Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2009 *Paper* 63

Targeted Genomic signature profiling with Quasi-alignment statistics

Rao Mallik Kotamarti* Douglas W. Raiford† Michael Hahsler‡
Yuhang Wang** Monnie McGee†† Maggie Dunham‡‡

*Southern Methodist University, rkotamarti@engr.smu.edu
†The University of Montana, douglas.raiford@mso.umt.edu
‡Southern Methodist University, mhahsler@lyle.smu.edu
**Southern Methodist University, yuhangw@engr.smu.edu
††Southern Methodist University, mmcgee@mail.smu.edu
‡‡SMU, mhd@lyle.smu.edu

http://biostats.bepress.com/cobra/art63

Targeted Genomic signature profiling with Quasi-alignment statistics

Rao Mallik Kotamarti, Douglas W. Raiford, Michael Hahsler, Yuhang Wang, Monnie McGee, and Maggie Dunham

Abstract

Genome databases continue to expand with no change in the basic format of sequence data. The prevalent use of the Classic alignment based search tools like BLAST have significantly pushed the limits of Genome Isolate research. The relatively new frontier of Metagenomic research deals with thousands of diverse genomes with newer demands beyond the current homologue search and analysis. Compressing sequence data into a complex form could facilitate a broader range of sequence analyses. To this end, this research explores reorganizing sequence data as complex Markov signatures also known as Extensible Markov Models. Markov models have found successful application in Biological Sequence analysis applications through small, but important extensions to the original theory of Markov Chains. Extensible Markov Model (EMM) offers a novel Quasi-alignment complement to the classic alignment based homologous sequence search methods like BLAST. EMM based BioInformatic analysis (EMMBA) incorporates automatic learning which allows the Markov chain creation dynamically. Oligonucletide or Genomic word frequencies form the core sequence data in alignment free methods. EMMBA extends the Karlin-Altschul statistics to bring forth an analogous E-Score statistical significance to the Quasi-alignment domain. By consolidating a community of sequences into a single searchable profile, EMM methodology further reduces the search space for classification. Through dynamic generation of the score matrix for each community profile, EMMBA fine tunes the score assignments. Each evaluation iteratively adjusts the profile score matrix to account for point probabilities of the query to ensure Karlin-Altschul assumptions are satisfied to derive meaningful statistical significance. The presence of multiple Quasi-alignments resembles multiple local alignments of BLAST. Quasi-alignments are scored based on a difference distribution

of Gumbel scores. Species signature profiles allow for statistical validation of novel species identification. Working in EMM transformation space speeds up classification and generates distance matrix for differentiation. The techniques and metrics presented are validated using the microbial 16s rRNA sequence data from NCBI.

Targeted Genomic signature profiling with Quasi-alignment statistics

Rao M. Kotamarti [1], Douglas W. Raiford, Michael Hahsler, Yuhang Wang, Monnie McGee and Margaret H. Dunham [1]

Abstract

Genome databases continue to expand with no change in the basic format of sequence data. The prevalent use of the Classic alignment based search tools like BLAST have significantly pushed the limits of Genome Isolate research. The relatively new frontier of Metagenomic research deals with thousands of diverse genomes with newer demands beyond the current homologue search and analysis. Compressing sequence data into a complex form could facilitate a broader range of sequence analyses. To this end, this research explores reorganizing sequence data as complex Markov signatures also known as Extensible Markov Models. Markov models have found successful application in Biological Sequence analysis applications through small, but important extensions to the original theory of Markov Chains. Extensible Markov Model (EMM) offers a novel Quasi-alignment complement to the classic alignment based homologous sequence search methods like BLAST. EMM based BioInformatic analysis (EMMBA) incorporates automatic learning which allows the Markov chain creation dynamically. Oligonucletide or Genomic word frequencies form the core sequence data in alignment free methods. EMMBA extends the Karlin-Altschul statistics to bring forth an analogous E-Score statistical significance to the Quasi-alignment domain. By consolidating a community of sequences into a single searchable profile, EMM methodology further reduces the search space for classification. Through dynamic generation of the score matrix for each community profile, EMMBA fine tunes the score assignments. Each evaluation iteratively adjusts the profile score matrix to account for point probabilities of the query to ensure Karlin-Altschul assumptions are satisfied to derive meaningful statistical significance. The presence of multiple Quasi-alignments resembles multiple local alignments of BLAST. Quasi-alignments are scored based on a difference distribution of Gumbel scores. Species signature profiles allow for statistical validation of novel species identification. Working in EMM transformation space speeds up classification and generates distance matrix for differentiation. The techniques and metrics presented are validated using the microbial 16s rRNA sequence data from NCBI.

Acknowledgements: The authors wish to acknowledge support received by the first author from T-Systems, Inc in the form of a graduate fellowship supporting his Ph.D. studies at SMU.

Southern Methodist University, rkotamarti@engr.smu.edu
Corresponding author: mhd@lyle.smu.edu

INTRODUCTION

Statistical analysis of Genomes often requires use of frequencies of letter patterns [1]. A Targeted Genomic Signature is an oligomer frequency distribution over a select section of an organismal Genome like 16s rRNA [2]. Quasi-alignment refers to a region wide alignment based on similar word frequencies between two sequence fragments. This is different from Classic alignment where the alignment is individual position based assessment using substitution matrices [3,4].

Supervised learning serves well in setting up libraries of models describing communities of sequences. This allows for easily determining the taxa of a sequence from a genomic sample as shown in [5]. Instead of merely searching for homologous sequences every time for every sample to determine possibly related genomes, targeted Data mining methods could allow for a more efficient search and organization of the known data [6].In this research, data mining versatility of different clustering metrics and a Markov model based classification are utilized to set up libraries of sequence community profiles.

As the amount of genomic data explodes to billions of molecular sequences, intelligent systems that can learn from and organize data into a compressed and consolidated form could improve efficiency. Learning framework based on Data Mining principles vastly improves such abstraction through creation of representative models [7] and improved models further extend the versatility of analyses.

Extensible Markov Model (EMM)

The theory of Markov Models is well known in Bioinformatics for its innate ability to represent sequence information [8] probabilistically with efficiency and unmatched sensitivity [9]. The extended forms of the model, such as the Hidden Markov Model [9] account for much of the successful application.

Main principle behind Markov modeling is that future state depends only on the current state or the immediate preceding ones depending on the order of the chain. First order Markov modeling, where a future state is based on the current state is by far the most prevalent in Bioinformatics applications. The Classic Markov model relies on fixed states that directly map the real world to symbols. However, more flexibility is useful when modeling dynamic biological systems.

The classic theory was extended to address the relatively newer field of dynamic data streams by Dunham et al [10]. The Extensible Markov Model (EMM) is the basis for the research presented here. EMMs allow addition, deletion and updating of states within a system.

The Extensible Markov Model [10] is a time varying first order Markov chain [11]. It is easy to think of a Markov chain as a directed graph where the nodes represent real world states and the arcs the transitions between them. Each arc A_{ij} is labeled with its cardinality as is each node n_i. Given an arc $A_{ij} =< n_i, n_j >$, the transition probability is calculated as $\frac{|A_{ij}|}{|n_i|}$. The salient features of the EMM are:

- The topology of the Markov chain varies including the number of nodes, the labeling of nodes, the number of arcs, and the labeling of the arcs. Algorithms are in place to insert new nodes, delete nodes, insert arcs, and delete arcs.
- Each node in the EMM corresponds to a cluster of real world states – as opposed to one real world state. EMM algorithms are able to use different clustering and similarity/distance measures.

EMMs have successfully been applied to many different applications including future state prediction [12] and rare event detection [13]. EMM is extended in this work to represent biological sequences in a more compact complex form useful for rapid classification and differentiation of organisms.

EMM based Bioinformatic Analysis (EMMBA) - Our Work

Goals: Classification using EMM allows us to ask the familiar questions like *how statistically significant is the association between an unknown sequence and a particular known community of sequences?* This deals with a MxN environment where there are N test sequences of interest that are to be classified across M profiles.

Differentiation, on the other hand, quantifies the distance by which, a community of sequences differ, in order to characterize intra-variability within a community. This deals with a NxN environment where there are N profiles that are evaluated all-against-all to generate a distance matrix which is subsequently used for Phylogenetic analysis.

Identification using EMM detects novel species. This is possible through assessment of an unknown Species EMM against a library of EMMs representing all known Species signature profiles.

Overview: EMMBA involves three successive steps: 1) Preprocessing sequence data, 2) building of model(s) to represent the community profiles and 3) evaluation of a sequence of interest to score its association with the communities.

As in all alignment free methods of sequence analysis, a word of fixed width is considered and its permutations are counted to create a frequency histogram or word statistics for a Genome [11]. The word width may be a 2 (di-mer), 3(tri-mer), 4(tetra-mer) and so on. The notation used is p-mer and it is found that beyond tetramer resolution, no significant benefit is observed in Quasi-alignment analyses.

Analysis of RNA sequences is possible by first transforming the RNA sequence to a numerical form using word frequencies and then generating the EMMs for further analysis. Similarly, DNA sequences may also be studied by first converting triplets to Amino Acids and then to a numerical form.

Scoring sequence comparisons: Assessment scores are determined by the product of probabilities associated with each correctly observed transition during the evaluation step of a sequence against a model. The incorrect transitions are allotted minimal probabilities. The missing transitions can be ignored to allow for partial sequence or fragment classification. However, this research only covers the complete sequence classification and defers fragment analysis to future Metagenomic research.

When analyzing sequence communities, it may be noted that some regions are more conserved than others and different regions contribute when establishing consensus. As such, score matrices are derived for each community making up a profile. Scoring of a test sequence against a model is highly sensitive to the word statistics of that community.

Related Work

Much of Markov model [14] applications in Bioinformatics deal exclusively with Hidden Markov concepts, to enable the notion that a symbol to state association is not necessarily fixed. HMMs have the advantage of accommodating multiple symbol outputs in a given state probabilistically. This allows for useful applications

such as Gene finding [15] and profile HMM [9]. However, HMMs require manual model creation initially which limits their rapid application. Similarly, profile HMM requires pre-alignment of multiple sequences prior to model generation which can be expensive. On the other hand, EMMs employ a learner to automatically build the model representing multiple sequences. Future research will address the inter-working between HMM and EMM where one can be converted to the other.

Covariance models differ from profile HMM in the aspect of possibility of intra-folding within a sequence according to Watson-Crick complementarities [16]. These models are useful for predicting structures. Usage of EMM for structure analysis is deferred to future research.

Our Research

This research continues with a formal presentation of EMM based Bioinformatics using the microbial 16s rRNA and then describes methods by which statistically significant classification and differentiation are accomplished. To substantiate claims about EMMBA versatility, three examples are illustrated as follows; the first is prediction of phylogenetic class using a 16s rRNA database of microbial organisms, the second is identification of organism and the third is demonstration of differential analysis by generating distance matrix and a phylogenetic tree for genomic sample of multiple species.

Creating new methods: Algorithmic approaches to problem solving are evident in classic alignment based Bioinformatics considering the predominant use of BLAST like similarity search tools [17]. Even in the Quasi-alignment space also, where oligomer or word statistics dominate, it is often necessary to develop new algorithms on the same old conceptual frameworks. For example, this is seen with Viterbi algorithm which uncovers the hidden state sequences [18]. However, in all such cases, the algorithmic adaptation is highly specialized and somewhat limited to the problem at hand. This research on the other hand presents a customizable algorithmic framework within the same modeling paradigm. Such framework can be used to answer many types of Bioinformatic questions against the same background of sequence library of profiles. For example, algorithms range from a simple transition sensitive match count aggregation usable for higher taxa classification to a more involved E-Score that scores according to extended Karlin-Altschul statistics. Other algorithmic adaptations presented here also include novel sequence recognizers as well as those that generate pair-wise distance matrix usable for phylogenetic analysis.

Ensuring Statistical Significance: While dealing with heuristics to address the massive sequence homology search issue, qualifying the results with sound statistical basis eliminates those reports that could occur by chance. This is done successfully by Karlin-Altschul statistics in various BLAST literature which uses a parametric model to characterize the statistical significance of each and every result [19,20]. Though the problem BLAST seeks to address is that of similarity search across an ever growing database of billions of residues using Classic alignment, the statistical principles offer extendable theoretical bases to Quasi-alignment solution space also. The primary distinction between the classic alignment based and Quasi-alignment domains can be isolated to the differences in the scoring of matches and sequences. This research explores extension of Karlin-Altschul statistics for the purpose of determining statistical significance for Quasi-alignment.

METHODS

Baseline EMM overview

The background for research outlined here is based on Extensible Markov Model. The EMM is built dynamically as input vectors are fed into the EMM Learner which either finds a matching state to cluster the new vector into or adds a new state. Whenever a new input vector is processed through an active EMM learner, the current state of the EMM changes to the state into which the new input vector is clustered. In case of a new state that is created as a result of not finding an existing matching state, the current state becomes the new state. Throughout the dynamic model building process, transitions are recorded and counted for calculating transition frequencies/probabilities as done typically in Markov chains except that in the classical Markov models, such information is known ahead of time; in case of EMM due to its dynamic nature, transition information is updated whenever a new input vector is fed to the EMM learner. Due to its automatic learning capability and the flexibility it offers for selecting various clustering techniques, EMMs are quite versatile with applications in many areas including Bioinformatics.

Bioinformatic extensions to EMM

Unlike traditional machine learning tools as well as EMM, bioinformatics deals with many long sequences of a select alphabet whether it be for representing DNA, RNA or protein. This format needs to be converted before EMM can be used to learn and build models. This is done by generating word statistics at various uniformly placed points along a sequence. The meaning of “word" is simply the sliding window of size varying between 2-4+ whose permutations are counted to generate frequency maps called Numerical Summarization Vectors. Process is discussed in the next section.

Once preprocessed, sequences take the familiar form of vectors which can be fed into an EMM learner which then either grows an existing state or grows the EMM itself by adding new states. As an EMM is built, an overall score matrix is also generated for it. The size of the score matrix is the same as the size of the input vector referred to as Numerical Summarization Vector (NSV) to be discussed in the next section. The entries in the score matrix match up with the frequency counts in the NSV. For example, a word length of 2 would have produced NSVs of size 16 for DNA which means for every pair of letters from the alphabet of 4 nucleotides, there is a score in the score matrix. Once all training data is processed and EMM creation is complete, the Score Matrix is finalized to take on a more symmetric form where the words sharing the same letters would be set to contain the same aggregate frequency. For example, a word formed from 2-mer may be AC or CA and in most cases such pairs would have different counts. The symmetric form would combine both values and assign it to both pairs. This is a prevalent practice in evolution studies, but alternate practices [21] are also in use. This research will include analysis of using symmetric as well as asymmetric score matrices because it may be possible to increase prediction accuracy by sacrificing elegance of symmetry.

The score matrix for an EMM is converted to a log-odds score matrix as is typically done in Bioinformatics. The Log-odds Score for a word variation like AG is defined as follows $LOG(f_{AG}/P(A).P(G))$ where f_{AG} is the frequency of letters A & G occurring together. P(A) and P(G) are the individual probabilities for letters A and G in the model composition itself. The logarithm is typically based on natural logarithm. The scores thus generated are then multiplied by a 10 and rounded off to generate whole numbers. As it will be seen later, the

score matrices have to follow certain criteria before they can be used for parametric analysis. This will be further discussed in the section for Statistical Significance.

Once augmented with score matrices, new sequences may be analyzed against these models to determine classification. In such assessments, statistical significance tests are performed to generate the match scores. Thus a number of extensions to the basic EMM frame work were necessarily added to prepare EMM for sequence analysis.

Community profiling

For sequence analysis, it helps to consolidate related sequences or the 16s rRNA sequences of the related organisms. The current classification has many levels of which Phylum is the highest and strain is the lowest. There are also several levels in-between such as Class and Genus. The communities at a level of interest can be consolidated into a compact model available for fast search that is statistically significant at some level. Such community consolidation is referred to community profiling. For example, a group of organisms within a class could be condensed into an EMM supplemented with a Score Matrix and a centroid vector in each state of the model. Subsequently, the centroid of the cluster that makes up a state becomes useful for assessing whether an NSV segment of a new sequence would best belong in one state cluster or another in the profile's EMM. Once the best possible match is determined, the match between state cluster's centroid and the test NSV segment can be scored and qualified with a significance level.

This is somewhat analogous to profile HMM in that the training sequences are known ahead of time to configure the HMM model prior to starting any biological analysis. In case of EMMBA also, all training sequences are consulted to derive the score matrix contents as well as individual sequence-letter probabilities. Both of these are used in deriving statistical significance while processing a query against the database of EMMs (models).

Formalization

Formal notation of Figure 1 for EMM based Bioinformatics is presented here along with explanations where needed for the theoretical portion. Statistical Significance related discussion is presented throughout.

