Given a large number of protein sequences from various organisms.Determine the smallest non-redundant representative set.

Shaik Rumaan
BIM2016004[1]

Abstract— **The objective of this project is given a large number of protein sequences from various organisms.Determine the smallest non-redundant representative set.**

Index Terms— **Kmer,Clustering, Frequency,Vector.**

I. INTRODUCTION

The 1980s and 1990s were a flourishing time for bioinformatics, where the emergence of sequence comparison algorithms revolutionized the computational and molecular biology fields.At that time, many computational biologists quickly became stars in the field by developing programs for sequence alignment, which is a method that positions the biological sequences building blocks to identify regions of similarity that may have consequences for functional,structural, or evolutionary relationships. Many successful alignment-based tools were created including sequence similarity search tools e.g BLAST,FASTA, multiple sequence aligners e.g ClustalW , Muscle, MAFFT , sequences profile search programs e.g PSI-BLAST these tools became game-changers for anyone who wanted to assess the functions of genes and proteins.All alignment-based programs, regardless of the underlying algorithm, look for correspondence of individual bases or amino acids or groups that are in the same order in two or more sequences. The procedure assumes that every sequence symbol can be categorized into at least one of two states match or mismatch although most alignment programs also model inserted/deleted states (gaps). However, as our understanding of complex evolutionary scenarios and our knowledge about the patterns and properties of biological sequences advanced, we gradually uncovered some downsides of sequence comparisons based solely on alignments.

II. MOTIVATION

As the size of public sequence databases doubles every two years. Searching the ever larger and more redundant databases is getting increasingly inefficient.Clustering can help to organize sequences into homologous and functionally similar groups and can improve the speed, sensitivity, and readability of homology searches.

III. ALGORITHM DESIGN

A. Background

The rationale behind this method is simple: similar sequences share similar words/k-mers (sub sequences of length k), and mathematical operations with the words occurrences give a good relative measure of sequence dissimilarity.

B. Methodology

We have divided this particular problem in 4 stages.

C. word frequency

1) Stage-1: First, the sequences being compared must be sliced up into collections of unique words of a given length.For example, two random sequences x = ATGTGTG and y =CATGTG and a word size of three nucleotides (3-mers)produces two collections of unique words: WX3 = ATG,TGT, GTG and WY3 = CAT, ATG, TGT, GTG. Because some words are often present in one sequence but not in the other sequence (i.e., CAT in y but not in x), we create a full set of words that belong to at least WX3 or WY3 to further simplify the calculations, resulting in the union set W3 = ATG, CAT, GTG, TGT.

2) Stage-2: The second step is to transform each sequence into an array of numbers (vector) (e.g., by counting the number of times each particular word (from W3) appears within the sequences). For sequences x and y, we identify two real-valued vectors:cX3 = (1, 0, 2, 2) and cY3 = (1, 1, 1, 1).

3) Stage-3: This step includes quantification of the dissimilarity between sequences through the application of a distance function to the sequence-representing vectors cX3 and cY3. This difference is very commonly computed by the Euclidean distance, although any metric can be applied. The higher the dissimilarity value(Threshold), the more distant the sequences; thus, two identical sequences will result in a distance of 0.

D. clustering

1) Stage-4: The final step includes clustering them based on the above dissimilarity measure and including the proteins to the representative set.For that, each sequence is compared with the other sequence . If the similarity with any sequence is above a given threshold, it is grouped into that cluster. Otherwise, a new cluster is defined with that sequence as the representative. For each sequence comparison, word frequency filter (stage1,stage2,stage3) above method is applied to the sequences to confirm whether the similarity is below

the clustering threshold.Here threshold means dissimilarity value.

Word-based alignment-free algorithms come in different colors and flavors, with methodological variations at each of the three basic steps. In the first step, one can try any resolutions of word lengths—it is important to choose words that are not likely to commonly appear in a sequence (the shorter the word, then the more likely it will appear randomly in a sequence). In practice, the word size (k) of 2–6 residues produces stable and optimal protein sequence comparisons across a wide range of different phylogenetic distances. As a rule of thumb, smaller k-mers should be used when sequences are obviously different (e.g., they are not related) whereas longer k-mers can be used for very similar sequences.Alternatively, protein alphabet can be reduced to a smaller number of symbols based on chemical equivalences. This procedure may increase the detection of homologous sequences that display very low identity. For example, the proteins can be represented by 5, 4, 3, or even 2 letters according to their different physical–chemical properties .

