Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny

Knut Rudi,^{1,2} Monika Zimonja¹ and Tormod Næs¹

Correspondence

Knut Rudi

knut.rudi@matforsk.no

¹MATFORSK Norwegian Food Research Institute, Osloveien 1, 1430 Ås, Norway

²Hedmark University College, 2306 Hamar, Norway

Alignment-independent phylogenetic methods have interesting properties for global phylogenetic reconstructions, particularly with respect to speed and accuracy. Here, we present a novel multimer-based alignment-independent bilinear mathematical modelling (AIBIMM) approach for global 16S rRNA gene phylogenetic analyses. In AIBIMM, jackknife cross-validated principal component analyses (PCA) are used to explain the variance in nucleotide n -mer frequency data. We compared AIBIMM with alignment-based distance, maximum-parsimony and maximum-likelihood phylogenetic methods, analysing taxa belonging to the *Proteobacteria* ($n=82$), *Actinobacteria* ($n=30$) and *Archaea* ($n=7$). These analyses indicated an attraction between the *Actinobacteria* and *Archaea* for the traditional methods, with the two taxa *Acidimicrobium* and *Rubrobacter* at the root of the tree. AIBIMM, on the other hand, showed that the *Actinobacteria* was tightly clustered, with *Acidimicrobium* and *Rubrobacter* within a distinct subgroup of the *Actinobacteria*. The application of AIBIMM was further evaluated, analysing full-length 16S rRNA gene sequences for 2818 taxa representing the prokaryotic domains. We obtained a highly structured description of the prokaryote diversity. Sample-to-model (Si) distances were also determined for taxa included in our work. We determined Si distances for models of the six major subgroups of taxa detected in the global analyses, in addition to nested subgroups within the *Alphaproteobacteria*. The Si-distance evaluation showed a very good separation of the taxa within the models from those outside. We conclude that AIBIMM represents a novel phylogenetic framework suitable for accommodating the current exponential growth of 16S rRNA gene sequences in the public domain.

INTRODUCTION

The majority of evolutionary/taxonomic work is based on aligning DNA sequences and then reconstructing the phylogeny based on predefined evolutionary models (Ludwig *et al.*, 1998; Woese, 1987; Kolaczowski & Thornton, 2004). Alignment-based approaches, however, are not well suited to analyses of large (global) datasets. The main reason is that the computer operation time required to align DNA sequences by dynamic programming increases exponentially with the number of taxa analysed. These problems, however, can be partly overcome using empirical alignment scaffolds. However, it is still difficult to identify correct alignment scaffolds and to perform statistical testing of alignments (Ludwig & Schleifer, 1999). Alignment-independent methods are now gaining increasing popularity

in global phylogenetic analyses, mainly due to their speed and the statistical tests available for these methods (Vinga & Almeida, 2002).

We present a novel two-step phylogenetic approach based on alignment-independent bilinear multivariate modelling (AIBIMM) for global analysis of 16S rRNA gene phylogeny. The first step in AIBIMM is to transform DNA sequence data into DNA n -mer frequencies. The n -mer frequency data are obtained by sliding a window of size n (i.e. n nucleotides) and identifying the base combination present for each step. The number of contributions for each of the 4^n combinations is then counted. The second step is based on using these data as the input in principal component analysis (PCA) (Mardia *et al.*, 1979). PCA is a multivariate statistical method that visualizes the main structures in data with redundancies, and it is particularly well suited to building models that can be used for classification. We used a classification based on individual PCA decompositions of subsets of the data (Wold, 1976). An outline of the AIBIMM approach, as exemplified by a multimer window of $n=5$, is shown in Fig. 1.

Abbreviations: AIBIMM, alignment-independent bilinear mathematical modelling; PCA, principal component analysis.

Details of the computer program used for AIBIMM and a step-by-step explanation of the method of PCA modelling are available as supplementary material in IJSEM Online.

Fig. 1. Schematic outline of the AIBIMM approach for a multimer of $n=5$. (a) A window of five nucleotides is slid along the DNA sequence one nucleotide at the time. (b) The frequency of the pentamers within is counted and stored in a table. (c) Bilinear models are then constructed based on principal component analyses (PCA) of the pentamer frequency data. (d) The PCA models can be used for S_i distance classification. The distance from each taxon to the classification model (S_i) is calculated and presented as a frequency histogram of taxa. The cut-off is set between the taxa within the target group [exemplified by an ellipsoid in (c)] and those outside so that all the taxa used to build the model are included in the target group. The accuracy of the classification is then evaluated based on how many additional taxa are within the target group. (e) Finally, dendrograms (or phylogenetic trees) can be constructed from the PCA data in (c). Jackknife cross-validation can be used to evaluate the stability of both the PCA plot (c) and the trees constructed (e).

