Zielezinski *et al. Genome Biology* (2017) 18:186
DOI 10.1186/s13059-017-1319-7

Genome Biology

REVIEW Open Access

Alignment-free sequence comparison: benefits, applications, and tools

Andrzej Zielezinski[1], Susana Vinga[2], Jonas Almeida[3] and Wojciech M. Karlowski[1*]

Abstract

Alignment-free sequence analyses have been applied to problems ranging from whole-genome phylogeny to the classification of protein families, identification of horizontally transferred genes, and detection of recombined sequences. The strength of these methods makes them particularly useful for next-generation sequencing data processing and analysis. However, many researchers are unclear about how these methods work, how they compare to alignment-based methods, and what their potential is for use for their research. We address these questions and provide a guide to the currently available alignment-free sequence analysis tools.

Introduction

The 1980s and 1990s were a flourishing time not only for pop music but also 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 [1], FASTA [2]), multiple sequence aligners (e.g., ClustalW [3], Muscle [4], MAFFT [5]), sequences' profile search programs (e.g., PSI-BLAST [1], HMMER/Pfam [6]), and whole-genome aligners (e.g., progressive Mauve [7], BLASTZ [8], TBA [9]); 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 thereof) 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—conserved/similar (match) or non-conserved (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.

Five cases where alignment-based sequence analysis might be troublesome

First, 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 number and order of genetic elements due to their high mutation rates, frequent genetic recombination events, horizontal gene transfers, gene duplications, and gene gains/losses [10]. These large-scale evolutionary processes essentially occur all the time in the genomes of other organisms. As a result, each genome becomes a mosaic of 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 [11, 12].

Second, the accuracy of sequence alignments drops off rapidly in cases where the sequence identity falls below a certain critical point. For protein sequences, there are 20

* Correspondence: wmk@amu.edu.pl
[1] Department of Computational Biology, Faculty of Biology, Adam Mickiewicz University in Poznan, Umultowska 89, 61-614 Poznan, Poland
Full list of author information is available at the end of the article

possible amino acid residues, and any two unrelated sequences can match at up to 5% of the residues. If gaps are allowed, then the percentage can increase to 25% [13]. Thus, in practical applications, the area of 20–35% identity is commonly regarded as the "twilight zone" [14], where remote homologs mix with random sequences. Below 20% identity, in the realm of the "midnight zone", homologous relationships cannot be reliably determined with plain pairwise alignments, often requiring more sophisticated alignment-based solutions, like profiles (e.g., PSI-BLAST) and hidden Markov models (e.g., HMMER). This failure is especially problematic in the annotation of protein superfamilies where the members retain structural kinship even when the average intersequence identity is 8–10% [15]. For nucleotide sequences, the accuracy of the alignments is even more disappointing. For instance, two randomly related DNA/RNA sequences can show up to 50% sequence identity when gaps are allowed, and the edge of the twilight zone can encompass nucleotide matches of up to 60–65% [16–18].

Third, alignment-based approaches are generally memory consuming and time consuming and thus are of limited use with multigenome-scale sequence data. The number of possible alignments of two sequences grows rapidly with the length of the sequences (for two sequences of length N there are $(2N)!/(N!)^2$ different gapped alignments [19], which results in about 10^{60} alignments for two sequences of length 100). Although there is a method, called dynamic programming, that guarantees obtaining a mathematically optimal (highest scoring) alignment without listing all possible solutions, it is also computationally demanding (time complexity is in the order of the product of the lengths of the input sequences) [20]. Therefore, despite the wealth of tools and more than 15 years of research [7, 21–25], the problem of long sequence alignment is not fully resolved [26]. In addition, available sequence evolutionary models may not directly apply to complete genomes, as recently implicated by the Alignathon project, where over 50% of the aligned positions—at the nucleotide level—were inconsistent between pairs of 13 tested methods [26]. Therefore, even the designers of the alignment algorithms and browsers do not claim that their results are correct at all sites across entire genomes [27].

Fourth, 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 [28]. However, the speed optimization does not come without "cost". These techniques rely on various shortcuts (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 crowdsourcing solutions (e.g., creating the online game Phylo to improve computer-created multiple sequence alignments) [29].

Finally, a sequence alignment depends on multiple a priori assumptions about the evolution of the sequences that are being compared. These various parameters (e.g., substitution matrices, gap penalties, and threshold values for statistical parameters) are somewhat arbitrary, which additionally strains Occam's razor to breaking point. Moreover, the scoring system is not consensual between applications, and many reports have shown that small changes in the input parameters can greatly affect the alignment [30]. Despite the awareness of the problem, how to choose alignment parameters may often cause problems and usually requires a trial and error approach. (i.e., if an alignment is not good enough, then one can tweak input parameters to get "better-looking" results). Furthermore, reference substitution matrices required for protein alignments (e.g., different series of BLOSUM and PAM) are often used without verifying whether they are representative of the sequences being aligned. Intriguingly, BLOSUM matrices, which are the most commonly used substitution matrix series for protein sequence alignments, were found to have been miscalculated years ago and yet produced significantly better alignments than their corrected modern version (RBLOSUM) [31]; this paradox remains a mystery.

What is alignment-free sequence comparison?

Alignment-free approaches to sequence comparison can be defined as any method of quantifying sequence similarity/dissimilarity that does not use or produce alignment (assignment of residue–residue correspondence) at any step of algorithm application. From the start, such restriction places the alignment-free approaches in a favorable position—as alignment-free methods do not rely on dynamic programming, they are computationally less expensive (as they are generally of a linear complexity depending only on the length of the query sequence [32]) and therefore suitable for whole genome comparisons [33–36]. Alignment-free methods are also resistant to shuffling and recombination events and are applicable when low sequence conservation cannot be handled reliably by alignment [37]. Finally, in contrast to alignment-based methods, they do not depend on assumptions regarding the evolutionary trajectories of sequence changes. Although these characteristics apply to all alignment-free methods, there are more than 100 techniques to consider [37].

