A Novel Quasi-Alignment-Based Method for Discovering Conserved Regions in Genetic Sequences

Anurag Nagar
Computer Science and Engineering
Southern Methodist University
Dallas, Texas 750205
Email: anagar@smu.edu

Michael Hahsler
Engineering Management, Information, and Systems
Southern Methodist University
Dallas, Texas 75205
Email: mhahsler@lyle.smu.edu

***Abstract*—This paper presents an alignment-free technique to efficiently discover similar regions in large sets of biological sequences using position sensitive p-mer frequency clustering. A set of sequences is broken down into segment and then a frequency distribution over all oligomers of size p (referred to as p-mers) is obtained to summarize each segment. These summaries are clustered while the order of segments in the set of sequences is preserved in a Markov-type model. Sequence segments within each cluster have very similar DNA/RNA patterns and form a so called quasi-alignment. This fact can be used for a variety of tasks such as species characterization and identification, phylogenetic analysis, functional analysis of sequences and, as in this paper, for discovering conserved regions. Our method is computationally more efficient than multiple sequences alignment since it can apply modern data stream clustering algorithms which run in time linear in the number of segments and thus can help discover highly similar regions across a large number of sequences efficiently. In this paper, we apply the approach to efficiently discover and visualize conserved regions in 16S rRNA.**

***Keywords*-DNA/RNA sequences; quasi-alignment; multiple sequence alignment; conserved sequences**

I. Introduction

Comparing and aligning sequences is a fundamental task in bioinformatics. Among its many applications is the identification of similar regions across large numbers of sequences. Such analysis can provide useful details about which sequences or regions are characteristic of a particular species and which parts of sequences are conserved across multiple species. These applications are typically based on computationally expensive procedures (e.g., BLAST [1], BAlibase [2], T-Coffee [3], MAFFT [4], MUSCLE [5], [6], Kalign [7] and ClustalW2 and ClustalX2 [8]) which find regions of high similarity across multiple sequences using expensive sequence alignment. Because of the high computational cost involved, discovering similar regions across thousands of sequences of a particular genome becomes practically impossible without the use of high performance and parallel computing.

Statistical signatures [9] created from nucleotide composition frequencies offer an alternative to using classic alignment. These alignment-free methods reduce processing time and look promising for whole genome phylogenetic analysis where previously used methods do not scale well [10]. However, these methods do not produce any alignment information. Another class of methods has tried to isolate highly repetitive short patterns across the entire genome by observation based methods [11], [12]. This technique does not scale well and can not be used to discover sequence patterns across multiple sequences.

Clearly, there is a need for an efficient method to discover similar areas across multiple sequences of a species or even an entire genome which harnesses the computational convenience of alignment-free methods while at the same time providing some, at least, approximate alignment information. Such methods can be used to detect coding regions of DNA sequences, genes with similar or related functions across genomes, discover phylogenetic relationships, and characteristic sequences or regions for different species.

Position specific p-mer frequency clustering (based on the work in [13]) combines the alignment-free approach with high-throughput data stream clustering techniques to efficiently produce so called quasi-alignments for large scale sequence data. In this paper, we discuss the application of position specific p-mer frequency clustering to identify regions of high similarity across multiple sequences.

II. Representing Genetic Sequences in a Compressed format

Position sensitive p-mer clustering is based on the idea of comparing sequences using p-mer frequency counts instead of computationally expensive alignment between the original sequences. This idea is at the core of so-called alignment-free methods [9]. However, in contrast to these methods we count p-mer frequencies in a position specific manner and then use high-throughput data stream clustering [14] to group similar sequence segments. This approach completely avoids expensive alignment of sequences and can be done in time linear to total number of bases in the data set. However, because of the clustering of like sequence segments, a probabilistic local quasi-alignment is automatically achieved, i.e.,

segments grouped in the same cluster are considered to be quasi-aligned.

The occurrences of letters or base compositions {A, C, T, G} in a genetic sequence provide frequency information. The occurrences of all patterns of bases of length p generates a p-mer frequency representation for a sequence. Instead of global frequencies, we count p-mer frequencies locally to retain positional information by first splitting the sequence into segments of a given size L. Within each segment we count the frequencies for all possible p-mers. We call this frequency profile a Numerical Summarization Vector (NSV). For example, suppose we have an input segment containing ACGTGCACG. If counting 2-mers, the NSV count vector would be

$$\langle 0, 2, 0, 0, 1, 0, 2, 0, 0, 1, 0, 1, 0, 0, 1, 0 \rangle$$

representing counts for the subpatterns
⟨AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT⟩

As we move down the input sequence, in each new segment p-mers are counted. Segment sizes may be varied and may or may not overlap. Also different values for p could be used within the same sequences. However, in the simple case used in this paper these parameters are fixed.

