BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2235–2241
doi:10.1093/bioinformatics/btq349

Sequence analysis

Advance Access publication July 12, 2010

Analyzing taxonomic classification using extensible Markov models

Rao M. Kotamarti[1,*], Michael Hahsler[1], Douglas Raiford[2], Monnie McGee[3] and Margaret H. Dunham[1,*]

[1]Department of Computer Science and Engineering, [2]Department of Computer Science, University of Montana, MT 59812 and [3]Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: As next generation sequencing is rapidly adding new genomes, their correct placement in the taxonomy needs verification. However, the current methods for confirming classification of a taxon or suggesting revision for a potential misplacement relies on computationally intense multi-sequence alignment followed by an iterative adjustment of the distance matrix. Due to intra-heterogeneity issues with the 16S rRNA marker, no classifier is available for sub-genus level, which could readily suggest a classification for a novel 16S rRNA sequence. Metagenomics further complicates the issue by generating fragmented 16S rRNA sequences. This article proposes a novel alignment-free method for representing the microbial profiles using extensible Markov models (EMMs) with an extended Karlin–Altschul statistical framework similar to the classic alignment paradigm. We propose a log odds (LODs) score classifier based on Gumbel difference distribution that confirms correct classifications with statistical significance qualifications and suggests revisions where necessary.

Results: We tested our method by generating a sub-genus level classifier with which we re-evaluated classifications of 676 microbial organisms using the NCBI FTP database for the 16S rRNA. The results confirm current classification for all genera while ascertaining significance at 95%. Furthermore, this novel classifier isolates heterogeneity issues to a mere 12 strains while confirming classifications with significance qualification for the remaining 98%. The models require less memory than that needed by multi-sequence alignments and have better time complexity than the current methods. The classifier operates at sub-genus level, and thus outperforms the naive Bayes classifier of the RNA Database Project where much of the taxonomic analysis is available online. Finally, using information redundancy in model building, we show that the method applies to metagenomic fragment classification of 19 *Escherichia coli* strains.

Availability and implementation: Source code and binaries freely available for download at http://lyle.smu.edu/IDA/EMMSA/, implemented in JAVA and supported on MS Windows.

Contact: mallik@kotamarti.com; mhd@lyle.smu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 29, 2010; revised on June 6, 2010; accepted on June 25, 2010

*To whom correspondence should be addressed.

1 INTRODUCTION

Many different techniques have been proposed for microbial classification. Most current approaches for classification are based on similarity between the organism to be classified and known correctly classified organisms. This similarity may be based on cell shape or structure, biochemical characteristics, G + C content, nucleic acid hybridization, etc. More recent techniques have been based on molecular sequence comparisons. Lilburn *et al.* (2006) presents an excellent review of available tools for microbial classification and emphasizes that currently available tools are not satisfactory. He argues for more algorithmic approaches.

Due to the ubiquitous nature of 16S rRNA, it is often used as the sequence for comparison. Sub-genus classification is considered a challenge due to lack of resolution if only the 16S rRNA is used. Use of alternate gene markers and whole genomes have been suggested (Case *et al.*, 2007; Dahllöf *et al.*, 2000), but 16S rRNA still retains its position as the most commonly used marker (Janda and Abbott, 2007). In fact, we show that only a small percentage of organisms or strains (2%) require markers and methods outside of the 16S rRNA.

A common characteristic of most sequence-based classification techniques is the requirement for sequence alignment that is extremely computationally expensive. Initial taxonomy evaluations were done with heat-maps generated using distance matrices from multiple sequence alignments (Lilburn and Garrity, 2004). The alignments and tools have been recently improved using RNA secondary structures and probabilistic models in Cole *et al.* (2009); however, reliance on subjective selection of gene copies and expensive sequence alignment may be avoided with alternative methods. This is exactly what we do in this article. Our approach uses all copies of 16S and completely avoids alignment while still achieving high classification accuracy at the species level.

Statistical signatures (Vinga and Almeida, 2003) created from base composition frequencies offer an alternative to using classic alignment. These alignment free methods reduce complexity and processing. Lempel-Ziv (LZ) complexity (Otu and Sayood, 2003) and EMM-based distance measures (Kotamarti *et al.*, 2010) look promising for whole genome phylogenies as the current methods do not scale well. However, it is desirable to avoid a resource intense *all against all distance measurement* in a fast growing taxonomy. Hidden Markov model-based profiles offer a solid probabilistic basis (Eddy, 1998) and their application is well tailored to DNA or protein analyses. However, the base model needed to be extended for RNA. The later extensions account for co-variability of base pairs in the RNA sequences due to Watson–Crick complementarity

(Eddy and Durbin, 1994). Eddy notes that the covariance models are more suited to small length RNAs (e.g. tRNAs) and as such for homologous search problems.

This article proposes the use of extensible Markov models (EMMs; Dunham *et al.*, 2004) to create compact classification models for sequence analysis as a more efficient alternative to any previous sequence-based technique. Use of machine learning allows EMM signatures to be automatically built. By completely avoiding alignment, the efficiency is greatly improved. The resulting EMM models (one per class) provide a compressed representative (signature) of the class.