The major difference between AIBIMM and other multimer-based phylogenetic analyses (Vinga & Almeida, 2002) is that AIBIMM utilizes cross-validation for objective determination of the dimensionality in data containing structured information. The influence of non-homologous multimer equalities (multimers that are equal but do not have the same evolutionary origin) is filtered as long as it does not introduce structure. The basic properties of AIBIMM have been adapted from methods for texture analysis of images (Kvaal *et al.*, 1998), which also attempts to eliminate the information of position.

The aim of our work was to evaluate the application of AIBIMM by comparison with traditional phylogenetic methods and by global analyses of 16S rRNA gene sequences for taxa covering the bacterial and archaeal domains (Ludwig & Schleifer, 1999; Acinas *et al.*, 2004; Stackebrandt & Goebel, 1994; Suau *et al.*, 1999; Weisburg *et al.*, 1991). Our conclusion from the AIBIMM evaluation is that we obtained a good, statistically based separation of the taxa analysed and that AIBIMM has an advantage over traditional phylogenetic methods (distance, maximum-parsimony and maximum-likelihood) for large datasets and for sequences that are difficult to align. AIBIMM is thus promising for future global 16S rRNA gene analyses (Curtis & Sloan, 2004).

METHODS

Dataset and transformation. We based our analyses on the representative 16S rRNA gene sequences for prokaryotes from the RDP-II database (Maidak *et al.*, 1997). These sequences were filtered based on size (minimum size allowed was 1200 bp) and accuracy (maximum 20 ambiguous sites allowed). These criteria were set empirically to avoid the noisiest data in the database. We found that 2818 of the 3324 taxa in the database satisfied our filtering criteria.

The sequences were transformed into multimer frequencies ($n=1$ to 6) by the computer program PhyloMode (described in Appendix 1 available as supplementary material in IJSEM Online; http://www.matforsk.no/web/sampro.nsf/downloadE/Microbial_community), developed in-house. The sizes of the multimer windows were chosen as a trade-off between detecting phylogenetic signals (homologous multimer equalities), avoiding base composition biases due to non-homologous multimer equalities and the requirements for computer operation time.

A given pair of multimers can either be equal due to a common ancestor (the same evolutionary origin) or equal due to mutational events (equal multimers with different evolutionary origin). The probability (P_{equal}) for homologous multimers equalities (equal multimers with the same evolutionary origin) was calculated by the formula $P_{\text{equal}} = (1 - x)^n$, where x is the frequency of nucleotide substitutions and n is the multimer window size. Basically, this formula describes the relationship between sequence and multimer divergence. The probability of non-homologous multimer equalities (equal multimers with different evolutionary origin) for the variable positions (P_{nonhom}) was calculated by $P_{\text{nonhom}} = 1 - [(4^n - 1)/4^n]^{xl}$, where l is the length of the sequences compared, x is the frequency of nucleotide substitutions and n is the multimer window size. This formula describes the probability that a pair of different multimers becomes equal due to mutational events. Given a multimer of size n , there are $(4^n - 1)$ alternative multimers to the given multimer. The probability that a random mutation would not generate a multimer equal to the given multimer is thus $(4^n - 1)/4^n$. Given a series of

random mutations r , this probability is $[(4^n - 1)/4^n]^r$. In our case, the number of random mutations is given by multiplying the length of the DNA sequence l by the frequency of nucleotide substitutions x . Finally, the probability that any of the random mutations r would lead to a multimer equal to the given multimer is one minus the probability that they would not.

Bilinear multivariate modelling. Bilinear multivariate modelling methods (see e.g. Martens & Næs, 1989) are designed for situations with many variables and possibly strong collinearities (correlations) among them, which is the case for the nucleotide multimer data. Redundant information in the original variables (multimers) is used to identify new latent variables (principal components) that reflect the most important underlying ('latent') structure of the data. The principal components are projections of the original variables onto the space of most important variability. In our data, the variability is due to differences among homologous or non-homologous multimer equalities.

