and the data matrix B (b) in CHEMTAX (solid line) and BCE (dashed line).

the over- and evenly determined problem, e.g., the LSEI algorithm (Lawson and Hanson 1995). Recently, a method has been devised to sample the solution space in case the problem is underdetermined (Van den Meersche et al. in prep).

If the input ratio matrix A is known exactly, the above outline suffices to estimate X. However, natural variations in biomarkers occur all the time: between species of the same taxon, between strains, or even between individuals of the same species. Physiological and/or environmental conditions influence the composition of an individual cell. Thus, the elements of the input ratio matrix A are themselves uncertain. One way to deal with this is to also treat the elements of the input ratio matrix A as unknowns, which have to obey certain constraints. This is the procedure adopted in the CHEMTAX method, where prior information about A is implemented as a uniform distribution with boundaries set by a limitation matrix, i.e., all the elements of A are assumed to be within minimum and maximum ranges (Fig. 1a). The CHEMTAX procedure then selects the optimal A and X matrices for which the discrepancy between the product XA and the data B is minimized in the least square sense, i.e., $\min\left(\|XA - B\|^2\right)$ while conforming to the inequalities A > Amin and A < Amax and the constraints for X.

Treating the uncertainty of the input ratio matrix A as a uniform distribution implies random ratios within the specified range. Literature values for the ratio matrix are not random but have a most likely value (the mean of the measurements) and some accuracy (the standard deviation).

To take into account uncertainties in the ratio matrix A, one can perform a least squares regression with uncertainties for independent variables A as well as dependent variables B, while X contains the regression parameters. An outline for this method can be found in Tarantola (2005). Alternatively, in the BCE algorithm, A is given a more realistic continuous probability distribution function with an average value and a standard deviation. Because both X and A are uncertain and have to be estimated, the problem is not linear.

In the formalism of Bayesian inference (Gotelli and Ellison 2004), we distinguish the model parameters X and A, a model M, which is the product XA, and a data series B. The rule of Bayes for the conditional probability density is generally valid:

$$p(X,A|B) = \frac{p(X,A,B)}{p(B)} = \frac{p(X,A)p(B|X,A)}{p(B)}$$

The left-hand side (probability of the parameters, given the data) is the probability of interest, i.e., the probability of both the elements of X and A, in view of the data B. In the right-hand side, the probabilities of X and A are independent. The probability of the data given the parameters p(B|X,A), doesn't depend on the parameters directly but only on the model result M = XA. Therefore, the rule of Bayes can be rewritten into:

$$p(X,A|B) = \frac{p(X)p(A)p(B|M)}{p(B)}$$

Generally, the only prior information on X is that all elements have to be positive, and the row sums have to be equal to 1. The prior probability of X can thus be considered constant. Also, the probability of the data in absence of the model, p(B), is constant. Thus, the combined posterior probability of the composition matrix X and the ratio matrix A given the data B, is proportional to the product of the prior probability of A, and the probability of B given the model outcome M = XA:

$$p(X,A|B) \propto p(A)p(B|M)$$

A prior probability distribution of A can be provided as any non-negative distribution with a given average and standard deviation. The probability of the data B given the model outcome, p(B|M), can also be estimated as a probability distribution depending on M.

Then we can calculate the posterior probability for each given set of X and A from the prior probabilities of A and B.

In BCE, the default implemented probability distribution for A as well as B is a γ distribution. There are two main reasons for this choice. First, all values of A and B must be positive, which excludes symmetric distributions such as the normal distribution. Two, the presence of zeros or near-zeros in the data, makes the lognormal distribution unsuitable, as it causes numerical problems with values that are near zero. γ distributions approach normal and lognormal distributions when standard deviations become smaller, and approach exponential distributions when the modus becomes smaller. They can also easily be expressed in terms of mean or modus and standard deviation. The assumed prior probability distributions of A and B in CHEMTAX and in the BCE are compared in Fig. 1.

The joint posterior probability distribution of the parameters X and A is sampled numerically using a random walk. The Metropolis-Hastings algorithm (Roberts 1996) generates samples of the parameter space (X and A) of which the distribution approximates this posterior probability distribution. The algorithm starts with a chosen initial set of parameters: the mean values of the input ratio matrix A and the least squares regression solution for X. This initial set can be viewed as one point in a multidimensional parameter space. From this parameter point (X_1,A_1) and every accepted point (X_i,A_i) thereafter, a new point is drawn randomly from a jump distribution that only depends on the last accepted point. This new point is either accepted or rejected based on the following criterion:

$$\text{if } r \le \frac{p(X_{i+1},A_{i+1}|B)\,\text{jmp}(X_i,A_i|X_{i+1},A_{i+1})}{p(X_i,A_i|B)\,\text{jmp}(X_{i+1},A_{i+1}|X_i,A_i)} \text{ accept } X_{i+1},A_{i+1}$$
$$\text{else accept } X_i,A_i$$