There are three distinct process domains in dealing with EMM based Bioinformatic Analysis: first is Numerical Summarization where the sequence data is converted to word statistics, the second is where the model is built based on the formatted training data from the first and finally the third called the evaluation step deals with using the model(s) built to analyze new sequence data.

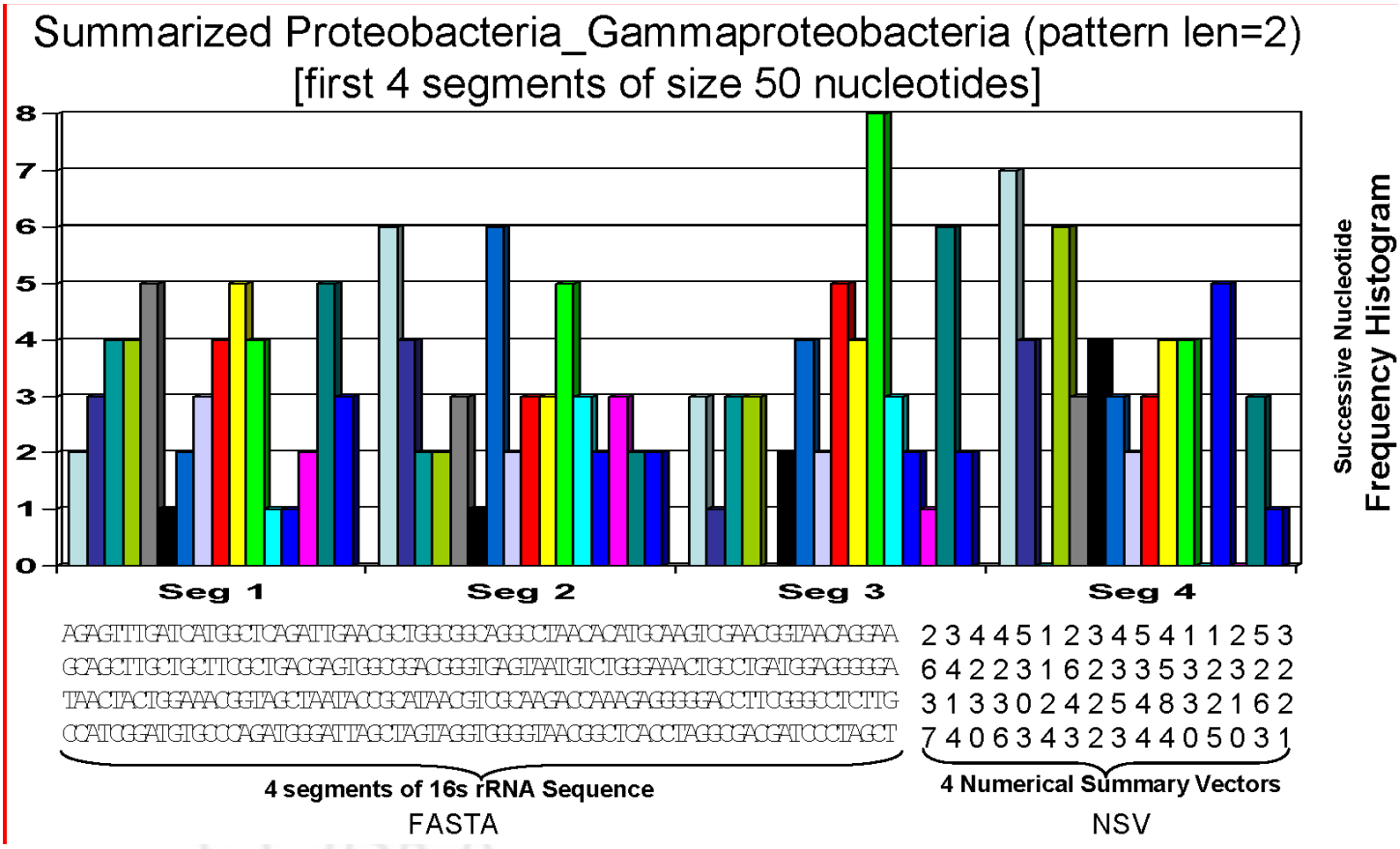
Preprocessing: Numerical Summarization:

Sequence-Words: Words in Bioinformatic sense are different from the linguistic sense in that spaces and punctuations do not separate them; they are simply defined as a sequence of characters of an alphabet of a given length. A sequence-word of length 3 is any consecutive 3-letter word that occurs in a sequence. Sequence-words may overlap; for example, a sequence of 4 letters would have 2 3-letter words. Once the words are accounted for, their frequencies provide a numeric representation. Usually there are several sequence files that need to be transformed into the appropriate numerical forms to create the desired EMMs.

Numerical Summarization Vectors (NSVs) are generated in the preprocessing step along with Individual Probability Vectors (IPVs). NSVs are used in building and evaluating EMMs where as IPVs are used in computing Score Matrices useful for computing statistical parameters. Procedure used for both is the same except that the word length is as desired by the model builder for the NSVs where as the word length is set to 1 for IPVs.

Sequence Transformation: Defining segment as a section of a sequence of size Z, there can be K segments in a sequence. Though segments can be of variable length, for simplicity, all segments of a sequence are assumed to be the same size z. Numerical Summarization is a transformation function acting on a sequence S of K segments generating K Numerical Summarization Vectors of length l. The transformation function itself is a counting function capturing the number of times a permutation of a pattern of length p occurs within a segment.

For example, a pattern width p of 2 would generate $4^2 = 16 = L$ within a segment where the base 4 is the size of the alphabet of nucleotides in an RNA sequence. For a coding DNA sequence, the alphabet consists of 20 amino acids and as such $L = 20^2$. For this work, we will use RNA sequence data only and analysis for DNA sequence data will be deferred to future publications. Continuing to formally define all the sub-components there-in, we get the following.



Numerical Summary Vectors contain a frequency histogram of counts for all permutations of the pattern of length p for each segment. In case of a sequence community, each sequence is separated from the previous with a start NSV which essentially represents a start state for the model. The NSV form of a sequence is used to either build models or to query against a set of already built models.

Figure 1: Numerical Summary Vector Representation

S	Sequence of letters of DNA alphabet or Amino Acids
s	Sequence fragment of a single segment
Z	Sequence Size
K	Number of segments in S.
	Number of Numerical Summarization Vectors (NSV) in S.
k	a segment or NSV number {0..K}
s_k	segment k
p	Pattern width [2 for a dimer, 3 for a trimer, 4 for a tetramer]
L	Length of an NSV (4^p for nucleotides or 20^p for amino acids)
	Number of permutations of p successive letters in an NSV
z	Length of a segment (N_s/K)
v_k	k^{th} NSV for a sequence S
V	Complete set of NSVs for a sequence S
$<s_1,s_2,..,s_k>$	Ordered set of all segments of Sequence = S
$<v_1,v_2,..,v_k>$	Ordered set of all NSVs of Sequence = V
c_L	count or frequency of L^{th} permutation in an NSV
$<c_1,c_2,..,c_L>$	Ordered set of all permutation-frequencies in an NSV
X	Score Matrix
X*	Adjusted Score Matrix (specific to each Query-Model pair)
f	Frequency
f'	Normalized Frequency
f*	Frequency after Symmetry
pp	Point Probability
IPV	Individual Probability Vector
NSV	Numerical Summarization Vector
M	Model
T	Test Sequence
V_t	Ordered NSV set of test sequence T
G	an EMM graph
E	number of edges in the EMM graph
NSV	number of nodes in the EMM graph
Ω	perfect graph of G where all nodes are inter-connected
Γ	number of edges in Ω
Δ	Γ - E
k'	matched node of the model
N	Length of Query
LP	Link Probability
SIM	Similarity Value
ζ (Zeta)	Similarity Value
Ψ	Distance Function
F	Function that returns Frequency of a word pattern
ξ	penalty associated with unsupported transition
K'	Number of matched states with supported transitions

This table describes the terminology to be used as reference for formalization used throughout this work.

Figure 2: Formal notation used in describing EMM based Bioinformatic Analysis

The notations S and s_k represent a sequence and a segment k respectively. $F(s_k)$ denotes the transformation function that converts the k^{th} segment in to a numerical summary form v_k.

Given K is the number of segments in S, $< s_1, s_2, .., s_K >$ represents set of all the segments in S. Similarly, $< v_1, v_2, .., v_K >$ represents the set of all Numerical Summarization Vectors (NSV) corresponding to sequence S.

$v_k = F(s_k)$ identifies the operation of transformation function on the k^{th} segment and $V = F(S)$ identifies transformation of the entire sequence with V representing the complete set of NSVs corresponding to S.

An NSV v_k is composed of a set of counts $C_k =< c_1, c_2, .., c_L >$ where L is the number of counts per an NSV. Counts $< c_1, c_2, .., c_L >$ are derived by counting the number of times a pattern of successive letters occurs within a segment. L, the number of counts per NSV, is derived from the number of possible permutations of a pattern of length p.

By increasing the pattern width p, the length of an NSV also increases as $L = 4^p$ for nucleotides and $L = 20^p$ in case of amino acids. The number of NSVs $K = |V|$ can also be adjusted for a given sequence by choosing segmentation size z for each segment. The equation $|V| = \lfloor |S|/z \rfloor$ implies a uniform segment length of z generating $|V|$ segments or NSVs.

At the end of transformation, each sequence is converted to a set of NSVs which are either used for model building or for evaluation to determine which model or which sequence community the set of NSVs belongs to. An example set of NSVs is shown in the Figure 2.

Resolution control is possible by adjusting the pattern width p and segment length z in the pre-processing stage. This will be discussed further in the Evaluation step. Pattern width also known as word length can have an effect on the statistical significance of the reported classification results. As it will be shown later, a minimum pattern width of 3 is known to offer higher significance levels in the 90% range while smaller values like p=2, though still accurate in classification in most cases, may offer significance well below 90%. The frequency histogram view of Figure 2 reflects the resolution of the summarization step and indicates the cases (not shown) where diminishing benefits occur when an unusually large value is used for pattern width p. In such cases, much of the frequency histogram will show zeros giving rise to other complexities when determining key parameters of the underlying distribution for scores.

Learning: Model Building (clustering measures): NSVs are used to build an EMM for a sequence or a community of sequences. A model is represented as a directed graph G containing N nodes and E arcs. Unlike a classical Markov Model, each $node$ is not bound to one symbol. In fact, each $node$ represents a cluster consisting of NSVs that are found to be similar by the model building process according to a similarity metric. The directed edges or $arcs$ are associated with additional information representing the relative probabilities of traversal assigned during the model building process.
It is important to observe that intra-similarity of a sequence is captured in nodes to the extent possible based on the segment length z and transitional information from one segment to next is condensed to its relative probability within the sequence. Since a segment is contiguous, its integrity is still preserved in the model inside nodes. Though the segments of a sequence may be distributed all over the graph, the transition among them is still available as weighted probabilities (frequency of arc traversals) in the model. It may be noted that state transitions can optionally influence during the sequence evaluation. This implies that missing transitions are penalized and only the supported transitions are counted. This in fact offers a tighter control on classifications; however, in biological sequences, it is well known that portions of sequences may relocate within and thus it

may not always be advisable to reject based on one or more missing transitions. In the analysis and data that is used in this research, the RNA sequence used is 16s rRNA for which such sub-sequence relocation is not found to be an issue.

The flexible arrangement of EMM is what makes HMM a subset special case of EMM with both sharing a common ancestor i.e. classic Markov model. While HMM requires prior knowledge and establishment of possible symbols generated by a model in a particular state, the EMM learns from the data itself. As data (NSVs) are fed into the model, the builder re-computes the dynamic definition of the state i.e. updated centroid of the state cluster thus decoupling a state and its static behavior of outputting a fixed symbol. Each EMM state could also be supplemented with frequencies of NSVs which helps maintain a probabilistic view of NSVs or symbols in HMM terminology. Such structure is useful for inter conversions between HMMs and EMMs.

Pursuing this idea more formally, an EMM is a graph G of N nodes & E arcs with each node associated with a cluster of NSVs making up the nodes and with each arc (possible transition) associated with a relative probability.

A model state search for a NSV of a new sequence would always produce a match in EMM frame work and the quality of the match is assessed based on a number of criteria (metrics); for example, a transition sensitive metric would consider the match complete, assigning it a probability value of 1, if the transition is supported in the Markov transition map of the model; otherwise, a small value ϵ. is assigned. Likewise, a transition agnostic metric like “Score" would assign a calculated score for the centroid of the best matched state. Thus, a state could output many types of “symbols" and different states may produce the “same symbols" also. In this sense, EMM is similar to HMM and in fact HMM becomes a subset of EMM where the symbols are predefined. On the other hand, in case of EMM, symbols are dynamic and metric dependant.

Details of model building are as illustrated in Figure 3. The very first vector defines the initial state while the second one causes the first model state (cluster) to be created. The third vector is checked against the first model state to see if its numerical composition is similar according to Jaccard similarity calculation (default). If it was similar, the third vector would have been added to the first model state’s cluster. Since it was not, a new model state is created with the third vector becoming the only member of the new state cluster. Process is repeated for all numerical summary vectors. It may be noted that each time a state is updated, its centroid vector is updated which is useful for generating an assessment score.

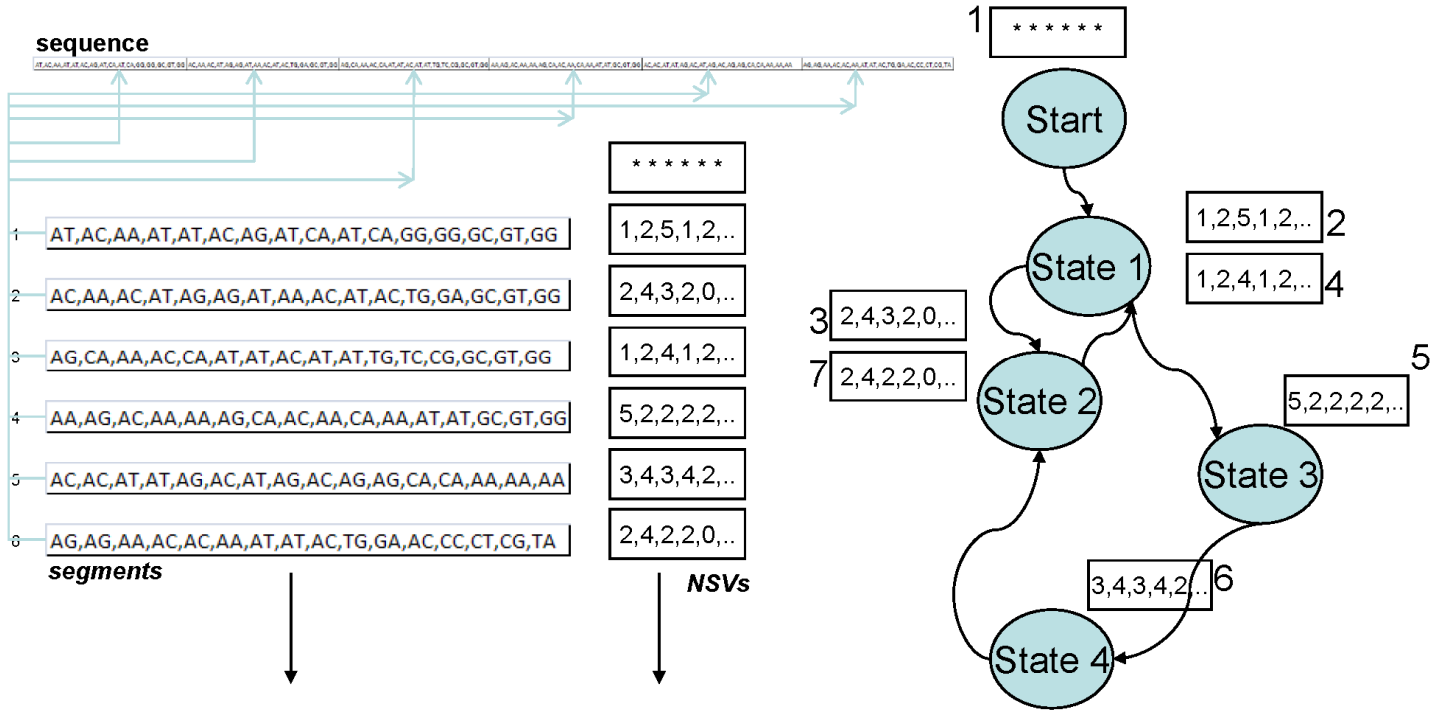
As a part of model building, a pre-Score Matrix is maintained to contain up-to-date overall frequency information for each word pattern. Similarly, individual frequency information is also updated using the IPVs to reflect the overall individual letter (for example, nucleotides) counts in a structure called Point Probability vector (PPV). This is done to facilitate subsequent Statistical Significance assessment of state matches. Thus, as each sequence is integrated into the EMM being built, the model’s pre-Score Matrix and Point Probability Vector are updated using the word counts of each NSV of each added sequence.