We used PCA with the NIPALS algorithm (Fisher & MacKenzie, 1923) for bilinear multivariate modelling (Mardia *et al.*, 1979). We used implementations from The Unscrambler software (CAMO Inc.) and the program PhyloMode developed in-house (see Appendix 1, available as supplementary material in IJSEM Online).

PCA models were tested using jackknife cross-validation (Unscrambler implementation) in order to determine the stability of the models and the dimensionality of the data (number of principal components needed to describe the redundant variance). This procedure is based on successively deleting one sample or a certain percentage of the observations from the data. The rest of the data are used to build the model. The model is then tested on the observations kept out of the computations, and the predicted residual variance is computed. This procedure is repeated until all samples have been deleted once. Finally, the total residual variance is determined by averaging the individual contributions from each segment. The square root of the residual predictive variance is the root mean square error of prediction (RMSEP). The maximum number of principal components needed to describe the redundant variance is determined from the RMSEP. The criterion is that the RMSEP should decrease for the principal components included. We used 281 segments with 10 samples per segment for the global analyses of all 2818 taxa, while full cross-validation was used for the rest of the analyses.

Principal component scores were plotted for visual inspection of the relationships between samples in the datasets. The so-called loadings, representing the contributions of the different original variables (multimers) on the final scores, were also plotted to determine which multimers are important for the principal components. A detailed description of the steps in the PCA analysis is given in Appendix 2, available as supplementary material in IJSEM Online.

Classification by distance to PCA models. In order to validate the group structures revealed by the PCA, a separate centred PCA model was built for each group identified to describe the main group structures. We used all taxa within the respective groups in constructing the PCA model. The reason for this was to obtain the best possible description of the group. Since the data have already been filtered for low-quality sequences, there will be no danger of overfitting the model to errors rather than true phylogenetic signals. The optimal number of principal components was selected separately for each model (according to the cross-validation results). The sample-to-model distances for the taxa to be classified (the so-called S_i value) were computed. For each model, S_i cut-off values were set so that they included all the taxa for which the model was built. The accuracy of the classification was then evaluated based on how many additional taxa had S_i values lower than the cut-off value. Note that this strategy has strong similarities to the SIMCA method (Wold, 1976) for classification.

AIBIMM tree construction. Trees were constructed to visualize the hierarchical structure in the AIBIMM data. The PCA coordinates were used as input for the cluster analyses. We used Euclidean distances since these represent the true distance between objects in orthogonal space.

We used centroid, single, complete and median linkage in the tree (dendrogram) construction. Statistical evaluation of tree stability was done using each of the cross-validation PCA coordinate data as the input for construction of separate trees. The branch frequencies (i.e. the number of times a given branch occurs among the jackknife trees) were subsequently used as stability indicators.

Traditional phylogenetic reconstruction. DNA sequences were aligned by the alignment software CLUSTAL W (Thompson *et al.*, 1997). The alignments were then corrected manually using the program BioEdit (Hall, 1999). Prealigned sequences (Maidak *et al.*, 1997) were used in addition to our own alignments.

Phylogenetic distance trees were constructed using Tamura–Nei distances and the minimum-evolution algorithm provided in the MEGA 3 software package (Kumar *et al.*, 2001). The interior branch test with 100 replicates was used for significance testing. Maximum-parsimony trees were also constructed using MEGA 3 with close-neighbour interchange (CNI) and bootstrap statistical testing (100 replicates). Maximum-likelihood analyses were run on the server <http://bioweb.pasteur.fr> using quartet puzzling and maximum-likelihood analyses for tree construction (Strimmer & von Haeseler, 1996).

RESULTS

Comparison of AIBIMM with traditional phylogenetic reconstruction

A subset ($n=119$) of the total data included in this study ($n=2818$) was analysed in a comparison between AIBIMM and traditional phylogenetic methods. We analysed taxa belonging to the *Proteobacteria* ($n=82$), *Actinobacteria* ($n=30$) and *Archaea* ($n=7$). These taxa were chosen because proteobacteria generally have lower G + C content than actinobacteria, while the outgroup taxa from the *Archaea* have high G + C content. Specifically, we included taxa from the two deep-branching subclasses, *Acidimicrobiidae* and *Rubrobacteridae*, within the *Actinobacteria*. These subclasses have previously been placed on the outskirts of the *Clostridia* by traditional phylogenetic reconstruction (Garritty *et al.*, 2003).