Where jmp is the jump distribution (see below), r is a random number sampled uniformly between 0 and 1, and p(X,A|B) is estimated as in Eq. 1. After many iterations, the distribution of the accepted points approaches the true posterior probability distribution of the parameters. The jump distribution can be any kind of distribution that selects a new point only depending on the last point. However, this jump distribution should be selected with care for optimal performance. For A, a normal jump distribution is used, and for X, a Dirichlet jump distribution. The Dirichlet distribution operates in a standard simplex, see e.g., Aitchison (1986), and therefore, ensures that every new randomly chosen X has a sum of 1 and all parameters > 0.

Good mixing of the MCMC random walk, meaning that the solution space has been thoroughly sampled, is an important criterion for the validity of the parameter estimates. Inspection of the MCMC output is a crucial step before accepting results. If the parameters vary randomly during the MCMC, without showing clear patterns, then the random walk can be considered well-mixed. This was double-checked by comparing runs with different initial conditions, which led to comparable results. Mixing improved a lot when a fraction of the ratio matrix standard deviation was used as jump length matrix. One-hundred thousand iterations proved to be sufficient for most model runs, whereas the longest model run contained 1,000,000 iterations.

The BCE algorithm is implemented as a function in the statistical environment R (http://www.r-project.org). It can be obtained from the authors upon request and, will be made available as an R-package from the R website. R is a free software implementation of the S statistical programming language. It offers a wide range of mathematical, statistical, and graphical techniques, and is comparable in power and applicability to major software packages such as Matlab and S-plus. Packages to assess convergence of the MCMC procedure are available in R and can be applied to the output of BCE.

Assessment

Two problems are frequently neglected in optimization approaches: identifiability and uncertainty. Uncertainty originates from the accumulated inaccuracies in data points and in the model. In Bayesian models, it is incorporated in the prior probability distributions. These then result in uncertainties in the parameter estimates, described in their posterior distributions. Identifiability has little to do with the quality of the data, but a great deal with the model: some parameters are difficult or impossible to estimate accurately, either because the model is underdetermined (see above), or because even small uncertainties in the data lead to large uncertainties in the result, due to the structure of the model.

Although Bayesian approaches have been incorporated into ecological textbooks (Gotelli and Ellison 2004), the approach is largely underexploited by the research community. To illustrate the advantage of the BCE for assessing identifiability, we will work out a simple case of two taxa and three biomarkers. A data sample (B) contains biomarker ratios observed in an experiment:

$$B=\begin{bmatrix}0.2 & 0.7 & 0.5\end{bmatrix}$$

The ratio matrix A_1 contains proposed biomarker ratios in the two taxa present in the sample:

$$A_1=\begin{bmatrix}0.3 & 0.7 & 0.4\\ 0.2 & 0.8 & 0.6\end{bmatrix}$$

Standard deviations are all set to 0.1. What is the expected taxonomic composition? In other words, what is the expected probability distribution of the elements of the solution matrix $X = [X_1\ X_2]$, with two components X_1 and $X_2 = 1 - X_1$?

The system is overdetermined when solving the equation $XA_1 = B$ and has only one best solution, bestX = [0.45 0.55]. However, due to the small differences in biomarker composition between the two taxa compared with the standard deviations, the resulting estimate for X is almost completely undetermined (Fig. 2a). The 66% prediction interval (equivalent to a mean ± standard deviation in a normal distribution) for X_1 is