Once all the sequences are processed i.e. once all the organisms in the training set are included in the profile (EMM), a final Score Matrix is prepared by first converting to a symmetric form and then to a log-odds (LOD) form.

It is a standard practice in a classic alignment framework, like BLAST, to use a symmetric score matrix and in fact, BLOcks of Amino Acid SUbstitution Matrix (BLOSUM) [3] and PAM [4] are both symmetric score matrices. This is primarily useful when considering evolutionary analysis where the substitutions are known to

occur and it is often convenient to assume that transition and transversion rates are the same for a pair of letters substituting for each other. BLAST searches for homologous sequences in the database and the homology could consist of well known substitutions [22]. Continuing this practice, the EMM score matrix is also made symmetrical by simply aggregating the frequencies of the words of similar base composition. For example, AG and GA frequencies are aggregated and used as the symmetric frequency for either pair.

However, unlike classic alignment framework, this Quasi-alignment framework does not deal with position by position letter-level matching and substitution. The latter deals with summary-data over a region, i.e. the frequency statistics of each possible word composition. As such, substitution information is not available for use in scoring.



Numerical Summary Vectors (NSV) constitute the numerical representations of equal sized segments along a 16s sequence which are used one at a time in building EMM model. Model building starts with a start Numerical Summary Vector (NSV); as each NSV is processed, it is compared to the existing states of the model. If the NSV is not found to be similar enough (per a Jaccard threshold T) 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 2). When all NSVs are processed, the model is said to be complete.

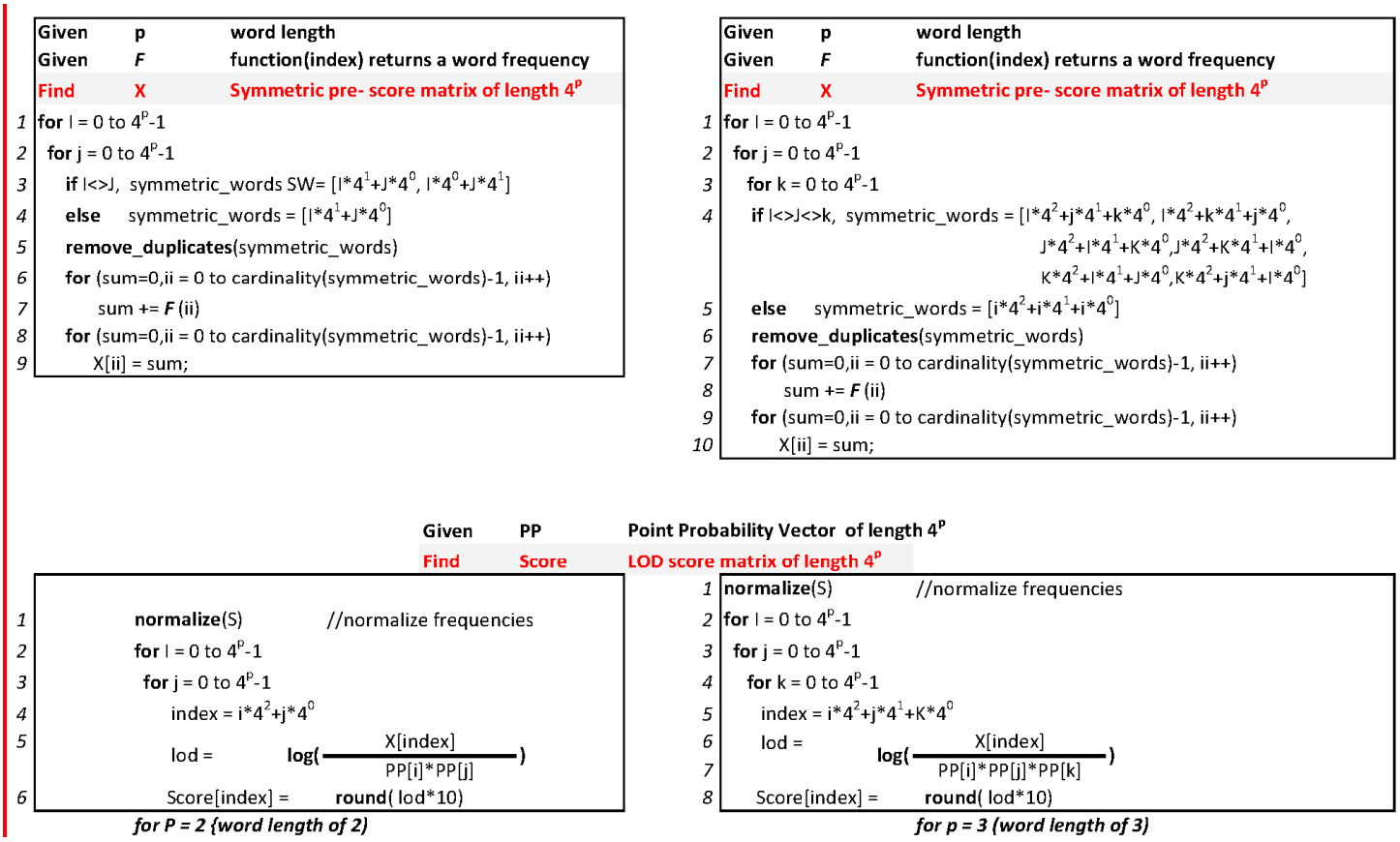
Figure 3: Phylogenetic Model Building process

The proposed method here is to score the target sequence itself to which a Quasi-alignment is found. This means that the base composition captured in the word frequency map of the model itself forms the basis for establishing a Score Matrix. Quality of the sequence segment matched is the score of the target segment itself. This implies that symmetry plays little role in this Quasi-alignment context as sequence pairs AG versus GA are indeed different from substitution pairs used in the classic alignment framework. Analysis will also be included based on asymmetric Score Matrix to highlight these facts.

Once symmetry is considered, the (Frequency populated) pre-Score Matrix is then normalized and converted to contain LOD scores [23]. LOD score for each word variant is created by computing $ln(f_{i,j}/[p(i).p(j))]$ where numerator is a word frequency (regardless of symmetry notion) and denominator is the product of individual

probabilities. The LOD scores are multiplied by 10 and rounded up to generate the final set of scores for each model.

A matrix is sufficient for representing a dimer (word length of 2) based EMM; however, longer word lengths i.e trimers and tetramers improve resolution and significance of results and require working with n-dimensional vectors. This is different from classic alignment framework where the substitution dominates which only deals with 2 dimensions i.e substitution between two nucleotides or amino acids. An extendable generalized algorithm is presented here to generate a Symmetric LOD Score Matrix in Figure 4.



Unlike classic alignment based framework which involves only 2-dimensional score matrix, Quasi-alignment framework deals with 3+ dimensions as well making symmetric score matrix generation complicated. The above shows how the algorithm can be extended from 2 to 3 and thus to any n. Once the symmetric score matrix is generated, it is then converted to Log-Odds (LOD) Score basis as shown.

Figure 4: symmetric score matrix generation

Application: Evaluating Sequences: Once models are built, they may be used for evaluating where new or sequences of interest belong. The sequences of interest may themselves be from organisms yet to be classified as known or unknown.

Test sequences are also converted into NSVs prior to evaluation; in fact, they themselves be converted to EMMs to speedup processing and in some cases actually improve accuracy. This is possible because an EMM can also be represented as a sequence of states where each state may be denoted by its centroid which is simply a vector of numbers like a typical NSV would be. However, care must be taken to be consistent in how such a sequence may be generated. This is easily achievable by using the order in which the states are created when building the EMM. This may still cause an issue if the input order of sequences used to build the model

changes. Future research for multi-sequence EMMs will specify a consistent integration function to build profiles from families of sequences. The integrity of order for this study is preserved by making sure that all sequences are processed in the same batch to generate EMM libraries.

Since the number of NSVs formed from states of an EMM typically would be less than the original set of NSVs, performance improves if EMM transformation of a sequence is used.

Since NSV is used one at a time in building the model and the transitions are also considered between states hosting the NSVs, we will consider NSV a basic observable unit in EMMBA formalization. In the process of formalizing the notation for Extensible Markov Model based sequence analysis, we will derive the Markov probability of a particular sequence of NSVs. In fact such probability is one of the proposed metrics discussed further in the subsequent sections.

We will call the state sequence, a path π. Since, the path follows a Markov chain, probability of a state depends only on the previous state. The i^{th} state in the path is calledπ_i. The chain is characterized by the parameters

$$a_{uv} = P(\pi_i = u | \pi_{i-1} = v)$$

Given a model M representing a sequence S, the probability of a test sequence T whose ordered NSV set is t is simply the product of transition probabilities associated with each NSV.

In other words, an NSV in a t is matched against one of the states and the probability of the arc from previous state to this new state due to the current NSV is considered the probability associated with it. In actuality, match is significant only if match value or the similarity value meets a certain threshold. Thus the symbol, as in established Markov terminology, is either 1 if a reasonable match value is found or 0 otherwise. Only in cases where symbol value is 1, a transition probability is computed as the relative frequency of the arc divided by the cardinality of the From-State; it is otherwise assigned a small value called ϵ.

Given $Node(t_{i-1})$ and $Node(t_i)$ respectively representing the state nodes that matched the previous NSV and the current NSV while satisfying the threshold(s), the transition of interest is between both state nodes and its Probability is represented as $P(t_i)$. However, such transition may or may not actually exist or be significant in the model itself. This is because, during the model building time, not necessarily every possible transition among states occurs.

Therefore, a transition probability associated with an NSV exists if and only if $match\_value > threshold$ and there is indeed a valid transition i.e a transition recorded while building the EMM. Using the usual notation $a_{i,j}$ to represent transition probability for the arc $A_{i,j}$ and e_i as the symbol value associated with state i, this may be represented as:

$$P(t_i) = \epsilon \text{ iff } a_{i,j} = 0 | e_i = 0$$
$$P(t_i) = a_{(i-1,i)} = P_i$$

Thus the probability of a test sequence T of K NSVs of V_t is:

$$P(T) = \prod_{i=0}^{K-1} P(t_i)$$

Unlike in HMMs where the symbol associated with a state is unknown requiring several algorithms to estimate the path, EMMs determine their symbol from the arc that reaches a particular state. This means that a state in EMMs can have a symbol 1 or 0 depending on the arc that leads to it. This is especially the case in case of evaluation of a test sequence where a transition between two successive NSVs may or may not be present in the model itself. In cases, where there is no such transition, the symbol value for the current state is 0; otherwise it is a 1. However, symbol value is only a component in computing the transition probability which will be included in the final product of transition probabilities. It may be noted that the symbol value could further change depending on Metric of interest being evaluated. For example, in case of a Score Metric, a matched state is further evaluated to derive a score which in effect a variable symbol that can be generated from the state.

Determining ϵ: Recall that ϵ is assigned to transitions, found during the evaluation step, but either absent in the model or correspond to a weak match,as indicated by a low similarity value, between an NSV v_k and the nearest (as in similarity) node k' of the model. Such ϵ vary in each case and are dynamically determined as follows.

At first, an EMM is built with N nodes and E arcs representing a graph G; from G a perfect graph Ω is extrapolated in which all nodes are connected to one another resulting in Γ total edges. The incrementally added arcs is $\Gamma - E = \Delta$ that are missing in the EMM graph. All arcs including the newly added Δ are assigned a pseudo count value of 1, according to Laplace rule [24]. Over the course of model building, the genuine transitions will have higher count values and therefore higher probabilities where as the missing ones will have a non-zero, yet very small values and therefore ε type small probabilities. A more sophisticated method for treating this is available [24] where the background data distribution adjusts the ε even further. Depending on whether amino acids or nucleotides are used in the sequence input, appropriate adjustment parameters values are used.

In the evaluation step, presence of ε influences its quality of membership in the community represented by the model. The more ε present in the evaluation of a test sequence, the more unlikely the membership. In fact, this is reflected in the metric $distance$ used in differentiation studies which will be discussed later in this section.

Adding Statistical Significance:

This research uses the Karlin-Altschul statistics [25,23] to derive statistical significance for every NSV match against a model. Once all NSVs of a sequence are processed against a model, some general characteristic of the overall significance may be concluded. Since Karlin-Altschul statistics were intended for an classic alignment framework used by BLAST [26,23], some explanation of how the theory applies here is presented here.

In a biological sequence, the letters are assumed to occur independently and any score matrix based on real biological sequence data inherits certain important characteristics [25]. For example, the scores conform to a Gumbel extreme value distribution [27-34,19,20] where the quality of a sequence with respect to its conformance to a score matrix or its associated model is represented by larger scores. Likewise, quality of an alignment can also be scored and analyzed using Gumbel distribution. The characterization leverages the parametric aspect of the Gumbel distribution to establish a baseline threshold for a given desired confidence level based on individual probabilities in a random sequence. In other words, so long as the computed score lies outside the threshold, it is considered “different from the norm" pertaining to an expected random distribution. This allows the search logic to quantify the quality of the “found local alignment" and consider the sequence as a possible homologue; in fact, yet another more subjective threshold is also used at the sequence level to filter out those with insufficient number of local alignments. The result of such rigor is reduction in search space and hence better performance.

BLAST algorithm seeds the search by first considering sliding trimers (i.e.3-letter words) as words for which an established (pre-computed) index is available for the target database of sequences. A partially matched word list is then scored using a score matrix to further screen which of these seed alignments to consider further. The chosen exceed some threshold and are bi-directionally extended as long as possible with increasing quality. Thus local alignment scoring beyond a threshold is a central aspect of finding the homologues pertaining to a query sequence.

EMM Evaluator also deals with a similar problem of finding the matching segments between a query sequence and those of a community of sequences. The difference is that the community of sequences have already been reduced to a consensus form i.e. a profile EMM. The consensus model has some states as clusters of related segments from various sequence members of the community while the other states reflect the unique segments that do not sufficiently belong to any other state clusters. Furthermore, both the query sequence and the model are already in a numerical summary form containing frequencies of word variants. However, the task remains the same as in classic alignment framework, where parts of query sequence (segments or NSVs) are matched (Quasi-aligned) against the states of a model to determine an assessment of membership. Finding the initial local alignment or the match for a query segment or NSV is simply done by searching for the state across the model with the maximum similarity value. This is computed by using the same similarity function used to make up the clusters and the build the model in the first place. For example, Jaccard is prevalently used as the similarity function [35] in EMMBA though other functions are available. The most likely match for an NSV is approximated using the similarity search using Jaccard measure again to narrow down the search space for a score based metric.

Taking classic alignment case, it is necessary to note that alignments tend to score higher than non-alignments because their substitutive value is close to their identity value i.e. if matching A to A would give the identity score, matching A to X (other than A) may still give a high score value depending on the substitution matrix. This means that a perfect alignment or near perfect alignment is expected to give a high score than a random alignment. If the higher score were to exceed a threshold, it would achieve the status of “extreme" and becomes an alignment to consider further.

In case of Quasi-alignment also, the alignment tends to have a high value and scoring the centroid of the matched state could provide a meaningful score. Once the match is found, the match can be scored to determine the quality of the match as well as quality of the target segment or the centroid of the model state matched. The centroid can be scored and compared against some threshold to determine its significance. However such score may not be discriminative enough since it does not differentiate between two similar matching NSVs. As such, p-value or significance measurements are not possible in the absence of non-positional alignment.

Alternately, the significance of the match may also be computed by considering the difference between the matched state centroid and the segment-NSV of the query sequence and then checking to see if the difference is smaller than some minimum threshold [36]. Thus each segment of the query will be associated with a score as well as significance which are available for further analysis. It may be seen that unlike classic alignment framework, the scoring here is done based on word statistics of the target or the differential as opposed to an alignment with substitution scoring.

As mentioned earlier, the score matrix for the Quasi-alignment framework will be based on the word and letter frequencies found in the sequence make-up of a model (EMM). The word variants are independent and

randomly distributed over the entire community of sequences. This satisfies the requirement for Gumbel extreme value distribution.