Figure 1 summarizes the model building process. NSVs representing segments are clustered and in addition the sequence information for the NSVs is preserved in a directed graph $G = (N, E)$, where $N = c_1, c_2, \ldots, c_N$ is the set of clusters and $E = e_1, e_1, \ldots, e_E$ is the set of transitions between clusters [15]. This graph can be interpreted as a Markov chain, however, unlike a classical Markov Model, each node is not bound to a single symbol but to a cluster representing similar segments, or, more precisely, segments with similar p-mer distributions.

Since several NSVs (i.e., segments) can be assigned to the same cluster, the resulting model compresses the original sequence (or sequences if several sequences are clustered into the same model). The directed edges preserve order information between segments by the probabilities of traversal from segment to segment during the model building process. Note that data stream clustering algorithms only need a single pass over the data which results in a time complexity linear in the number of bases in the dataset. This also makes adding new sequences to an existing model fast since it only depends on the number of bases in the new sequences. For details see [15].

The similarity between NSVs used for clustering can be calculated using several measures. Measures suggested in the literature to compare sequences based on p-mer counts (alignment-free methods) include Euclidean distance, squared Euclidean distance, Kullback-Leibler discrepancy and Mahalanobis distance [9]. Recently, for Simrank [16] an even simpler similarity measure, the number of matching p-mers (typically with $p = 7$), was proposed for efficient search of very large databases.

The string edit distance [17] is also related to alignment and is computed in a similar way using dynamic programing. In the area of approximate string matching Ukkonen proposed to approximate the expensive computation of the edit distance between two strings by using q-grams (analog to p-mers in sequences) [18]. First, q-gram profiles are computed and then the distance between the profiles is calculated using Manhattan distance. The Manhattan distance between two p-mer NSVs, x and y, is defined as:

$$d_{\text{Manhattan}}(x, y) = \sum_{i=1}^{4^p} |x_i - y_i| \quad (1)$$

Manhattan distance also has a particularly straightforward interpretation for NSVs. The distance counts the number of p-mers by which two sequences differ which gives the following lower bound on the edit distance between the original sequences s_x and s_y:

$$d_{\text{Manhattan}}(x, y)/(2p) \leq d_{\text{Edit}}(s_x, s_y) \quad (2)$$

This relationship is easy to prove since each insertion/deletion/substitution in a sequences destroys at the most p p-mers and introduces at most p new p-mers. Although, we can construct two completely different sequences with exactly the same NSVs (see [18] for a method to create such strings), we are typically interested in sequences of high similarity in which case $d_{\text{Manhattan}}(x, y)/(2p)$ gets closer to the edit distance. Note, however, that position sensitive p-mer frequency clustering is not restricted to using Manhattan distance, it can use any distance/similarity measure defined on the frequency counts in NSVs.

A p-mer frequency cluster model can be created for a single sequence or a group of sequences. The advantage of this approach is that it compresses the sequence information first by creating NSVs and then reduces the number of NSVs by clustering. Typically, we will create a cluster model for a whole family of sequences by simply adding the NSVs of all sequences to a single model following the procedure in Figure 1. This will lead to even more compression since many sequences within a family will share NSV clusters stemming from similar sequence segments.

III. Implementation using Open Source R Package QuasiAlign

To make the described procedure accessible, we are developing an R package called QuasiAlign [19][1]. The package is built on top of a data stream clustering package also developed by one of the authors [15], [20]. The QuasiAlign package provides various methods for handling and storing genetic sequences in a database and for creating and visualizing GenModels. The process used in this paper is as follows:

[1]The QuasiAlign package is available at http://r-forge.r-project.org/projects/mmsa/

