Sequence transformation to a complex signature form for consistent Phylogenetic tree using Extensible Markov Model

Kotamarti Rao M, Hahsler Michael, Raiford Douglas W and Dunham Margaret H

Abstract— **Phylogenetic tree analysis using molecular sequences continues to expand beyond the 16S rRNA marker. By addressing the multi-copy issue known as the intra-heterogeneity, this paper restores the focus in using the 16S rRNA marker. Through use of a novel learning and model building algorithm, the multiple gene copies are integrated into a compact complex signature using the Extensible Markov Model (EMM). The method clusters related sequence segments while preserving their inherent order to create an EMM signature for a microbial organism. A library of EMM signatures is generated from which samples are drawn for phylogenetic analysis. By matching the components of two signatures, referred to as quasi-alignment, the differences are highlighted and scored. Scoring quasi-alignments is done using adapted Karlin-Altschul statistics to compute a novel distance metric. The metric satisfies conditions of identity, symmetry, triangular inequality and the four point rule required for a valid evolution distance metric. The resulting distance matrix is input to PHYologeny Inference Package (PHYLIP) to generate phylogenies using neighbor joining algorithms. Through control of clustering in signature creation, the diversity of similar organisms and their placement in the phylogeny is explained. The experiments include analysis of genus Burkholderia, a random microbial sample spanning several phyla and a diverse sample that includes RNA of Eukaryotic origin. The NCBI sequence data for 16S rRNA is used for validation.**

I. Introduction

16S rRNA, a part of ribosomal RNA, is an essential and ubiquitous gene sequence and it is commonly collected and used for microbial identification [1] and classification [2]. However, an organism may have more than one copy of the sequence and though rare, these copies may be results of lateral transfers from others organisms [3]. Selecting the right copy is not always obvious though some methods are known in the literature[4]. Since all approaches may not necessarily select the same copy, some inconsistencies are unavoidable[5]. Extensible Markov Models (EMM)[6] can be used to create a unique representation in such cases [7] from all sequences.

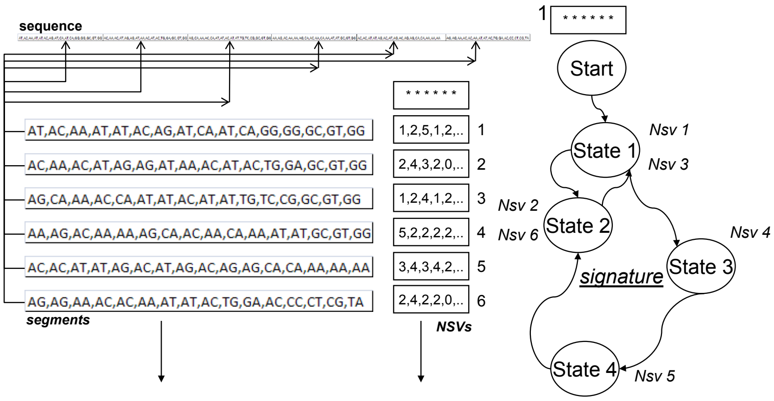
EMM is a time varying Markov chain or a directed graph with nodes representing the states containing the clusters of related dynamic event data and arcs representing the transitions unique to the order of events. Such modeling has many applications in a variety of fields including future state prediction [8] and rare event detection [9]. Though several alternative single copy markers are in use specifically for diagnostic identification purposes, we explored a way to extend the well established role of 16S rRNA in defining the overall microbial taxonomy [10] to lower taxa analysis.

As a Bioinformatic adaptation of an Extensible Markov Model, EMM Bioinformatic Analysis (EMMBA) transforms m molecular sequences to a single EMM signature of M states. It can be considered a representation of sequence data with states representing clusters of similar sequence segments and inter-state transition probabilities representing the implicit order within the sequences. It has recently been reported that EMM signature representation of sequence data is computationally more efficient than the native standard format used in genome libraries [7]. Any multi-feature nominal or numeric data can be learned and stored in a compact signature form useful for efficient mining. In addition to the applications in sequence transformation, EMMBA has important applications in Bioinformatic machine learning, such as Metagenomic sequence classification, differentiation and species identification. As the number of genome sequences becomes large or as it becomes necessary to process data in real time, statistics based heuristics are needed. For these reasons, There have been many attempts to design both classic alignment based, as well as, alignment free methods to perform large scale sequence searches and analyses [11, 12]. Unlike the traditional sequence analyses that directly utilize the raw homologous sequence data, this paper presents a complex Markov signature representation for sequence analysis based on machine learning with time complexity of $O(\frac{mL}{K} + m^2)$ as opposed to $O(mL + L^2 + m^2 log(m))$ of ClustalW+clustering method where m, L and K represent the number of sequences, the length of one such sequence and the number of equal sized segments in each sequence respectively. By including all the available sequence information as opposed to selecting a single representative one, consistency in phylogeny is made possible by this approach.

The rest of this paper is organized as follows: Section II derives an equivalent Markov formulation of sequence learning that is suitable for the sequence signature design. Section III includes related work and introduces the comparative sequence analysis using all-to-all distance computation and algorithmic framework. Some Phylogenetic study results are presented in Section IV to illustrate the performance of EMM signature differentiation. Finally, Section V discusses the results and concludes the paper.

Kotamarti Rao M, Hahsler Michael and Dunham Margaret H are with the Department of Computer Science and Engineering, Southern Methodist University, Dallas and Raiford Doug is with University of Montanna (email: {rkotamarti, mhd}@lyle.smu.edu).