To determine λ, the equation $\Sigma P_i.EXP(\lambda^{Si}) = 1$[25]where P_i is the probability associated with a unique score S_i. Both P_i and S_i are derived such that for every unique score there is a normalized probability. As it will be seen later, this can be quite involved and often requires numerical method based fine tuning. The core algorithm is presented here along with the iterative fine tuning algorithm in Figure 5 which follows:

Compute Scores&Probs Function

```
Given   p       word length
Given   Score   LOD score matrix of length 4^p
Given   Q       Query:Point Probability vector of length 4
Find    S       Set of Unique Scores
Find    P       Set of score associated probabilities
1 for (next,I = 0 ; I < 4^p; I++)
2   for j = 0 to 4^p-1
3     index = i*4^2+j*4^0
4     if unique(Score[index]) // remembers unique scores
5       P[Score[index]]  += Q[I]*Q[J]
6       S[next++]  = Score[index]
7 normalize(P )
8 return P & S
```

for P = 2 {word length of 2)

```
Given   I       word length
Given   F       word frequency vector of length 4^p
Given   P       Point Probability vector of length 4^p
Find    S       Set of Unique Scores
Find    P       Set of score associated probabilities
1 for (next,I = 0 ; I < 4^p; I++)
2   for j = 0 to 4^p-1
3     for k = 0 to 4^p-1
4       index = i*4^2+j*4^1+K*4^0
5     if unique(Score[index]) // remembers unique scores
6       P[Score[index]]  += Q[I]*Q[J]
7       S[next++]  = Score[index]
8 normalize(P )
9 return P & S
```

for P = 3 {word length of 3)

Prepare Set of unique Scores and Probabilities

```
        Given   Score   LOD score matrix of length 4^p
        Given   P       Set of score associated probabilities
        Given   S       Set of Unique Scores
        Find    P & S   such that E < 0
alpha = 1.0
1 While (E (P,S) > 0))    //E=∑Pi*Si
2   if (alpha*E ) < 0.5 alpha =  1.0
3   Decrease_scores(Score, alpha*E)          //subtract from old scores to create new scores
4   P&S <-  ComputeScores&Prob(Score, S)      //recompute unique scores & probs

5 return P & S  //set of scores & prob are now ready
```

In order to determine the statistical significance of the seed matches (local alignments) which are high similarity scoring NSV-state pairs, Karlin-Altschul statistics can be used; this requires that the expected sum of product of unique scores and their probabilities be less than zero. The algorithms here show how to derive set of unique scores and associated normalized probabilities from a model wide score matrix and query based point probabilities. Variants for word sizes are presented to demonstrate algorithm's simple extendability to multiple dimensions. A fine tuning numerical method is also presented to recover from the case where the expected sum is positive which automatically adjusts the baseline score matrix and re-derives the set of unique scores and their probabilities.

Figure 5: Iterative algorithm for generating valid set of unique scores & probabilities

Karlin-Altschul statistics requires that $E = \Sigma P_i * S_i < 0$. In case, this is found not to be the case i.e. in case the expected score E is > 0, Altschul et al [25] propose that scores be adjusted using $S_i* = S_i - \alpha$*E where $\alpha > 0$. We found that while it is necessary to adjust the scores, it is not sufficient to maintain the integrity of scores & probabilities without re-computing probabilities. Once the probabilities are recomputed and normalized, the expected sum should be rechecked and the process is repeated until convergence. This added refinement is also shown in the algorithm presented above.

Given set of scores *S* and probabilities *P*, λ and K parameters are estimated using publicly available Karlin-Altschul subroutines [37] which have been ported into Java. Thus, λ, K and adjusted Score Matrix *S** are computed for every query.

Given λ, K and S*, Karlin-Altschul theorem I [25] allows computation of Threshold by formula $T_M = -ln((1/NK) * ln(1/c))/\lambda$ where N is the length of the query-segment and c is statistical significance. But this threshold is not applicable for Quasi-alignment where it is not possible to compare position by position as in Classic alignment.

Each selected NSV-state match is scored by scoring the matched-state's centroid which is a vector of mean-frequencies. Since there is no substitution context, question arises whether centroid scoring alone is sufficient to establish any kind of alignment. In order to address this, we instead propose exploring the differential i.e. difference between the centroid of the matched state and the segment-NSV of the query and continuing to use it in an Extreme value distribution framework as far as scores are concerned. Difference based analysis is also noted in biological literature [36,38]. Since difference of zero means perfect match, the smaller the value the better and more significant the match. However, unlike typical Extreme Value distribution where the extreme maximum and extreme minimum values and thresholds are involved, what exists as a distribution in case of difference distribution may be very different.

Here the difference score needs to be less than the threshold T_d. The difference score distribution may be analyzed by examining the mean and the variance. Since the scores being compared come from the same Gumbel distribution, their difference possibly follows the same structure as difference normal distribution where the mean is zero and variance, in case of a standard normal distribution, is 2. It may be noted that mean is zero because difference normal distributions have the new Mean as the difference between the two Means [39].However, the difference between the Means would be zero in our case since the two items being compared are drawn from the same distribution. This is because the Score Matrix is constructed from the model as well as the query and thus forms a common distribution to draw score assignments from. As a next step, we propose the following conjecture.

Conjecture: Difference distribution formed from closely related samples selected uniformly from the same Gumbel distribution may be approximated using a Difference Normal distribution.

The above conjecture considers the fact that the values being compared come into play only when their corresponding centroid vectors are already known to be highly similar; recall that a similarity function is used to establish the matching between an NSV and a model state. This conjecture is experimentally shown to hold true as shown in the Results section.

Since, Gumbel difference is also expected to be zero in case of perfect match or Quasi-alignment,a Mean of zero is applicable. However, the variance derivation for Gumbel difference cannot be possibly be 2 as in Difference standard normal distribution because Gumbel extreme value distribution's cumulative probability function is quite different from that of standard normal distribution.

For a Gumbel distribution, it is given that variance is $\beta\pi/6$ where β is reciprocal of λ [40]. Thus we now have mean and variance for our approximate Normal difference distribution as 0 and $\pi/(6*\lambda)$ respectively. Since the absolute value of the difference is considered in the Difference distribution, another scaling factor defined by $\sigma^2{}_* = \sigma^2 * (1-\sqrt{2}/\pi)$ is applied according to rules of Half Normal distribution [41]. With Difference Distribution's $\sigma^{'}$ and Mean $\mu^{'}$, the p-value for a Quasi-alignment may be expressed as follows:

$$Pvalue = \frac{1}{2} + \frac{1}{2} * ERFC(\frac{x-\mu^{'}}{\sigma^{'}})[?]$$

where ERFC is the complementary Error Function [42]. This p-value represents the significance for the NSV match with a model state. However, the match itself may or may not be significant enough for consideration. A threshold is required to determine if a difference score is significant.

c	σ Multiple		λ						
	Normal	Score	*0.01*	*0.05*	*0.08*	*0.1*	*0.15*	*0.18*	*0.2*
0.01	0.012533	0.013704	1.370381	0.274076	0.171298	0.137038	0.091359	0.076132	0.068519
0.05	0.062707	0.068562	6.856214	1.371243	0.857027	0.685621	0.457081	0.380901	0.342811
0.1	0.125661	0.137395	13.73952	2.747904	1.71744	1.373952	0.915968	0.763307	0.686976
0.9	1.644854	1.798445	179.8445	35.9689	22.48056	17.98445	11.98963	9.991362	8.992226
0.95	1.959964	2.142979	214.2979	42.85959	26.78724	21.42979	14.28653	11.90544	10.7149
0.99	2.575829	2.816352	281.6352	56.32705	35.2044	28.16352	18.77568	15.6464	14.08176
			Difference Threshold						

Derivation

For a Normal Difference distribution between Normal distributions of μ_1, μ_2 and σ_1,σ_2,

the resulting $\mu = \mu_1+\mu_2$

the resulting $\sigma = sqrt(\sigma_1^2+\sigma_2^2)$

if samples are from the same distribution,

$\mu = 0$

$\sigma = sqrt(2\sigma_1^2)= \sigma_1*sqrt(2)$

Following the same for Score Difference distribution

$sigma_{difference} = \sigma_1*sqrt(2)$

$= ((pi/sqrt(6))/\lambda)*sqrt(2)$

$= (pi/sqrt(3))/\lambda = \frac{1.813799}{\lambda}$

Since $|\sigma_{difference}|$ will be used, a scaling factor needs to be applied as per Half normal distribution

where $\sigma^{2*} = \sigma^{2*}(1-2/pi)= \sigma^{2*} 0.36338$

$= \sigma^{2*} 0.60281^2$

Now

$sigma_{difference} = \frac{1.813799}{\lambda} *0.60281= \frac{1.093377}{\lambda}$

multiplying $\sigma_{difference}$ with chosen n-multiple gives the Threshold at which significance can be ascertained.

The table gives the reverse relation of sigma multiples corresponding to a few often used values for the area under the normal curve. These values are useful to approximate (asymptotic) confidence intervals of the specified levels based on Normal distributed (or asymptotically Normal) estimators. The distribution basis here is score difference distribution which is approximated as a difference normal distribution. This is used to establish a threshold close to zero to determine the significance of matches found between a query sequence and a model profile of a sequence community.

Figure 6: Sigma multiples for Gumbel Confidence Intervals

According to normal distribution [43], the 95% interval falls at inverse Error function mapping for 0.95 i.e. ERFINV for 0.95 at σ' which is standard deviation for a difference distribution where mean is zero and λ is 1. The Figure 6 tabulates the σ multiples to use and also presents the derivation for the sigma multiple for score

difference distribution. The σ multiples define the threshold below which a difference score is considered significant since we are dealing with a difference distribution.

Since we are really interested in the values close to zero which is also the mean, we choose the multiple associated with 0.05% which includes only the top 5% of the matches that have a difference close to zero. For example, to establish the 95% confidence difference threshold for the Gumbel difference, we use the following according to Figure 6.

$$T_d = \frac{0.07}{\lambda}$$

Where λ is a scaling factor and is dependant on combination query (sequence being evaluated) as well as the fine tuned score matrix for the model (or EMM). T_d is the difference threshold below which the difference scores are considered significant. The number of matches or the number associated with the top 5% matches is used in computing the sequence level significance.

In summary, the difference threshold used will be based on a much sharper Gumbel difference threshold as derived above.

Sequence Significance and E-Score: With difference threshold in place, the number of significant Quasi-alignments m can be determined by simply comparing the score difference for each-matched state pair against the threshold. The ones that fall below the difference threshold are flagged as meeting the significance criterion.

Karlin-Altschul statistics proposes using the following formula for determining the significance of multiple local alignments thus producing an overall P-value.

$$P - Value = 1 - e^{-y} * \Sigma_{i=0}^{m-1} y^i/i!$$

Where , $y = KNe^{-\lambda S}$. For m=1, the above simplifies to the original form for the P-value of single local alignment. However, this is not applicable in our case of difference distribution which is considered a normal difference distribution. We propose that the sequence level significance be computed as the average significance of all significant matches.

Given m* the number of significant matches, the sequence level significance is computed as follows

$$Significance_S = \frac{\Sigma_0^{m*-1} Significance_{quasi-alignment}}{m*}$$

E-score “E" calculation as proposed by Karlin-Altschul statistics i.e. $Pvalue = 1 - e^{-E}$ applies equally well for the Quasi-alignment case presented here as follows:

$$Significance_S = 1 - e^{-E}$$

Word length versus significance: In all the above formulations, extreme value distributions require query lengths to be long and it is to be expected that word lengths of 3+ should achieve higher significance since word length of 2 yields NSV length of only $4^2 = 16$ which is much smaller than word length 4 that yield NSV of 256.

NSV length versus Significance: As the word length increases and segment length over which Numerical Summarization is performed, an NSV may contain zero values for some of the word (variation) counts. This tends to effect the quality of λ and therefore the accuracy of significance.

Karlin-Altschul statistics addresses similar issue with the inability to account for edge effect where alignment may not have sufficient room to complete toward maximum score. The recommended length correction is given by $n' = n - ln(K * n)/H$ where H is information entropy which is derived at the same time other parameters (λ and K) are derived [25].

Since, presence of zeroes for some counts in an NSV also does present loss of information, similar correction can also be attempted for Quasi-alignment framework. However, since λ values can get quite small resulting in small values for H, the length correction is not always applicable. As such, the length correction will be applied only where it is needed and meaningful. This is determined by checking to see if the resulting correction still maintains the effective length with in acceptable range i.e. $0 << n' < n$.

Alternately, using EMM form of sequences under evaluation reduces the complexity since the NSVs corresponding to the state centroids tend to hold statistics averaged over a state's cluster thus eliminating much of the zero counts issue.

In summary, extension of Karlin-Altschul statistics should be possible to the Quasi-alignment framework proposed by EMMBA as follows:

- LOD score matrix concept is applicable so long as it is derived for each model and iteratively adjusted to meet the Karlin-Altschul assumption for negative expected score. Symmetric property may or may not be necessary. In addition, the scores are not based on evolutionary substitution statistics, but on simple sequence community relevant word statistics.
- A set of probabilities is associated with a unique set of scores from the score matrix which is adjusted based on point probabilities from the query; however, these also need iterative fine tuning if the scores are adjusted to meet the negative expected score requirement.
- Seeding to determine the most likely match to consider further may done by using the similarity metric. The maximum similarity producing NSV-Model state is considered the equivalent of High Scoring match Pair (HSP) for which significance may be evaluated.
- Score difference is considered a better indicator for determining the significance since there is no way to compute consensus score based on substitution in a non-classic alignment situation. The difference between the scores of matched state's centroid and the NSV (word frequencies of a sequence-segment) is assumed to approximately follow a Normal difference distribution.
- Difference Normal distribution analysis is extended to Difference Gumbel distribution to determine the distribution parameters such as the Mean and Standard Deviation. Similar extension is also done to derive the sigma multiples for confidence intervals. Empirical observation as well as formal Normality tests based on Anderson-Darling [44] appears to confirm these assumptions.
- A threshold is established by scaling the desired sigma multiple associated with confidence using Gumbel parameter λ which is computed for every query-model pair.
- Quasi-alignment is considered significant if associated score difference is below the established difference threshold.
- Number of significant Quasi-alignments is computed and used in deriving the sequence level significance as an extension from Karlin-Altschul statistics for multiple local alignments.

- Standard E-score to P-value computation from Karlin-Altschul statistics is used.

Methods for Applications of EMMBA

Unknown Sequence Identification: To identify the species of a 16s rRNA sequence family, the following method is to be used. First, a complete library of signature EMM profiles for all Species is to be generated using all available 16s rRNA sequence data for training. Next, the sequence information from an unknown origin, presumably still microbial, is converted to NSV form and then to an EMM. The EMM representation of the unknown origin may then be evaluated against each and every one of the Species in the library. The maximum scoring signature profile with acceptable significance becomes the identification. The same technique is used for novel species identification as well in which case since there is model, the significance level is expected to be lower.

Known 16s rRNA Sequence Sample Classification: Sequence classification refers to determining the taxonomy of a given sequence given that its class already is known and available ahead of time. EMMs could be built for various taxonomic levels based on available sequence data in the NCBI database. The models can even be fine tuned by experimenting with different parameter settings, resolution levels and validated using rigorous 10x cross validation. Once finalized, a final set of models can be built into a library of EMMs which is then available for quick classification of lab sample of sequences belonging to several different organisms. The EMMBA evaluator can be configured to process such samples and find a likely classification. Since classification choices are ranked, additional information is available to allow for analyzing the related classes to a sequence of interest. It will be possible to answer questions like 1) what is the statistical significance that the sequence belongs to a certain class or 2) which classifications are most likely for a sequence at 90% statistical significance. This functionality is most useful for Metagenomic classification [45] where only partial sequence data may be available as opposed to complete genomes of the microbes found. The research assumes that targeted Metagenomics is used i.e. the sequence samples contain 16s rRNA header.

Sequence Sample Differentiation: Community differentiation allows for study of phylogeny where the smaller differences group some in one branch and the others in different branches. This is also referred to as all-against-all analysis to establish a distance matrix for further analysis. In this case, analysis extends to the many 16s rRNA sequences an organism may have. This helps in studying how easy or difficult it is to differentiate species or genera. For example, it is known for some time that it is difficult to resolve the Bacilli strains. This can be analyzed using the differentiation function. Another application is for Metagenomic differentiation [45] where different Metagenomic samples are to be compared. In this case, each sample is subjected to all-against-all analysis deriving some metric (discussed in the Metrics section) of intra-sample diversity which is used in comparison.

Metrics

All sequence evaluations are quantified using various metrics. The metrics are categorized into two groups. The basic group provides metrics at individual organism-model pair level and are useful in establishing a rank order. The aggregate group offers metrics to characterize a whole sample of organisms useful for comparative genomic analysis.

Basic

There are four basic metrics proposed to assign a numeric assessment to evaluation of an organism against a model(EMM). In all cases, the EMM or the model that is associated with the largest metric value is considered

the most likely class. Two of the metrics SumScore and DiffScore are derived using Karlin-Altschul statistics while the remaining two are derived using traditional methods.

SumScore: SumScore is computed as

$$\Sigma_{i=1}^{K}\Sigma_{j=0}^{L-1} log(X_j^* * C_{i,j}] * a_{j-1,j} * \zeta_j$$

Where K is the number of segments in a query sequence, L is 4^p with p being the word length (length of the NSV), X_j^* is the adjusted score for j^{th} count from the score matrix and $C_{i,j}$ is the j^{th} count in the matched-state's centroid for the i^{th} NSV. ζ is the similarity value found for the match at j^{th} state and $a_{j-1,j}$ is the Markov Transition Probability. The model scoring the highest SumScore is considered the most likely class for a query sequence.