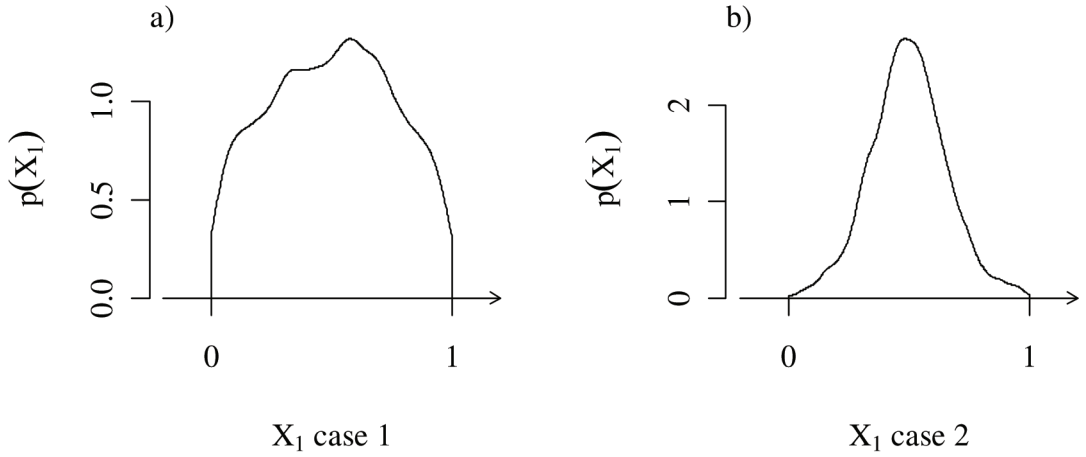

Fig. 2. Identifiability: Posterior probability density function of the composition estimates of the first of two taxa with similar (a) and clearly distinct (b) biomarker ratios. Identifiability problems occur when taxa have similar ratios (a).

[0.21 0.78], while the 66% prediction interval of a uniform distribution between 0 and 1 would be [0.17 0.83] (the distribution of X_2 is of course the mirror image of the distribution of X_1). This demonstrates that if estimated errors on the data are large compared with the differences in biomarker ratios (e.g., pigment composition) of the considered taxon, a solution can remain unidentifiable, even for an overdetermined system.

Now consider a different input ratio matrix, A_2, where differences in biomarker compositions between the two taxa are more pronounced:

$$A_2=\begin{bmatrix}0.3 & 0.7 & 0.1\\ 0.2 & 0.8 & 0.9\end{bmatrix}$$

The solution with the highest probability is the same as for previous case: bestX=[0.45 0.55]. However, the uncertainty ranges are much smaller (Fig. 2b). The 66% prediction interval for X_1 is now [0.35 0.64].

Mackey et al. (1996) presented an input ratio matrix for a Southern Ocean phytoplankton community, mainly based on quantitative data from algal cultures. We use this same ratio matrix to estimate compositions with the BCE, which should ease comparison with results obtained with the CHEMTAX method. An artificial data matrix B was produced by taking the product of a random compositional matrix X, and the ratio matrix A, and adding small perturbations (<5%). The original compositional matrix is the expected result of the analysis.

Figure 3 shows pairwise scatter plots of the random walk solutions of one sample (below diagonal), together with the marginal probability distributions of the taxa (on the diagonal). Bayesian methods like BCE output a multidimensional probability distribution. Marginal distributions and summary statistics can then be inferred from this distribution. Groups of organisms that cannot be properly quantified can also be identified. It also illustrates another aspect of uncertainties in a composition matrix (with a constant row sum of 1). The MCMC random walk shows that while the estimates of the proportion of most taxa are uncorrelated, the estimates for the relative proportion of Cryptophytes and Haptophytes are inversely correlated. These two taxa make up the bulk of the composition of the sample, and have a larger range of uncertainty. Because all other taxa are well-constrained and the sum of all taxa is one, also the sum of Cryptophytes and Haptophytes is well-constrained, and their distributions are inversely correlated. Note however that this inverse correlation originates from the model and does not imply an ecological trade-off between these two taxa.

Posterior distributions of the ratio matrix (A) and the composition matrix (X) are determined by their prior distributions and the distribution of the data. For the remainder of this section, we discuss the consequences of prior distribution selection and sample size on the result. All samples are generated based on the Southern Ocean phytoplankton community data of Mackey et al. (1996), and using variable standard deviations of the ratio matrix and data matrix.

In Fig. 4, we compare uninformative (uniform) and informative prior distributions (γ distributions with a relative standard deviation of 0.2) for the input ratio matrix. When uninformative priors for the ratio matrix are used, the posterior distribution of the ratio matrix and the composition matrix are solely estimated based on the data. This is a situation that is comparable with the CHEMTAX algorithm. The graphs in 4b show the results for runs including one sample. In this case, large uncertainties in the composition estimates (X) appear, even for small uncertainties on the data. This is best understood by considering the number of unknowns (all elements of X and all non-zero elements of A) compared with the number of data points. There are simply not enough constraints, and the problem remains underdetermined. The inclusion of several samples in one run will raise the number of constraints and make the problem overdetermined. When 40 samples are included (Fig. 4d), the compositional uncertainties are lower and with reliable data (small