This work was supported by the NSF Net-Centric IUCRC/T-SYSTEM Inc. industrial memberships in the form of Graduate studies sponsorship for the first author.


sequence
1
******
Start
******
AT,AC,AA,AT,AT,AC,AG,AT,CA,AT,CA,GG,GG,GC,GT,GG
AC,AA,AC,AT,AG,AG,AT,AA,AC,AT,AC,TG,GA,GC,GT,GG
AG,CA,AA,AC,CA,AT,AT,AC,AT,AT,TG,TC,CG,GC,GT,GG
AA,AG,AC,AA,AA,AG,CA,AC,AA,CA,AA,AT,AT,GC,GT,GG
AC,AC,AT,AT,AG,AC,AT,AG,AC,AG,AG,CA,CA,AA,AA,AA
AG,AG,AA,AC,AC,AA,AT,AT,AC,TG,GA,AC,CC,CT,CG,TA
segments
1,2,5,1,2,.. 1
2,4,3,2,0,.. 2
1,2,4,1,2,.. 3
5,2,2,2,2,.. 4
3,4,3,4,2,.. 5
2,4,2,2,0,.. 6
NSVs
Nsv 1
Nsv 3
State 1
Nsv 2
Nsv 6
State 2
Nsv 4
State 3
signature
State 4
Nsv 5

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

Fig. 1. The Model Building Process

II. Formulations

The letters or base compositions of an RNA sequence $\{A,\ C,\ U,\ G\}$ provide frequency information. The co-occurrences of a pattern of bases of length l generates an l-mer frequency representation for a sequence ([12]).

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

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

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

$$\begin{aligned} F(S) &=< F'(S_1), F'(S_2), ..., F'(S_m) > \\ F'(S_i) &=< F^*(s_{(i,1)}), F^*(s_{(i,2)}), ..., F^*(s_{(i,k)}) > \\ F^*(s_{(j,k)}) &=< v_{(j,k,1)}, v_{(j,k,2)}, ..., v_{(j,k,n)} > \end{aligned}$$

Where the functions F, F' and F* are Numerical Summarization Functions to convert molecular sequences to oligomer frequency form. The function F collects m sequences being transformed and applies the F' transformation function. The function F' converts a single sequence to k equal sized segments and applies the function F^* on each segment. The function F^* converts a sequence segment to a vector of frequencies with each frequency representing one of the oligomer variants. Such vectors are referred to as Numerical Summarization Vectors (NSV) and the size of an NSV is given by 4^l where l is oligomer length.

Finally, $\uplus$ extends the directed graph i.e. *sequence signature* as shown in Figure 1 by clustering each NSV generated by the F function into a new or an existing node of the graph and appropriately updating the arc information. The initial signature graph is $empty\ graph$. The function $\uplus$ can be further expressed by the following algorithm:

For each NSV v_i,

1) Find the closest match, i.e. the nearest node to the NSV v_i.
2) if match not close enough, create a new node.
3) add arc from current node to the matched/new node.

where the closest match is defined as the node whose centroid is at a minimal Euclidean distance from the NSV v_i and adding an arc involves updating the arc probability. The state of the graph (current node) is always the last matched node. The state resets to the start state for every new sequence.

Fig. 1 shows the EMM build operation graphically. In this section, we formulated the EMM build operation as a transformation problem, which is suitable for compressing sequences into a signature. The transformation funtion of $\uplus$ is equivalent to a compression function. This is explained as follows:

Since S representing all sequences of an organism can be thought of as a matrix of m rows and k columns with all its elements as sequence fragments of equal size, the resulting graph signature G will have no more than k states. Since several similar segments are clustered into fewer states, number of nodes is never larger than the number of segments. This is first order compression. Since there are several similar sequences and their segments are also clustered into the signature graph, the final number of nodes is also never larger than the total number of segments. In fact, if the sequences are indeed similar, number of states is surely smaller than the number of segments. Since a cluster is represented by its centroid only, a node would have only one vector though several similar vectors were clustered into it. This is the second order compression. Since segments are represented as oligomer frequency vectors whose size depends on the chosen oligomer length, the compression should exist so long as the NSV size is less than the segment size. With the added benefit of compression, we can design a library of sequence signatures that are compact due to space reduction and informative due to arc probabilities preserving the segment order statistics. By reducing the number of similar sequence fragments into a single numerical vector form, the quantity and the complexity of comparisons is reduced while also avoiding the processing to select a candidate gene copy for analysis.

III. Comparative Sequence Analyses

In recent decades, several effective sequence analysis techniques have been proposed and in use. Karlin and Altschul developed [11] a fast heuristic algorithm based on statistical significance of basic local alignment. Results of BLAST algorithm are used to-date as a step in phylogeny analysis. Lilburn et al presented a heat map visualization technique [4] to examine and correct for any inconsistencies in the generation of Phylogeny. For traditional Phylogenetic analyses, the process involves performing a multi-sequence alignment and finding a reliable distance between all sequence pairs based upon some specific conserved fragment. Due to rise in the next generation sequencing projects, whole genomes are also at one's disposal, but this adds to the complexity by requiring yet another sequence selection process. The prevalent method by the authoritative microbial classification resource is described by Lilburn et al in [4] using heatmap visualization. The general method for generating Phylogeny is as follows:

1) Pick the longest 16S rRNA for a genome with the most conserved homologue positions.
2) Perform Multiple Sequence Alignment for all the representative 16S rRNA sequences at those positions.
3) Pick a pairwise distance estimation method and create a distance matrix.
4) Apply Hierarchical clustering to generate dendrogram attached heat map visualization for the resulting phylogenetic tree.

The main drawback of this method is that the approach must pick one of several 16S rRNA copies in a genome and also that different pair-wise distance measurements generate different Phylogenetic trees. Having to perform multiple sequence alignment is also time consuming. Thorne et al suggest [13] that freeing phylogenies from the artifacts of alignment could help improve topological accuracy. We propose using the EMM signatures of the 16S sequence profiles of organisms to compute the distance matrix. Since signature profiles contain all available 16S rRNA copies, no information is sacrificed. This process is described using a differentiation operation between the two EMM signatures involved. When all sequence pairs are evaluated, a distance matrix is generated.