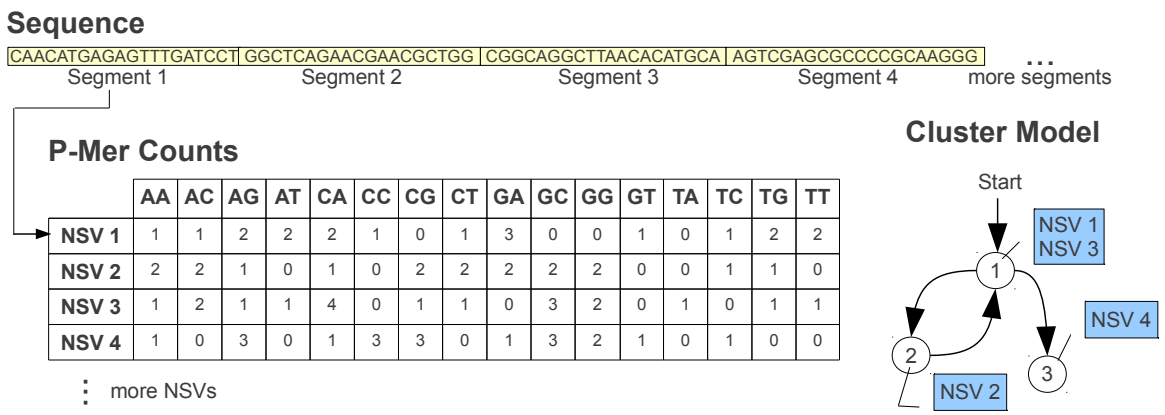

Figure 1. The Model Building Process. The sequence is split into several segments. For each segment a Numerical Summary Vectors (NSV) is calculated by counting the occurrence of p-mers (2-mers in this case). Model building starts with an empty cluster model. As each NSV is processed, it is compared to the existing clusters of the model. If the NSV is not found to be close enough (using a distance measure on the NSVs) a new cluster is created. For example Cluster 1 (circle) is created for NSV 1 and Cluster 2 for NSV 2. NSV 3 was found close enough to NSV 1 and thus was also assigned to Cluster 1. Finally, Cluster 3 is created for NSV 4. In addition to the clusters also the transition information between the clusters (arrows) is recorded. When all NSVs are processed, the model building process is finished.

1) Load the sequences into the database.
2) Create Numerical Summarization Vectors (NSVs) from a selected subset of sequences.
3) Use the NSVs to create position sensitive p-mer frequency clustering models called GenModel.
4) Prune or trim the model to remove noise (noise is represented by sparse clusters).
5) Visualize the model to identify highly similar regions across sequences.

There exists a function for each of the above tasks in the QuasiAlign package. Several parameters can be controlled. For selecting sequences from the database we can select sequences from a given phylogenetic rank (e.g., rank is "Phylum" and name is "Firmicutes" selects all sequences available for Firmicutes). To create NSVs we need to specify the segment size (default is 100 bases), if there is overlap between the segments (default is no overlap) and the value for p (default is 3-mers). Finally, for model building we need to specify the clustering measure (default is Manhattan) and the clustering threshold. The default threshold is 30 which means that we approximately require an edit distance of less than 5 to cluster two segments together (see Equation 2). For an in-depth description of all available parameters we refer the reader to the documentation of the QuasiAlign package [19].

IV. Experiments

A. Dataset

For the experiments presented in this paper, we used a set of approximately 400,000 16S rRNA sequences from the Greengenes project [21]. These sequences are widely used for classification and phylogenetic analysis of microorganisms. The QuasiAlign package contains an import routine for FASTA files from Greengenes which automatically places them into a relational database. The routine also extracts the available phylogenetic classification information and makes it available for querying. Each sequence from the Greengenes project has a unique identifier which is used as the primary key in our database.

B. Illustration

GenModels provide vital information about the similarity of segments and regions which are highly conserved across multiple sequences can be easily identified. Such areas are likely to be responsible for a particular function or provide a needed structural characteristic.