EMM (Dunham *et al.*, 2004) is a time-varying Markov chain, which can be viewed as a directed graph with nodes representing clusters of real-world events and arcs representing the ordering of the associated events. EMMs have been used to model many different applications in a variety of fields including future state prediction (Meng and Dunham, 2006) and rare event detection (Meng *et al.*, 2006). As a bioinformatic adaptation of an EMM, EMM bioinformatic analysis (EMMBA) transforms m molecular sequences to a single EMM signature of m' states. It can be considered a representation of sequence data with states representing clusters of similar sequence segments and inter-state transition probabilities representing the implicit order within the sequences.

The salient features of our classification approach include:

- High classification accuracy [based on Bergey's Manual (Garrity *et al.*, 2005)] even at the species level.
- Creation of a new EMM to create Profile graphs (Kotamarti and Dunham, 2010) for each species using all of its 16S rRNA copies.
- No alignment of sequences is required. A quasi alignment is obtained as a direct byproduct of the EMM graphs themselves.

The quasi-alignment technique also makes EMMs suitable for classification using metagenomic sequences.

The rest of this article is organized as follows: Section 2 derives an equivalent Markov formulation of sequence learning. Section 3 describes the method used to confirm or question the classification of an organism and introduces an extension to the Karlin–Altschul statistics in order to incorporate a log-odds (LODs) score matrix to aid in statistical defense of a classification. In Section 4, 10-fold cross-validation tests, comparison to RNA database project (RDP) classifier and sub-genus classification results are presented to illustrate the performance of EMM. Finally, Section 5 presents future direction for EMM and concludes the article.

2 MODEL CONSTRUCTION

The occurrences of letters or base compositions *{A, C, U, G}* of an RNA sequence provide frequency information. The occurrences of all patterns of bases of length l generates an l-mer frequency representation for a sequence (Vinga and Almeida, 2003).

We use the notation $F(S)$ to represent a transformation function acting on m 16S rRNA sequences of an organism and $G=(V,E)$ representing a directed graph of V nodes and E edges. The vertices are also referred to as nodes and states of the EMM graph to improve readability. Mathematically, EMM generation can be expressed as

$$G' = G \biguplus F(S) \tag{1}$$

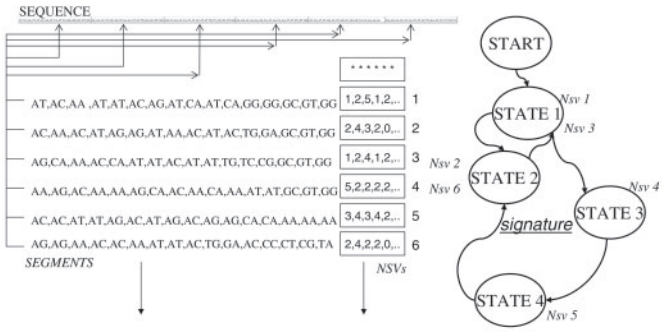

Fig. 1. The model building process. NSVs constitute the numerical representations of equal-sized segments along a 16S sequence, which are used in building an EMM signature. Signature building starts with a start state; as each NSV is processed, it is compared to the existing states of the model. If the NSV is not found to be close enough (per a squared Euclidean threshold) as in the case of NSV 1, a new state (1) is created with the new NSV as its first cluster member; otherwise, the new NSV (as in the case of NSV 3) is simply added to the matching cluster state node (state 1). When all NSVs are processed, the model building process is finished.

Where $\biguplus$ is the operator to integrate new sequence segments into a model being built. $F(S)$ is further expressed in terms of nested functions F' and F^* as follows:

$$F(S) = <F'(S_1), F'(S_2), ..., F'(S_m)> \tag{2}$$

$$F'(S_i) = <F^*(s_{(i,1)}), F^*(s_{(i,2)}), ..., F^*(s_{(i,k)})> \tag{3}$$

$$F^*(s_{(j,k)}) = <v_{(j,k,1)}, v_{(j,k,2)}, ..., v_{(j,k,n)}> \tag{4}$$

In Equations (2)–(4), the functions F, F' and F^* are numerical summarization functions that convert molecular sequences to oligomer count form. The function F collects m 16S rRNA sequences S being transformed and applies the F' transformation function. The function F' converts a single sequence to k equal-sized segments and applies the function F^* on each segment S_i. The function F^* converts a sequence segment to a vector v of counts with each count representing one of the oligomer variants. Such vectors are referred to as Numerical Summarization Vectors (NSVs) and the size of an NSV is given by 4^l, where l is oligomer length.

Finally, $\biguplus$ extends the directed graph, i.e. *sequence signature* as shown in Figure 1, by clustering each NSV generated by the F function into a new or an existing node of the graph and appropriately updating the arc information. Initially, the signature graph is *empty*. A squared Euclidean distance function *dist* and a threshold τ are utilized to determine clustering of NSVs in a node. The function $\biguplus$ can be further expressed by the following algorithm:

given that the V vertices of EMM graph also represent the nodes of clusters of similar NSVs and with V_c representing the current state, for each NSV v_i in $F(S)$,

(1) Find the closest match, i.e., the nearest node V_j to the NSV v_i.