Given two EMMs, e_1 and e_2, the distance between them is:

$$D(e_1, e_2) = d(e_1, e_2) + d(e_2, e_1)$$

where $d(e_i, e_j)$ measures how far e_j is from the e_i as below

$$d(e_i, e_j) = \sum_{k=0}^{|e_j|} \omega^*(s_k, s'_k)$$

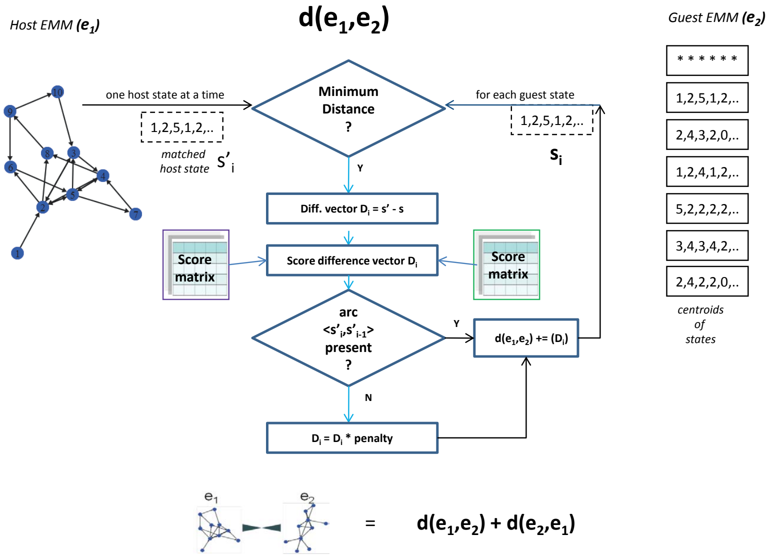
where $\omega^*(s_k, s'_k)$ scores the matched pair (k, k') of states

Since both graphs are independent with their own sets of vertices and weighted edges, the distance between them is computed using a two step approach. First, one of the two EMM graphs is fixed as the host while the other is made a guest. Guest is evaluated against the host EMM graph to determine its distance from the host. Prior to evaluation, the guest EMM is converted to an ordered list of NSVs where each NSV is the centroid vector of a node in the graph. Since EMM states are numbered in the order of their creation, the same order is used in generating the NSV list for the guest EMM. Each guest NSV is then searched against the host EMM graph of nodes to find the closest match based on minimum Euclidean distance. Comparison is done between the guest NSV and the centroids of host nodes. Once the closest match is determined, the pair i.e. guest NSV and the host node are said to be in quasi-alignment. The quasi-alignment is then scored and aggregation of all quasi-alignment scores produces distance between the guest and the host. The process is repeated by switching the roles of guest and host to derive distance in the other direction also. Both uni-directional distances are aggregated to derive a symmetric distance between both EMM graphs. The details of scoring quasi-alignments and their aggregation are discussed next.

The distance evaluation function $d(e_i, e_j)$ scores the quasi-alignments of e_j against e_i by taking the difference between the two matched states. The difference between two EMM states is the difference between their centroids. The centroids in EMM signatures are vectors of mean frequencies of oligomer patterns occurring across all sequence segments that make up the state. The difference between the centroids generates a difference centroid which is scored. Kotamarti et al have proposed a LOD score matrix for EMM signatures [7] by extending Karlin-Altschul statistics [11]. Given X^{e_a} as a scoring array for EMM e_a, and $\overrightarrow{c}$ as the vector difference in the centroids of quasi-aligned states of both EMMs, the scoring function $\omega^*(s_k, s'_k)$ may be expressed as follows:

$$\omega^*(s_k, s'_k) = \sum_{y=0}^{n-1} X_y^{e_i} \overrightarrow{c}_y + \sum_{y=0}^{n-1} X_y^{e_j} \overrightarrow{c}_y$$

Where y loops over all the n elements in the centroid vector. Since both EMMs are Markov models, the state transition differences between the two are also considered when scoring the quasi-alignments between them. In the notation $d(e_1, e_2)$, e_1 is called the host and the e_2 is called the guest. As each state from e_2 is searched across e_1 to find the nearest match, the state transition validity is checked.



Distance between two EMM signatures is determined by adding the results of $d(e_1, e_2)$ and $d(e_2, e_1)$. The operation of $d(e_1, e_2)$ is described in the figure where e_1 is the host EMM and e_2 is the guest EMM. For each state s_i from e_2, the nearest (Euclidean based) state s'_i from the e_1 is determined. The difference in centroid vectors of the state pair (s_i, s'_i) is considered for scoring. In case there is no valid transition between the current host state s'_i and the previously matched s'_{i-1}, a penalty is applied to the difference score between the current match or quasi-alignment. When all states of e_2 are processed, all the difference scores are added to derive a one way distance value. The process is repeated with e_2 as the host and e_1 as the guest deriving distance in the other direction. Both distances are added to derive the final distance between both EMM signatures.

Fig. 2. Algorithmic operation of the EMM differentiator

Given i' as the matched host node for a guest NSV i, a state transition $(i' - 1, i')$ is valid if there is an edge or arc present between the currently matched host node i' and the previously matched host node $i' - 1$. In essence, it is not sufficient for a guest NSV to match a host node, but it should also follow the arc present in the host EMM graph. If a transition implied by the quasi-alignment does not occur in the host EMM, its score is penalized [7]. The penalty, due to unsupported transitions, has the effect of increasing the distance between the two EMMs. It is computed as $-log(\frac{1}{\epsilon_i})$ where $\epsilon = \frac{1}{|V|}$ where $|V|$ represents the size or the number of nodes in the host EMM. In cases where the transition is valid in the host, the score is weighted by the probability of the arc $a(i' - 1, i')$. In summary, with V_h and V_g as the number of nodes in the host EMM and the guest graph respectively, the distance d between host h and guest g is computed as:

$$d(h, g) = \Sigma_{i=0}^{|V_g|-1}(weight'_i \times score_q)$$

$$weight'_i = -log\left(\frac{1}{\epsilon'_i}\right) = -log(|V_h|) \text{ if unsupported arc}$$

$$OR$$

$$= a_{(i'-1,i')} \text{ for supported arc}$$

Where $score_q$ is the score of the quasi-alignment. The distance measure thus computed is in fact a weighted Manhattan distance measure [14] which is found to obey the conditions for a distance metric. Though not required of a distance metric, we also validated the the four point rule over a large number of datasets as it is required for a evolutionary distance measure. The proof in its basic form is presented in the appendix. A complete proof for the metric is quite involved due to the different types of weights such as scores, probabilities and penalties that are applied. As such, complete proof is deferred as future work. The time and space complexity of distance evaluation function can be determined from the following algorithm:

1) Select the host state nearest to the guest state - this produces a quasi-alignment.
2) Derive the difference of centroids into a centroid vector.
3) Score each element in the difference centroid vector and compute their sum.
4) Multiply each sum score with a penalty if the quasi-alignment leads to a state without a supported transition or arc.
5) Take the sum of the sum scores (adjusted for penalty) to derive the one way distance.

Given *number of nodes* as $|V|$ and $|V'|$ for both, time complexity is $O(2|V||V'|)$ and space complexity is $O(|V|+|V'|)$. More generally, these may be expressed as $O(|V^2|)$ and $O(2|V|)$ where $|V|$ is the maximum of the number of nodes in the two EMM graphs. Since this method is an all-against-all distance evaluation based and is not based on alignment, its time complexity compares favorably with the current Multiple Sequence Alignments. There is no doubt that an optimal multi sequence alignment is more desirable though it is NP-hard [15]. More practical methods such as ClustallW [16, 17] use progressive alignment methods for which the time complexity is still $O(mL + L^2)$ to which clustering adds an additional term of $(m^2log(m))$ where m and L

represent the number and length of the aligned sequences respectively. While multi-sequence alignment basis offers much in the way of understanding closely related microbial taxa, analysis of similarity in a more diverse setting is the goal of the method presented here. Indeed Metagenomic samples present such diversity as well as the whole of NCBI microbial taxonomy when required. The architecture of distance evaluation function $d(e_i, e_j)$ for EMMs may be best described with Figure 2.

Karlin-Altschul proposed [11] a log-odds (LOD) Score for scoring alignments along with a theorem supporting Gumbel extreme value distribution of LOD scores. Extending the Karlin-Altschul statistics, quasi-alignments can also be scored using LOD scores by use of dynamically built Score Matrices for each EMM signtaure [7]. This has become necessary since the alignment basis used in [11] deals with a substitutive environment where one base is substituted with another in an alignment. The substitution matrices such as BLOSUM [18] and PAM [19] are widely used in BLAST literature[20, 21]. Since EMMs use the frequency form of bases and not the bases directly, its quasi-alignment context deals with comparing numerical vectors of oligomer frequencies and the EMM score matrix reflects the LOD scores for occurrences of specific oligomer patterns. Using the oligomer score matrices, each evaluation of an EMM generates a series of quasi-alignments and the corresponding difference scores which forms a difference distribution of Gumbel scores [7]. In fact, statistical significance for each quasi alignment can be computed using the difference score d. Averaging the significance values across all quasi alignments of an evaluation generates a *general significance* [7]. Such measure can be used in validating phylogeny as well in cases of unexpected associations found between unrelated taxa, referred to as *mis-associations*.

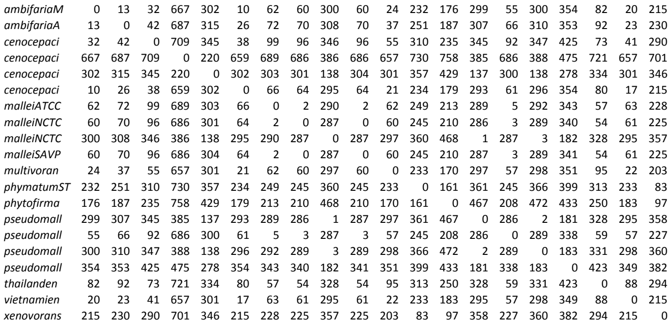
IV. Experiments and Results

The 16S rRNA Database utilized in this analysis is derived from the NCBI Microbial Complete Genome Database [22]. The sequences were extracted from the annotated whole genome sequence files using keyword searches. From this data, a new database was built that consists of individual files, one per microbial organism, in FASTA format. The original dataset was derived from the NCBI as of August 2009 and consists of 782 organisms each with multiple 16S sequences where applicable.

The FASTA header for each file contains five pieces of information: phylum, class, genus, species and organism name. We found that some of the header information in the NCBI database was missing in some cases. There were several cases of missing genus or even class information. Since this type of information is used for verifying the topological accuracy, such data is excluded from analysis. The final database consisted of 676 organisms.

As a general process, the organism sequences of interest are transformed into EMMs prior to further analysis. To show the effectiveness and efficiency of the proposed sequence differentiation, three experiments are performed. In the first experiment, the genus Burkholedria is selected for its known diversity [23] to study its phylogeny based on EMM differentiation. The 16S rRNA sequences of organisms within the genus are first transformed into separate EMM signatures and then an all-to-all evaluation is performed resulting in a distance matrix D. Given D, P, Q, R, S as a distance matrix with P,Q,R & S as the organism indices, the validity of a distance metric can be proved if the following conditions are satisfied[24].

$$\begin{aligned} D(i,j) &= 0 \text{ for all } i = j \\ D(i,j) &= D(j,i) \\ D(P,S) &<= D(P,Q) + D(Q,S) \text{ and} \\ D(P,Q) + D(R,S) &= max(D(P,R) + D(Q,S), \\ &\quad D(P,S) + D(R,Q)) \end{aligned}$$