Alignment-free approaches can be broadly divided into two groups [38, 39]: methods based on the frequencies of subsequences of a defined length (word-based methods) and methods that evaluate the informational content between full-length sequences (information-theory based methods). There are also methods that cannot be classified in either of the groups, including those based on the length of matching words (common [40], longest common [41], or the minimal absent [42, 43] words between sequences), chaos game representation [44], iterated maps [45], as well as graphical representation of DNA sequences, which capture the essence of the base composition and distribution of the sequences in a quantitative manner [46, 47].

All of the alignment-free approaches are mathematically well founded in the fields of linear algebra, information theory, and statistical mechanics, and calculate pairwise measures of dissimilarity or distance between sequences. Conveniently, most of these measures can be directly used as an input into standard tree-building software, such as Phylip [48] or MEGA [49].

How do word frequency-based methods work?

The rationale behind these methods is simple: similar sequences share similar words/*k*-mers (subsequences of length *k*), and mathematical operations with the words' occurrences give a good relative measure of sequence dissimilarity. The method is also tightly coupled with the idea of genomic signatures, which were first introduced for dinucleotide composition (e.g., GC content) [50] and further extended to longer words. This process can be broken into three key steps (Fig. 1).

First, the sequences being compared must be sliced up into collections of unique words of a given length. For example, two DNA sequences x = ATGTGTG and y = CATGTG and a word size of three nucleotides (3-mers) produces two collections of unique words: W_3^X = {ATG, TGT, GTG} and W_3^Y = {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 W_3^X or W_3^Y to further simplify the calculations, resulting in the union set W_3 = {ATG, CAT, GTG, TGT}.

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 W_3) appears within the sequences). For sequences x and y, we identify two real-valued vectors: $c_3^X = (1, 0, 2, 2)$ and $c_3^Y = (1, 1, 1, 1)$.

The last step includes quantification of the dissimilarity between sequences through the application of a distance function to the sequence-representing vectors c_3^X and c_3^Y. This difference is very commonly computed by the Euclidean distance, although any metric can be applied [51]. The higher the dissimilarity value, the more distant the sequences; thus, two identical sequences will result in a distance of 0.

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

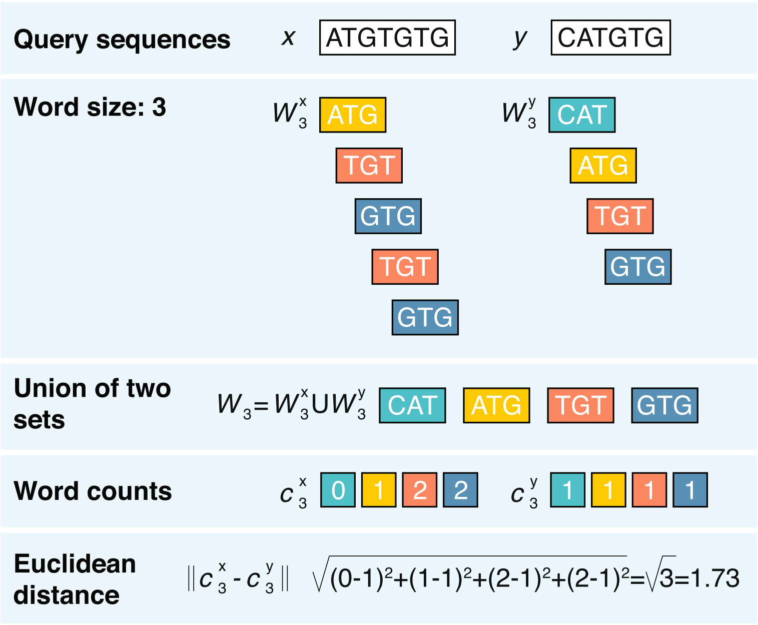


Fig. 1 Alignment-free calculation of the word-based distance between two sample DNA sequences ATGTGTG and CATGTG using the Euclidean distance

optimal protein sequence comparisons across a wide range of different phylogenetic distances [52, 53]; in nucleotide sequence analyses, *k* can safely be set to 8–10 for genes or RNA [54], 9–14 bases for general phylogenetic analyses [34, 55, 56], and up to 25 bases in case of comparison of isolates of the same bacterial species [33, 57]. 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 [55, 58]. Alternatively, DNA/RNA or 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 [53]. For example, the four-letter DNA alphabet can be distilled to two-letter purine–pyrimidine encoding [55], and proteins can be represented by 5, 4, 3, or even 2 letters according to their different physical–chemical properties [52]. 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 and clustering [32]. Additionally, because word-based methods operate on vectors, their mathematical elegance allows the employment of more than 40 functions other than the Euclidean distance, such as the Pearson correlation coefficient [38], Manhattan distance, and Google distance [59].

How do information theory-based methods work?

Information theory-based methods recognize and compute the amount of information shared between two analyzed biological sequences. Nucleotide and amino acid sequences are ultimately strings of symbols, and their digital organization is naturally interpretable with information theory tools, such as complexity and entropy.