(2) If $dist(V_j, v_i) > \tau$,

  (a) add a new node $V_i = v_i$,

  (b) add a new arc from V_c to V_i and

  (c) update EMM current state $V_c = V_i$, else

(3) Add NSV v_i to the node $V_j = V_j \cup \{v_i\}$, update the frequency of the arc V_c to V_j and set the EMM current state as $V_c = V_j$.

where the closest match, referred to as *quasi alignment*, which is defined to exist when the node whose centroid is at a minimal distance from the NSV v_i is found. The threshold τ is selected such that segments which are very similar are clustered together. The current state of an EMM graph is always the last-matched node. The current state changes when a new NSV matches to another state or results in creation of a new state. An arc is added/updated whenever the current state changes. Adding an arc involves updating the arc probability. The state resets to the start state for every new sequence.

Figure 1 shows the EMM build operation graphically. In this section, we formulated the EMM build operation as a transformation problem, which is suitable for compressing sequences into a signature. The transformation function $\biguplus$ is a compression function. Considering m sequences of k segments each, there are mk data points. Suppose an EMM is built with L states. Due to clustering, $mk \leq L$. Typically, we choose clustering criterion such that $mk \ll L$, which results in compression.

Aggregation of EMMs into a single EMM called *profileEMM* is how the signatures for higher taxa such as *genus* or *class* are created. For each EMM that is to be aggregated to a profileEMM, the centroids of nodes, which are simply the average vectors of the NSVs assigned to them, are used as input NSVs for the model building. For example, if a species EMM has 20 states, there will be 20 average vectors used as NSVs for building a higher level, i.e. genus model.

3 METHODS

In recent decades, several effective taxonomic analysis techniques have been proposed and are in use. Lilburn and Garrity, 2004 present a heat-map visualization technique that examines and corrects for any inconsistencies in the microbial taxonomy. The process has been recently improved in the areas of alignment and sequence selection (Cole *et al.*, 2009). The central process for generating the heat-map is described as follows:

(1) Pick the longest 16S rRNAs for a genome with the most conserved homologue positions.

(2) Perform multiple sequence alignment for all the representative 16S rRNA sequences at those positions.

(3) Pick a pairwise distance estimation method and create a distance matrix.

(4) Apply hierarchical clustering to generate heat-map visualization for the resulting phylogenetic taxonomy.

Alignment-free methods exist that can compute a reasonable distance matrix (Kotamarti *et al.*, 2010; Otu and Sayood, 2003). The former requires more computational resources to exhaustively determine short unique substrings, but the latter is faster, works more efficiently, avoids the single copy selection issue by combining them and also provides a valid distance metric. Once a distance matrix is available, the algorithm *Self Organizing Self Correcting (SOCC)* can automate the process of spotting and correcting for taxonomic errors; however, this is easily done in case of EMM by directly considering the ranked list of models.

In later publications (Garrity and Lilburn, 2005), the manual process of visual examination for taxonomy errors was substituted with a more automatic, iterative distance matrix manipulation. The algorithm SOSC uses a statistical basis to shuffle elements of the distance matrix in a multi-pass scheme to reorganize sequences into similar clusters. The final step of heat-map generation and visualization is replaced with an iterative distance matrix manipulation algorithm. Using the data mining direction taken by Garrity and Lilburn (2005), we propose building classification models instead of relying on computing an all-against-all distance matrix. By utilizing EMM signatures of the 16S sequence profiles of organisms, we prevent discarding potentially useful copies of the 16S rRNA thus eliminating the subjective sequence selection process. Furthermore, we propose a Markov transition probability-based classification rank score aided by reports of statistical significance to assess the taxonomy. The task of assessing membership involves considering each member, measuring its degree of membership in a potential host community represented as a profileEMM, and then calculating its statistical significance. A member may either be an individual organism with all its copies of 16S rRNA or a higher taxon such as *genus*. Highlights of the proposed method are:

(1) Evaluate a taxon against all of the communities to which it could potentially belong, which generates a rank order.

(2) Consider the highest ranking community as the best classification for the taxon.

(3) A taxonomic anomaly is identified if the highest ranking community is not consistent with the observed placement in the published microbial taxonomy (Garrity *et al.*, 2005).

Next, formal notation will be given for the above. Given an EMM e_q as a member taxon signature to be classified against one of T communities represented by EMMs e_t, where $t = 1, 2, ..., T$, the evaluation of e_q against a profileEMM of e_t is given by the function $r(e_q, e_t)$, where r computes a rank measure. The profileEMM $e_{t'}$ with the highest rank is considered the community to which e_q belongs. We will introduce the terms *difference score* and *transition presence probability* to explain classification. The former is derived as the sum of differences in individual LOD scores of states when a match is found between a query state and a model state. The latter implies a non-zero value if the model includes a transition arc that connects matched state and the current state.