As an illustration, Figure 2 shows a GenModel plot of sequences from the phylum *Dictyoglomi*. This very small phylum consists of 17 16S rRNA sequences which vary in size between 1400 and 1500 bases. These sequences have been broken down into segments of size 100 bases, aggregated using 3-mers and clustered using the default settings. The resulting GenModel contains 41 clusters. The plot shows each of the clusters as circles where the number is just an id, but the circle size represents how many segments were assigned to it. The arrows represent the order in which the segments occurred in the original sequences. A stronger arrow indicates that more sequences have the two adjacent segments in the same order. For example, Figure 2 shows that one of the common transition paths is the cluster sequence $1 \rightarrow 2 \rightarrow 3 \rightarrow \cdots \rightarrow 14 \rightarrow 15$ indicating that most sequences are extremely similar as one would expect in a set of sequences of a single phylum. In addition the plot shows that almost all sequences go through a few

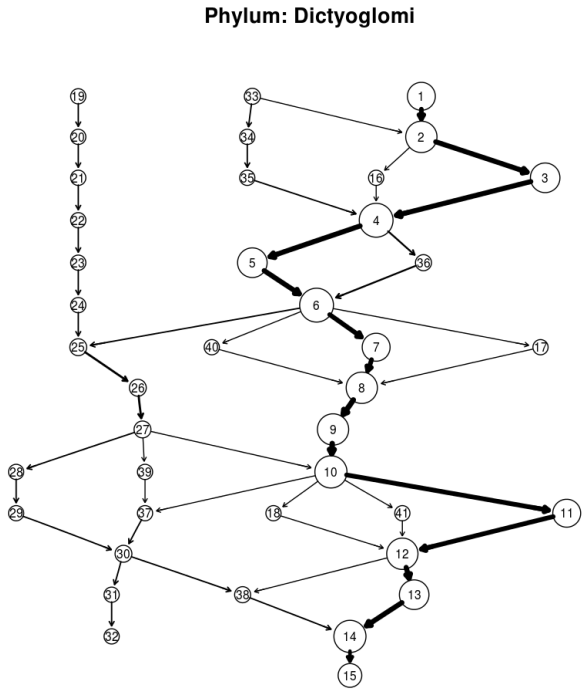

Figure 2. GenModel plot from 16S rRNA sequences from the phylum Dictyoglomi represents the sequences top-down. Circles indicated clusters of segments and arrows show the order of segments in the original sequences.

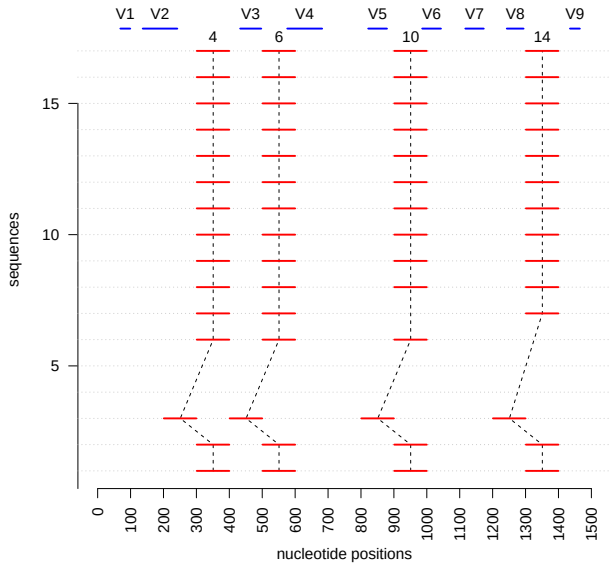

Figure 3. Visualizing the position of the segments for the four largest clusters of the GenModel of Dictyoglomi. These segments indicate well preserved regions in the sequences.

clusters (e.g. 4 and 6) which represent candidates for highly conserved regions. Interesting in Figure 2 is the almost completely separate path starting with cluster number 19. This indicates that a few sequences are very different from the majority of the sequences in the set. This might indicate a possible novel strain or might even point to a classification error which needs to be verified by taxonomists.

Each cluster in Figure 2 consists of members that are short segments of a larger sequence. The clusters can be also visualized in terms of common areas in sequences. Figure 3 shows the approximately 1500 bases (x-axis) for the 17 sequences (y-axis). The segments grouped into the 4 largest clusters of the GenModel are shown by red horizontal lines. In this model all red horizontal lines are exactly 100 bases long because a segment length of 100 was chosen. The segments that are part of the same cluster are joined by vertical dotted lines and the cluster id from Figure 2 is shown on top. We see that the well preserved segments are found in clusters 4, 6, 10 and 14 which corresponds to the largest clusters where almost all sequences join in Figure 2. We can make two more interesting observations in Figure 3. Sequence 3 has all matching segments shifted to the left. This indicates that the sequence is not complete and there are bases missing at the beginning of the sequence. Also we see that sequences 4 and 5 do not contain similar segments which is consistent with our observation from Figure 2 where a few sequences form a separate path starting with cluster 19.