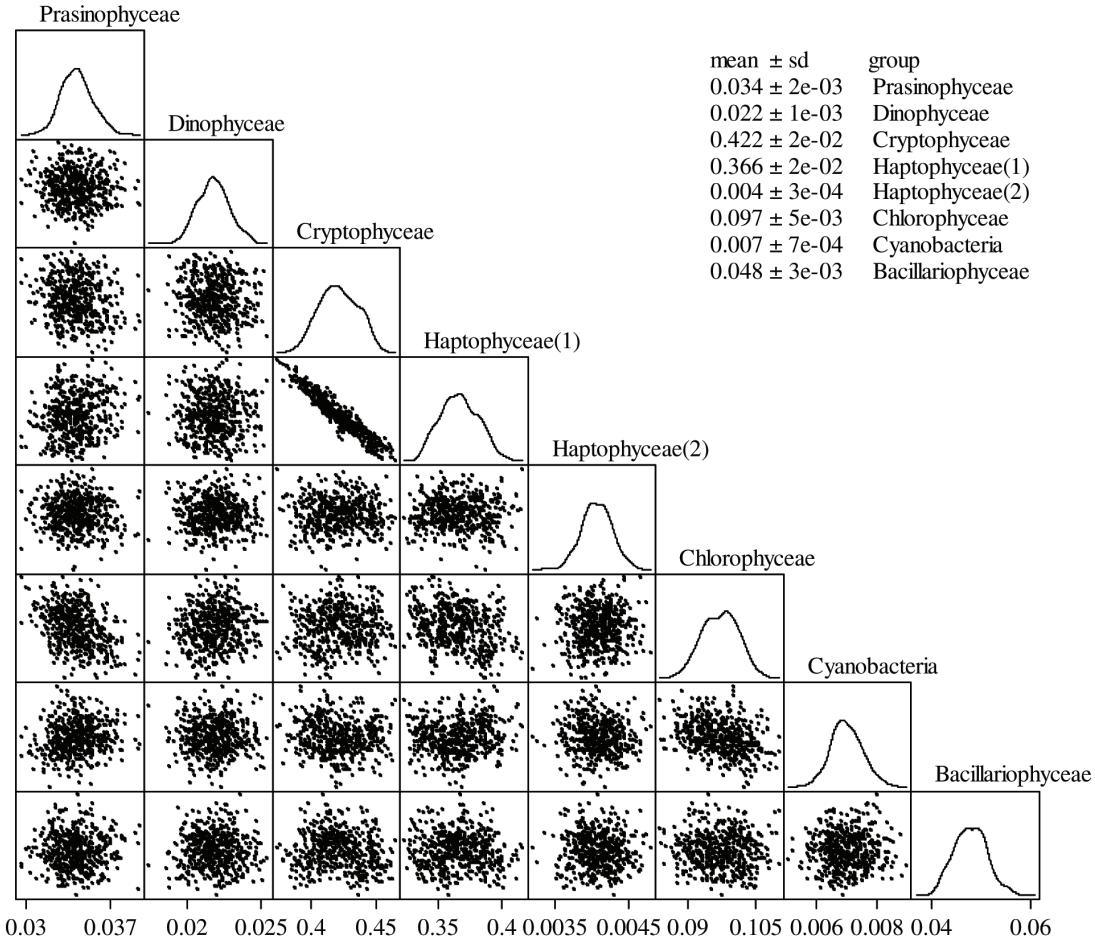

Fig. 3. Results of the MCMC random walks in 1 sample of the Southern Ocean artificial data. Note the correlated distributions for cryptophytes and haptophytes. Only 500 out of 5000 iterations are shown for clarity.

standard deviation), it is possible to infer compositions with reasonable precision.

When informative priors are used for the input ratio matrix A, and 40 samples are included per run (Fig. 4c), we observe little or no difference in the posterior distribution of the compositional matrix X compared with uninformative priors for A (compare Fig. 4c and 4d). This is due to the fact that the results are completely determined by the data matrix B. The huge amount of data (40 samples) weigh so heavily on the result that the ratio matrix has only minor impact.

It is usually not advisable to include this many samples in one analysis. By combining samples from different communities, the obtained variability in the output ratio matrix will reflect spatiotemporal variability in the community properties. In other words, when including one sample in the analysis, the posterior probability in the input ratio matrix will reflect the uncertainty for this particular sample. When more samples from a larger area are combined, the uncertainty will also include the spatial and temporal heterogeneity. This is illustrated in Fig. 4a and 4c. When including 40 samples per run (Fig. 4c), uncertainty is high. When we include only one (the first) sample in the analysis, and use an informative prior probability distribution of the ratio matrix A (Fig. 4a), we see that the composition matrix X can be estimated with good accuracy. The taxonomic composition is now a compromise between samples and knowledge on biomarker ratios.