The classification is represented by the notation t' as follows:

$$t' = \text{argMax } \{ t \epsilon \{1, 2, ..., T\} | r(e_q, e_t) \} \tag{5}$$

where $r(e_q, e_t)$ measures how similar e_q is to the e_t as below

$$r(e_q, e_t) = \frac{1}{1 + \sum_{k=0}^{|e_q|-1} \left(S(s_k, s'_k) ln(P(s_k, s'_k)) \right)} \tag{6}$$

where $S(s_k, s'_k)$ and $P(s_k, s'_k)$ represent the difference score and the transition presence probability, respectively, for the quasi alignment $< s_k, s'_k >$ which is defined as a pair of member-model states with the least-squared Euclidean distance. The value for a transition probability is 1, if the arc $< s'_{k-1}, s'_k >$ is present in the profileEMM e_t. If there is no such arc, a penalty value $-ln(1/|e_t|)$ is used. The denominator computes the score adjusted to highlight the difference in a quasi alignment and the rank measure function generates higher values for smaller difference scores. Scoring quasi alignments and computing P-values are described next.

Karlin and Altschul (1990) proposed a LODs score for scoring alignments along with a theorem supporting a Gumbel extreme value distribution of LOD scores. Extending the Karlin–Altschul statistics, quasi alignments can also be scored with LOD scores by use of dynamically built score matrices for each EMM signtaure (Kotamarti *et al.*, 2009). This has become necessary since the alignment basis used in Karlin and Altschul (1990) deals with a substitutive environment, where the aligned base pairs that are different constitute a substitution. Since EMM uses the count form of bases and not the bases directly, its quasi-alignment context deals with comparing numerical vectors of oligomer counts and the EMM score matrix reflects the LOD scores for occurrences of specific oligomer patterns. Kotamarti *et al.* (2009) describes an algorithm for dynamically building a score matrix for each profileEMM and adjusting it in the context of a member taxon to compute the key Gumbel distribution parameters such as λ and K. The algorithm is quite involved and the reader is referred to earlier work by the authors (Kotamarti *et al.*, 2009); a brief summary of it is presented here for completeness:

(1) Create LOD score matrix for each model.

(2) Adjust the scores when assessing membership using the individual probabilities of the bases in a member's sequence(s).

(3) Fit a Gumbel distribution by computing the parameters λ and K using numerical methods at `http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/tools/blast.c`.

Let v_k and v'_k be the numerical vectors of a quasi alignment (closest squared Euclidean match) $< s_k, s'_k >$. Let X be a score matrix, i.e. a score array of size $|v_k|$. The scoring of quasi alignments is computed as a difference score $d_q[k]$, which is the weighted Manhattan distance, where $k = 1, 2, ..., |e_q|$ and $|e_q|$ is the number of states in the member EMM e_q, as follows:

$$d_q[k] = \sum_{j=1}^{|v_k|} \left(X_j | ((v_k)_j - (v'_k)_j) | \right) \tag{7}$$

where j loops over all oligomer combinations (64 for a 3-mer basis) and X_j is the LOD score for j-th pattern. For example, for a 3-mer basis, a pattern like '*ACU*' is associated with a numeric value in the vectors as $(v_k)_j$, $(v'_k)_j$ and X_j.

The difference scores obtained from scoring quasi alignments are then assessed for significance. Once again, referring to the Karlin–Altschul statistics, P-value is computed using the Gumbel distribution framework. However, unlike in classic alignment scoring of local alignments where perfect alignments generate maximum scores, quasi-alignment statistics of using difference scores generate scores around zero for near alignments. Using μ and σ to represent mean and SD, when samples are drawn from two different distributions (μ_1, σ_1) and (μ_2, σ_2), the difference distribution formed from the differences of samples is characterized by $(\mu_d = \mu_1 - \mu_2$ and $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2})$ (Yates and Moore, 2007). When samples are drawn from the same distribution, the resulting mean of the difference distribution becomes zero and the variance is doubled in value as characterized by $(\mu_d = 0, \sigma_d = \sqrt{2\sigma^2})$. Karlin–Altschul statistics for classic alignment are interested in the one-sided tail probabilities because perfect alignments do not follow the null hypothesis of randomness. In fact, the high scoring local alignments fall outside a confidence threshold in the right-side tail. However, in our case, difference scores could fall on either side of zero and the tail probability supporting the alternate hypothesis becomes one minus the area under the cumulative distribution function corresponding to the interval $-d$ and $+d$, where $d = |d_q[i]|$ of the difference distribution. Adapting Karlin–Altschul P-value, given K, λ' as the distribution parameters and $|e_q|$ representing the number of quasi alignments, we get the following:

$$P'\text{-value} = 1 - (e^{K|e_q|e^{-\lambda'(+d)}} - e^{K|e_q|e^{-\lambda'(-d)}}) \text{ where } \lambda' = \lambda/\sqrt{2}$$

The smallest difference score among those of all quasi alignments is the best base line P-value for the member. This, however, can be further improved to take into account other significant alignments as well. Given a 95% confidence interval, the quasi alignments with $d \leq d_t$ are considered significant where d_t corresponds to the 5% threshold in the difference distribution characterized by $\mu = 0$ and λ'. We propose the following adaptation of the Karlin–Altschul statistics to obtain a P-value for the difference distribution.

$$P'\text{-value} = 1 - (e^{-y_2} \Sigma_{i=0}^{h-1} y_2^i/i! - e^{-y_1} \Sigma_{i=0}^{h-1} y_1^i/i!) \tag{8}$$

where $y_2 = K|e_q|e^{-\lambda'(+d)}$ and $y_1 = K|e_q|e^{-\lambda'(-d)}$. h is the number of quasi alignments that are significant. If $h = 1$ indicating that there is only one significant quasi alignment, the P-value reverts to Equation (1). Though normal distributions are most popular, we opted to use the difference Gumbel distribution due to the conservative P-values it appears to generate (Kotamarti *et al.*, 2009).