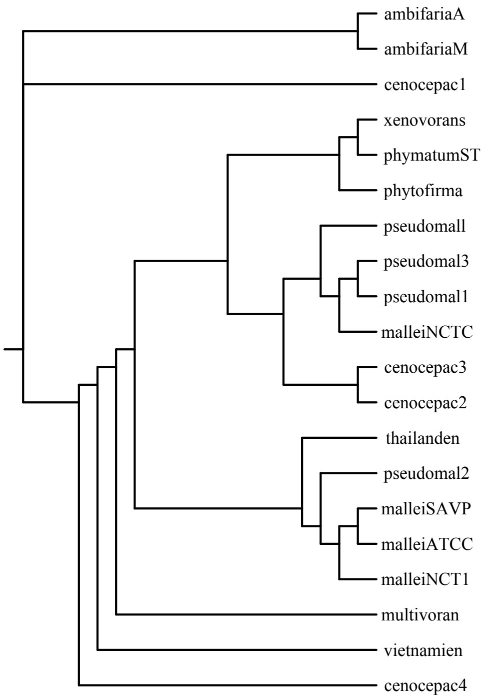
ambifariaM	0	13	32	667	302	10	62	60	300	60	24	232	176	299	55	300	354	82	20	215
ambifariaA	13	0	42	687	315	26	72	70	308	70	37	251	187	307	66	310	353	92	23	230
cenocepaci	32	42	0	709	345	38	99	96	346	96	55	310	235	345	92	347	425	73	41	290
cenocepaci	667	687	709	0	220	659	689	686	386	686	657	730	758	385	686	388	475	721	657	701
cenocepaci	302	315	345	220	0	302	303	301	138	304	301	357	429	137	300	138	278	334	301	346
cenocepaci	10	26	38	659	302	0	66	64	295	64	21	234	179	293	61	296	354	80	17	215
malleiATCC	62	72	99	689	303	66	0	2	290	2	62	249	213	289	5	292	343	57	63	228
malleiNCTC	60	70	96	686	301	64	2	0	287	0	60	245	210	286	3	289	340	54	61	225
malleiNCTC	300	308	346	386	138	295	290	287	0	287	297	360	468	1	287	3	182	328	295	357
malleiSAVP	60	70	96	686	304	64	2	0	287	0	60	245	210	287	3	289	341	54	61	225
multivoran	24	37	55	657	301	21	62	60	297	60	0	233	170	297	57	298	351	95	22	203
phymatumST	232	251	310	730	357	234	249	245	360	245	233	0	161	361	245	366	399	313	233	83
phytofirma	176	187	235	758	429	179	213	210	468	210	170	161	0	467	208	472	433	250	183	97
pseudomall	299	307	345	385	137	293	289	286	1	287	297	361	467	0	286	2	181	328	295	358
pseudomall	55	66	92	686	300	61	5	3	287	3	57	245	208	286	0	289	338	59	57	227
pseudomall	300	310	347	388	138	296	292	289	3	289	298	366	472	2	289	0	183	331	298	360
pseudomall	354	353	425	475	278	354	343	340	182	341	351	399	433	181	338	183	0	423	349	382
thailanden	82	92	73	721	334	80	57	54	328	54	95	313	250	328	59	331	423	0	88	294
vietnamien	20	23	41	657	301	17	63	61	295	61	22	233	183	295	57	298	349	88	0	215
xenovorans	215	230	290	701	346	215	228	225	357	225	203	83	97	358	227	360	382	294	215	0

The distance matrix shows the all-to-all evaluations of all organisms in the Burkholderia genus. The distance metric used is based on EMM signature analysis of sequence data. The conditions required for a distance metric[24] are 1)zeroes on the diagonal 2)symmetry 3) Triangle inequality and 4) Four point condition [25]. The distance matrix and therefore the proposed metric satisfy all conditions of a distance metric.

Fig. 3. Distance Matrix using EMM Signature derived metric

The distance values in the matrix shown in Figure 3 were verified to satisfy all the conditions for a distance metric and thus presents a valid input for phylogeny. The phylogeny is derived and shown in Figure 4 which is generated using an improved version [26] of the neighbor joining method included in the Phylogeny Inference Package (PHYLIP) [27, 28]. Due to space constraints, the organism name excludes the prefix *Burkholderia* and includes only the first 10 characters of the remainder. For example, the organism *Burkholderia-ambifaria-MC40-6* is written as *ambifaria*.

As shown in Figure 4, the phylogeny for genus Burkholderia, is in general, topologically accurate. This is indicated by the fact that “similar organisms” are grouped together though some exceptions do exist. It seems there is an unexpected association between the species prefixed by *Burkholdier-mallei, Burkholdier-pseudomallei and Burkholdier-cenocepacia*. Examining the distance matrix does indicate that some strains are nearer to strains of other species than their own. We investigated this to see if this was an artifact of how clustering was employed in creating the EMM signature. First, we noticed that the statistical significance value is very high



The phylogeny of Burkholdier, generated using our distance metric, is shown. The topological accuracy can be analyzed by verifying the placement of similar organisms. With the exception of a few organisms, the topology is generally correct.

Fig. 4. Phylogeny of Burkholdier

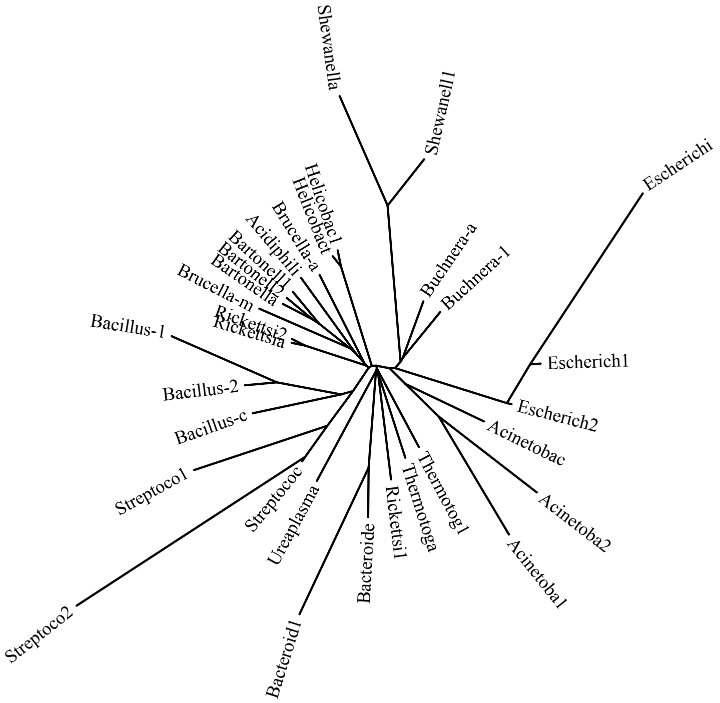
between the pairs where an unexpected association appeared to exist. To investigate this phenomenon, the clustering level was increased in the EMM signature generation process by increasing the Euclidean-threshold. The reason for this was to encourage more clustering of near-similar sequence segments while clearly separating the distinct ones. The resulting distance matrix (not shown) provided much clarification of the three genera. We further confirmed by building a multi-sequence alignment and generating a distance matrix using a modified Kimura-2 parameter model at the GreenGenes web site [29] as shown in Figure 5.