Using the R-based package QuasiAlign, we can also inspect the p-mer distribution in different clusters. Figure 4 shows a barplot of the p-mer distribution for cluster 4 which represents a well preserved segment. The error bars show the variation of the counts of different p-mers in all segments that were grouped in the cluster. The plot shows the absence of certain 3-mers such as AGT, ATA, ATT, etc. In fact, out of the 4^3 or 64 possible 3-mers 14 are completely absent in this cluster across all the sequences.

C. Identifying Conserved Regions in 16S rRNA

It is well known that 16S rRNA contains hypervariable regions that are highly dissimilar between different species and are generally thought of as being characteristic of a particular species [22], [23]. It has also been reported that the hypervariable regions are flanked by highly conserved regions on both ends [24], [25]. These conserved regions have many possible applications such as PCR amplification using universal primers [22].

Nine identified hypervariable regions in 16S rRNA consist of nucleotides number 69–99, 137–242, 433–497, 576–682, 822–879, 986–1043, 1117–1173, 1243–1294 and 1435–1465 and are denoted by V1 through V9 respectively. GenModels can cluster similar segments and can therefore reveal which segments are highly similar across multiple sequences. We can use this fact to identify highly similar or conserved regions in a sample containing diverse species. These regions

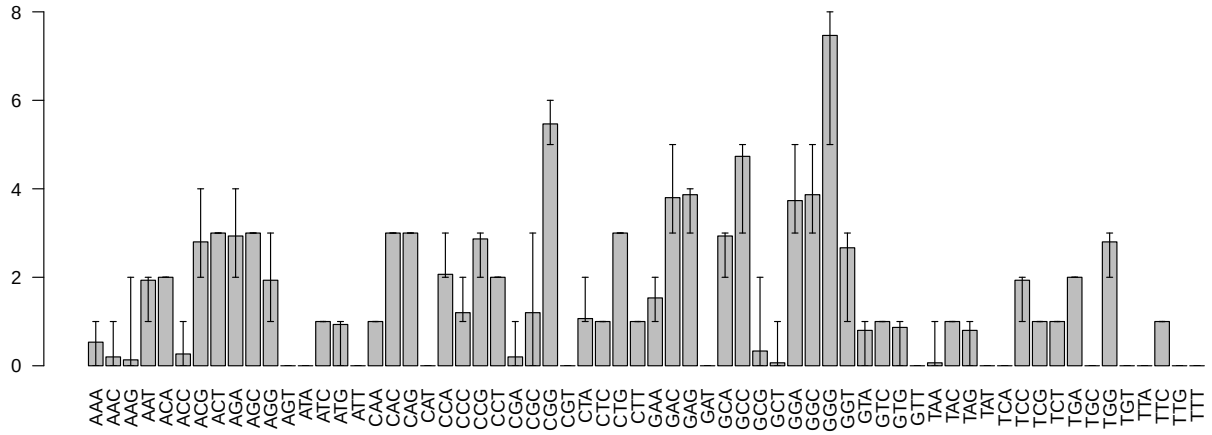

Figure 4. 3-mer distribution of segments grouped into cluster 4 for the GenModel for the phylum Dictyoglomi.

should also contain the known highly conserved regions which are the ones flanking the hypervariable regions.

The GenModel for the phylum Dictyoglomi contains several species and so we would expect greater variations in the hypervariable regions than in the flanking preserved regions. Figure 3 shows the 4 largest clusters found for the phylum Dictyoglomi. The positions of the hypervariable regions are shown as blue lines with labels V1 through V9 at the top of the plot. It is very clear that our algorithm identifies the regions that flank the hypervariable regions. For example, cluster 4 is to the immediate right of the hypervariable region V2 and to the immediate left of the hypervariable region V3. It thus covers the mentioned conserved areas. Similar arguments can be presented for clusters 6, 10, and 14.

D. Large-scale Experiments

The QuasiAlign package allows efficient large-scale experimentation, analysis, and visualization. We processed the entire Greengenes database consisting of over 400,000 sequences using the default settings and analyzed it for interesting patterns and clusters.