Each evaluation by rank measure function generates several data points for closer analysis in the event of unexpected classification that is required for revisions in taxonomy. The values reported include $1-P$-values, E-Score, BitScore, inverted difference score, the best quasi alignment that is the alignment with the smallest difference score and the total number of quasi alignments that exceed 95% confidence threshold. The E-Score and BitScore are analogs for the $1-P$-values for an outcome (Ian Korf *et al.*, 2003).

Our proposed method for taxonomy verification requires the use of all available relevant genomic data, which is typical of a taxonomic classifier; as such, to eliminate bias, we will include a 10-fold cross-validation experiment in Section 4 which systematically separates the training and the test sets in assessing the predictive power of EMM. Besides confirming a taxonomic placement with a statistical significance declaration, our proposed method identifies ambiguities in taxonomy.

4 EXPERIMENTS AND ANALYSIS

In this section, we comparatively analyze the introduced methods using the 16S rRNA sequences. Since the distance matrix method is the primary basis on which the Bergey's manual is maintained, we will exclude this method in the comparison to eliminate bias in classifier accuracy. Any deviations from the Bergey's manual are identified as possible candidates for revision. Due to space limitations, only the highlights of genus and sub-genus level EMM LODs classifier are described; authors present a detailed 10-fold cross-validation example for EMM-based classification of microbial strains into phylogenetic classes in Kotamarti *et al.* (2009). The 16S rRNA database utilized in this analysis is derived from the NCBI Microbial Complete Genome Database at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. The sequences were extracted from the annotated whole genome sequence files using keyword searches. From these data, a new database was built that consists of individual files, one per microbial organism, in FASTA format. The original dataset was derived from the NCBI as of August 2009 and consists of 782 organisms each with multiple 16S sequences where applicable. There were several cases of missing genus or even class information. Since this type of information is used for verifying the topological accuracy, such data are used only for testing unknown sequence detection. The final database consisted of 676 organisms. For comparison, we considered the legacy clustering-based method of using the heat-map/SOCC method and a leading classification method based on a naive Bayes classifier. Figure 2 summarizes the three methods including EMM with regards to operations performed in preprocessing, training and analysis (confirm/deny) phases. The space and time complexities are summarized in Figure 3. Our approach only takes $O(MK^2)$ for space and $O(N+NK+M^2)$ for time, N is the number of sequences, M the number of models (or classes) and K the number of equal-sized segments. These are much less than the prevailing methods. Details are provided in Supplementary Material.

We performed five experiments, using $l = 3$, i.e. 3-mer pattern for generating counts, to validate our approach. Selection of $l = 3$ was found to be sufficient for taxonomic verification and triplets are convenient when working with DNA sequences as well, as they map to amino acids. The experiments are: (i) perform 10-fold cross-validation for classification of organisms into genera of the phylum *Proteobacteria* and comparatively analyze classification errors for known/unknown sequences; (ii) classify all genera and species of the existing microbial taxonomy; (iii) classify strains of Bacilli; (iv) classify *E. coli* fragments; and (v) compare EMM and LZ complexity methods for phylogeny. The results and descriptions follow:

Experiment 1: 10-fold Classification of *Proteobacteria* strains. We chose the largest *phylum Proteobacteria* that consists

of 5 *classes*, 129 genera and 351 organisms. EMM signatures are first generated for all the genera of the phylum, and then also for all the organisms using all the available 16S rRNA copies. As expected, space efficiency is achieved as shown in Figure 4. Second, the dataset is randomly divided into 10 partitions. Third, leaving one partition at a time as the test partition, the remaining nine partitions are used as the training data from which models for genera are generated. Next, the test partition, i.e. the organismal signatures from the test partition are evaluated against all the models and the best classification is generated. Misclassifications are counted and those for which no models exist are discarded. We found the classifications to match 84% on average for genus and 100% for class. Figure 5 shows the accuracy breakdown by each of the 10-fold runs. Also shown is the comparative analysis between the naive Bayes classifier (Wang *et al.*, 2007) and the EMM classifier in Figure 5 using three partitions representing worst, best and moderate performance by EMM. The partitions were structured into multiple files of required format prior to using the naive Bayes classifier implementation of the RDP 2.2 classifier. Unlike EMM, the original distance matrix and the RDP classifier methods require selection of a single 16S rRNA copy among the many found in some strains. To the best of our knowledge, these selection procedure details are not published; as such, we used a random selection when selecting a single copy. The misclassifications for both initially also counted the cases where there is no model for the genus being classified. Lack of models for some is due to low cardinality of a particular genus as it is very common to encounter a genus with a single member only. Therefore, if a single member genus is in a test partition, there would be no corresponding training model and the best a classifier could do is to find the closest one which would still count as a misclassification. Errors due to lack of models are adjusted and the remaining classification errors, also shown,