We used prealigned sequences from the RDP-II database in our traditional phylogenetic reconstruction (Maidak *et al.*, 1997). Our own alignment with CLUSTAL W (Thompson *et al.*, 1997) was clearly not correct (results not shown). The difficulties in aligning the sequences are probably due to a high degree of divergence in the variable regions of 16S rRNA for the taxa analysed.

The AIBIMM tree (Fig. 2a) showed tight clustering, within a distinct subgroup of the *Actinobacteria*, for taxa belonging to *Rubrobacter* (*Rubrobacteridae*) and *Acidimicrobium* (*Acidimicrobiidae*). The distance (Fig. 2b), maximum-parsimony (Fig. 2c) and maximum-likelihood (Fig. 2d) methods, on the other hand, showed an unresolved deep branching of the *Actinobacteria*. The *Actinobacteria* appeared more closely

Fig. 2. Comparison of AIBIMM with traditional phylogenetic methods. Only topologies where there are major differences between the methods are shown. The other branches presented as unresolved. Reconstructions were done using AIBIMM with a multimer window $n=6$ (a), distance methods (b), maximum-parsimony methods (c) and maximum-likelihood methods (d). In (a), jackknife cross-validation support (percentage of 20 trees) is shown for each branch in the AIBIMM tree constructed from centroid-coordinate linkage (identical tree topologies were also obtained with single, complete and median linkage; not shown). In (b), numbers at nodes for the distance tree show the interior branch confidence as a percentage, while the bar shows the genetic distance. In (c), the frequency of bootstrap trees supporting a given branch in the tree is shown for the maximum-parsimony tree, while the bar indicates genetic steps. Finally, in (d), no jackknife or bootstrap supports were determined for the maximum-likelihood tree due to the extensive computer operation time required for these analyses.

related to the *Archaea* than to the *Proteobacteria* in these trees. In particular, the placements of *Rubrobacter* and *Acidimicrobium* were inconsistent.

Evaluation of AIBIMM multimer window size

We conducted a theoretical evaluation of the multimer window size $n=1$ to 6 with respect to the probability for homologous and non-homologous multimer equalities. We evaluated the contribution of these two forms of multimer equality given different sequence divergence (Fig. 3). We determined that the probability of non-homologous multimer equalities increased with decreasing multimer window size and increasing divergence. The probability of homologous multimer equalities, on the other hand, decreased with increasing multimer size and divergence.

We analysed the entire dataset, consisting of 2818 taxa, empirically using nucleotide multimer windows from $n=1$ to 5. We did not evaluate hexamers ($n=6$) for the global data because these analyses were computationally too demanding. Basically, what was determined empirically was that the shorter multimers, $n < 5$, revealed a structuring related to the 16S rRNA gene G + C content. For instance, *Archaeoglobus fulgidus* (*Archaea*) and *Slackia exigua* (*Actinobacteria*) both have G-homopolymer tetramer frequencies of 27. These two taxa clustered tightly together using AIBIMM for the tetramer frequency data (results not shown). This clustering, however, is clearly not phylogenetically relevant, but rather a result of non-homologous multimer equalities. The $n=5$ multimer, on the other hand, showed a main structuring in accordance with the

phylogeny of the taxa (Garrrity *et al.*, 2003), indicating structuring according to homologous multimer equalities. A window size of $n=5$ (pentamer) was therefore used through the remainder of this work in evaluating AIBIMM for the global dataset.

Global AIBIMM model

The global data are presented as PCA plots because of the difficulties of representing large sets of data as trees. We determined a global AIBIMM model based on nucleotide pentamers for all 2818 taxa included in this work (Fig. 4a). The first three principal components explained 23 % of the variance in the data. The variance explained by cross-validation was almost identical to the modelled variance, strongly supporting the validity of the AIBIMM model. The model separated the *Archaea* and *Bacteria* as two major groups. The *Bacteria* were further separated into *Firmicutes*, *Actinobacteria*, *Alpha-/Delta-/Epsilonproteobacteria*, *Beta-/Gammaproteobacteria* and a miscellaneous unseparated group, as indicated by the ellipsoids.

Basically, the archaea were separated along the third principal component (z-axis), the actinobacteria along the first component (x-axis) and the rest of the bacteria along the second component (y-axis). The variance for the miscellaneous group, however, was not well explained by the first three principal components.