Probability: Sum of Log Probability $\sum_1^K log(a_{i-1,i})$ is calculated by summing over all transition probabilities greater than a specified Threshold. The ones that fall below the threshold are given an extremely low score ε, but still added in to the metric. The model scoring the highest probability metric is considered the most likely class.

Propensity: Sum of Log Propensities $\sum_1^n log(a_{i-1,i}) * \zeta_i$ is calculated by summing over the products of transition probabilities and the associated similarity quantifier. The ones with no transition support in the model are assigned a very low score ε to the product element before adding into the metric. The model scoring the highest propensity metric is considered the most likely class.

Diffscore: DiffScore is computed as

$$\Sigma_{i=1}^{r*}\Sigma_{j=0}^{L-1} X_j^* * log([k'_{c(i,j)} - v_{i,j}]] * a_{i-1,i}$$

Where r^* is the number of segments in a query sequence with difference scores less than the minimum difference threshold, L is 4^p with p being the word length, X_j^* is the adjusted score for j^{th} count from the score matrix, $k'_{c(i,j)}$ and $v_{i,j}$ are the j^{th} counts in the matched-state's centroid and the segment-NSV in context. $a_{i-1,i}$ is the associated transition probability. The model scoring the smallest differential Score is considered the most likely class for a query sequence.

Apparent Distance Ψ*:* Distance between any two EMMs *e1* and *e2* is defined as $\Psi_{e1}e2$+$\Psi_{e2}e1$ where $\Psi_{e1}e2$ evaluates EMM 2 against the host model EMM 1. Apparent distance is limited to evaluation from one side only.

Distance is computed as a result of mutual evaluation of each other's EMM where the comparables are all in EMM form. In such evaluations, each EMM measures its apparent distance in terms of score differentials in all matches depending on whether each match is followed by a supported Markov transition in the host EMM. In cases where such transition is found, the score differential between the matched states is taken and where such isn't found, a penalty is applied to the score differential of the match pair. The penalty ξ is computed as the $log(1/\varepsilon)$ which is applied as a multiplicative factor to the score differentials with transition. The long formula for this metric may be expressed by

$$\Sigma_{i=0}^{K'}\alpha + \Sigma_{i=0}^{K-K'+1}\beta \text{ where}$$
$$\alpha = \Sigma_{j=0}^{n-1} X_j^{**} * log([k'_{c(i,j)} - v_{i,j}])$$

Which aggregates the differential scores of matches supported with a valid transition known to exist in the host EMM and

$$\beta = \Sigma_{j=0}^{K-K'+1} X_j^{**} * log([k'_{c(i,j)} - v_{i,j}]) * \varepsilon]$$

Which, scores the differential scores of matches with unsupported transitions i.e. the transition from the previous state to the current matched state is not one that is known to exist in the host model (EMM).

Notation: K is 4^p with p being the word length, K' is the number of segments with supported transition and (K-K') is the number of segments with unsupported transition. $X_j^{*}*$ is the joint score for j^{th} count from the score matrix, $k'_{c(i,j)}$ and $v_{i,j}$ are the j^{th} counts in the matched-state's centroid and the segment-NSV in context. ξ represents the penalty factor to account for increased differential due to unsupported transitions. This metric is used for building the distance matrix from which Phylogeny can be inferred or recovered.

Aggregate Metrics

On an overall assessment of a sample evaluation against a library of models, there are four different aggregate metrics proposed. Of these, two are c-score [46] based though these are derived based on the third aggregate metric - classification accuracy. Another new metric called Delta value is more applicable for differentiation which involves distance matrices.

Classification Accuracy: Based on each basic metric, once each organism is assigned its most likely class or model, the overall accuracy of such assignments can be reported. This is possible in cases where the organism name itself holds the key to what the right class should be and the cases or the metric buckets, where this correct model is not chosen contribute to error for that metric. Correctly classified percentage of organisms for each basic metric is reported as the classification accuracy.

Compatibility-Score (c-Score): c-Score measures and reports the difference in non-trivial splits as a metric [46]. Though it usually does this by examining a phylogenetic tree, it may be derived just as easily using the classification accuracy metric itself as shown in Figure 7.

Thus c-score can be computed by $1 - 2E/N_s$ where $E = 1 - classification\_accuracy$ and N_s is the number of phyla and classes with more than one organism.

weighted c-Score: The weighted c-Score, proposed here, takes into account that all error are not the same and that some are more trivial than others. By assigning severity weights to the type of error, a more conservative error estimate is indicated. For example, a misclassification resulting in a phylum level error is assigned a severity weight of 1.5 where as the same at Species level is given a value of 1.

The formula is $1 - 2E^*/N_s$ where E^* is computed based on the type of error and the level of classification desired. For example, if reporting classification accuracy at the class level, the weighted error would include contributions from class and phylum level errors only with Phylum level error having a larger weight.

Suppose EMMBA classifies taxa up to Genus level which means that it places the OTUs (Operational Taxonomic Units) in the appropriate Genus of appropriate Class, Phylum and Domain.

EMMBA classification tree can be up to 4 levels deep.

A non trivial split is a partitioning of the tree where both partitions contain more than one member.
Number of non trivial splits in a Genus differentiation taxonomy tree may be computed as:

Number of Genera with more than one member (OTU)+
Number of Classes with more than one OTU +
Number of Phyla with more than one OTU+

or alternately as:

$N_s = \sum|G,C,P>1|$ where G, C, P represent Genera, Classes and Phyla respectively.

A misplaced OTU causes L number of invalid orincompatible non-trivial-splits or alternately expressed as:

$|incompatible| = E*(L-1)$ where L is the depth of classification tree [4 for a Genus tree]
E is the number of misclassifications

c-Score is defined as (number of compatible non-trivial splits)/(total number of non-trivial splits) or alternately:

$|c\text{-}Score| = |compatible|/N_s$
$=(Ns - |incompatible|)/Ns$
$=1- E*(L-1)/N_s$ where E is the number of misclassifications
L is the depth of the tree (=4 for Genus tree)
N_s is the num of Genera, classes & phyla with >1 organism
$= 1- 3E/N_s$

cScore [46,2] is intended to compare two trees to check for differences in terms of non-trivial splits. Taking NCBI classification as the gold standard, a new classification's tree is measured using cScore. Since, EMMBA methodology already uses the NCBI classification as implied in the name of an organism itself; its metric for measuring classification accuracy already captures pertinent information useful for computing c-score without generating a phylogenetic tree. The derivation for this is shown in the above analysis for a Genus level classification. To extend this for a class level tree, subtracting 1 from the number of levels will suffice.

Figure 7: Deriving cScore from Classification accuracy without generating Phylogenetic tree

Distance Matrix: is an aggregate representation of the inter-distances among all the sequences being differentiated. For example, if n sequences belonging to n organisms are being differentiated, a distance matrix would contain nxn elements with each element containing a distance value. There are two types of distance matrices possible, one that is asymmetric which records the apparent distance between sequences and the other that is symmetric which records the true distance i.e. summation of relevant apparent distances. The matrix may then be uploaded to any Phylogeny inference package to generate phylogenetic tree such as the PHILIP [47].

Delta Value δ: δ measures [48] the tree likeness of a distance matrix, so this measure is more applicable in case of differentiation where such matrix is the output. It is computed by taking one quartet (four points) at a time from the distance matrix and computing the δ over it and then taking the mean of all such measures. The formula is given by

$$\delta_m = \Sigma_1^{\binom{N}{4}}[(d_{xv|yu} - d_{xu|yv})/(d_{xv|yu-d_xy|uv})]/\binom{N}{4}$$

Where, the notation of $d_{xv|yu}$ means $d_{xv} + d_{yu}$ and the notation d_{xv} is simply the distance between x and v points in the distance matrix. The δ_m lies between 0 and 1. Larger the value less tree like the sample is. This is useful in Metagenomic differentiation studies where the sample characteristics are compared. Different values could mean more or less diverse mix of microbes in the sample.

Software Tools

R is a free software environment for statistical computing and graphics [49] used for generating the NSV and IPV files of this research. A Java program called EMMBA is used to build EMMs to perform classification, differentiation and identification experiments; the program is not published externally outside the Southern Methodist University as of Nov 12, 2009.

AGATE Statistical Analysis is an EXCEL program available on the web [44] for Normality test using Anderson-Darling method.

AISEE Graph Visualization is a commercial software [50] for generating EMM network graphs shown in Figure 9.

PHYlogenetic Inference Package (PHYLIP) is a website hosting the server for Phylogenetic tree generation using Neighbor-Joining Method [47].

RESULTS

Implementation

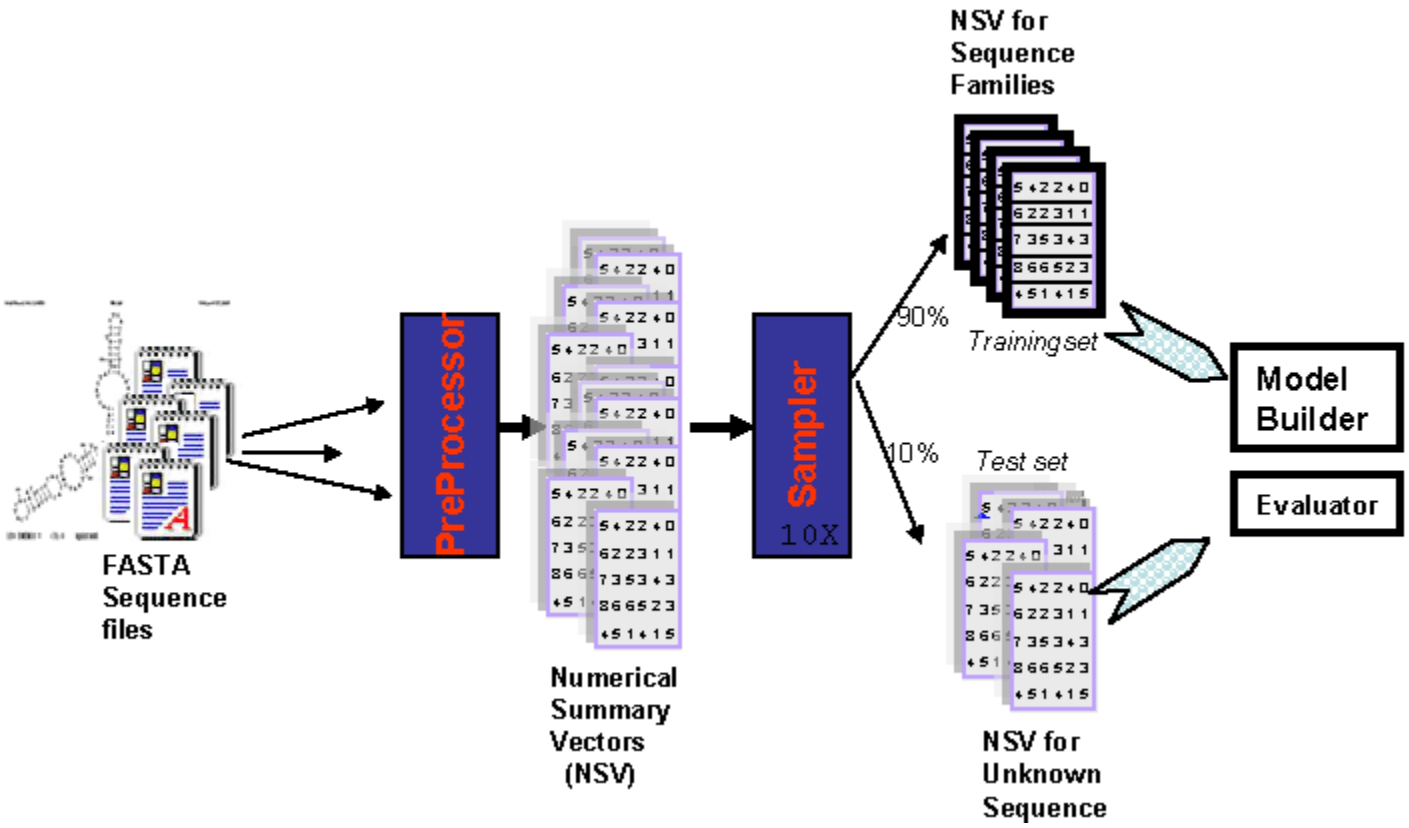
The 16s rRNA Database utilized in this analysis is derived from the NCBI FTP site [51]. The database consists of individual files, one per microbial organism, in FASTA format. The format uses improved headers for subsequent data preparation processing prior to using them in Extensible Markov Modeling and evaluation steps.

The original dataset called ORG, which was derived from the NCBI as of August 2009 and consists of 782 organisms each with multiple 16s sequences where applicable. The FASTA header for each file contains five pieces of information: Phylum, Class, Genus, Species and Organism name as used in NCBI FTP Database. The Database is used in generating the training communities of sequences as well as a set of randomly selected test organisms to be used in subsequent classification experiments.

Preprocessing 16s Sequences

We found that some of the header information in the NCBI organism was missing in some cases. There were several cases of missing Genus or even Class information. Since this type of information is used for automatically verifying classification results, such data is excluded from analysis. The final database consisted of 676 FASTA files with one per organism.

The ORG Dataset is first pre-processed and then separated to facilitate multi-step (10x method) model building and model validation as shown in Figure 8.



Once FASTA formatted multi-copy 16s sequence files for each organism are converted to their Numerical summary equivalents, they are then randomized, divided into 10 non-overlapping partitions with each containing 10 %. Following the 10x methodology, one of the partitions is selected as test candidates while the remaining 9 are used for training i.e to build models (EMMs). The process is repeated by selecting a different partition for test and the remaining other 9 for training. When all partitions finish taking turns as the test partition, the 10x cross validation is said to be complete and the results are averaged over the 10 runs. This method is done to add the necessary rigor and eliminate any bias in the results.

Figure 8: overall process for preparing training & test datasets from 16s rRNA database

Modeling sequence communities with the Extensible Markov Model follows conversion of nominal sequence data into numeric form. More appropriately, the sequence data is actually converted to Numerical summary form i.e. to count vectors where counts represent the number of times a particular nucleotide pattern of certain length occurs within a segment of a predetermined length in given sequence of nucleotides. For this study, we opted to divide the 16s sequence into equal size segments of varying sizes (20, 80) with a pattern width of 3. Since a 16s rRNA sequence on average is approximately 1542 nucleotides long, the number of segments ranges from 77 to about 192. For a pattern width of 3 if chosen, there will be 4^3 = 64 different types of counts for each segment corresponding to 64 variations of nucleotides.

Numerical Summarization: For a pattern width of 3, each 16s Sequence is thus converted into several Numerical Summarization Vectors (NSV) of size 64. In case of multiple 16s sequences for an organism, the vectors for each sequence are captured serially with a start count vector at the beginning.

Sequence selection for training & testing: Once all organisms' 16s sequences are converted to their numerical summary vector representations, they are then further organized into training and test sets as shown in Figure 8.

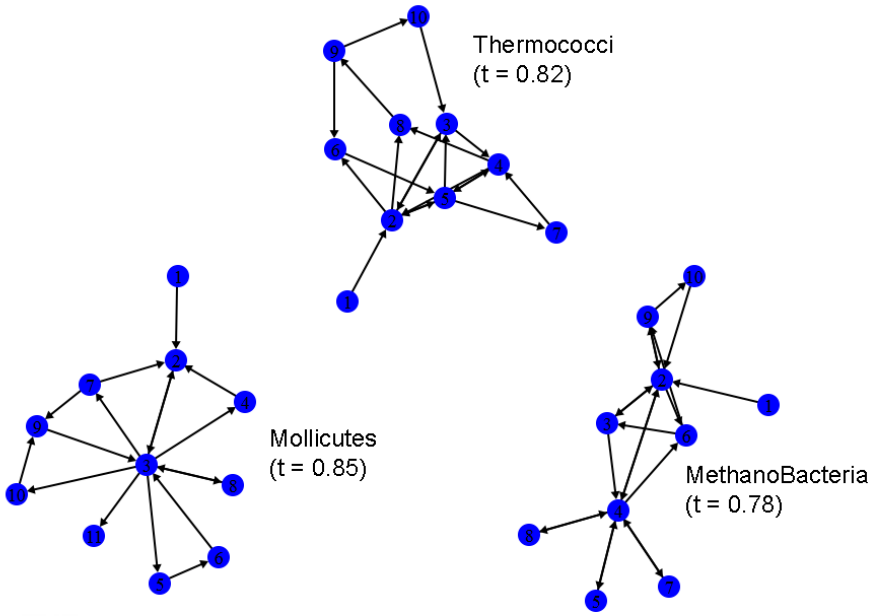
Data files are sampled to select one tenth (default) of available organisms for test leaving the remainder for training in a 10x cross validation scheme [52] the scheme allows for averaging the results over 10 independent analyses of the same data by dividing the data into 10 equal partitions each of which takes turn in being used as

a test partition. For each run, once the test partition is set aside, the remaining nine partitions become the training pool. The training set is used to build aggregated models at a required granularity. For example, models can be built at Class level, Genus level or even at the Species level. In case of an aggregation at Phylogentic Class level, thus each Class will have all its organisms with their 16s sequence data in numerical summary form except for those selected for testing.