To analyze the combined effect of the informative prior distributions on the posterior distribution in more detail, a series of model runs were performed with the same mean value for ratio matrix and data matrix, but with different relative standard deviations (Fig. 5). One sample was included. Increases in standard deviation in the ratio matrix as well as in the data matrix are reflected in increased standard deviations in the estimated composition matrix. However, these spreads have little effect on the mean composition. The BCE mean composition is very similar to the composition which resulted from a CHEMTAX analysis using the same data and ratio matrix. Also the median of the ratio matrix did not deviate a lot from the original ratio matrix. The slight discrepancy is due to a

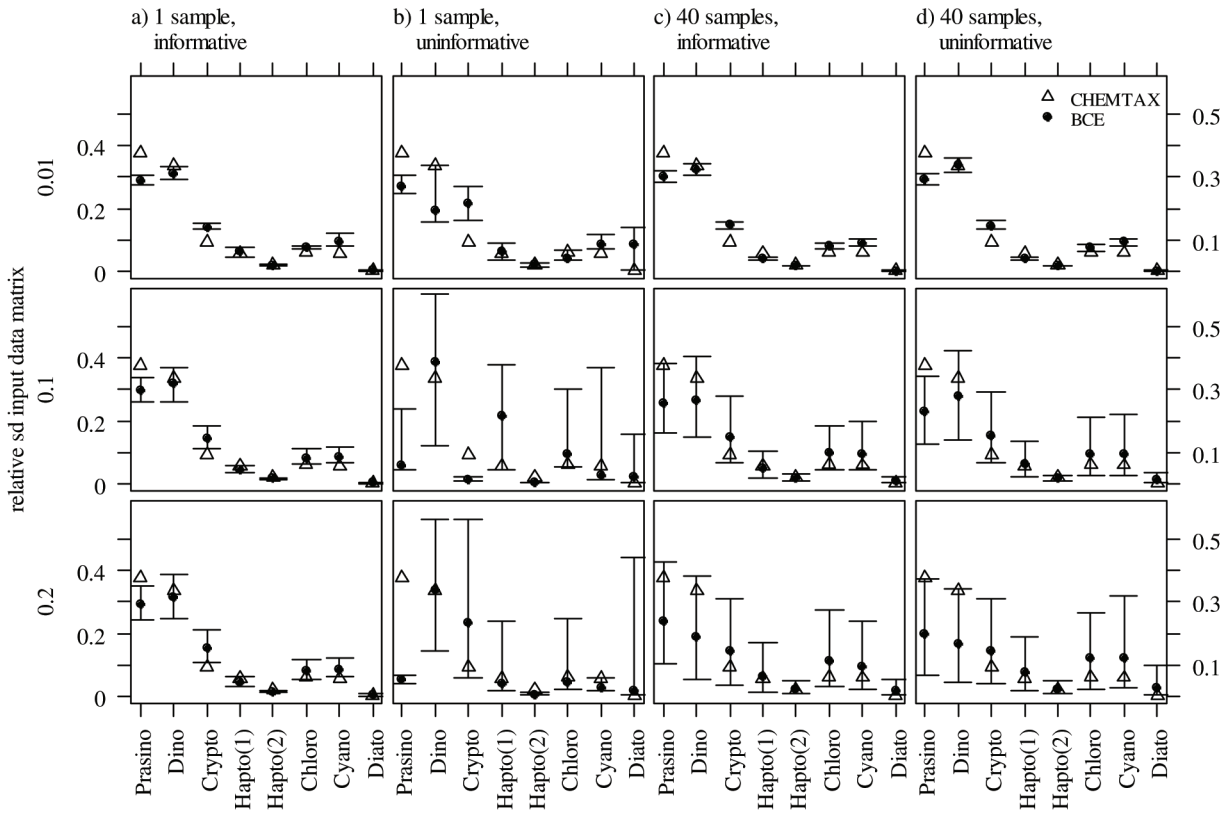

Fig. 4. Comparison of uninformative (uniform, b, d) and informative (a, c) prior distributions for the input ratio matrix. CHEMTAX results are presented for comparison. Informative priors are γ distributions with a relative standard deviation of 0.2. The results obtained when including one sample (a,b) and 40 samples (c,d) are compared for different relative standard deviations for the prior distributions of the data matrix. Error bars indicate the 10% and 90% quantiles.

difference in implementation. CHEMTAX uses rescaled input matrices for data and biomarker ratios. BCE doesn't rescale, and therefore, can be applied with a combination of different types of biomarkers. All the columns in the BCE data matrix are thus independent, which is not the case in CHEMTAX. Because CHEMTAX defines a normal distribution on rescaled, dependent values and BCE defines distributions on the non-rescaled, independent values, results are not exactly the same, even when using the same distribution types.