There was low or no structure (different clusters) of the loading for the pentamer variables in the global AIBIMM model (Fig. 4b). This is in stark contrast to the high structure found in the score plot (Fig. 4a). There are therefore

Fig. 3. Evaluation of multimer window size with respect to probability of homologous (●) and non-homologous (▲) multimer equalities. The evaluation was based on a 1300 bp sequence with 0 to 20% divergence. We evaluated window sizes of $n=1$ to $n=6$ (a–f).

no particular structures for the pentamers that are more important than others for explaining the variance in the score plot. This suggests that the structures are mainly explained by homologous multimer equalities, since non-homologous multimer equalities would result in structures in the loading plot such as, for example, gradients of G + C content.

AIBIMM models for the major subgroups

The support for the six subgroups presented in Fig. 3 was also evaluated by Si distance classification (Fig. 5). The classification based on Si showed that all the models, except the model for the miscellaneous group (Fig. 5b), gave a good separation of the target and non-target taxa.

The *Actinobacteria* and the *Euryarchaeota/Crenarchaeota* were the best-separated groups, with 99.2 and 99.0%, respectively, of taxa classified correctly within the Si cut-off distance. In contrast, only 3.6% of the taxa within the cut-off Si distance for the miscellaneous group were classified correctly. Thus, this group is clearly not monophyletic. The group probably contains several different evolutionary lineages. The conclusion from the global Si distance classification, however, is that there are phylogenetic signals in the DNA sequence data that enable the separation of the major bacterial groups.

The AIBIMM models used for classification based on Si distances also showed structuring of the taxa within each subgroup (Fig. 6). The variance explained by cross-validation, compared with the modelled variance, was significant to at least the eighth principal component for all the subgroups, except in the miscellaneous group (this group was significant until the fifth principal component). This further supports the conclusion that the data are highly structured. The *Euryarchaeota/Crenarchaeota* were separated into three tightly clustered groups with x , y centroid coordinates of (−14, 12), (−17, −14) and (20, 0), respectively (Fig. 6a). The first group contained methanogenic taxa, the second halophilic taxa and the third thermophilic taxa. There was no clear separation of *Euryarchaeota* and *Crenarchaeota* for the thermophilic taxa. These groups seemed continuous with the highest score for the first principal component for the *Crenarchaeota*. The loadings, however, showed that there has been a selection within *Crenarchaeota* for structures contained in the pentamer 5′-GGGGG-3′. The *Aeropyrum pernix* strains are the most extreme, with a mean frequency of 20 for these G-tract pentamers, while the mean value for the entire dataset of bacteria and archaea is 3.

The miscellaneous group (Fig. 6b) contains bacteria belonging to the genus *Prevotella* for taxa with positive scores for both the first and the second principal components. There was no or low structure, however, for the other taxa. The

Fig. 4. Global AIBIMM model for all prokaryotes. A cross-validated model was generated as described in Methods. (a) Score plot for the first three principal components, respectively explaining 12, 6 and 5% of the variance. The major groups are indicated in this plot. (b) Corresponding loading plot for the variables (nucleotide pentamers).

scores for these taxa were also unstable with respect to cross-validation. These are seen as relatively large errors in the cross-validation results. The reason for the instability is probably the lack of redundant information for proper classification due to the inclusion of relatively few taxa for each group.

There are relatively few tight clusters of taxa within the *Firmicutes* (Fig. 6c). The cross-validation results, however, were stable, with low errors, showing that redundant data gave the same pattern. It is therefore likely that this pattern is due to the evolution of the *Firmicutes* and not instability or uncertainty of the model. The *Staphylococcaceae*,

however, formed a tight cluster with centroid coordinates (−15, −2.5). The *Mycoplasmataceae* also formed a relatively distinct cluster of taxa (centroid coordinates 2.3, −14). There was also an apparent structuring of aerotolerant taxa (negative score for first principal component) and anaerobic taxa (positive score for first principal component).