As an example, we present our analysis of the phylum Fusobacteria. This phylum was chosen because of its abundance in the human gut [26]. The Greengenes database contains around 1100 sequences of the phylum Fusobacteria. It has two major genera—Fusobacterium (with 482 sequences) and Cetobacterium (with 367 sequences). We have performed the analysis at the genus level this time to discover highly conserved areas within a given genus and also in a combined sample of the two genera to inspect conserved regions across genera. Since sequences from the same genus are closely related to each other, we would expect to discover more similarities in the hypervariable regions than across genera.

Figure 5 highlights the 5 largest clusters from the genus Fusobacterium. From the plot, it can be seen that the conserved segments obtained by our model are in close agreement with the known hypervariable regions. The top 5 identified regions cover most of the hypervariable regions V4, V5, V6, and V8. These regions were also found to be very similar when we ran a multiple sequence alignment on them using Clustal [8]. Greater coverage and more specificity can easily be obtained by using a smaller segment size and larger number of clusters which is omitted here for space restrictions.

Similarly, Figure 6 shows the 5 most conserved regions in the genus Cetobacterium. Here also a clear overlap can be seen with the hypervariable regions V1, V2, V5, V6, and V8.

In the next step, we combine the two genera (Fusobacterium and Cetobacterium) and try to find the most similar regions across the two genera. Since we have now a mixture of species in this sample, we expect more variation in the hypervariable regions. This is confirmed when we visualize the 5 most conserved regions (largest clusters) of the combined model in Figure 7. The most conserved clusters identified are now the regions flanking hypervariable regions V2–V7. This indicates that the model is able to efficiently identify conserved areas across multiple species.

E. Performance Analysis

The standard way to check for similar regions across a set of sequences is using multiple sequence alignment (MSA), which is known to be computationally expensive. Quasi-alignment can find similar regions across a large set of sequences easily and efficiently. Here we compare the

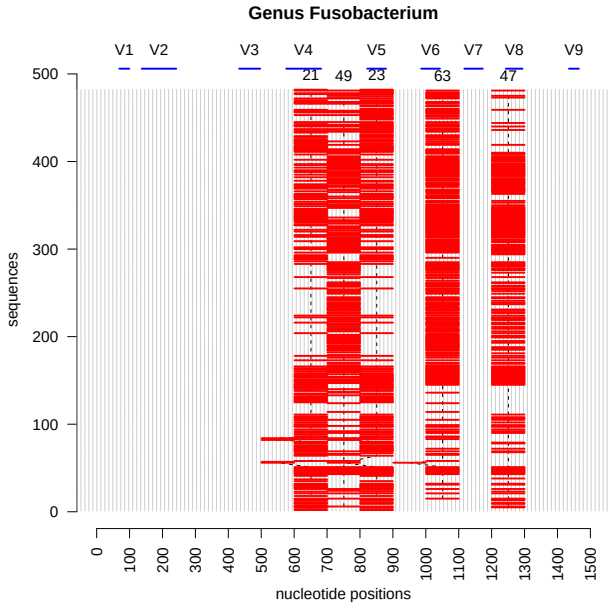

Figure 5. Segment similarity plot of the 5 largest clusters from the Genus Fusobacterium.

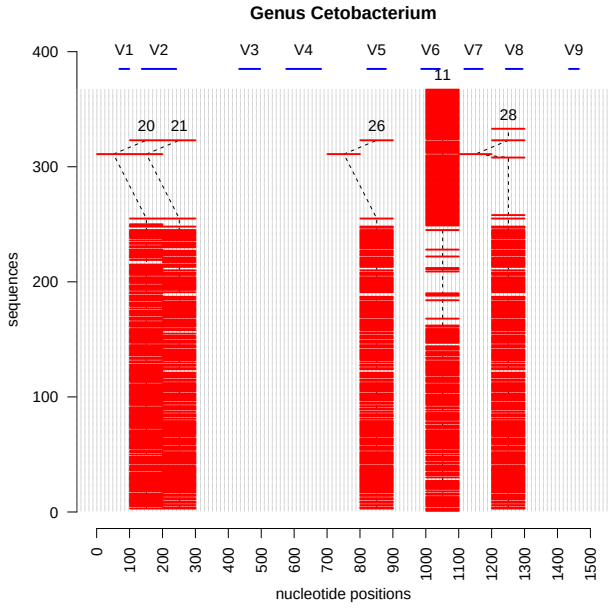

Figure 6. Segment similarity plot of the 5 largest clusters from the Genus Cetobacterium.