Discussion

Traditional (or frequentist) and Bayesian statistics primarily differ in the use of prior information. Ecology is one of the fields where Bayesian inference only recently started gaining popularity (Clark 2005). The lack of "plug and play" software packages and the apparent complexity of Bayesian applications often discourage researchers. However, ecological data are notorious for not meeting assumptions for frequentist models and for their complexity that goes beyond the power of these simple models. The power of Bayesian approaches lies exactly therein: fewer assumptions have to be met, prior knowledge can be incorporated at will and the hierarchical structure of Bayesian models allows analysis with virtually any complexity. While unidentifiability in classical approaches forces the use of over-simplified models, there is no such need in Bayesian modeling (Omlin and Reichert 1999). As illustrated in the assessment, this turns out to be a relevant feature when estimating compositions. The BCE treats uncertainties and unidentifiabilities properly and incorporates them into the solution. Monte-Carlo simulation based correlation plots as shown in Fig. 3 guide the user in identifying the groups of organisms that cannot be properly quantified from a certain data set because of identifiability problems. The BCE also provides uncertainty estimates around the best estimates and thus fully embraces the variance inherent to natural communities.

The BCE has a great flexibility in its choice of prior probability distributions. If one is uncomfortable with defining prior probabilities, uninformative (e.g., uniform) priors are an attractive solution. Uniform priors for the ratio matrix in combination with normal distributions for the data matrix produce results that are equivalent to the results obtained with a least squares method such as CHEMTAX. There are some important consequences when using uniform distributions for the input ratio matrix. As shown in the assessments, the final result for biomarker ratios and sample compositions

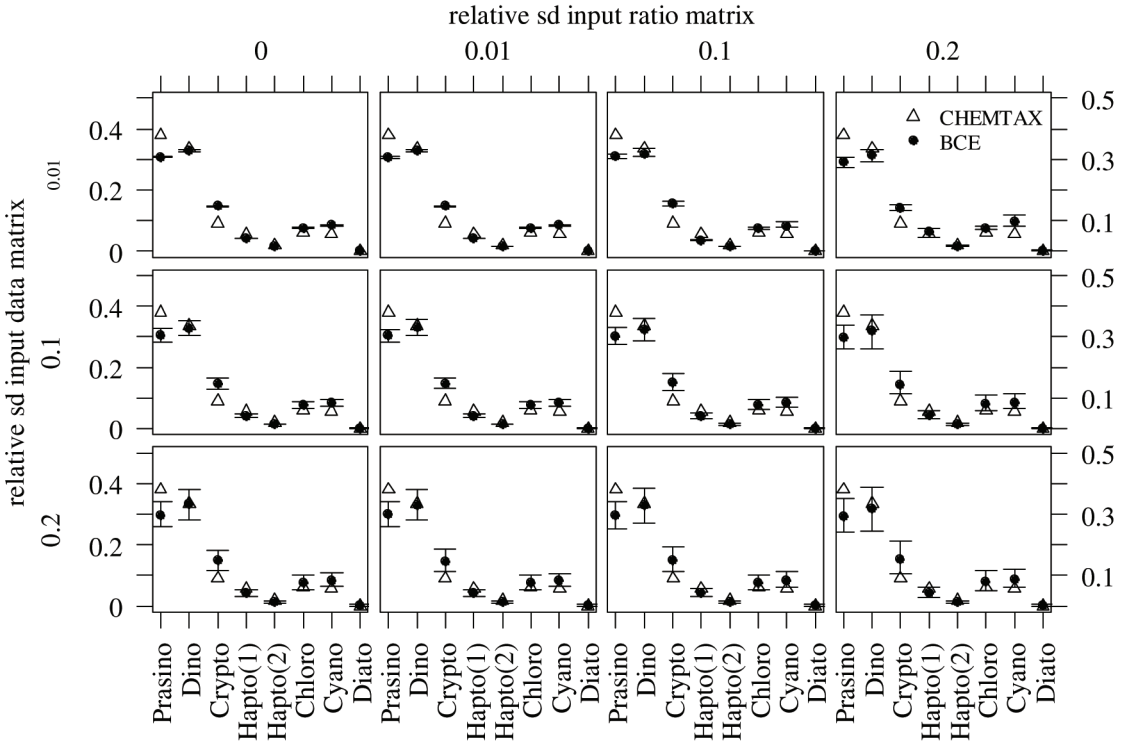

Fig. 5. Estimated compositions based on CHEMTAX and BCE using different relative standard deviations for prior distributions of ratio matrix and data matrix. Means were the same for all distributions. Absolute standard deviations were set to 0.001. Relative standard deviations were varied. Error bars indicate the 10% and 90% quantiles.