For example, the Kolmogorov complexity of a sequence can be measured by the length of its shortest description. Accordingly, the sequence AAAAAAAAAA can be described in a few words (10 repetitions of A), whereas CGTGATGT presumably has no simpler description than specification nucleotide by nucleotide (1 C, then 1 G and so on). Intuitively, longer sequence descriptions indicate more complexity. However, Kolmogorov did not address the method to find the shortest description of a given string of characters. Therefore, the complexity is most commonly approximated by general compression algorithms (e.g., as implemented in zip or gzip programs) where the length of a compressed sequence gives an estimate of its complexity (i.e., a more complex string will be less compressible) [60]. The calculation of a distance between sequences using complexity (compression) is relatively straightforward (Fig. 2). This procedure takes the sequences being compared (*x* = ATGTGTG and *y* = CATGTG) and concatenates them to create one longer sequence (*xy* = ATGTGTG-CATGTG). If *x* and *y* are exactly the same, then the complexity (compressed length) of *xy* will be very close to the complexity of the individual *x* or *y*. However, if *x* and *y* are dissimilar, then the complexity of *xy* (length of compressed *xy*) will tend to the cumulative complexities of *x* and *y*. Of course, there are as many different information-based distances as there are methods to calculate complexity. For example, Lempel–Ziv complexity [61] is a popular measure that calculates the number of different subsequences encountered when viewing the sequence from beginning to end (Fig. 2). Once the complexities of the sequences are calculated, a measure of their differences (e.g., the normalized compression distance [62]) can be easily computed. Many DNA-specific compression algorithms are currently being applied to new types of problems [63].

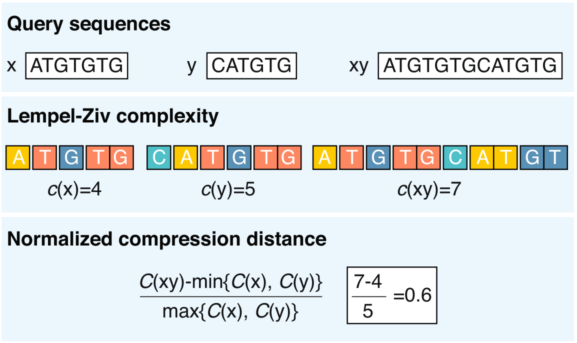


Fig. 2 Alignment-free calculation of the normalized compression distance using the Lempel–Ziv complexity estimation algorithm. Lempel–Ziv complexity counts the number of different words in sequence when scanned from left to right (e.g., for *s* = ATGTGTG, Lempel–Ziv complexity is 4: A|T|G|TG). Description of compression algorithms in alignment-free analysis has been reviewed extensively [63]

Another example of an information measurement often applied to biological sequences is entropy. This measurement is not similar to the entropy referenced in thermodynamics. Reportedly, Claude Shannon, who was a mathematician working at Bell Labs, asked John von Neumann what he should call his newly developed measure of information content; "Why don't you call it entropy," said von Neumann, "[…] no one understands entropy very well so in any discussion you will be in a position of advantage […]" [64]. The concept of Shannon entropy came from the observation that some English words, such as "the" or "a", are very frequent and thus unsurprising. Thus, these words are redundant because the message can probably be understood without them. The real essence of the message comes from words that are rare, such as "treasure" or "elixir". Therefore, Shannon developed a formula to quantify the uncertainty (entropy) of finding a given element (word) in an analyzed sequence (text). Using Shannon's concept, Kullback and Leibler [65] introduced a relative entropy measure (Kullback–Leibler divergence, KL) that allowed for a comparison of two sequences. The procedure involves the calculation of the frequencies of symbols or words in a sequence and the summation of their entropies in the compared sequences (Additional file 1: Figure S1).

Both information-theory concepts (complexity and entropy) have a clear association despite their methodical differences. For instance, a low-complexity sequence (e.g., AAAAAAAAA) will have smaller entropy than a more complex sequence (e.g., ACCTGATGT). The application of information theory in the field of sequence analysis and comparison has exploded in recent years, ranging from global (block entropies and coverage) to local genome analyses (transcription factor binding sites, sequences as time-series and entropic profiles) [39]. Additionally, retrieving higher-level correlations in gene mapping and protein–protein interaction networks and the striking resemblance with communication systems is attracting research attention to this field.

How are alignment-free methods used in next-generation sequencing data analysis?

The data volume of samples sequenced so far (estimated to be only 10^{-20}% of the total DNA on Earth [66]) is already challenging the storage and processing capacities of modern computers. In particular, the amount of data generated via next-generation sequencing is swiftly outpacing analytics capabilities, mainly due to the computationally intensive multiple alignment step. Alignment-free methods not only provide a significant increase in speed over primary next-generation sequencing applications (e.g., expression profiling [67–70], genetic variant calling [71–74], de novo genome assembly [75–77], phylogenetic reconstruction [78–81], and taxonomic classification in metagenomic studies [82–86]) (Table 1), but also offer ways to obtain biologically meaningful information directly from raw next-generation sequencing data.

For example, alignment-free tools for transcript quantification (Kallisto [69], Sailfish [67], Salmon [70]) show that most of the information provided by aligners is not necessary for high-quality estimation of transcript levels. These tools build an index of *k*-mers from a reference set of transcripts and then calculate the expression by matching them to each sequencing read directly. Such “pseudoalignment” [69] describes the relationship between a read and a set of compatible transcripts. Grouping pseudoalignments belonging to the same set of transcripts allows one to directly infer the expression of each transcript model. This approach to quantify gene/transcript expression levels from RNA sequencing reads is both 10–100 times faster than any of the alignment-based methods and at least as accurate as best-performing alignment-based workflows (e.g., TopHat-Cufflinks) [87, 88].

Another major application of next-generation sequencing technologies includes profiling of genomic variabilities, such as single nucleotide/variant polymorphisms. These genomic alterations are typically detected by genotype calling on mapped reads (e.g., Samtools mpileup [89] and GATK HaplotypeCaller [90]). However, alignment-free tools (FastGT [73] and LAVA [71]) allow for genotyping of known variants directly from next-generation sequencing data, based on *k*-mer analysis. Since these methods are 1–2 orders of magnitude faster than traditional mapping-based detection, they seem to be ideally suited for clinical applications, where sequencing data from a large number of individuals need to be processed in a timely manner. For example, MICADo analyzes third-generation sequencing reads for each patient sample within the context of the data of the whole cohort in order to capture patient-specific mutations [72] and ChimeRScope predicts fusion transcripts with potential oncogenic functions, based on the *k*-mer profiles of the RNA-seq paired-end reads [74].