The training classes are then used to build EMM models. The models are built by processing numerical summary vectors of each 16s sequence belonging to a training class, referred to as sequence community.

Building Phylogenetic Sequence community Models

When all sequences along with their numerical summary vectors are thus processed, the model is said to be complete and representative of the sequence community or the corresponding phylogenetic class. Throughout the building process, a Score Matrix is also updated which is subsequently used for deriving a match score and a statistical significance level. The scores are in log-odds (LOD) form and reflect the values to be assigned for a word or pattern count.



Phylogenetic Class models built with EMM are directed weighted graphs with controllable cluster similarity threshold for visualization. For example, by adjusting the Jaccard threshold T, the model graphs are generated for the Phylogenetic classes - Mollicutes, MethanoBacteria and ThermoCocci. These are for visualization purposes only. For classification analysis, a baseline threshold like 95% is used across all model building. Typically, higher the threshold, larger the size of the model graph and more the information content useable for species differentiation. The graphs are generated using AISEE software package using file output by the EMM model builder.

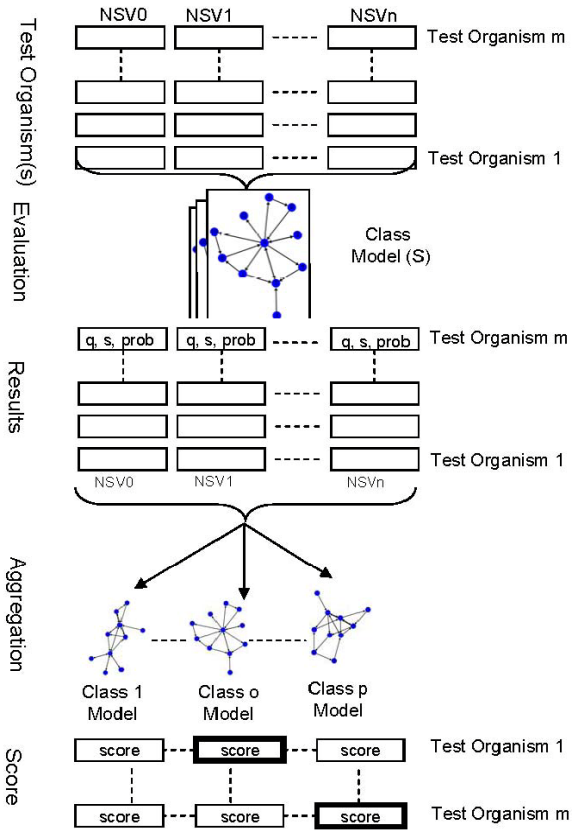
Figure 9: Phylogenetic class model visualization

The class model thus built can be interrogated for transition probabilities of any arc in the model graph. As described earlier, the arc probabilities are derived simply by dividing the number of times the arc is traversed by the total number of times the *"from-State"* is matched. These transition probabilities are used to filter out segments of member sequence that do not follow the expected transitions in a host community model. These probabilities in fact bring the specificity that is required to establish valid membership in a community of sequences.

Structure of EMM Models & Visualization: Extensible Markov Models are directed connected graphs with weights on each arc and nodes. They can also be used for visualization of an entire Phylogenetic class. For visualization purposes, it would be preferable to limit the number of state cluster nodes of the graphs to a manageable number (less than 20). This is possible by experimenting with the Jaccard similarity threshold at which the NSVs are processed to match existing state clusters in the model graph. The Figure 9 shows equal size graphs for Phylogenetic Classes Mollicutes, Thermococci and Methanobacteria.

Evaluation/Classification of Test Organisms

The FASTA formatted test files are then converted to their Numerical Summary Vector (NSV) form as illustrated in Figure 2. After sampling shown in Figure 8 and model building shown in Figure 3 the test-set is ready for evaluation against all the models. For each model, each test organism in NSV format is compared one NSV at a time recording the most matched state and the transition to it from the previous matched state. The evaluation process is as shown in Figure 10.



The Test organisms are processed one at a time starting with the first numerical summary vector (NSV). Each NSV is searched against all states of the model looking for most similarity. The matching state (s) with the highest similarity, the similarity value (q) and the transition probability (prob) are recorded for each NSV. Once all NSVs are processed with all result triplets recorded, next organism and its NSVs are then evaluated. Once all test organisms are thus processed, the results will be used subsequently to derive model scores. The overall score for test sequence is computed for each class model according to the metric used. Once all test sequences are thus processed, for each organism, the model with the best score becomes the chosen model which will be subsequently used in determining the classification accuracy.

Figure 10: Evaluation of Test Organisms to derive model scores

Evaluator first records, for each NSV of the test organism, the most similar state, the similarity quantifier and the transition probability from previously matched model state to the current most similar state. Once all NSVs are processed, metrics are computed for ultimate rank calculation as shown in Figure 10 that determines the best

class model for the test organism. The metrics calculated are *Sum of Log Probabilities and Sum of Log Propensities*.

Classification Experiments & Results

Validating the Difference Distribution: As discussed in the Results section, the difference distribution formed from taking the difference between an NSV segment and a matched state of the model is assumed to approximately follow Normal distribution. The Figure 11 is output of a Darling-Anderson test [44,39].

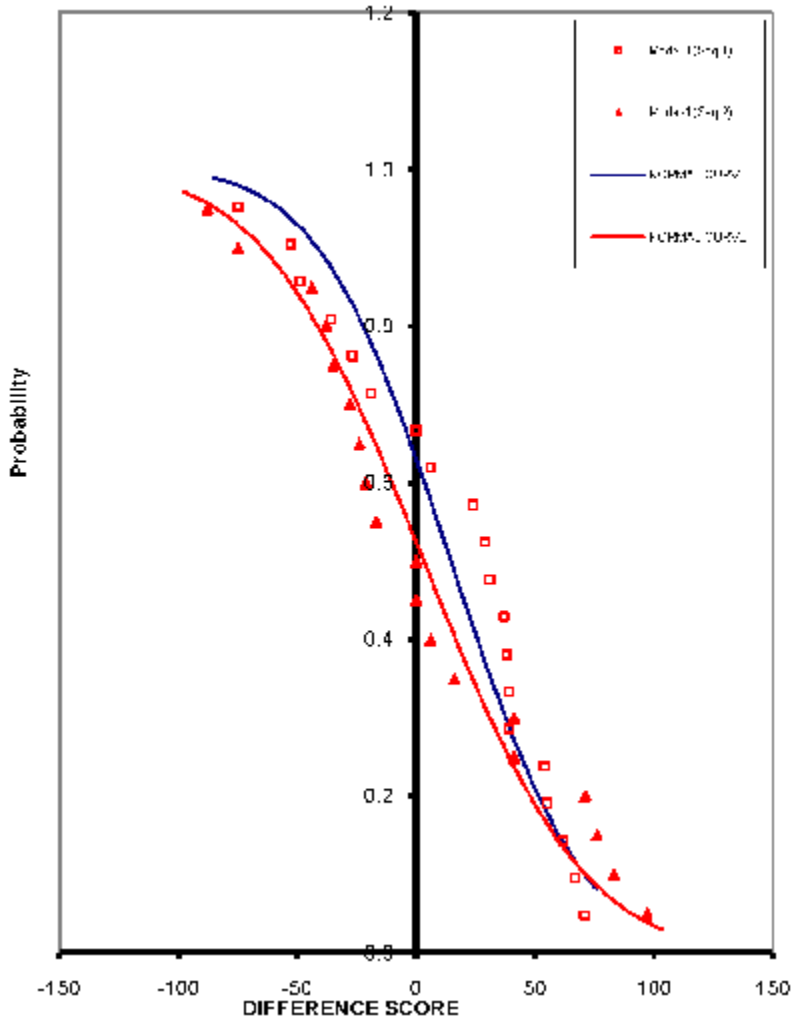
DATA SUMMARY		
	QUASI-ALIGNMENT	
STATISTIC	<MODEL 1 SEQ 1>	<MODEL 2 SEQ 2>
Sample Size	20	19
No. of Batches	1	1
Mean	14.65	3.21
Std.dev	43.84	53.04
% Co. Variation	299.23	1652.05
Minimum	-75.00	-88.00
Maximum	71.00	97.00

Anderson-Darling Test for Normality(α=0.01)

O.S.L	0.0725	0.3776
Normality is	Acceptable	Acceptable
O.S.L. for pooled data is	NA	

Check for Normality based on graphical method

Pearson Coefficient r	0.9678	0.9851
Normality is	Acceptable	Acceptable
r for pooled data is	NA	



The difference distributions from two sequence evaluations against a model are tested for Normality using Anderson-Darling test [Anderson1952,i]. Results show that Normality is acceptable.

Figure 10: 10x cross validation results and aggregate metrics

Metrics Used: Sum of Log Probability, Sum of log Propensities, SumScore and DiffScore are computed and reported for each pair of model & test sequence. The model scoring the maximum for each metric class assigned to a test sequence becomes the classification result. Once all test sequences are evaluated and classification determined, aggregate metrics are used to analyze the overall performance of the classifier.

Experiments: Sampling the entire 16s genomic database excluding those skipped due to missing header information resulted in 10 partitions of 67 organisms each. These partitions are then used for 10x cross validation. Building sequence models involves selection of sequences from all partitions except the one used as test partition. First author's earlier experience with microbial classification [5] indicates that the best performance of a classifier is achieved when

classification is attempted at a sub-class i.e. at a level lower than the desired classification level. Since Genus is a more granular level than Phylogenetic Class, it will be used as the classification level which will provide a more targeted classification. Since there are 246 sub-classes i.e. Genera and only 33 phylogenetic classes in the dataset, this approach makes sense in achieving the desired goal.

Each classification is evaluated against the intended label at Phylum and class levels. If the match occurs, classification is considered successful. The match at Genus level is not considered here since our interest in classifying the organisms into correct phylogenetic classes. As such, a mismatch at Genus level may still put the classification in the correct phylogenetic class as demonstrated in [5] by the first author.

10x Cross Validation Results Summary:

		SumScore	∑log Prob	∑log Prop	DiffScore			SumScore	∑log Prob	∑log Prop	DiffScore
PHYLUM	*orphans+*	94%	98%	94%	93%	**PHYLUM**	*orphans-*	94%	99%	95%	93%
CLASS	*orphans+*	92%	97%	93%	90%	**CLASS**	*orphans-*	92%	98%	94%	91%

Word Length: 3 **Segment Size: 80** **c-Score 0.90** **Weighted c-Score 0.88**

Individual Runs are:

		SumScore	∑log Prob	∑log Prop	DiffScore			SumScore	∑log Prob	∑log Prop	DiffScore
PHYLUM	*orphans+*	96%	99%	96%	93%	PHYLUM	*orphans-*	97%	100%	97%	94%
PHYLUM	*orphans+*	94%	94%	91%	88%	PHYLUM	*orphans-*	97%	97%	94%	91%
PHYLUM	*orphans+*	96%	100%	96%	93%	PHYLUM	*orphans-*	96%	100%	96%	93%
PHYLUM	*orphans+*	91%	99%	96%	91%	PHYLUM	*orphans-*	91%	99%	96%	91%
PHYLUM	*orphans+*	90%	96%	91%	91%	PHYLUM	*orphans-*	90%	96%	91%	91%
PHYLUM	*orphans+*	90%	99%	97%	96%	PHYLUM	*orphans-*	90%	99%	97%	96%
PHYLUM	*orphans+*	93%	100%	90%	90%	PHYLUM	*orphans-*	93%	100%	90%	90%
PHYLUM	*orphans+*	96%	97%	93%	94%	PHYLUM	*orphans-*	96%	97%	93%	94%
PHYLUM	*orphans+*	96%	100%	97%	96%	PHYLUM	*orphans-*	96%	100%	97%	96%
PHYLUM	*orphans+*	94%	96%	96%	91%	PHYLUM	*orphans-*	94%	96%	96%	91%
CLASS	*orphans+*	96%	97%	96%	93%	CLASS	*orphans-*	97%	98%	97%	94%
CLASS	*orphans+*	93%	93%	88%	81%	CLASS	*orphans-*	95%	95%	91%	83%
CLASS	*orphans+*	94%	99%	94%	91%	CLASS	*orphans-*	94%	99%	94%	91%
CLASS	*orphans+*	87%	97%	94%	88%	CLASS	*orphans-*	87%	97%	94%	88%
CLASS	*orphans+*	87%	93%	90%	88%	CLASS	*orphans-*	87%	93%	90%	88%
CLASS	*orphans+*	87%	99%	96%	90%	CLASS	*orphans-*	87%	99%	96%	90%
CLASS	*orphans+*	91%	100%	88%	88%	CLASS	*orphans-*	91%	100%	88%	88%
CLASS	*orphans+*	88%	96%	90%	90%	CLASS	*orphans-*	89%	97%	91%	91%
CLASS	*orphans+*	94%	99%	94%	93%	CLASS	*orphans-*	94%	99%	94%	93%
CLASS	*orphans+*	90%	96%	96%	91%	CLASS	*orphans-*	91%	97%	97%	92%

The summary of the 10X cross validation runs and the individual run's results clearly show that EMMBA classifies the microbial organisms at the Phylogenetic class reasonably well. In general, class prediction is seen to be more effective when classified at a lower sub-class level such as genus in this case. Kotamarti et al [5] have previously demonstrated that Sub-class level classification yields better performance. Since there are 33 classes and only 12 of them are over 10 members each as opposed to 246 sub-class or Genera, the technique is expected to perform better. The high success rate at 90% or higher confirms the success of the approach. In addition to the accuracy metrics, also reported are aggregate metrics c-Score and Weighted c-Score that measure compliance of classification results with NCBI taxonomy [46]. Weighted c-Score penalizes more for errors at higher taxa than lower taxa.

Figure 11: 10x cross validation results and aggregate metrics

10x cross validation was performed on the overall dataset at Genus level and the results of the 10 independent experiments are recorded in Figure 12. Orphan qualification against the results reflects the fact that cardinality controlled 10x sampling sometimes results in skipping the creation of certain community sequence models. For example, there are 246 Genera out of which approximately 60% have no more than one organism; if a single member

Genus is in the test partition, there will be none in the training set for building a model. When there is no model for a test organism, it is called an orphan. However, classification results are shown with and without accounting for orphans. In the Figure 12 $Orphan+$ refers to results where accuracy rate is lower because there are no models for some test organisms to classify into. It makes sense to consider results for which orphans are accounted for only. These results are identified by "*orphans-*".

Organism	SumScore	Probability	Propensity	DiffScore
Chlamydophila-pneumoniae-TW-183	PCGS(1.00,12)	PCGS(1.00,12)	PCgs(0.47,1)	PCgs(0.47,1)
Bifidobacterium-longum-subsp--infantis-ATCC-15697	PCGS(0.93,3)	PCGS(0.93,3)	PCGS(0.93,3)	PCGS(0.93,3)
Borrelia-burgdorferi-B31	PCGs(0.96,3)	PCGs(0.96,3)	PCGs(0.96,3)	PCGs(0.96,3)
.....				
....				
Legend:				
P implies correct classification of Phylum	**p** implies incorrect classification of Phylum			
C implies correct classification of Class	**c** implies incorrect classification of Class			
G implies correct classification of Genus	**g** implies incorrect classification of Genus			
S implies correct classification of Species	**s** implies incorrect classification of Species			
(X,Y) where **X**= p-Value or Statistical Significance	**Y**=E-Score			

Successful classification is indicated by the presence of an uppercase letter for a given level; for example, P implies that Phylum level classification is successful and that a lowercase p implies an unsuccessful classification. The numbers in the bracket represent the significance in terms of sequence p-value and E-Score respectively. Only three organisms are shown as examples, but infact, there were 670 organisms being classified into 33 phylogenetic classes.

Figure 12: Classification (partial) success with statistical significance and E-score

The output for individual classification of an organism is shown in Figure 13. As seen in the Figure 13, classification (for output of 3 randomly selected organisms) is shown to be successful if the corresponding level is shown is uppercase. For the levels not classified correctly, the corresponding letters are shown in lowercase. Such result of classification is shown for each Metric. Also shown is Statistical Significance for the classification as observed by score differential assessed against a difference threshold. E-Score is also shown next to the significance level. Both P-value and E-Score are at sequence level and are derived using Karlin-Altschul statistics [25] which utilizes the presence of multiple matches (local alignments) between a query sequence and a model sequence. Where the significance is high, the result reported is statistically justified; otherwise, more information may be required to arrive at a conclusion regarding the classification.