will then mainly depend on the data. This implies that one needs excellent quality data and a sufficiently large number of samples to include in the run. Otherwise identifiability problems will occur, uncertainty will be unbearably large, and there will be poor recovery of the true composition in the samples. Moreover, these samples need to be independent and have distinct compositions. When there is a strong dependence between samples, identifiability can remain an issue. There is, however, an important ecological caveat for the use of different independent samples in one run. When combining samples, one assumes that the taxonomical groups present in each sample have exactly the same biomarker composition. Composition of biomarkers, notably pigments, is not uniform within higher taxonomical groups. Besides interspecific variation, compositions may also vary within one species due to differences in environment (temperature, salinity, light, nutrients, etc.) (Jeffrey et al. 1997). Hence, data from time series or transects should not automatically be lumped together. For example, samples coming from different salinities in an estuary will contain different communities with unique species compositions and thus unique biomarker ratios within the groups. In most cases, it is therefore advisable to only consider a few samples per run. This has the additional consequence that attempts to estimate biomarker ratios from field data should be discouraged, unless this is the goal of the study and sampling is performed accordingly. In that case, one may also include well-constrained priors for the taxonomical compositions using microscopy or other techniques. These priors are then used as input for the model, and the biomarker ratios are estimated from biomarker data and taxonomical compositions.

In general, we have more information on biomarker ratios than simple min-max bounds; prior information is available as mean ± standard deviation in particular for pigment or lipid composition of phytoplankton species (Ahlgren et al. 1990, 1992; Bourdier and Amblard 1987; Jeffrey et al. 1997; Reuss and Poulsen 2002; Vera et al. 2001; Volkman et al. 1989). One can formulate input probability distributions of biomarker ratios from literature or from experimental data. This approach avoids the identifiability problems that occur with uninformative priors and small sample sizes. At the same time, it makes good sense statistically. Instead of first pretending that no prior information is available and then adapting the prior distribution post-analysis to obtain an "acceptable" result (the approach adopted by CHEMTAX users), it is statistically more sound to use all available information prior to analysis. The result is then a probability distribution of the solution, reflecting the available statistical knowledge. If the information provided for the biomarker ratios is not compatible with the data, the maximum likelihood result will find a compromise between the probability of the ratio matrix and the probability of the data. The deviation of this maximum

likelihood solution from the prior biomarker ratios and the data can be used to verify post-hoc the proposed standard deviations in the prior probability distributions.

Comments and recommendations

There is one important challenge in the use of Bayesian inference, involving the Markov Chain Monte Carlo algorithm. This algorithm uses predefined probability distributions to find new random solutions. These jump probabilities determine the efficiency of the algorithm. It is often necessary, also for BCE, to adjust parameters for these distributions in order to limit calculation time. For most problems, we envisage that the default settings will be adequate; however, for applications that differ significantly from the default, we suggest the procedure proposed by Raftery and Lewis (1996) as a starting point.

The BCE has been developed to estimate microbial taxonomy based on pigment or lipid compositional data, or on a combination of both. However, the approach is generic and can be used for other compositional estimation problems as well. For instance, combined with isotope labeling of pigments or lipids, it can be used to estimate the growth rate of phytoplankton groups. Many organic contaminants are analyzed with chromatographic techniques and certain pollution sources have characteristic ratios. The BCE may prove useful to estimate the contribution of different potential pollution sources to aquatic systems or organisms. Nucleic acid–based fingerprinting methods such as T-RFLP and DGGE are used extensively in microbial ecology. Provided that the output patterns of these techniques can be linked quantitatively to the studied microbial communities and taxa, there might be a use for BCE in quantifying compositions of bacterial communities.

References

Ahlgren, G., I. B. Gustafsson, and M. Boberg. 1992. Fatty-acid content and chemical-composition of fresh-water microalgae. J. Phycol. 28:37-50.

———, L. Lundstedt, M. Brett, and C. Forsberg. 1990. Lipid-composition and food quality of some fresh-water phytoplankton for Cladoceran zooplankters. J. Plankton Res. 12:809-818.

Aitchison, J. 1986. The statistical analysis of compositional data. Blackburn.

Boschker, H. T. S., and J. J. Middelburg. 2002. Stable isotopes and biomarkers in microbial ecology. FEMS Microbiol. Ecol. 40:85-95.

Bourdier, G., and C. Amblard. 1987. Fatty-acid composition of lacustrian phytoplankton (Pavin Lake, France). Int. Rev. Ges. Hydrobiol. 72:81-95.

Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. Ecol. Lett. 8:2-14.

Damsté, J. S. S., and others. 2002. Linearly concatenated cyclobutane lipids form a dense bacterial membrane. Nature 419:708-712.

Dijkman, N. A., and J. C. Kromkamp. 2006. Phospholipid-derived fatty acids as chemotaxonomic markers for phytoplankton: application for inferring phytoplankton composition. Mar. Ecol. Prog. Ser. 324:113-125.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis, 2 ed. Chapman & Hall.