Conventional next-generation sequencing computation came of age with the emergence of the MapReduce functional pattern to orchestrate parallelization of order-free operations [90]. It is, therefore, of no surprise that it would be advantageous to implement alignment-free methods for the same pattern. Such a solution comes naturally to the word counting implementation of *k*-mer analysis and may have further reaching implications for the molecular applications discussed in the previous paragraph, and it is also found to be a natural fit to scale-free approaches to alignment-free methods [91]. This solution was successfully put to the test in the simultaneous screening of 20 *Streptococcus pneumoniae* genomes for shared suffixes in a volunteer distributed computing implementation of that alignment-free MapReduced implementation [92].

One of the most demanding tasks in today's biology includes assembly of the newly sequenced genomes. In standard applications, it requires an error correction step and construction of the genome scaffold based on read similarity (sequence overlaps). Several alignment-free tools have been created to correct sequencing reads (e.g., Quorum [93]), designed mainly to be fast and memory efficient (e.g., Lighter [94] using sampling of *k*-mers instead of counting), as well as highly accurate (e.g., Trowel [95] using quality threshold rather than coverage cut-off in order to extract trusted *k*-mers).

The advent of third-generation sequencing technologies (PacBio and Oxford Nanopore) provides an opportunity to study new genomes with unprecedented speed and quality. However, the noisy nature of sequencing

Table 1 Alignment-free sequence comparison tools available for next-generation sequencing data analysis

Category	Analysis	Tool	Primary features	Implementation	Reference	URL
Mapping	Transcript quantification	kallisto	Transcript abundance quantification from RNA-seq data (uses pseudoalignment for rapid determination of read compatibility with targets)	Software (C++)	[69]	https://pachterlab.github.io/kallisto/
		Sailfish	Estimation of isoform abundances from reference sequences and RNA-seq data (*k*-mer based)	Software (C++)	[67]	http://www.cs.cmu.edu/~ckingsf/software/sailfish/
		Salmon	Quantification of the expression of transcripts using RNA-seq data (uses *k*-mers)		[70]	https://combine-lab.github.io/salmon/
		RNA-Skim	RNA-seq quantification at transcript-level (partitions the transcriptome into disjoint transcript clusters; uses *sig*-mers, a special type of *k*-mers)	Software (C++)	[68]	http://www.csbio.unc.edu/rs/
	Variant calling	ChimeRScope	Fusion transcript prediction using gene *k*-mers profiles of the RNA-seq paired-end reads	Software (Java)	[74]	https://github.com/ChimeRScope/ChimeRScope/wiki
		FastGT	Genotyping of known SNV/SNP variants directly from raw NGS sequence reads by counting unique *k*-mers	Software (C)	[73]	https://github.com/bioinfo-ut/GenomeTester4/
		Phy-Mer	Reference-independent mitochondrial haplogroup classifier from NGS data (*k*-mer based)	Software (Python)	[157]	https://github.com/danielnavarrogomez/phy-mer
		LAVA	Genotyping of known SNPs (dbSNP and Affymetrix's Genome-Wide Human SNP Array) from raw NGS reads (*k*-mer based)	Software (C)	[71]	http://lava.csail.mit.edu/
		MICADo	Detection of mutations in targeted third-generation NGS data (can distinguish patients' specific mutations; algorithm uses *k*-mers and is based on colored de Bruijn graphs)	Software (Python)	[72]	http://github.com/cbib/MICADo
	General mapper	Minimap	Lightweight and fast read mapper and read overlap detector (uses the concept of "minimazers", a special type of *k*-mers)	Software (C)	[77]	https://github.com/lh3/minimap
Assembly	De novo genome assembly	MHAP	Produces highly continuous assembly (fully resolved chromosome arms) from third-generation long and noisy reads (10 kbp) using a dimensionality reduction technique MinHash	Software (Java)	[76]	https://github.com/marbl/MHAP
		Miniasm	Assembler of long noisy reads (SMRT, ONT) using the Overlap-Layout Consensus (OLC) approach without the necessity of an error correction stage (uses minimap)	Software (C)	[77]	https://github.com/lh3/miniasm
		LINKS	Scaffolding genome assembly with error-containing long sequence (e.g., ONT or PacBio reads, draft genomes)	Software (Perl)	[75]	https://github.com/warrenlr/LINKS/
	Read clustering	afcluster	Clustering of reads from different genes and different species based on *k*-mer counts	Software (C++)	[158]	https://github.com/luscinius/afcluster
		QCluster	Clustering of reads with alignment-free measures (*k*-mer based) and quality values	Software (C++)	[159]	http://www.dei.unipd.it/~ciompin/main/qcluster.html
	Reads error correction	Lighter	Correction of sequencing errors in raw, whole genome sequencing reads (*k*-mer based)	Software (C++)	[94]	https://github.com/mourisl/Lighter
		QuorUM	Error corrector for Illumina reads using k-mers	Software (C++)	[93]	https://github.com/gmarcais/Quorum
		Trowel		Software (C++)	[95]	