The *Alpha-/Delta-/Epsilonproteobacteria* are clearly divided into three separate clusters (Fig. 6d). The *Alphaproteobacteria* consists of the largest group, with 227 taxa, with centroid coordinates of (3.5, 0). The *Epsilonproteobacteria* gave the most negative score for both principal components

Fig. 5. Si distance classification for the major prokaryotic groups. Si distance classifications were done for all 2818 taxa analysed based on the PCA models for the groups marked in Fig. 3. A histogram is shown for the $\log_{10}$ frequency distribution (a frequency of 0 is coded as 1 to allow the $\log_{10}$ transformation) of taxa relative to the sample-to-model distance (Si). Dashed lines represent thresholds for whether the taxa are within or outside the target group. Percentages represent the actual fraction of taxa for which the model was built. The number of principal components used to construct the model was determined from the cross-validation data. The numbers of principal components used were 8, 5, 12, 12, 14 and 10, respectively, for the models in (a)–(f).

(centroid coordinates $-22, -12.5$). The *Deltaproteobacteria* gave a negative score for the first principal component and a positive score for the second (centroid coordinates $-9.5, 12$).

The *Beta-/Gammaproteobacteria* were divided in several clusters (Fig. 6e). The cluster of the *Betaproteobacteria* gave the highest negative score for both principal components (centroid coordinates $-12, -10$). The *Gammaproteobacteria*, on the other hand, was separated into several different clusters. This class is relatively diverse, with centroid coordinates ranging from *Pseudomonas* ($-9, 12.5$) to the *Enterobacteriaceae* ($12, -6$).

The structures within the *Actinobacteria* resemble those of the *Firmicutes*, with a few tight clusters (Fig. 6f). The only

relatively distinct clusters identified were corynebacteria and mycobacteria, with respective centroid coordinates of $(-10, -5)$ and $(-11.5, 6)$. There was also a separation of two major clusters along the first principal component. Both corynebacteria and mycobacteria gave negative scores, while there was a wide diversity of taxa giving positive scores.

Nested AIBIMM approach

We analysed the 227 taxa belonging to the *Alphaproteobacteria* (outlined in Fig. 6d) as an example of the nested AIBIMM approach. This analysis revealed four main clusters (Fig. 7a). The tight cluster with centroid coordinates $(-2, -9)$ consisting of 63 taxa belonging to the *Rhizobiales* was further analysed by the nested approach. When

Fig. 6. Separate AIBIMM models for the major prokaryotic groups. Separate AIBIMM models were constructed for each of the groups evaluated by Si distance classification in Fig. 3. The major groups are marked with dashed ellipses. The explained variances (%) for principal components 1 and 2 were 30 and 9, 16 and 9, 10 and 7, 13 and 7, 13 and 8, and 11 and 9, respectively, for (a)–(f).

analysing this group alone, new cluster patterns were revealed (Fig. 7c). A cluster of 17 taxa belonging to the genus *Bartonella* (centroid coordinates 11.7, 2.1) was further analysed using the nested approach. Analyses of this genus alone gave three new clusters (Fig. 7e). Two of these clusters consisted of single species, *Bartonella bacilliformis* (centroid coordinates -10, 1.5) and *Bartonella quintana* (centroid coordinates 0, -10). The last group,

with centroid coordinates (2.5, 1.5), was more diverse and consisted of several species. However, as evaluated by the cross-validation results, further separation of the taxa was not possible (results not shown). Finally, the Si distance classification revealed that the different PCA models gave excellent separation of the taxa within the nested models from those outside. There were no misclassifications in any of the models for the 2818 taxa evaluated (Fig. 7b, d, f).

Fig. 7. Nested AIBIMM approach. In (a), the *Alphaproteobacteria* group indicated in Fig. 6d was analysed separately by PCA. In (c), the *Rhizobiales* group marked in (a) was analysed by the nested approach. Finally, in (e), the *Bartonella* group marked in (c) was analysed by PCA. The S_i distance classification for the groups marked with thick dashed ellipses is shown in (b), (d) and (f) (frequencies of 0 coded as 1 to allow the $\log_{10}$ transformation). The explained variances (%) for principal components 1 and 2 were 11 and 11, 24 and 13, 30 and 18 for (a), (c) and (e). The numbers of principal components used for the S_i distance classification were 9, 2 and 1, respectively, in (b), (d) and (f).

There was also structuring for the taxa that were not used to build the models. The peak of taxa closest to the *Rhizobiales* model (Fig. 7b) represents the other alphaproteobacteria and the more distant major peak represents the other bacteria, while the most distant taxa belong to the *Archaea*. The same pattern was observed for the *Bartonella* model, with an additional peak closest to the model, which are the taxa belonging to the *Rhizobiales* (Fig. 7d). The pattern was even conserved for a model based on a subgroup of the *Bartonella*, with the other *Bartonella* taxa closest to the model (Fig. 7f).