Gieskes, W. W. C., and G. W. Kraay. 1983. Dominance of Cryptophyceae during the phytoplankton spring bloom in the central north-sea detected by HPLC analysis of pigments. Mar. Biol. 75:179-185.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter [eds.]. 1996. Markov chain Monte Carlo in practice. Chapman & Hall.

Gillan, F. T., R. B. Johns, T. V. Verheyen, J. K. Volkman, and H. J. Bavor. 1981. Trans-monounsaturated acids in a marine bacterial isolate. Appl. Environ. Microbiol. 41:849-856.

Gotelli, N. J., and A. M. Ellison. 2004. A primer of ecological statistics. Sinauer Associates.

Jeffrey, S. W., R. F. C. Mantoura, and S. W. Wright. 1997. Phytoplankton pigments in oceanography: guidelines to modern methods. UNESCO Publishing.

Latasa, M. 2007. Improving estimations of phytoplankton class abundances using CHEMTAX. Mar. Ecol. Prog. Ser. 329:13-21.

Lawson, C. L., and R. Hanson, J. 1995. Solving least squares problems. Society for Industrial and Applied Mathematics.

Li, W. K. W., and others. 1983. Autotrophic picoplankton in the tropical ocean. Science 219:292-295.

Lionard, M., K. Muylaert, W. Vyverman, and D. Van Gansbeke. 2005. Influence of changes in salinity and light intensity on growth of phytoplankton communities from the Schelde river and estuary (Belgium/The Netherlands). Hydrobiologia 540:105-115.

Llewellyn, C. A., J. R. Fishwick, and J. C. Blackford. 2005. Phytoplankton community assemblage in the English Channel: a comparison using chlorophyll a derived from HPLC-CHEMTAX and carbon derived from microscopy cell counts. J. Plankton Res. 27:103-119.

Mackey, M. D., D. J. Mackey, H. W. Higgins, and S. W. Wright. 1996. CHEMTAX-A program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. Mar. Ecol. Prog. Ser. 144:265-283.

Muylaert, K., and others. 2006. Spatial variation in phytoplankton dynamics in the Belgian coastal zone of the North Sea studied by microscopy, HPLC-CHEMTAX and underway fluorescence recordings. J. Sea Res. 55:253-265.

Omlin, M., and P. Reichert. 1999. A comparison of techniques for the estimation of model prediction uncertainty. Ecol. Model. 115:45-59.

Platt, T., D. V. S. Rao, and B. Irwin. 1983. Photosynthesis of picoplankton in the oligotrophic ocean. Nature 301:702-704.

Raftery, A. E., and S. M. Lewis. 1996. Implementing MCMC, p. 115-130. *In* W. R. Gilks, S. Richardson, and D. J. Spiegelhalter [eds.], Markov Chain Monte Carlo in practice. Chapman & Hall.

Reuss, N., and L. K. Poulsen. 2002. Evaluation of fatty acids as biomarkers for a natural plankton community. A field study of a spring bloom and a post-bloom period off West Greenland. Mar. Biol. 141:423-434.

Roberts, G. O. 1996. Markov Chain concepts related to sampling algorithms, p. 45-58. *In* Gilks, W. R., S. Richardson, and D. J. Spiegelhalter [eds.], Markov chain Monte Carlo in practice. Chapman & Hall.

Rodriguez, F., M. Varela, and M. Zapata. 2002. Phytoplankton assemblages in the Gerlache and Bransfield Straits (Antarctic Peninsula) determined by light microscopy and CHEMTAX analysis of HPLC pigment data. Deep-Sea Res. II 49: 723-747.

Sargent, J. R., R. J. Parkes, I. Mueller-Harvey, and R. J. Henderson. 1987. Lipid biomarkers in marine ecology. *In* M. A. Sleigh [ed.], Microbes in the sea. Ellis Horwood Limited.

Tarantola, A. 2005. Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics.

Van Den Meersche, K., J. J. Middelburg, K. Soetaert, P. Van Rijswijk, H. T. S. Boschker, and C. H. R. Heip. 2004. Carbon-nitrogen coupling and algal-bacterial interactions during an experimental bloom: Modeling a C-13 tracer experiment. Limnol. Oceanogr. 49:862-878.

Vera, A., C. Desvilettes, A. Bec, and G. Bourdier. 2001. Fatty acid composition of freshwater heterotrophic flagellates: an experimental study. Aquat. Microb. Ecol. 25:271-279.

Volkman, J. K., S. W. Jeffrey, P. D. Nichols, G. I. Rogers, and C. D. Garland. 1989. Fatty acid and lipid composition of 10 species of microalgae used in mariculture. J. Exp. Mar. Biol. Ecol. 128:219-240.

Submitted 17 July 2007
Revised 16 October 2007
Accepted 13 November 2007