Burkholderia-MSA	**mallei**				**pseudomallei**		
mallei-ATCC-23344	0	0.000007	0.002347	0.000007	0.000778	0.003132	0.003132
mallei-NCTC-10229	0.000007	0	0.002347	0.000007	0.000778	0.003132	0.003132
mallei-NCTC-10247	0.002347	0.002347	0	0.002347	0.003132	0.000778	0.000778
mallei-SAVP1	0.000007	0.000007	0.002347	0	0.000778	0.003132	0.003132
pseudomallei-1710b	0.000778	0.000778	0.003132	0.000778	0	0.002347	0.002347
pseudomallei-668	0.003132	0.003132	0.000778	0.003132	0.002347	0	0.000007
pseudomallei-K96243	0.003132	0.003132	0.000778	0.003132	0.002347	0.000007	0

This distance matrix is generated from a multi-sequence alignment using the Green-Genes web site [29]. This matrix also shows that one of the *B.psedumallei* and *B.mallei* organisms have larger distances from their own kind. As such, the distinct ones are phylogenetically shown to be farther from the rest. For example, the *B.pseudomallei-1710b* and *B.mallei-NCTC-10247* are shown separated from their own kind. This validates our method and the proposed metric.

Fig. 5. Clarifying distances by comparing to a Distance Matrix generated using Multiple Sequence Alignment

Having shown a reasonable topological view of the diverse *Burkholderia* genus, another experiment was performed using a randomly selected sample of organisms. This experiment includes 5 phyla, 7 classes and 13 genera in the dataset. The phylogeny is shown in Figure 6.

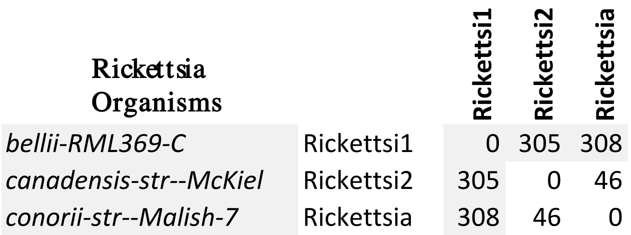


A random sample representing 5 phyla, 7 classes and 13 genera is used to perform an EMM signature differentiation. The distance matrix is then used to generate Phylogeny based on BIONJ [26] algorithm using the PHYlogeny Inference Package [27] on the web [28]. Except for *Rickettsi* group, all groupings are fully recovered. The *Rickettsi1 or Rickettsia-bellii-RML369-C* organism is known to have diverged out of the more common *Rickettsi* genera for spotted fever and typhus [30] and hence its relative disassociation.

Fig. 6. Phylogeny for a random sample of microbia

It is evident that similar organisms are grouped together in the phylogeny shown in Figure 6. For example, the organisms for *Buchnera, Shewanella, Helicobact, Bartonell, Bacillus, Streptococci, Actinoba, Thermotoga, Bacteroide* are fully recovered. The exceptions are the *Rickettsi and Brucella*. The *Brucella* organisms actually belong to two different genera and, indeed are distinct from each other by more than the *Bartonella* group and hence the placement and they are shown to share a common ancestor correctly. The *Rickettsi* on the other hand isolates one organism from the other two. Stothard et al proved [30] that the isolated organism *Rickettsi1 or Rickettsia-bellii-RML369-C* has diverged out of the other genera of *Rickettsi* much earlier. The distance matrix for this group alone (Figure 7) clearly shows the distinction in the *Rickettsi* sample.

Having shown the efficacy of EMM signature differentiation of a random sample of microbial taxa in Figure 6, a final experiment was performed to differentiate a more diverse sample that includes microbia as well as organisms of eukaryotic origin. This experiment adds an RNA sequence (obviously not of 16S rRNA) of a brown dog tick *Rhipicephalus-sanguineus-synganglion*, sand fly *Normalized-Phlebotomus-papatasi* and mouse obtained from the NCBI Refseq sites to a microbial sample consisting of 7 phyla, 8 classes and 10 genera. The phylogeny is shown in Figure 8 which clearly shows the proximity of all *Bacillus* and