The second step (mapping sequences onto vectors) is by far the most customizable instead of using vectors of word counts or word frequencies, there are many other ways to create vectors, which range from weighting techniques to normalization. Additionally, because word-based methods operate on vectors, their mathematical elegance allows the employment of many functions other than the Euclidean distance, such as the Pearson correlation coefficient,Manhattan distance, and Google distance.

IV. ANALYSIS

A. Analysis

1) Stage-1: In the Stage 1 The complexity is the size of kmer(k) multiplied by number of protein sequences(n) means O(n*k).

2) Stage-2: In the Stage 2 The complexity is number of element in the union of the set of kmers multiplied with number of kmers of a protein sequence.

3) Stage-3: In the Stage 3 The complexity is the number of element in the union of the set of kmers of both sequences

4) Stage-4: In the Stage 4 The complexity is the number of protein sequences multiplied by number of protein sequences multiplied by time of (stage1+stage2+stage3).

V. DISCUSSIONS

The method discussed above is one of the ways to solve the problem but they are many other methods like the prominent open source soft wares that were developed like CD-HIT,UCLUST,KCLUST these are the bench mark methods and these methods can handle data sets of very large size

Alignment producing programs assume that homologous sequences comprise a series of linearly arranged and more or less conserved sequence stretches. However, this assumption, which is termed collinearity, is very often violated in the real world. A good example is viral genomes, which exhibit great variation in the num ber and order of genetic elements due to their high mu tation rates, frequent genetic recombination events, horizontal gene transfers, gene duplications, and gene gains/losses . These large-scale evolutionary pro cesses essentially occur all the time in the genomes of other organisms. As a result, each genome becomes a unique lineage specific segments i.e regions shared with a subset of other genomes. Furthermore, the alignment approach may often overlook rearrangements on an even smaller scale for instance, the linear and modular organization of proteins is not always preserved due to frequent domain swapping, or duplication or deletion of long peptide motifs .

The computation of an accurate multiple- sequence alignment is an NP-hard problem, which means that the alignment cannot be solved in a realistic time frame. This situation explains why more than 100 alternative faster methods have been developed over the past three decades . However, the speed optimization does not come without cost. These techniques rely on various shortcuts means heuristics that do not guarantee the identification of the optimal and highest scoring alignment and often result in inaccuracies that limit the quality of many downstream analyses e.g phylogenetic. The complexity of the sequence alignment problem even calls for crowd sourcing solutions.

VI. CONCLUSIONS

The rationale behind this method is simple: similar sequences share similar words/k-mers (sub sequences of length k), and mathematical operations with the words occurrences give a good relative measure of sequence dissimilarity.The final step includes clustering them based on the above dissimilarity measure.

With the rapidly growing numbers of protein sequences from genome sequencing projects,methods that can cluster huge sequence sets in a reasonable time are in great need. Tools that are sensitive enough to cluster sequences together down to 30 sequence identity will be particularly useful, since at that similarity protein domains usually still have the same or very similar molecular functions, in particular if their domain architecture is conserved.

Most existing sequence clustering methods rely on BLAST or FASTA for calculating the similarities between the sequences to be clustered,which makes them too slow for many clustering tasks ahead.As sequencing technology becomes less expensive and more ubiquitous, the computational challenges of sequence analyses will become even more prominent. This issue pushes the current focus of development towards faster alignment-independent solutions.They do have some skeletons in their closet. For example, using long k-mers in word-based methods may impose a substantial memory overhead. Some of these issues have already been

addressed for example, recent reports demonstrate the reasonable memory usage of word-based approaches with long 25-mers for phylogenetic reconstruction of more than 100 bacterial genomes.Nevertheless, alignment free algorithms are rapidly extending the range of their applications and answering previously intractable questions in phylogenomics and horizontal gene transfer, population genetics, evolution of regulatory sequences, and links between the genome and epigenome . Disadvantages of nextgeneration sequencing data processing and analysis seem to be particularly well addressed by the alignment-free methods. The currently dominant kmer approaches are bound to novel measures for biological applications e.g Google distance.