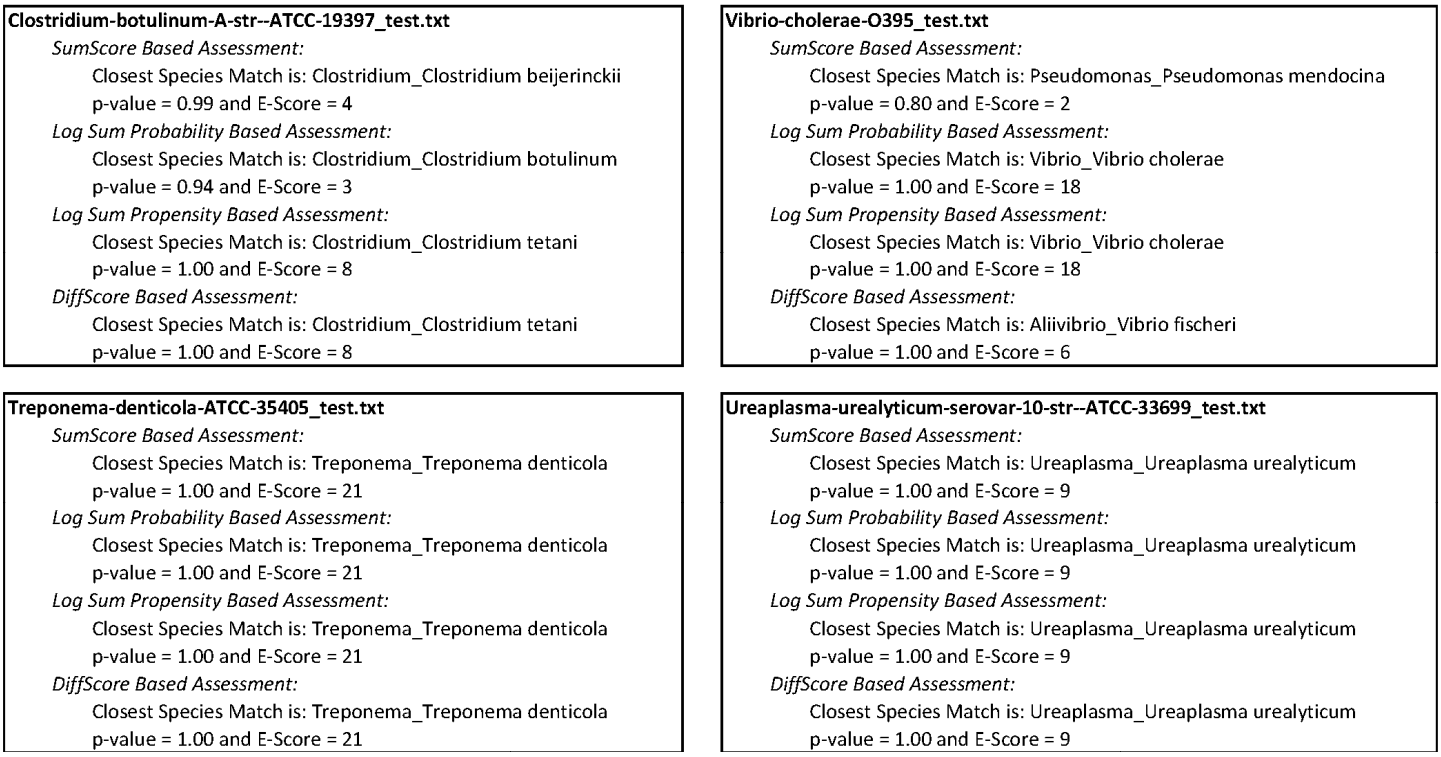
With orphan results accounted for, the summary results show that classification accuracy is well above 85% for all both Phylum and Class levels. The score metrics performed equally well in general with the $sumlogprobability$ metric outperforming others; however, the score based metrics offer additional information such as significance with acceptable success level of 90% accuracy. Metrics c_Score shows a well above 85% value indicating the reasonable success of the classifier; however, the weighted c_Score shows a relatively lower score reflecting the degree of error in misclassifications. Both are reported against the *Sum Log Probability* metric based classification results only.

Identification Experiments & Results

A species level EMM library is created. The library is of size 418 EMMs at the time of this paper. Some organisms are arbitrarily selected to be the ones to be identified. These are suffixed with “_test" which are

automatically picked up by the Evaluator to test against the library of EMMs. The classifier output is examined to see if the test organisms are identified correctly.

As shown in Figure 14, the four test organisms *Clostridium_botulinum, Treponema_denticola, Vibrio_Cholerae and Urealplasma_urealticum* were easily identifiable by the best performance metric i.e. *Sum Log Probabilities* with significance greater than 90%.



After the species library of 418 is created, four randomly selected organisms were subjected to a search test to see if they would be correctly identified. The organisms shown demonstrate a clear and certain identification with significance reported in the 90% range. The best performing metric *Sum Log Probabilities* provides consistent conclusion here.

Figure 13 : Identifying Organisms with the Species library of EMMs

To assess the performance of our identifier for novel organisms, we introduced three foreign RNA sequences belonging to Sand Fly, Brown dog tick and Mouse stem cells. These were mixed with the other test sequences and the experiment was rerun. Figure 15 shows the much lower significance scores implying that the sequences are perhaps novel organisms.

However, the Figure 16 shows the cases where identity remains a mystery. This is so because there is no clear pattern across the four metrics regarding the identifiability of the three organisms *Chalmydia_trachomatis, Clostridium_perifringens*. This will be further explored in the Resolution Control section.

Differentiation of species with 16s rRNA

As discussed in the methods section, the “Apparent Distance Metric" is used to create a distance matrix. Since Bacillus strains are harder to differentiate [53,54,55], they are selected to set up a distance matrix. Three foreign (non microbial) RNA introduced earlier in the document are also added to the group to analyze distance sensitivity in EMMBA.

The distance matrix shown in Figure 17 is a symmetric distance matrix where Apparent Distances are replaced with True Distance values. The matrix is then applied to a Phylogeny Inference Program called PHYLIP [47] at a server website [56] to obtain a phylogenetic tree using the Neighbor Joining method [57] as shown in Figure 18.

It may be interesting to note the delta values shown in Figure 17 as they tend to indicate the diversity.

Rhipicephalus-sanguineus-synganglion-_test.txt Brown Dog Tick

SumScore Based Assessment:
Closest Species Match is: Mycobacterium_Mycobacterium gilvum
p-value = 0.36 and E-Score = 0

Log Sum Probability Based Assessment:
Closest Species Match is: Aeropyrum_Aeropyrum pernix
p-value = 0.05 and E-Score = 0

Log Sum Propensity Based Assessment:
Closest Species Match is: Sulfolobus_Sulfolobus acidocaldarius
p-value = 0.00 and E-Score = 0

DiffScore Based Assessment:
Closest Species Match is: Sulfolobus_Sulfolobus acidocaldarius
p-value = 0.00 and E-Score = 0

Normalized-Phlebotomus-papatasi-_test.txt Sand Fly

SumScore Based Assessment:
Closest Species Match is: Bartonella_Bartonella quintana
p-value = 0.45 and E-Score = 1

Log Sum Probability Based Assessment:
Closest Species Match is: Borrelia_Borrelia afzelii
p-value = 0.13 and E-Score = 0

Log Sum Propensity Based Assessment:
Closest Species Match is: Bartonella_Bartonella henselae
p-value = 0.48 and E-Score = 1

DiffScore Based Assessment:
Closest Species Match is: Bartonella_Bartonella henselae
p-value = 0.48 and E-Score = 1

NIA-Mouse-Hematopoietic-Stem-Cell_test.txt MOUSE

SumScore Based Assessment:
Closest Species Match is: Sphingopyxis_Sphingopyxis alaskensis
p-value = 0.42 and E-Score = 1

Log Sum Probability Based Assessment:
Closest Species Match is: Gluconobacter_Gluconobacter oxydans
p-value = 0.79 and E-Score = 2

Log Sum Propensity Based Assessment:
Closest Species Match is: Sphingopyxis_Sphingopyxis alaskensis
p-value = 0.42 and E-Score = 1

DiffScore Based Assessment:
Closest Species Match is: Elusimicrobium_Elusimicrobium minutum
p-value = 0.11 and E-Score = 0

Three foreign RNA sequences belonging to *Mouse*, *Brown Dog Tick and Sand Fly* were introduced to see if identification would give some indication of novel cases unseen before for capture in the EMM libraries. The significance values are so low across the metric space that the novelty of the foreign RNA is clearly established.

Figure 14 : Novel sequence detection with Species Library of EMMs

Chlamydia-trachomatis-A-HAR-13_test.txt

SumScore Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia trachomatis
p-value = 0.00 and E-Score = 0

Log Sum Probability Based Assessment:
Closest Species Match is: Chlamydophila_Chlamydophila abortus
p-value = 0.69 and E-Score = 1

Log Sum Propensity Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia trachomatis
p-value = 0.00 and E-Score = 0

DiffScore Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia muridarum
p-value = 0.89 and E-Score = 2

Clostridium-perfringens-str--13_test.txt

SumScore Based Assessment:
Closest Species Match is: Clostridium_Clostridium beijerinckii
p-value = 1.00 and E-Score = 10

Log Sum Probability Based Assessment:
Closest Species Match is: Clostridium_Clostridium perfringens
p-value = 0.65 and E-Score = 1

Log Sum Propensity Based Assessment:
Closest Species Match is: Clostridium_Clostridium beijerinckii
p-value = 1.00 and E-Score = 10

DiffScore Based Assessment:
Closest Species Match is: Clostridium_Clostridium tetani
p-value = 1.00 and E-Score = 8

These are some cases where the identification is uncertain since there is no clear pattern in the metric space. In such cases, it may be necessary to use a finer resolution to better differentiate. This is explored further in Figure 17.

Figure 15: Examples of uncertain identification which requires finer parametric resolution

Using the Neighbor Joining Method [57] tree construction, the Figure 18 shows the proximity of closely related strains like Bacillus while clearly separating the foreign RNA belonging to *Mouse, Sand Fly and Brown Dog Tick*. Considering no

alignment was needed ahead of time, this differentiation output of distance matrix seems reasonable for Metagenomic analysis.

Resolution Control: Occasionally, coarse settings for key parameters of EMMBA improve response time at the cost of accuracy. Whenever the accuracy is found to be insufficient, fine tuning of the parameters is necessary.

RNA identification: source Organism	Normalized-Phle	NIA-Mouse-Hem	Rhipicephalus-sa	Acholeplasma-la	Aster-yellows-wi	Rickettsia-bellii-R	Buchnera-aphidic	Bacillus-clausii-K	Bacillus-halodura	Bacillus-licheniformis	Staphylococcus-a	Chlamydophila-p	Chloroherpeton-t	Rhodopirellu	Ther
Normalized-Phlebotomus-papatasi-(Genus Rana:Frog)	0	29	30	55	58	48	52	59	56	64	58	56	62	53	56
NIA-Mouse-Hematopoietic-Stem-Cell(Genus Mouse)	29	0	33	53	49	50	56	58	53	58	53	48	51	50	53
Rhipicephalus-sanguineus-synganglion-(Gen. Canis:Dog)	30	33	0	61	62	61	65	67	56	66	66	62	61	59	61
Acholeplasma-laidlawii-PG-8A	55	53	61	0	63	82	75	84	76	83	72	76	74	77	76
Aster-yellows-witches--broom-phytoplasma-AYWB	58	49	62	63	0	82	78	73	71	81	74	75	88	81	73
Rickettsia-bellii-RML369-C	48	50	61	82	82	0	90	86	78	90	82	78	87	84	84
Buchnera-aphidicola-str--5A--Acyrthosiphon-pisum-	52	56	65	75	78	90	0	85	77	76	73	77	86	83	77
Bacillus-clausii-KSM-K16	59	58	67	84	73	86	85	0	44	49	65	77	86	87	70
Bacillus-halodurans-C-125	56	53	56	76	71	78	77	44	0	56	56	73	89	82	81
Bacillus-licheniformis-ATCC-14580	64	58	66	83	81	90	76	49	56	0	54	73	81	79	79
Staphylococcus-aureus-subsp--aureus-USA300	58	53	66	72	74	82	73	65	56	54	0	75	82	83	75
Chlamydophila-pneumoniae-AR39	56	48	62	76	75	78	77	77	73	73	75	0	82	81	73
Chloroherpeton-thalassium-ATCC-35110	62	51	61	74	88	87	86	86	89	81	82	82	0	81	83
Rhodopirellula-baltica-SH-1	53	50	59	77	81	84	83	87	82	79	83	81	81	0	79
Thermodesulfovibrio-yellowstonii-DSM-11347	56	53	61	76	73	84	77	70	81	79	75	73	83	79	0
Delta Score	0.98														

The matrix is a result of inter evaluation of a pair of EMMs where one EMM evaluates another EMM as a test sequence and vice versa. Each evaluation produces an apparent distance. Results of both evaluations i.e. apparent distances are added to derive the true distance and used in the Distance Matrix to achieve symmetry. The distance value not only reflects the state transition differences but also uses EMM score matrices to quantify. The delta score computes the divergence or the tree-likeness of a distance matrix. Here the value of 0.98 indicates significant divergence as expected.

Figure 16: Distance Matrix generation using inter-evaluation of EMMs

Chlamydia-trachomatis-A-HAR-13_test.txt

SumScore Based Assessment:
Closest Species Match is: Chlamydophila_Chlamydophila felis
p-value = 1.00 and E-Score = 30
Log Sum Probability Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia trachomatis
p-value = 1.00 and E-Score = 9
Log Sum Propensity Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia trachomatis
p-value = 1.00 and E-Score = 9
DiffScore Based Assessment:
Closest Species Match is: Chlamydia_Chlamydia trachomatis
p-value = 1.00 and E-Score = 9

Clostridium-perfringens-str--13_test.txt

SumScore Based Assessment:
Closest Species Match is: Clostridium_Clostridium perfringens
p-value = 1.00 and E-Score = 36
Log Sum Probability Based Assessment:
Closest Species Match is: Clostridium_Clostridium perfringens
p-value = 1.00 and E-Score = 36
Log Sum Propensity Based Assessment:
Closest Species Match is: Clostridium_Clostridium perfringens
p-value = 1.00 and E-Score = 36
DiffScore Based Assessment:
Closest Species Match is: Clostridium_Clostridium perfringens
p-value = 1.00 and E-Score = 36

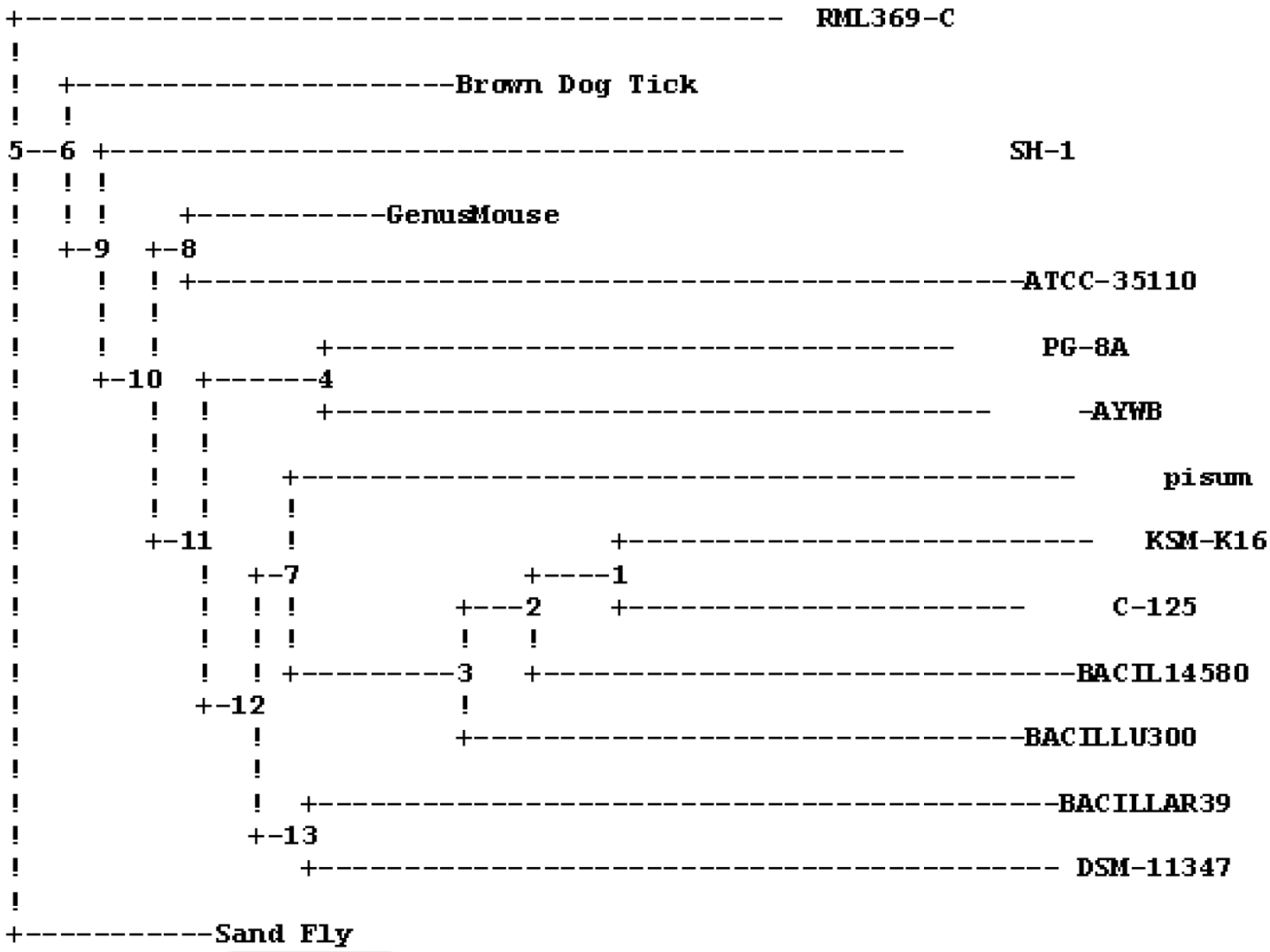
Increasing the granularity of numerical summarization by reducing the segment size from 80 to 20 significantly improves the prediction accuracy. The two organisms were previously unidentifiable at a segment size of 80, but with new reduced segment size, the metrics clearly show the prediction.