DISCUSSION

Resolution of phylogenetic clusters by AIBIMM

Our results showed that there are structures common to the *Actinobacteria* which were revealed in the AIBIMM tree (Fig. 2a), but not by traditional phylogenetic methods (Fig. 2b, c). These differences could be due to over-estimation of the structure by AIBIMM and/or lack of resolution of the traditional methods. However, since we used a multimer window of $n=6$ in this comparison, the

probability of non-homologous multimer equalities should be low (see Fig. 3f). Thus, it is not likely that the structures determined by AIBIMM are due to non-homologous multimer equalities. Furthermore, non-homologous multimer elements were not identified in the loading plot.

A challenge with alignments is the analysis of positions containing noise, such as alignment errors, or positions that have lost phylogenetic signal due to mutation saturation. Such positions lower the resolution of the deep branches in phylogenetic trees. Evolutionary models (Kolaczowski & Thornton, 2004; Felsenstein, 1981; Hillis & Huelsenbeck, 1992; Holder & Lewis, 2003; Huelsenbeck *et al.*, 2001; Brochier & Philippe, 2002) or separate analyses of conserved regions (Ludwig *et al.*, 1998; Brochier & Philippe, 2002) have been used to increase the resolution.

The distinct clustering of the *Actinobacteria* detected by AIBIMM is an interesting example with respect to analyses of deep phylogenetic branches. The AIBIMM results showed that the *Actinobacteria* is a relatively tightly clustered group separated along the first principal component in our global model. All other bacteria were separated along the second principal component (see Fig. 3). Thus, the *Actinobacteria* are distinct among the bacteria according to our AIBIMM model. In contrast, a recent state-of-the-art, alignment-based phylogenetic reconstruction placed the *Actinobacteria* in a divergent unresolved group between *Beta-/Gammaproteobacteria* and *Alpha-/Delta-/Epsilonproteobacteria* (Acinas *et al.*, 2004), which is not in accordance with standard taxonomy (Garrity *et al.*, 2003). The reason is probably the lack of resolution in the phylogenetic methods applied (Kolaczowski & Thornton, 2004).

AIBIMM and traditional phylogenetic methods have different fields of application. Given a correct alignment, relatively few taxa and a known evolutionary model, alignment-based methods represent the most direct phylogenetic reconstruction. These methods try to reconstruct the actual evolutionary steps.

The advantage of AIBIMM, however, is the possibility to analyse large global sets of data. The global analysis of the *Actinobacteria*, for instance, indicated that this bacterial group is distinct from other bacteria. Such global structures are difficult to reveal using alignment-based phylogenetic methods. As we see it, the basic problem with current phylogenetic approaches is that phylogenetic signals can not be separated from signals due to incorrect alignments or mutation saturation (noise) in global sets of data. These problems, however, are addressed in AIBIMM by using cross-validated PCA analyses. The importance of using PCA here lies primarily in its ability to compress information onto the most important dimensions of variability in the dataset and, in this way, to focus only on valid common structures in the taxa considered. The more subtle parts of the genetic information are filtered by the jackknife cross-validation, such as unstructured data due to non-homologous multimer equalities. As a consequence,

the combination of cross-validated PCA with an optimal multimer window size selects parts of the data that contain phylogenetic structure.

Size effect of nucleotide multimer window

The optimal size of the nucleotide multimers applied to the AIBIMM transformation is dependent on the variance (length and divergence) of the sequences. Theoretical considerations of equalities due to homologous or non-homologous multimers can be used in selecting an optimal window size (see Fig. 3).

For instance, for window size $n=4$, the crossing-point at approximately 15 % divergence between the probability of homologous and non-homologous multimer equalities (Fig. 3d) indicates that non-homologous multimers can introduce structure in the $n=4$ data. Empirically, this can be seen as structures such as G+C content become predominant in the variance among the taxa. The effect is even more pronounced in the empirical data for the shorter multimers ($n=1$ to 3). Theoretically, hexamer analyses give the best separation between homologous and non-homologous multimer equalities (see Fig. 3f). There are, however, $4^6 = 4096$ combinatorial possibilities of hexamers, making these analyses less efficient with respect to computer operation time when analysing large sets of data.