	Preprocessing	Training	Confirm/Deny
Heatmap/SOSC	Sequence Selection + ClustalW	Hierarchical C.L Clustering	Self Organizing Self Correcting (SOSC)
Naïve Bayes	Sequence Selection + Counting	Training Models	Classification
EMM	Counting	Building EMMs	Classification

Fig. 2. Overview of operations for taxonomic analysis under various methods: operations for each method depend on the type of method used. The legacy clustering method requires sequence selection and multi-sequence alignment in the preprocessing stage, while the other methods require count generation of l-mers. For the naive Bayes classifier 8-mer is used, and for the EMM method 3-mer is used . The sequence selection step is not required for EMM because all sequence copies are utilized in creation of an EMM signature, whereas in other cases only select copies that pass select criteria are used.

	Time Complexity			Space Complexity
Method	Preprocessing	Training	Confirm/Deny	
Heatmap	N+ NL+L^2	N^2logN	NLogN	NL+L^2
Naïve Bayes	N+M	((L-8)/8)M to 4^8M*	M^2	((L-8)/8)M
EMM	N	NK	M^2	MK^2

Fig. 3. Comparison of complexities for taxonomic analysis under various methods: complexity is detailed for each method by major operation where L is the length of a sequence, N is the number of sequences, M is the number of models (or classes) and K is the number of equal-sized segments. Multiple sequence alignment is by far the most expensive operations in space and time. Naive Bayes classifier (Wang *et al.*, 2007) uses an 8-mer, which could theoretically require 4^8 feature attributes, but optimal design is assumed due to lack of detailed information. Reduced space complexity of $O(MK)$ for EMM is possible in case of classification algorithms since transition information is not used.

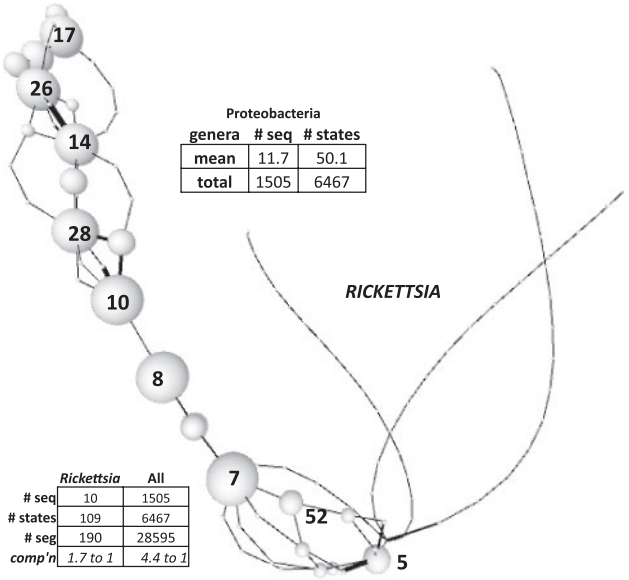

Fig. 4. EMM visualization showing compression of sequences. EMM achieves sequence compression by clustering similar segments of sequences into nodes. EMM for genus *Rickettsia* is shown to compress 190 sequence segments of 80 nt into 109 nodes of NSV with a numerical vector of size 64. Likewise, all genera models are compressed achieving a ratio 4.4 to 1 space efficiency.

Proteobacteria Classification

phyla	classes	genera	species	strains
1	5	129	219	371

Run #	1	2	3	4	5	6	7	8	9	10
Accuracy	76%	88%	96%	77%	88%	79%	88%	76%	88%	88%

taxon level	set 1	set 2	set 3
genera	35	28	32
Method	*mis-classifications*		
Naïve Bayes	11	12	9
EMM	15	13	8
unclassified	11	12	9

genera classification errors			
mis-classified	#	Naïve Bayes	EMM
Buchnera	1	Citrobacter	Serratia
Geobacter	1	Geobacter	Pelobacter[1]
Haemophilus	2	Haemophilus	Pausteurella[1]
Rhizobium	2	Rhizobium	Ochrabactrum[2]

[1] DNADIST on global alignments confirms greater proximity than the current placement.

[2] Several exact matching 80 bps matches found makes this an interesting revision.