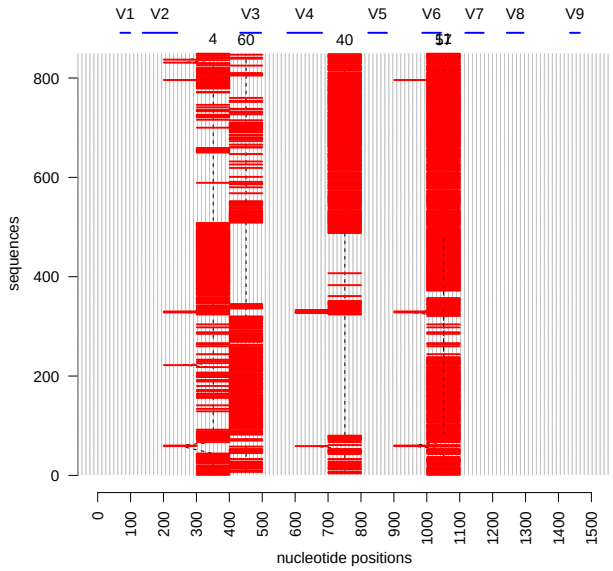

Figure 7. Segment similarity plot of the 5 largest clusters from the Genera Fusobacterium and Cetobacterium combined.

run times of our quasi-alignment method with progressive MSA, a popular MSA method. We randomly sampled between 10 and 200 16S rRNA sequences from the over 400,000 sequences available in the Greengenes project [21]. We performed both, MSA and our quasi-alignment based algorithm, on the sampled subset of sequences and then compared the run times in Figure 8. For the experiments, we used for progressive MSA the open source tool ClustalW2 [8], which is written in C++ and available for download at http://www.clustal.org/clustal2/. For quasi-alignment we used again the R package QuasiAlign with default settings. The experiments were run on OS X with a 2.6 GHz Intel Core i7 processor with 8 GB of main memory (both programs only use a single core).

Figure 8 shows that ClustalW2, which uses progressive MSA, has a run time polynomial in the number of sequences (which is much better than the exponential run time needed for dynamic programming based MSA). However, quasi-alignment is very fast and the run time only grows linearly with the number of sequences (and number of found NSV clusters) making it much more suited for large sets of sequences. It even enables us to do interactive analysis where the quasi-alignment is produced in a few seconds on the fly. Of course, MSA gives us much more alignment information, however, quasi-alignment can be used on a huge data set first and then MSA can be used for further analysis of how segments in a cluster align, reducing the number and size of sequences to be aligned significantly.

V. Discussion and Future Work

In this paper we have presented how sensitive p-mer frequency clustering, an efficient alignment-free method, can be used to discover highly conserved regions.

We have performed experiments on 16S rRNA sequences obtained from the Greengenes project and used out implementation of the R package QuasiAlign to demonstrate

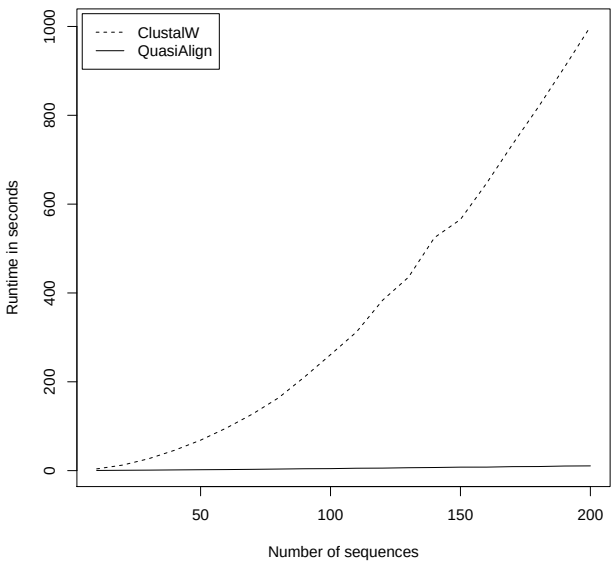

Figure 8. Comparing the run times of our quasi-alignment method with traditional multiple sequence alignment.

how several visualization methods can be used to efficiently identify conserved regions. The identified conserved regions lie in the immediate vicinity of known hypervariable regions which is known to be highly conserved.