The multimer window of $n=5$ is quite efficient with respect to computer operation time, and the probability of homologous multimer equalities is higher than for non-homologous equalities for sequences with less than approximately 25 % divergence. Comparison between the empirical AIBIMM data and prior knowledge about the relatedness between the taxa analysed indicates that homologous multimer equalities dominate the structure for the multimer window of $n=5$. These are the reasons why $n=5$ was chosen for the global analyses in this work.

Non-homologous multimer structures

It is not possible to remove all effects of non-homologous multimers in the AIBIMM analyses. However, if the non-homologous structure is known, such as G+C content, it can be avoided in the AIBIMM analyses. It is possible to evaluate whether the loading for a given principal component follows for example a multimer G+C gradient. If the loading follows a gradient, the given principal component can be excluded from the AIBIMM analysis in order to obtain a more accurate phylogenetic description.

Codon usage biases in protein-encoding genes are another example of non-homologous multimer structure that could represent a particular problem with respect to AIBIMM. This problem, however, could probably partly be avoided by omitting the third, wobble position in the AIBIMM analyses. The resulting dinucleotides should then be used as the unit in the multimer analyses (analysing windows of $2n$). The principal component loadings can also be

evaluated with respect to codon-usage gradients as for the G + C content biases described above.

A novel coordinate-based AIBIMM classification system

Our view is that it would better to classify bacteria according to an AIBIMM coordinate system rather than by traditional alignment- and/or tree-based systems (Garritty *et al.*, 2003). Coordinate-based classification systems are better suited for storage and retrieval from databases than are tree-based systems. AIBIMM also enables a more objective and statistically more soundly based classification, since alignments are avoided. Furthermore, the classification system should be adaptable to very large sets of data. There are, for instance, no technical barriers to implementing such classification systems that accommodate the 1 to 2 million bacterial species assumed to exist on Earth (Curtis & Sloan, 2004).

ACKNOWLEDGEMENTS

Basic finances from MATFORSK Norwegian Food Research Institute and Hedmark Sparebank supported this work. We thank Ragnhild Norang for carefully reading and commenting on the paper.

REFERENCES

- Acinas, S. G., Klepac-Ceraj, V., Hunt, D. E., Pharino, C., Ceraj, I., Distel, D. L. & Polz, M. F. (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554.
- Brochier, C. & Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417**, 244.
- Curtis, T. P. & Sloan, W. T. (2004). Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Curr Opin Microbiol* **7**, 221–226.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376.
- Fisher, R. A. & MacKenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *J Agric Sci* **13**, 311–320.
- Garritty, G. M., Bell, J. A. & Lilburn, T. G. (2003). *Taxonomic outline of the prokaryotes*, release 4. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn. New York: Springer. doi:10.1007/bergeysoutline
- Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.
- Hillis, D. M. & Huelsenbeck, J. P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *J Hered* **83**, 189–195.
- Holder, M. & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4**, 275–284.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.
- Kolaczekowski, B. & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984.
- Kumar, S., Tamura, K., Jakobsen, I.-B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.
- Kvaal, K., Wold, J. P., Indahl, U. G., Baardseth, P. & Næs, T. (1998). Multivariate feature extraction from textural images of bread. *Chemometrics Intell Lab Syst* **42**, 141–158.
- Ludwig, W. & Schleifer, K.-H. (1999). Phylogeny of bacteria beyond the 16S rRNA standard. *ASM News* **65**, 752–757.
- Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M. & Schleifer, K. H. (1998). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568.
- Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J. & Woese, C. R. (1997). The RDP (Ribosomal Database Project). *Nucleic Acids Res* **25**, 109–111.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.
- Martens, H. & Næs, T. (1989). *Multivariate Calibration*. Chichester: Wiley.
- Stackebrandt, E. & Goebel, B. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.
- Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13**, 964–969.
- Suau, A., Bonnet, R., Sutren, M., Godon, J. J., Gibson, G. R., Collins, M. D. & Dore, J. (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65**, 4799–4807.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.
- Vinga, S. & Almeida, J. (2002). Alignment-free sequence comparison – a review. *Bioinformatics* **19**, 513–523.
- Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* **173**, 697–703.
- Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51**, 221–271.
- Wold, S. (1976). Pattern recognition by means of disjoint principal component models. *Pattern Recognit* **8**, 127–139.