Fig. 5. Genera classification for the strains of *Proteobacteria*: The 10-fold cross-validation of genus classifier for the phylum *Proteobacteria* is shown with the percent of correct classifications. The incorrect classifications may also be due to ambiguous placements in the reference taxonomy. The comparative analysis for three test partitions includes misclassifications due to single member genera that are selected as test with no counterpart in the training set. Except for *Buchnera*, other misclassifications identify potential areas of revision or further study. Thus, EMM offers improved time and space complexity while identifying potential areas of concern in the existing taxonomy.

are analyzed further as follows. Figure 5 also shows the genera that are found as misclassified by either method or both. It is interesting to note that EMM finds a more reasonable classification, justifiable using DNADIST-based distance, for both *Geobater* and *Haemophilus* strains than suggested by the RDP classifier as well as Bergey's manual. It is equally interesting to point out that there are significant local alignments between *Rizhobium* strains and the near cousin *Ochrobactrum* though a global-aligned DNADIST analysis confirms Bergey's manual. Since any taxonomy is simply based on the convention chosen and not necessarily reflective of a particular biological fact of interest, the current separation of the two *Rizhobium* strains and *Ochrobactrum* remains interesting. Unlike, the RDP classifier, normalized LOD scores used by EMM in identifying such areas of interest is perhaps useful. We will consider how both methods handle unclassified sequences next which is useful for classifying sequences from strains for which the complete genome is yet to be determined.

We added 17 test strains from several archaea and bacteria phyla to analyze classifier response. The EMM classifier reported a *zero base P′-value* for 12 of them indicating immediate conclusion that the sequences are of unknown origin. When the same was tested with the RDP naive Bayes classifier, only the best possible matching taxa are reported since the training set does not include any of the test phyla. This means that RDP classifier operation did not appear to detect novel sequences. On the contrary, the EMM LODs classifier provides a modified rank function output which detects the novel sequences 100% though the *P*-value derivative appears to detect 70% only. However, when they were input to the online RDP classifier, the correct classifications were determined. This is most probably due to the fact that the training set for the online version already includes all known and available sequences. From a taxonomic classification stand point, it is a common practice to build models with all the available genomic information and provide a query service for sequences known or unknown as input by a user. This is exactly the case with the RDP classifier that works based on the existing classifications and the models. However, it does not handle sub-genus classification. We take this useful concept to sub-genus level in the following experiments where all the available 16S rRNA information is used in building the classification models.

Experiment 2: classification of all genera and species. Since sub-genus levels are considered difficult to differentiate and the naive Bayes classifier available online at Ribosomal Database Project 2 does not classify below genus level, we extended our test to verify species level in addition to the genus of the published microbial taxonomy. The EMMs are built first for all genera, species and strains. Evaluation of each strain against all available species models generated high-ranking classifications. Of these, all strains were classified correctly for genus level. With the exception of 12 strains, all strains were also classified under the correct species. For all correct classifications, we found them to be significant at 95%. Table 1 shows that the 12 misclassifications are, in fact, due to presence of exact copies of the 16S rRNA sequences. For example, the strains *Bordetella parapertussis 12822* and *B. bronchiseptica RB50* have the identical set of 16S rRNA sequences and as such both species models attained the first rank position for them resulting in ambiguity.

Experiment 3: classification of *Bacillus* strains. Of all microbial classifications, identifying the exact species of a strain is the most difficult. Typically, wet-lab techniques and/or multi-locus methods are used. However, we extended our experiment to classification of *Bacillus* strains and found that it was successful for 17 out of 20 *Bacillus* strains. The remaining 3 are of *Bacillus anthracis* type: Ames-Ancestor, Ames and Sterne, which share one or more exact copies of the 16S rRNA making identification ambiguous at the strain level.

Table 1. Genus and species classification results: evaluation of all strains confirmed genera placement at 100%

Classifier accuracy of genus and species			
Taxa level	Count	Accuracy (%)	P'-value (%)
Genera	248	100	95
Species	455	98.2	95

Species with common 16S rRNA sequences	
Brucella abortus	Francisella novicida
Brucella canis	Francisella tularensis
Brucella ovis	
Brucella suis	
Bordetella bronchiseptica	Mycobacterium bovis
Bordetella parapertussis	Mycobacterium tuberculosis

However, all species could not be confirmed due to presence of exact copies of the 16S rRNA in 12 species (∼2% of total). There are four groups as shown in the figure that share one or more exact 16S rRNA sequences making their classification ambiguous.

Experiment 4: classification of sequence fragments. In all experiments so far, we have used all available 16S rRNA sequence data. This test involves testing a single fragment of one of the gene copies of the 16S rRNA. Since EMMs are built using segmented sequences of fixed size, flexible segment boundaries that allow overlapping was utilized. Overlapped segments result in redundant counting of l-mers since two successive segments are the same except that the second one is shifted by a base pair. This helps in finding the quasi alignment between a fragment starting at any position and an EMM during the evaluation step. We believe this to be ideal for handling the well-known binning problem in metagenomics. However, true alignment-free metagenomics will require EMMs built for all the available relevant sequence data. This is deferred as future research. For this test, strain models are re-generated as parent models with overlapped segments. The fragments are created by taking a 16S rRNA sequence of each of the 19 *E.coli* organisms and discarding the first 500 bp of the first 16S rRNA sequence. The remaining ∼1000 bp are segmented as 80 bp and multiple of these 80 bp segments are used to create a test fragment. These fragments were evaluated and all were correctly classified against their species, i.e. *E.coli*. Depending on the location of the fragment, the accuracy ranged between 70% and 100%.