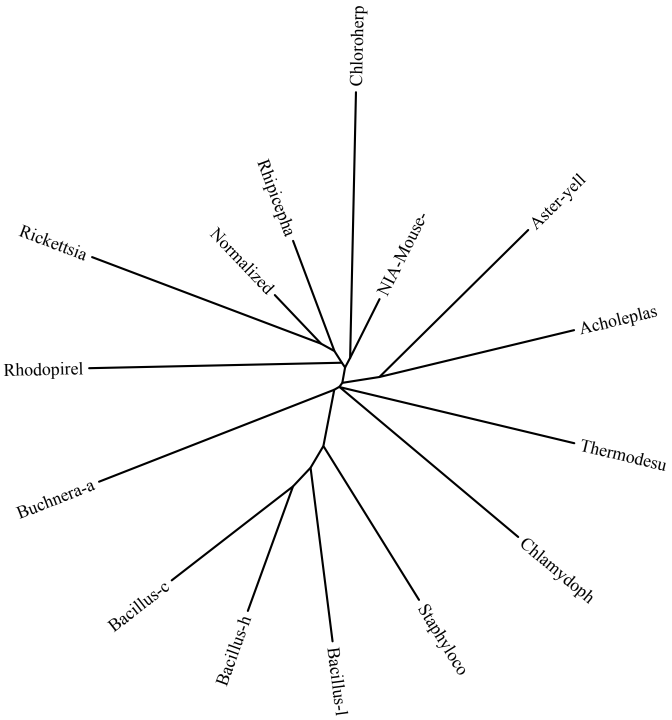


Rickettsia Organisms		Rickettsi1	Rickettsi2	Rickettsia
bellii-RML369-C	Rickettsi1	0	305	308
canadensis-str--McKiel	Rickettsi2	305	0	46
conorii-str--Malish-7	Rickettsia	308	46	0

The distance matrix for the *Ricketts*i sample of three organisms clearly shows the separation of *Rickettsia-bellii-RML369-C* from the other two belonging to the spotted fever category. This accounts for diversity shown in the phylogeny of Figure 6.

Fig. 7. Diversity of Rickettsi

the nearest organism *Staphylococcus* while separating the eukaryotic species *brown dog tick, sand fly and mouse* to the out edges.



The experiment shows phylogenetic relationship of the sample consisting of three Eukaryotic RNA from *tick (Rhipicephalus-sanguineus-synganglion), fly (Normalized-phlebotomus-papatasi) and mouse* and 12 microbia collection of 7 phyla, 8 classes and 10 genera. The close grouping of Bacillus and the distantly co-located eukaryotic species reflects the effectiveness of using EMM signature differentiation for phylogeny.

Fig. 8. Phylogeny of a diverse sample of microbia and some eukaryota

V. Discussion and Conclusion

Though 16S rRNA is the marker of choice for the majority of microbial classification, its heterogeneity due to multiple copies of the gene causes difficulties in differentiating organisms when traditional methods are employed. We developed a new sequence analysis approach using all the available sequence copies of 16S rRNA of a microbial organism. First, a compact model that includes the complex intra-sequence order is generated using an Extensible Markov Model (Figure 1). Such compact models are called EMM signatures and transforming organisms with all their 16S rRNA sequences to EMM signatures creates a *signature library*. The library can then be used for comparative genomic analysis such as all-to-all differentiation. Evaluation of an EMM signature against another is proposed to derive a unique and novel distance metric (Figure 2). The distance metric obeys the criteria proposed in [24] to ensure that the requirements of identity, symmetry, triangular inequality and the four point rule [25] are satisfied. Using the distance metric, we showed how to generate distance matrices in a format compatible for analysis with PHYlogeny Inference Package (PHYLIP) [27] available on the web [28].

The NCBI sequence database for the 16S rRNA [22] is used to build an EMM signature library for 500 organisms. Four experiments were performed to verify the proposed method and the metric. First, a diverse genus such as *Burkholderia* of class *Betaproteobacteria*[23] was used to generate the distance matrix (Figure 3). The phylogeny for *Burkholderia* genus was derived using an improve neighbor joining algorithm [31] called BIONJ [26] available as part of the PHYLIP package [27] at [28]. The phylogeny shown in Figure 4 was analyzed. The diversity of the genus is explored by examining the statistical significance of the mis-associations in cases of *B.mallei, B.pseudomallei and B.cencepia*. The distance matrix also confirmed the unusually high significance levels for the mis-associations. We investigated this further by increasing the clustering to further compress the signatures to highlight the differentiation. The resulting partial distance matrix in Figure 5 clearly showed that one of the organisms from both *B.mallei and B.pseudomallei* are indeed distinct and the proximity of a pair of organisms from both is indicative of their co-evolution. Second, another experiment was performed using a random sample consisting of 5 phyla, 7 classes and 13 genera and the phylogeny shown in Figure 6 confirms topological accuracy. The only exception seen is due to *Rickettsia-bellii* which is found to have diverged much earlier from the related organisms [30] and hence its isolation from the rest in the figure. Third, another experiment with a sample consisting of three eukaryotic RNA sequences from brown dog tick, mouse and sand fly as well as 15 organisms of microbial origin was performed. The resulting phylogeny [Figure 8] clearly isolated the three eukaryotic sequences and also showed the proximity of the bacillus and related organisms.

Several approaches for phylogeny are in existence, but all of them either require multi-sequence alignment or do not use all of the available sequence information. Yet, new approaches are emerging proposing to use the whole genomes and other markers such as *RecA, RPOB, 23s rRNA* [3, 32] etc. Though we have shown the EMM method using the multi-copy 16S marker, it is equally applicable to single copy markers where the signature form representation and alignment-free processing could improve space and time complexities. Applying the EMM signature concept to profiling sequence communities, Metagenomic classification and diagnostic identification of strain level sequence data will be included as future research areas.

References

[1] J. E. Clarridge, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases." *Clin Microbiol Rev*, vol. 17, no. 4, pp. 840–62, table of contents, Oct 2004. [Online]. Available: http://dx.doi.org/10.1128/CMR.17.4.840-862.2004

[2] W. G. Weisburg, S. M. Barns, D. A. Pelletier, and D. J. Lane, "16S ribosomal DNA amplification for phylogenetic study." *J Bacteriol*, vol. 173, no. 2, pp. 697–703, Jan 1991.

[3] R. J. Case, Y. Boucher, I. Dahllf, C. Holmstrm, W. F. Doolittle, and S. Kjelleberg, "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies." *Appl Environ Microbiol*, vol. 73, no. 1, pp. 278–288, Jan 2007. [Online]. Available: http://dx.doi.org/10.1128/AEM.01177-06

[4] T. G. Lilburn and G. M. Garrity, "Exploring prokaryotic taxonomy." *Int J Syst Evol Microbiol*, vol. 54, no. Pt 1, pp. 7–13, Jan 2004.