Table 1 Alignment-free sequence comparison tools available for next-generation sequencing data analysis *(Continued)*

						https://sourceforge.net/projects/trowel-ec/
Metagenomics	Assembly-free phylogenomics	AAF	Phylogeny reconstruction directly from unassembled raw sequence data from whole genome sequencing projects; provides bootstrap support to assess uncertainty in the tree topology (*k*-mer based)	Software (Python)	[78]	https://github.com/fanhuan/AAF
		kSNP v3	Reference-free SNP identification and estimation of phylogenetic trees using SNPs (based on *k*-mer analysis)	Software (C)	[80, 81]	https://sourceforge.net/projects/ksnp/files/
		NGS-MC	Phylogeny of species based on NGS reads using alignment-free sequence dissimilarity measures d_2* and d_2^S under different Markov chain models (using *k*-words)	R package	[79, 160]	http://www-rcf.usc.edu/~fsun/Programs/NGS-MC/NGS-MC.html
	Species identification/ taxonomic profiling	CLARK	Taxonomic classification of metagenomic reads to known bacterial genomes using *k*-mer search and LCA assignment	Software (C++)	[84]	http://clark.cs.ucr.edu/
		FOCUS	Reports organisms present in metagenomic samples and profiles their abundances (uses composition-based approach and non-negative least squares for prediction)	Web service Software (Python)	[161]	http://edwards.sdsu.edu/FOCUS/
		GSM	Estimation of abundances of microbial genomes in metagenomic samples (*k*-mer based)	Software (Go)	[162]	https://github.com/pdtrang/GSM
		Mash	Species identification using assembled or unassembled Illumina, PacBio, and ONT data (based on MinHash dimensionality-reduction technique)	Software (C++)	[163]	https://github.com/marbl/mash
		Kraken	Taxonomic assignment in metagenome analysis by exact *k*-mer search; LCA assignment of short reads based on a comprehensive sequence database	Software (C++)	[83]	https://ccb.jhu.edu/software/kraken/
		LMAT	Assignment of taxonomic labels to reads by *k*-mers searches in precomputed database	Software (C++/Python)	[82]	https://sourceforge.net/projects/lmat/
		stringMLST	*k*-mer-based tool for MLST directly from the genome sequencing reads	Software (Python)	[86]	http://jordan.biology.gatech.edu/page/software/stringMLST
		Taxonomer	*k*-mer-based ultrafast metagenomics tool for assigning taxonomy to sequencing reads from clinical and environmental samples	Web service	[164]	http://taxonomer.iobio.io/
	Other	d2-tools	Word-based (*k*-tuple) comparison (pairwise dissimilarity matrix using d2S measure) of metatranscriptomic samples from NGS reads	Software (Python/R)	[56, 165]	https://code.google.com/p/d2-tools/
		VirHostMatcher	Prediction of hosts from metagenomic viral sequences based on ONF using various distance measures (e.g., d_2)	Software (C++)	[153]	https://github.com/jessieren/VirHostMatcher
		MetaFast	Statistics calculation of metagenome sequences and the distances between them based on assembly using de Bruijn graphs and Bray–Curtis dissimilarity measure	Software (Java)	[166]	https://github.com/ctlab/metafast

The up-to-date list of currently available programs can be found at http://www.combio.pl/alfree/tools/. Accessed 23 August 2017

LCA lowest common ancestor, *NGS* next-generation sequencing, *SNP* single-nucleotide polymorphism, *SNV* single-nucleotide variant data demands dedicated solutions to access more complex genomes. The MinHash Alignment Process was designed for this task employing probabilistic, locality-sensitive hashing. Integration of the MinHash Alignment Process with the Celera Assembler enabled reference-grade de novo assemblies of several eukaryotic genomes [76]. Another example includes currently developed Miniasm de novo assembler [77], which uses an overlap-layout-consensus approach [96]. Miniasm requires all-versus-all read self-mappings as input, which can be obtained by the alignment-free Minimap tool. Finally, LINKS [75] is a genomic tool designed for scaffolding genome assemblies with long reads (including draft genomes). The major advantage of this method is the use of paired *k*-mers from variable long sequence sources without a need of read correction.

Metagenomics, the study of genomic sequences obtained directly from the environment (e.g., aquatic ecosystems, human body), has become a primary application of alignment-free methods, in particular programs designated for fast and precise profiling of microbial communities. For example, Kraken [83] and CLARK [84] are top-performing tools designed for this task—they assign taxonomic labels to individual reads in large datasets with near perfect accuracy (precision > 99%), even in the presence of unknown organisms. These programs perform metagenomic classification of next-generation sequencing reads based on the analysis of shared *k*-mers between an input read and each genome from a precomputed database. Kraken additionally assigns each *k*-mer to the lowest common ancestor of all organisms whose genomes contain corresponding *k*-mers (Additional file 1: Figure S2). The evaluation of the accuracy and speed of 14 widely used metagenome analysis tools [97] showed that Kraken and CLARK are top state-of-the-art tools with the highest speed, accuracy, and sensitivity (i.e., the fraction of reads that is correctly classified).

The alignment-free techniques are continuously being applied to new next-generation sequencing based solutions, for example, phylogenomics (reviewed in [57]), where advances have facilitated construction of high-quality phylogenies directly from raw, unassembled genome sequence data, bypassing both genome assembly and alignment. Assembly and alignment-free phylogenetic tools are already available on the market (AAF [78], NGS-MC [79], and kSNP [80, 81]) and although algorithmically different (e.g., based on single-nucleotide polymorphism calls or various dissimilarity measures), all of them are capable of phylogeny reconstruction of non-model species even in cases of low sequence coverage or lack of a reference genome. In addition, the AAF program provides bootstrap support to assess the confidence of tree topology and addresses problems of homoplasy, sequencing error, and incomplete coverage.

Where else can alignment-free sequence comparison methods be applied?

Progress over the past two decades has led alignment-free research from bioinformatics “curiosities” to a broadening range of successful applications that accompany mainstream biology [37].

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 [98]. Another rising trend for the use of word-based alignment-free methods is the detection of functional and/or evolutionary similarities among regulatory sequences (e.g., promoters, enhancers, and silencers) to estimate their in vivo activities in different organisms (flies and mammals, including humans) [99–103].