Experiment 5: alignment-free phylogeny. LZ complexity-based phylogeny (Otu and Sayood, 2003) and the EMM counter part (Kotamarti *et al.*, 2010) methods were compared using 20 strains (92 16S rRNA sequences) of the most diverse phylum *Burkholderia*. While both methods generate the same phylogeny, the EMM method completed in <5 min, whereas the LZ complexity method took over 8 min to generate the distance matrix. Whole genome level comparison remains to be done.

5 CONCLUSION

In this work, we expanded on the alignment-free alternative to using classic alignment methods and statistics in sequence analysis. With the goal of improving data representation and complementing classification with significance, we explored profile models for predicting microbial taxonomy.

Though 16S rRNA is the marker of choice for the majority of microbial classification, its heterogeneity, due to multiple copies of the gene, usually requires a method to select some and discard others based on some criteria. Our proposed sequence analysis approach uses all the available sequence copies of 16S rRNA of a microbial organism by reducing similar sequence regions into fewer nodes in the model. In this work, we showed that a compact model that includes the complex intra-sequence order can be built using an EMM (Fig. 1). Such compact models are called EMM signatures and transforming organisms with all their 16S rRNA sequences to EMM signatures creates a *signature library*. The library can then be used for homology assessment within each taxa hierarchy to derive member significance.

Four experiments were carried out to demonstrate the space compression and classification accuracy down to strain level. In all, except the lowest classification levels, i.e. species and strain classifications, the current classifications and those by our method are in perfect agreement. We showed, for species level, 12 ambiguities that occur due to the heterogenity issues of the 16S rRNA; however, achieving 98% accuracy (Table 1) with identification of ambiguities that require multi-locus methods is reasonable. Furthermore, same technique applied at strain level, i.e. when diagnostic identification is attempted, either identifies a single strain or a small number of strains that share one or more exact copies of the 16S rRNA. The rank function output offers feedback for immediate correction in case of taxonomic placement errors, which could auto-update profile models thus eliminating the need for a distance matrix-based correction. Future work will address this.

Several approaches for curative analysis of microbial taxonomy based on the 16S rRNA are in existence, but all of them either require extensive multi-sequence alignment and/or do not use all of the available sequence information. Yet, new approaches are emerging proposing the use of the whole genome (Otu and Sayood, 2003) and/or other markers such as *RecA, RPOB, 23s rRNA,* etc., (Case *et al.*, 2007; Dahllöf *et al.*, 2000) in order to improve clarity across species. We have shown in our article that by using improved statistical signature methods, classification, significance analysis and homology assessment can be performed effectively with less memory and computation.

Funding: NSF Net-Centric IUCRC/T-SYSTEM Inc. industrial memberships in the form of graduate studies sponsorship (to R.M.K.).

Conflict of Interest: none declared.

REFERENCES

Case,R.J. *et al.* (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol*., **73**, 278–288.

Cole,J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

Dahllöf,I. *et al.* (2000) rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl. Environ. Microbiol.*, **66**, 3376–3380.

Dunham,M.H. *et al.* (2004) Extensible Markov model. In *Proceedings of the Fourth IEEE International Conference on Data Mining ICDM '04*, IBFI, Schloss Dagstuhl, Germany, pp. 371–374.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Garrity,G. *et al.* (2005) *Bergey's Manual of Systematic Bacteriology, Second Edition.* Vol. 2, 2nd edn. Springer, New York.

Garrity,G.M. and Lilburn,T.G. (2005) Self-organizing and self-correcting classifications of biological data. *Bioinformatics*, **21**, 2309–2314.

Ian Korf *et al.* (2003) *BLAST*. O'rielley, Sebastopol, CA.

Janda,J.M. and Abbott,S.L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.*, **45**, 2761–2764.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Kotamarti,R.M. and Dunham,M.H. (2010) Alignment-free sequence analysis using Extensible Markov Models. In *9th International Workshop an Data Mining in Bioinformatics (BIOKDD'10)*.

Kotamarti,R.M. *et al.* (2009) Targeted genomic signature profiling with Quasi-alignment statistics. *COBRA Preprint Series*.

Kotamarti,R.M. *et al.* (2010) Sequence transformation to a complex signature form for consistent phylogenetic tree using extensible Markov model. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2010 IEEE Symposium on*, IEEE Explore, Washington DC, USA, pp. 1–8.

Lilburn,T.G. and Garrity,G.M. (2004) Exploring prokaryotic taxonomy. *Int. J. Syst. Evol. Microbiol.*, **54**, 7–13.

Lilburn,T.G. *et al.* (2006) Computational aspects of systematic biology. *Brief Bioinform.*, **7**, 186–195.

Meng,Y. and Dunham,M.H. (2006) Online mining of risk level of traffic anomalies with user s feedbacks. In *Granular Computing, 2006 IEEE International Conference*, IEEE Computer Society, Washington, DC, USA, pp. 176–181.

Meng,Y. *et al.* (2006) Rare event detection in a spatiotemporal environment. *Granular Computing, 2006 IEEE International Conference*, Academy Publisher, Finland, pp. 629–634.

Otu,H.H. and Sayood,K. (2003) A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, **19**, 2122–2130.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.

Yates,D.S. and Moore,S. (2007) *The Practice of Statistics 3rd edition*. W.F. Freeman & Co., New York, New York.