[5] D. Moreira and H. Philippe, "Molecular phylogeny: pitfalls and progress." *Int Microbiol*, vol. 3, no. 1, pp. 9–16, Mar 2000.

[6] M. H. Dunham, Y. Meng, and J. Huang, "Extensible markov model," in *Proc. Fourth IEEE International Conference on Data Mining ICDM '04*, 1–4 Nov. 2004, pp. 371–374.

[7] R. M. Kotamarti, D. W. Raiford, M. Hahsler, Y. Wang, M. McGee, and M. H. Dunham, "Targeted genomic signature profiling with quasi-alignment statistics," *COBRA Preprint Series*, November 2009. [Online]. Available: http://biostats.bepress.com/cobra/ps/art63

[8] Y. Meng and M. H. Dunham, "Online mining of risk level of traffic anomalies with user s feedbacks," *Granular Computing, 2006 IEEE International Conference*, pp. 176–181, 2006.

[9] Y. Meng, M. H. Dunham, M. Marchetti, and H. Jie, "Rare event detection in a spatiotemporal environment," *Granular Computing, 2006 IEEE International Conference*, pp. 629–634, 2006.

[10] M. T. Bergey's, Ed., *Manual of Systematic Bacteriology, Vol. 15 (20012009) 2nd edn*, 2nd ed. New York: Springer Verlag, 2009.

[11] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." *Proc Natl Acad Sci U S A*, vol. 87, no. 6, pp. 2264–2268, Mar 1990.

[12] S. Vinga and J. Almeida, "Alignment-free sequence comparison-a review." *Bioinformatics*, vol. 19, no. 4, pp. 513–523, Mar 2003.

[13] J. L. Thorne and H. Kishino, "Freeing phylogenies from artifacts of alignment." *Mol Biol Evol*, vol. 9, no. 6, pp. 1148–1162, Nov 1992.

[14] R. Shahid, S. Bertazzon, M. Knudtson, and W. Ghali, "Comparison of distance measures in spatial analytical modeling for health service planning," *BMC Health Services Research*, vol. 9, no. 1, p. 200, 2009. [Online]. Available: http://www.biomedcentral.com/1472-6963/9/200

[15] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment." *J Comput Biol*, vol. 1, no. 4, pp. 337–348, 1994.

[16] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res*, vol. 22, no. 22, pp. 4673–4680, Nov 1994.

[17] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." *BMC Bioinformatics*, vol. 5, p. 113, Aug 2004. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-5-113

[18] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks." *Proc Natl Acad Sci U S A*, vol. 89, no. 22, pp. 10 915–10 919, Nov 1992.

[19] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, no. suppl 3, pp. 345–351, 1978.

[20] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[21] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep 1997.

[22] [Online]. Available: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

[23] L. Vial, M.-C. Groleau, V. Dekimpe, and E. Dziel, "Burkholderia diversity and versatility: an inventory of the extracellular products." *J Microbiol Biotechnol*, vol. 17, no. 9, pp. 1407–1429, Sep 2007.

[24] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction." *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, Nov 2003.

[25] P. Buneman, "The recovery of trees from measures of dissimilarity," *Mathematics the the Archeological and Historical Sciences*, p. 387395, 1971.

[26] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Mol Biol Evol*, vol. 14, no. 7, pp. 685–695, Jul 1997.

[27] J. Felsenstein, "PHYLIP (phylogeny inference package), version 3.57 c," *Seattle: University of Washington*, 1995.

[28] B. Nron, H. Mnager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery, and C. Letondal, "Mobyle: a new full web bioinformatics framework." *Bioinformatics*, vol. 25, no. 22, pp. 3005–3011, Nov 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp493

[29] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." *Appl Environ Microbiol*, vol. 72, no. 7, pp. 5069–5072, Jul 2006. [Online]. Available: http://dx.doi.org/10.1128/AEM.03006-05

[30] D. R. Stothard, J. B. Clark, and P. A. Fuerst, "Ancestral divergence of Rickettsia bellii from the spotted fever and typhus groups of Rickettsia and antiquity of the genus Rickettsia." *Int J Syst Bacteriol*, vol. 44, no. 4, pp. 798–804, Oct 1994.

[31] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Mol Biol Evol*, vol. 4, no. 4, pp. 406–425, Jul 1987.

[32] I. Dahllf, H. Baillie, and S. Kjelleberg, "rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity." *Appl Environ Microbiol*, vol. 66, no. 8, pp. 3376–3380, Aug 2000.

Appendix: Distance $D(e_1, e_2)$ is a metric

The distance function in its simplest form only considers the number of unsupported transitions when evaluating a guest EMM against the host EMM. This is because when the quasi alignments are associated with transitions that exist in the host, the difference centroid tends to be zero or very close to zero. On the other hand, when the quasi alignments occur with no existing arc between the currently matched host state and the previously matched state, a penalty equal to $-log(|V_h|)$ is applied as a weight. When quasi alignment is established between a guest NSV and a host node, the difference between their centroids generates a difference vector. For scoring, each element in the vector is multiplied by a score and all the scored elements are aggregated to derive score. The component distance $d(e_1, e_2)$ is a sum of all scored quasi-alignments multiplied again by a penalty of fixed value. If a score of unity is used, the score of each quasi alignment becomes a standard Manhattan distance multiplied by a constant penalty value. Thus $d(e_1, e_2)$ is also a Manhattan distance multiplied by a constant value. Applying the same for $d(e_2, e_1)$, the final distance $D(e_1, e_2)$ is derived by $d(e_1, e_2)+d(e_2, e_1)$ which is again a Manhattan distance multiplied by a constant penalty value. Since Manhattan distance is a known valid distance metric and adding two Manhattan distances or multiplying a constant value still upholds the metric attribute, our distance function $D(e_1, e_2)$ is also a distance metric.