Sequence rearrangements are particularly well handled by alignment-free sequence analyses. Recent studies described the mosaic structure of viral and bacterial genomes (e.g., by characterizing the recombination break points in HIV-1 strain and *Escherichia coli* genomes). This analysis provides new evidence for the long-held suspicion that animal *E. coli* pathogens can also infect humans [104]. Another study [105] discovered a clear signal for a pair of *E. coli* genomes that had undergone an engineered 125-kb horizontal gene transfer 20 years ago. Alignment-free measures were also applied to detect domain shuffling signatures in proteins [106] and to identify the members of complex multidomain proteins, such as kinases [107].

Horizontal gene transfer strongly complicates the task of reconstructing the evolutionary history of genes and species, and alignment-free methods have also proved to be helpful in this field. For example, in a comprehensive study of bacterial genomes, the authors used oligonucleotides as genomic signatures and showed that horizontal gene transfers accounted for 6% of the genomes on average [108]. Furthermore, the statistical relationships between genomic signatures among several thousand species provided information about possible donor taxa for the identified foreign sequences. In other studies [109, 110], alignment-free approaches were applied to the genomes of the human pathogen *Staphylococcus aureus* and recovered regions of lateral origin that corresponded to genes involved in transport, antibiotic resistance, pathogenicity, and virulence.

Whole-genome phylogeny [111] is another area where alignment-free methods play an increasing role. Many studies [34, 112–118] addressed the phylogenetic reconstruction of prokaryotes, such as the whole-genome phylogeny of *E. coli* O104:H4, which was the strain that caused the 2011 outbreak in Germany. The analysis revealed a direct line of ancestry leading from a putative typical enteroaggregative *E. coli* ancestor through the 2001 strain to the 2011 outbreak strain [113]. The alignment-free based phylogeny of almost a hundred Zika virus strains suggested that this mosquito-borne flavivirus originated from Africa and then spread to Asia, the Pacific islands, and throughout the Americas [119]. Alignment-free methods have recently been applied to infer phylogenetic relationships among eukaryotic species (fungi [120], plants [121], and mammals [35]); the resulting trees were extremely similar to the species trees created by the manually curated NCBI taxonomic database, which reflects the current taxonomic consensus in the literature.

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 [122], clustering expressed sequence tag sequences with full-length cDNA data [123], and aggregating gene and protein sequences into functional families [124–126]. 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 [127, 128] based on the statistical transformation of protein sequences into uniform vectors with various amino acid properties showed an impressive prediction accuracy of up to 89% in discriminating positive and negative sets of bacterial, viral, and tumor antigen datasets. 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 barcodes [129–133].

The available alignment-free-based software for general sequence comparison are listed in Table 2. For convenience, we categorized the listed programs into basic research tasks, such as small scale pairwise/multiple sequence comparisons, whole genome phylogeny (from viral to mammalian scale), BLAST-like sequence similarity search, identification of horizontally transferred genes and recombination events, as well as annotation of long non-coding RNAs and regulatory elements.

How to use alignment-free methods for research purposes

Among programs listed in Table 2, CAFE is an example of a general purpose alignment-free software that allows exploration of relationships among multiple DNA sequences through a graphical user interface. The tool integrates 28 dissimilarity measures based on *k*-mer analysis, including ten conventional (e.g., Euclidean, Manhattan, d_2), 15 based on presence/absence of *k*-mers (e.g., Jaccard and Hamming distances) and three state-of-the-art measures based on background adjusted *k*-mer counts (i.e., CVTree, d_2^* and d_2^S). The resulting pairwise dissimilarities among the sequences form a distance matrix, which can be directly saved in a standard PHYLIP format. In addition, CAFE presents pairwise dissimilarity measures in a form of different visualizations, including dendrogram (i.e., tree illustrating the clustering of the sequences), heatmap, principal coordinate analysis, and network display.

Most of the listed tools, including CAFE, are standalone programs (only a few were implemented as web services) and therefore may require some specific installation procedures. In this summary article, we have launched a novel, publicly accessible web application for alignment-free sequence comparisons/phylogeny, in a way that anyone can give it a try without any programming deployment effort (no expertise required). The web application (http://www.combio.pl/alfree) uses 38 popular alignment-free methods to calculate distances among given nucleotide or protein sequences. By default, running an analysis is a “one-step process”—after providing the input sequences the server will execute the alignment-free analysis in a fully automated mode without the need for further user intervention. The results are reported as a consensus phylogenetic tree that summarizes the agreement between various individual methods’ trees, thus allowing users to assess the reliability of given phylogenetic relationships across different methods (Fig. 3). Users can also browse trees obtained by individual methods as well as inspect distance measures for any pair of query sequences by using interactive heat maps and tables.

How well do alignment-free methods work?

The performance of alignment-free methods has improved greatly since the introduction of the first alignment-free measure exactly 30 years ago [134]. The challenge today, however, is not a lack of alignment-free algorithms (there are almost 100 published methods), but the number of benchmarking approaches to alignment-free sequence comparison—once a new method is published, a new evaluation procedure and/or selected dataset is also introduced. For example, the majority of algorithms have been evaluated using various sets of simulated DNA sequences [54, 135, 136], primate/mammalian mitochondrial genomes [40, 61, 137, 138], whole prokaryotic genomes/proteomes [117, 139], selected plant genomes [121, 140], small subsets of homologous genes [141, 142], and different combinations thereof [36, 139].

Giving the heterogeneity of testing procedures, it has been quite an achievement by four independent studies to evaluate several classic distance measures for their application under different scenarios of sequence evolution.