In this paper we have only presented initial experiments. Future work we will build models for lower phylogenetic ranks down to the species level. This will allow us to identify ultra-conserved regions. Also, we will conduct comprehensive experiment with various parameter such as segment size, p-mer length, etc. Going to smaller segment sizes will allow us to find smaller conserved areas more accurately.

The most significant contribution of our work is its computational efficiency since it provides an approximation to alignment (a quasi-alignment) without requiring expensive multiple sequence alignment. The model can be created and inspected on a standard PC without the requirement of high performance computing tools.

ACKNOWLEDGMENT

This research was partially supported by the National Human Genome Research Institute under grant no. R21HG005912 and by the National Science Foundation under grant no. IIS-0948893.

REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[2] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, Mar. 1981.

[3] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, Sep. 2000.

[4] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[5] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, 2004.

[6] R. Edgar, "Muscle: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, no. 1, pp. 113+, Aug. 2004.

[7] T. Lassmann and E. L. Sonnhammer, "Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment." *Nucleic Acids Research*, vol. 34, July 2006.

[8] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, pp. 2947–2948, November 2007.

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

[10] J. D. Thompson, F. Plewniak, and O. Poch, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs," *Bioinformatics*, vol. 15, no. 1, pp. 87–88, Jan. 1999.

[11] R. K. Moyzis, J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. Meyne, R. L. Ratliff, and J. R. Wu, "A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes," *Proceedings of the National Academy of Sciences*, vol. 85, no. 18, pp. 6622–6626, Sep. 1988.

[12] D. L. Grady, R. L. Ratliff, D. L. Robinson, E. C. McCanlies, J. Meyne, and R. K. Moyzis, "Highly conserved repetitive DNA sequences are present at human centromeres." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 5, pp. 1695–1699, 1992.

[13] R. M. Kotamarti, M. Hahsler, D. Raiford, M. McGee, and M. H. Dunham, "Analyzing taxonomic classification using extensible Markov models," *Bioinformatics*, vol. 26, no. 18, pp. 2235–2241, 2010.

[14] C. Aggarwal, Ed., *Data Streams – Models and Algorithms*. Springer, 2007.

[15] M. Hahsler and M. H. Dunham, "rEMM: Extensible Markov model for data stream clustering in R," *Journal of Statistical Software*, vol. 35, no. 5, pp. 1–31, 2010. [Online]. Available: http://www.jstatsoft.org/v35/i05/

[16] T. DeSantis, K. Keller, U. Karaoz, A. Alekseyenko, N. Singh, E. Brodie, Z. Pei, G. Andersen, and N. Larsen, "Simrank: Rapid and sensitive general-purpose k-mer search tool," *BMC Ecology*, vol. 11, no. 1, Apr. 2011.

[17] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, 1966.

[18] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theoretical Computer Science*, vol. 92, no. 1, pp. 191–211, 1992.

[19] A. Nagar and M. Hahsler, *QuasiAlign: Infrastructure for Quasi-Alignment of Genetic Sequences*, 2012, R package version 0.0-3. [Online]. Available: http://r-forge.r-project.org/projects/mmsa/

[20] M. Hahsler and M. H. Dunham, *rEMM: Extensible Markov Model for Data Stream Clustering in R*, 2012, R package version 1.0-3. [Online]. Available: http://CRAN.R-project.org/package=rEMM

[21] "Greengenes website – 16S rRNA gene database," 2012, Accessed: 05/2012. [Online]. Available: http://greengenes.lbl.gov

[22] S. Chakravorty, D. Helb, M. Burday, N. Connell, and D. Alland, "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria." *Journal of Microbiological Methods*, vol. 69, no. 2, pp. 330–339, 2007.

[23] Y. Van de Peer, S. Chapelle, and R. De Wachter, "A quantitative map of nucleotide substitution rates in bacterial rRNA," *Nucleic Acids Res*, vol. 24, no. 17, pp. 3381–91+, 1996.

[24] G. C. Baker, J. J. Smith, and D. A. Cowan, "Review and re-analysis of domain-specific 16s primers." *Journal of Microbiological Methods*, vol. 55, no. 3, pp. 541–555, 2003.

[25] K. M. McCabe, Y. H. Zhang, B. L. Huang, E. A. Wagar, and E. R. McCabe, "Bacterial species identification after dna amplification with a universal primer pair." *Molecular Genetics and Metabolism*, vol. 66, no. 3, pp. 205–211, 1999.

[26] M. Arumugam et al., "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, May 2011.