VII. Result

To check performance of method it was tested on different data sets.Some of them are small sized and other are medium sized.The non-redundant representative set varies with the size of kmer and threshold.In practice, the word/kmer size (k) of 2–6 residues produces stable and optimal protein sequence comparisons across a wide range of different phylogenetic distances. As a rule of thumb, smaller k-mers should be used when sequences are obviously different e.g they are not related whereas longer k-mers can be used for very similar sequences.
First Case:
input: is given fasta file
k=5 and threshold 36
output: the final representative set protein ids
5E5-RAT,41-DROME,5HT2A-DROME,110KD-PLAKN,11S2-SESIN,4ET-HUMAN, 104K-THEPA, 104K-THEAN, 41-HUMAN,194K-TRVSY.
Second Case:
input: is given fasta file
k=6 and threshold 36
output:the final representative set protein ids
104K-THEAN, 194K-TRVSY, 11S2-SESIN, 41-HUMAN, 4ET-HUMAN, 41-DROME, 110KD-PLAKN, 104K-THEPA
Third Case:
input: is given fasta file
k=3 and threshold 30
output:the final representative set protein ids
5HT2B-DROME,108-SOLLC,1A1C-DIACA,2AAA-YEAST,41-CANFA, 3BP1-HUMAN,110KD-PLAKN,3S1-FAGES, 2A5D-YEAST, 104K-THEPA, 41-XENLA,5NTC-XENLA,5E5-RAT,4ET-HUMAN,194K-TRVSY,41-DROME, 104K-THEAN,5HT2A-DROME

VIII. Applications

Distantly related, remote sequences that evolve beyond recognizable similarity are one of the most classic applications of alignment-free mastering. For example, alignment-free approaches were successfully employed in functional annotation of unknown G-protein-coupled receptor integral cell membrane proteins that play a key role in transducing extracellular signals and have great relevance for pharmacology sequences that could not be assigned to any previously known receptor family

Another rising trend for the use of word-based alignment-free methods is the detection of functional and/or evolutionary similarities among regulatory se- quences e.g promoters, enhancers, and silencers to es- timate their in vivo activities in different organisms flies and mammals, including humans.

Sequence classification is another field that might benefit from bringing together different alignment-free approaches, such as grouping expressed sequences tags that originate from the same locus or gene family , clustering expressed sequence tag sequences with full- length cDNA data , and aggregating gene and pro- tein sequences into functional families Alignment-free methods are also used to recognize and classify antigens that are encoded in a sequence in a subtle and recondite manner that is not identifiable by sequence alignment. A recent approach based on the statistical transformation of protein sequences into uniform vectors with various amino acid properties showed an impressive prediction accuracy of up to 89in discriminating positive and negative sets of bacterial, viral, and tumor antigen data sets. Another common use of alignment free methods is the classification of species based on a short DNA sequence fragments that can act as true taxon bar codes.

References

[1] Sims GE, Jun S, Wu GA, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci USA.2009;106:2677–82.
[2] Wu T-J, Huang Y-H, Li L-A. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bio informatics.2005;21:4125–32.
[3] Höhl M, Ragan MA. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol.2007;56:206–21.
[4] Höhl M, Rigoutsos I, Ragan MA. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online. 2006;2:359–75.
[5] Alignment-free sequence comparison benefits, applications, and tools Andrzej Zielezinski1, Susana Vinga2 ,Jonas Almeida3 and Wojciech M. Karlowski1
[6] Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis.Brief Bioinform. 2014;15:890–905.
[7] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8.
[8] Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep. 2014;4:6504.
[9] Lee JC, Rashid NA. Adapting normalized google similarity in protein sequence comparison. International Symposium on Information Technolnology. September 2008. p. 1–5.
[10] Suwa M.Bioinformatics tools for predicting GPCR gene functions. In: Filizola M, editor. G protein-coupled receptors – modeling and simulation.Springer:Netherlands; 2014. p. 205–24.