Table 2 Alignment-free sequence comparison tools available for research purposes

Category	Name	Features	Implementation	Reference	URL
Pairwise and multiple sequence comparison	ALF	Calculation of pairwise similarity scores (using N2 measure) for sequences in fasta file	Software (C++)	[101]	https://github.com/seqan/seqan/tree/master/apps/alf
	Alfree	25 word-based measures, 8 IT-based measures, 3 graph-based measures, W-metric	Web service Software (Python)	This article	http://www.combio.pl/alfree
	decaf + py	13 word-based measures, Lempel–Ziv complexity-based measure, average common substring distance, W-metric	Software (Python)	[52, 53]	http://bioinformatics.org.au/tools/decaf+py/
	multiAlignFree	Multiple alignment-free sequence comparison using five word-based statistics	R package	[167]	http://www-rcf.usc.edu/~fsun/Programs/multiAlignFree/
	NASC	Non-aligned sequence comparison: four word-based measures and 2 IT-based measures	Matlab framework	[38]	http://web.ist.utl.pt/susanavinga/NASC/
Whole-genome phylogeny	ALFRED ALFRED-G	Phylogenetic tree reconstruction based on the average common substring approach	Software (C++)	[168, 169]	http://alurulab.cc.gatech.edu/phylo
	andi	Computation of evolutionary distances between closely related genomes by approximation of local alignments (*k*-mer based d_a measure); scalable to thousands of bacterial genomes	Software (C)	[170]	https://github.com/evolbioinf/andi/
	CAFE	Alignment-free analysis platform for studying the relationships among genomes and metagenomes (offers 28 word-based dissimilarity measures)	Software (C)	[171]	https://github.com/younglululu/CAFE
	CVTree3	Phylogeny reconstruction from whole genome sequences based on word composition	Web service	[172, 173]	http://tlife.fudan.edu.cn/cvtree3
	DLTree	Automated whole genome/proteome-based phylogenetic analysis based on alignment-free dynamical language method	Web Service	[174]	http://dltree.xtu.edu.cn
	FFP	Feature frequency profile-based measures for whole genome/proteome comparisons (from viral to mammalian scale)	Software (C/Perl)	[34, 55, 112]	https://sourceforge.net/projects/ffp-phylogeny/
	jD2Stat (JIWA)	Generation of the distance matrix using D_2 statistics to extract *k*-mers from large-scale unaligned genome sequences	Software (Java)	[54]	http://bioinformatics.org.au/tools/jD2Stat/
	kr	Efficient word-based estimation of mutation distances from unaligned genomes	Software (C)	[175]	http://guanine.evolbio.mpg.de/cgi-bin/kr2/kr.cgi.pl
	FSWM/kmacs/Spaced	Three tools for alignment-free sequence comparison based on inexact word matches	Software (C++) Web service	[36, 176]	Software currently unavailable Software currently unavailable Software currently unavailable
	SlopeTree	Whole genome phylogeny that corrects for HGT	Software (C++)		http://prodata.swmed.edu/download/pub/slopetree_v1/
	Underlying Approach	Phylogeny of whole genomes using composition of subwords	Software (Java)	[139]	http://www.dei.unipd.it/~ciompin/main/underlying.html
Sequence similarity search tool	RAFTS3	Searches of similar protein sequences against a protein database (>300 times faster than BLAST)	Matlab	[177]	https://sourceforge.net/projects/rafts3/
Annotation of long non-coding RNA	FEELnc	Prediction of lncRNAs from RNA-seq samples based word frequencies and relaxed open reading frames	Software (Perl/R)	[178]	https://github.com/tderrien/FEELnc
	lncScore	Identification of long non-coding RNA from assembled novel transcripts	Software (Python)	[152]	https://github.com/WGLab/lncScore
Horizontal gene transfer	alfy	Alignment-free local homology calculation for detecting horizontal gene transfer	Software (C)	[104, 109]	http://guanine.evolbio.mpg.de/alfy/
	rush	Detection of recombination between two unaligned DNA sequences	Software (C)	[105]	http://guanine.evolbio.mpg.de/rush/
	Smash	Identification and visualization of DNA rearrangements between pairs of sequences	Software (C)	[179]	http://bioinformatics.ua.pt/software/smash/

Table 2 Alignment-free sequence comparison tools available for research purposes *(Continued)*