Figure 17: 10x cross validation results and aggregate metrics

For example, all experiments were conducted so far using a segment size of 80 and a word or pattern width of 3. Reducing the segment size improves the accuracy of prediction as there will be more granularity in the data.

When segment size is reduced to 20 and identification is reattempted for the two organisms of Figure 16 it is clearly seen from Figure 18 that prediction accuracy significantly improves.

In general, there are several parameter controls for EMMBA to adapt to different applications. This includes various clustering metrics that can be used when building EMMs as well.



Phylogenetic tree is generated using PHYlogeny Inference Package [47] available on the web [56]. The method selected is *Neighbor Joining Method*. As seen in the Figure, the foreign RNA belonging to *Sand Fly, Brown Dog Tick and Mouse* are shown on the outer branches of the tree indicating remoteness rest of the microbial sample. The Figure also places the Bacilli strains closer though they belong to different Genera.

Figure 18: Phylogenetic Tree of EMM differentiation output of Distance Matrix

DISCUSSION

This research formally (Figure 1) explored reorganization of the biological sequences to a more compact statistically equivalent Extensible Markov Model form (Figure 9). The EMMs created (Figure 3), were then organized further to form profiles of related organisms. The organism level sequence data was obtained from NCBI as FASTA files which is preprocessed to a word frequency form prior to further usage.

The effectiveness of using EMMs for biological sequences has been demonstrated through three distinct domains of bioinformatics of today and they are *Classification, Identification and Differentiation*. Classification

refers to the prediction of taxonomy for a given organism. Identification pertains to determining the possible taxa of a sequence as confidently as possible and Differentiation refers to recovery of phylogeny from all-against-all oriented distance matrix reports. The research presented here performed effectively in all three areas as discussed further here.

Grouping EMMs of sequences by taxonomic level made it possible to build libraries of models which were used for classification and phylogenetic analysis. Using the notion of sequence communities, the research explored classification of organisms into appropriate phylogenetic classes using a novel evaluation method(Figure 10).

In order to assess statistical significance of such classification, we have extended the Karlin-Altschul statistics [25] derive scores and the threshold levels. Extensions to Karlin-Altschul statistics were necessary because of differences in how the sequences are compared. In a classic alignment domain which is the basis of BLAST [23] sequences are aligned and scored. In a Quasi-alignment domain such as our approach, sequence comparison is done at word frequency level. New algorithms were proposed (Figures 5, 4) for building symmetric score matrices in an alignment free context such as that described here. Since no substitution matrices could be used, score differentials and usage of difference distributions (Figure 6) were derived and subsequently shown to be effective. Four criteria were proposed as pseudo metrics to determine the classification each of which is reported with statistical significance and E-score.

10x classification was used to verify the accuracy of classification (Figure 8). By targeting to predict sub-class level such as Genus, we have achieved phylogenetic class level prediction accuracy well above 90% as shown in Figure 12. A four level classification for each organism evaluation is obtained by matching the classification result against the expected labels. This was found to be useful for determining at what level of the taxonomy a particular test sequence would be classifiable. For each classification, four different criteria or pseudo metrics were assessed and reported along with significance and E-score values as shown in Figure 13. Matching the signature of a model and a query sequence is a non-trivial process. Of the four criteria or pseudo metrics used, the most effective one appears to be *Sum Log Probabilities* which is very sensitive to intra-sequence Markov transitions. However, the other metrics are also reasonable in their effectiveness and help provide a sanity check for the leading metric when assessing the overall results. It is to be noted that using extended Jaccard for clustering is not ideal when working with centroids. In fact, when Euclidean measure is used in place of extended Jaccard, the accuracy across all metrics improved except for the *Sum Log Probabilities* as shown in Figure 20.

It is often the case that classification at higher taxa is more successful than at lower taxa. This can be explained by the fact that there is simply more training data available at higher taxa due to large overall membership in terms of organisms. Knowledge of knowing the distribution at each level helps determine which level to work with to obtain successful classification. In our case, we found that the number of Genera is reasonably large at 246 compared to Phylogenetic Class pool size of only 33. Since the average number of genera per Class is much greater than 1, chances for predicting the Class are improved if Genus is used as the target class. However, this is not the only method for successful classification. Experimenting with granularity and making sure that the folds are balanced could achieve the same success at any desired level.

Results of overall classification are also computed using c_Score [46] which measures compliance with NCBI's view of correct taxonomy by measuring the number of non-trivial splits. A weighted c_Score measure is proposed to account for severity of missed classifications and thus provide a more conservative view of compliance. We showed that it is not necessary to build a phylogenetic tree to compute nontrivial splits required

for c_Score determination. Derivation of c_Score from the classification errors is shown in Figure 7. The c_Score values of 90% and a weighted c_Score values of over 80% confirms the reasonable performance of our multilevel Phylogenetic Classifier.

Metric Test	SumScore	∑log Prob	∑log Prop	DiffScore
EXT. JACCARD	97%	98%	97%	94%
EUCLIDEAN	97%	97%	97%	97%

Clustering measure used when building models can affect the performance of classification especially for the score related metrics. This is demonstrated by this table where the Euclidean based metric outperforms Jaccard in three of the four metrics.

Figure 19: 10x cross validation results and aggregate metrics

Since it is possible to consolidate related organisms or strains into a single complex model, a library of EMMs at the granularity of Species is created. Microbial Identification explores the possibility of readily determining the taxa of an unknown organism by assessing the strength and significance of its membership against each EMM in the library. When a segment size of 80 and a word length of 3 were used, identification was as expected and unambiguous as shown in Figure 14 for organisms *Clostridium_botulinum, Treponema_denticola, Vibrio_Cholerae and Urealplasma_urealticum.* However, more granularity was needed to identify *Chalmydia_trachomatis* and *Clostridium_perifringens*. Reducing the segment size helped disambiguate the results. The before and after segment size reduction are shown in Figures 16 and 19. Though all four criteria or pseudo metrics are generally effective, we found that *Sum Log Probabilities* provides more consistent results. Resolution control aspect of EMMs is useful in adjusting the system responsiveness and the level of accuracy required. This is expected to find application in Metagenomic classification where the sequence information is fragmented.

Phylogenetic trees are generated from carefully built distance matrices. Usage of a proper distance metric may be verified by attempting to compute delta value [48]. Our research explores usage of delta scores to report the diversity of a distance matrix which is useful in case the sample is of Metagenomic origin. Computation of delta score first confirms that each quartet of a distance matrix satisfies the four point condition [48] to ensure triangle inequality required of a true metric. We found that our four criteria used in classification and identification do not conform to the rules of a true metric. By defining distance as a sum of inter-EMM evaluations, we have achieved a true distance metric. By selecting a variety of organisms combined with three foreign RNA from *Mouse, Tick and Sand Fly*, a distance matrix was built (Figure 17). The distance matrix achieved a delta score > 0.9 indicating high degree of diversity. The matrix was then input to a Neighbor Joining Method [57] of the PHYlogenetic Inference Package [47] on the web [56] to generate a phylogenetic tree (Figure 18).

The phylogenetic relationships indicated in the Figure 18 clearly separate the non-microbial organisms from the microbial ones. Furthermore, the tree also shows the close proximity of the Bacillus strains though they all belong to different Genera. Our research here confirms the effectiveness of using EMM transformations in analyzing the microbial diversity of a collection of organisms which may be found in a metagenomic context.

The new transformational space offered by EMMs is useful for re-examining traditional sequence analysis issues and for exploring structure predictions in the future. Metagenomics and the Human Microbiome facilitate complex landscape for dealing with multitude of genomes all at once. This places huge demands on traditional

classic alignment methods. Effective Metagenomic Classification requires complex representations of taxa and Metagenomic Diversity analysis benefits from the issues of multiple sequence alignment. The future research will extend EMMBA methods to classify sequence fragments and differentiate Metagenomes from different times and/or places.

Using word statistics or counting short pattern sequences has been known [1] which explored different statistical distance measures for clustering. To the best of our knowledge, automatic learning to build models dynamically has not been explored in the literature. Similarly, clustering related segments of a single or multiple sequences to form a compact Markov model equipped with transition probabilities has also not been found in the literature though some derivatives may be assumed in profile HMMs [9].Using Extreme Value Distributions in an alignment free context where there are no substitution matrices to derive scores is a natural extension from Karlin-Altschul statistics [25] though its application toward a difference distribution is novel from our perspective. The fact that statistical signature libraries can be created from individual sequences or communities of sequences which can be used to classify, identify and differentiate combined with significance reporting is useful for Metagenomic Bioinformatics.

REFERENCES

1 Vinga S, Almeida J. Alignment-free sequence comparison-a review. 2003 Mar:513-523.

2 Auch AF, Henz SR, Holland BR, Göker M. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. 2006:350.

3 Henikoff SHaJG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992:10915-10919.

4 Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. 1978:345-351.

5 Kotamarti RM, Raiford DW, Raymer ML, Dunham MH. A data mining approach to predicting phylum for microbial organisms using genome-wide sequence data. 2009.

6 Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. 2004 Oct:2479-2481.

7 Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, et al. Machine learning in bioinformatics. 2006 Mar:86-112.

8 Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. 1986 Jul:5155-5159.

9 Eddy SR. Profile hidden Markov models. 1998:755-763.

10 Dunham MH, Meng Y&HJ. Extensible Markov model. In: Proc. Fourth IEEE International Conference on Data Mining ICDM '04; 2004. p. 371-374.

11 Jain AU. The Use of Time-Varying Markov Model to Study the Effect of Weather on Asthma. 2007.

12 Meng Y, Dunham MH. Mining Developing Trends of Dynamic Spatiotemporal Data Streams. 2006:43-50.

13 Meng Y, Dunham MH, Marchetti MF, Jie H. Rare Event Detection in a Spatiotemporal Environment. 2006:629-634.

14 Freedman D. Markov Chains. Holden-Day; 1975.

15 Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. 1998 Jan:544-548.

16 Eddy SR, Durbin R. RNA sequence analysis using covariance models. 1994 Jun:2079-2088.

17 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. 1990:403-410.

18 Golod D, Brown DG. A tutorial of techniques for improving standard hidden markov model algorithms. 2009 Aug:737-754.

19 Altschul SF, Gish W. Local alignment statistics. 1996:460-480.

20 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. 1997:3389-3402.

21 Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. 1980 Dec:111-120.

22 Eddy SR. Where did the BLOSUM62 alignment score matrix come from? 2004 Aug:1035-1036.

23 Ian Korf MY. BLAST. O'Reilly; 2003.

24 Durbin R. Biological Sequence Analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.

25 Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. 1990 Mar:2264-2268.

26 Smith TF, Waterman MS. Identification of common molecular subsequences. 1981:195-197.

27 Altschul. National Center for Biotechnology Information. [Internet]. 2005 [cited 2009 November]. Available from: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html.

28 Gumbel. Statistics of the extremes. Columbia University Press; 1958.

29 Mott. Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. 1992:54-59.

30 Waterman M, Vingron M. Sequence Comparison Significance and Poisson Approximation. 1994:367-381.

31 Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. 1985 Jan:645-656.

32 Collins JF, Coulson AF, Lyall A. The significance of protein sequence similarities. 1988 Mar:67-71.

33 Waterman MS, Vingron M. Rapid and accurate estimates of statistical significance for sequence data base searches. 1994 May:4625-4628.

34 Pearson WR. Empirical statistical estimates for sequence similarity searches. 1998 Feb:71-84.

35 Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. 1901:547-579.

36 Greenberg DA, Berger B. Using lod-score differences to determine mode of inheritance: a simple, robust method even in the presence of heterogeneity and reduced penetrance. 1994 Oct:834-840.

37 Karlin. NCBI toolkit crossreference. [Internet]. 2009 Available from: http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/tools/blast.c.

38 Blaisdell BE. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. 1989 Dec:538-547.

39 Weisstein E. "Normal Difference Distribution." From MathWorld--A Wolfram Web Resource. [Internet]. 2009 [cited November]. Available from: http://mathworld.wolfram.com/NormalDifferenceDistribution.html.

40 Weisstein EW. Weisstein, Eric W. "Gumbel Distribution." From MathWorld--A Wolfram Web Resource. [Internet]. 2009 Available from: http://mathworld.wolfram.com/GumbelDistribution.html.

41 Weisstein EW. "Half-Normal Distribution." From MathWorld--A Wolfram Web Resource. [Internet]. 2009 Available from: http://mathworld.wolfram.com/Half-NormalDistribution.html.

42 Weisstein EW. "Erfc." From MathWorld--A Wolfram Web Resource. [Internet]. 2009 Available from: http://mathworld.wolfram.com/Erfc.html.

43 Weisstein EW. "Confidence Interval." From MathWorld--A Wolfram Web Resource. [Internet]. 2009 Available from: http://mathworld.wolfram.com/ConfidenceInterval.html.

44 Anderson TW, Darling DA. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. 1952:193-212.

45 Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. 2009 Sep:673-676.

46 Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. Whole-genome prokaryotic phylogeny. 2005 May:2329-2335.

47 Felsenstein J. PHYLIP (phylogeny inference package), version 3.57 c. 1995.

48 Holland BR, Huber KT, Dress A, Moulton V. Delta plots: a tool for analyzing phylogenetic distance data. 2002 Dec:2051-2059.

49 CoreTeam R. R development core team. [Internet]. 2005 Available from: http://www.R-project.org.

50 company A. AISEE Graph Layout software. [Internet]. 2009 [cited April]. Available from: http://www.aisee.com/.

51 Unknown. Major resources available by ftp (NCBI). [Internet]. [cited 2009 August]. Available from:

http://www.ncbi.nlm.nih.gov/Ftp/.

52 Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995:1137-1143.

53 Gao L, Qi J, Sun J, Hao B. Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. 2007 Oct:587-599.

54 Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. 2007 Sep:2761-2764.

55 Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. 2007 Jan:278-288.

56 Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C. Mobyle: a new full web bioinformatics framework. 2009 Nov:3005-3011.

57 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. 1987 Jul:406-425.

58 Mering Cv, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. Quantitative phylogenetic assessment of microbial communities in diverse environments. 2007 Feb:1126-1130.

59 Meng Y, Dunham MH. Online Mining of Risk Level of Traffic Anomalies with User s Feedbacks. 2006:176-181.

60 Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. 1995 Jul:283-290.

61 Karlin S, Bucher P. Correlation analysis of amino acid usage in protein classes. 1992:12165-12169.

62 Karlin S, Brendel V. Chance and statistical significance in protein and DNA sequence analysis. 1992:39-49.

63 Isaksson C, Dunham MH. A Comparative Study of Outlier Detection. 2009.

64 Dunham MH, Quick D, Wang Y, McGee M, Waddle J. Visualization of DNA/RNA Structure using Temporal CGRs. 2006:171-178.

65 Dunham MH, Ayewah N, Li Z, Bean K, Huang J. Spatiotemporal Prediction using Data Mining Tools. Idea Group; 2005.

66 Dunham MH. Data Mining Introductory and Advanced Topics. Prentice Hall; 2003.

67 Charlie Isaksson YM, Margaret H. Dunham “L. Risk Leveling of Network Traffic Anomalies. 2005:258-265.

69 University WS. AGATE Statistical Analysis Program (Anderson-Darling Normality Test). [Internet]. Available from: www.niar.wichita.edu/coe/NCAMP./Programs/ASAP_2008_v1.0.xls.

70 Altschul. Statistics of psi-BLAST scores. [Internet]. Available from: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html.