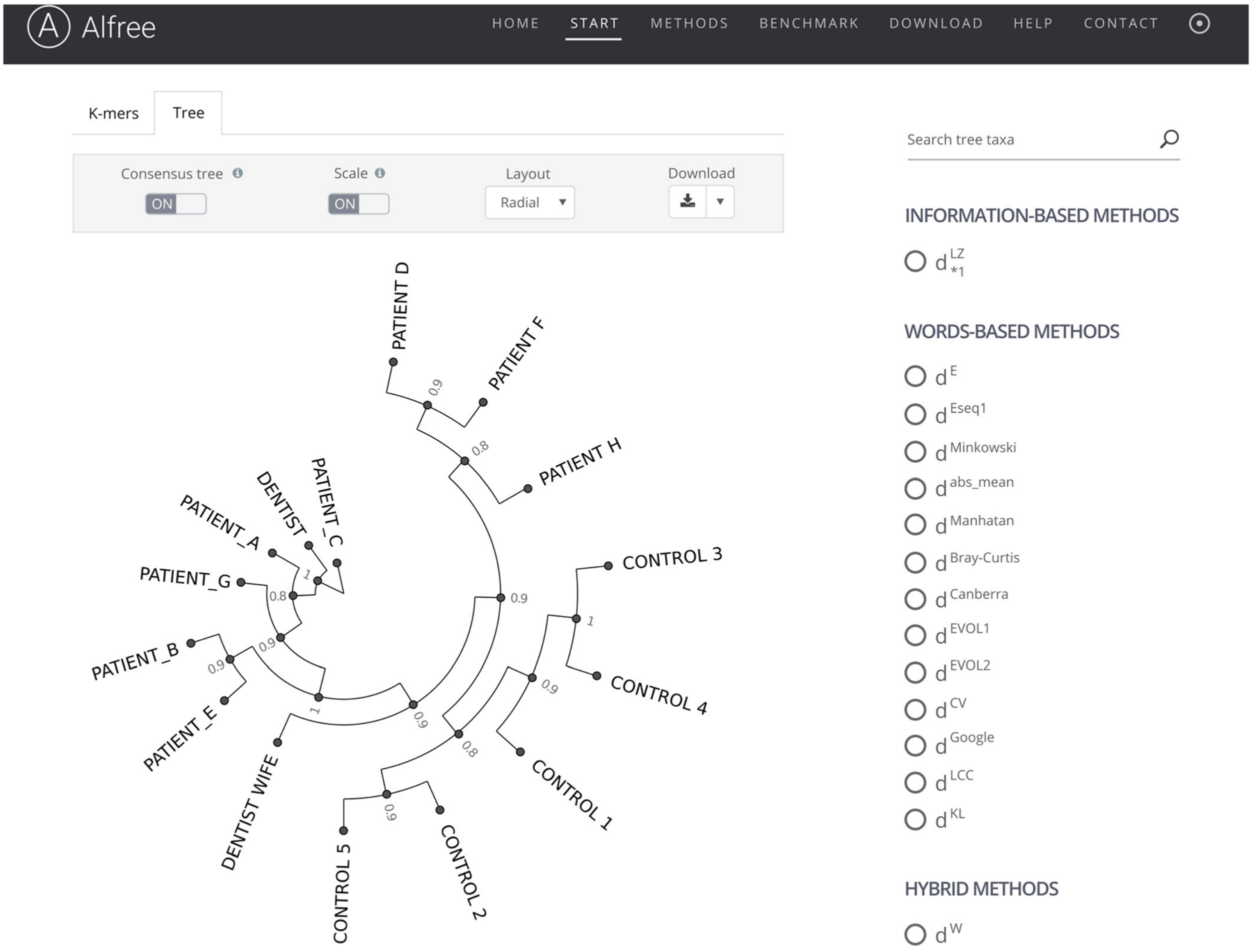
	TF-IDF	Detection of HGT regions and the transfer direction in nucleotide/protein sequences	Software (C++)	[110, 180]	https://github.com/congyingnan/TF-IDF
Regulatory elements	D2Z	Identification of functionally related homologous regulatory elements	Software (Perl)	[102]	http://veda.cs.uiuc.edu/d2z/
	MatrixREDUCE	Prediction of functional regulatory targets of TFs by predicting the total affinity of each promoter and orthologous promoters	Software (Python)	[181]	https://systemsbiology.columbia.edu/matrixreduce
	RRS	Detection of functionally similar group of enhancers and their regions	Software (Perl/C)	[182]	http://goo.gl/7gW578
Sequence clustering	d2_cluster	Word-based clustering EST and full-length cDNA sequences	Software (C)	[123]	https://github.com/shaze/wcdest/
	d2-vlmc	Word-based clustering of metatranscriptomic samples using variable length Markov chains	Software (Python)	[183]	https://d2vlmc.codeplex.com/
	mBKM	Clustering of DNA sequences using Shannon entropy and Euclidean distance	Software (Java)	[124]	https://github.com/Huiyang520/DMk-BKmeans
	kClust	Large-scale clustering of protein sequences (down to 20–30% sequence identity)	Software (C++)	[125]	https://github.com/soedinglab/kClust
Other	COMET	Rapid classification of HIV-1 nucleotide sequences into subtypes based on prediction by partial matching compression	Web service	[184]	https://comet.lih.lu/
	PPI	Identification of protein–protein interaction by coevolution analysis using discrete Fourier transform	Software (Python)	[185]	https://github.com/cyinbox/PPI
	VaxiJen	Antigen prediction based on uniform vectors of principal amino acid properties	Web service	[127]	http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html

The up-to-date list of currently available programs can be found at http://www.combio.pl/alfree/tools/. Accessed 23 August 2017
HGT horizontal gene transfer, *IT* information theory

The first benchmark, by Vinga and Almeida (2004) [143], compared the accuracy of six word-based methods in recognition of structurally and evolutionary relationships among proteins. Höhl and Ragan [52, 53] tested the accuracy of nine alignment-free methods in the construction of phylogenetic trees using homologous proteins representing a wide range of phylogenetic distances. Both research groups showed that, in general, tested alignment-free methods can be as good as alignment algorithms and, as reported in [52], may perform even better in case of protein sequences that underwent domain shuffling events. Dai and colleagues (2008) [99] confronted nine alignment-free distance measures and two alignment-based approaches (Needleman–Wunsch and Smith–Waterman alignment methods) in annotation of functionally related regulatory sequences in human and fly. Almost all tested alignment-free methods detected statistically relevant similarities in sequence compositions in contrast to alignment-based methods that showed only limited correspondence recognizable by alignments. In a recent benchmark, Bernard and colleagues (2016 [33]) used simulated and empirical microbial genomes to test the sensitivity of nine alignment-free methods under different evolutionary schemes. All approaches generated biologically meaningful phylogenies—alignment-free methods were most sensitive to the extent of sequence divergence, less sensitive to low and moderate frequencies of horizontal gene transfer, and most robust against genome rearrangements.

We extended the benchmark of Vinga and Almeida (2004) to test 33 popular alignment-free methods (as well as the Smith–Waterman algorithm—the most accurate algorithm for sequence alignments) in the classification of structural and evolutionary relationships between protein sequences from the SCOPe/ASTRAL database [144]. This resource provides a high-quality structural classification of proteins at four levels: class, folds, superfamilies, and families (for details see Additional file 2: Table S1). As in the previous study [143], we used a representative subset of the SCOPe database (containing proteins sharing less than 40% identity) as a reference to test 25 word-based and eight information theory-based alignment-free methods along with different combinations of their input parameters, such as word size (from 1 to 4) and vector type (e.g., counts, frequencies, etc.). The performance of each method was assessed using AUC statistics (area under the receiver operating curve; for details about methods see Additional file 3: Supplementary methods).

The alignment-based algorithm (Smith–Waterman algorithm) was outperformed at all SCOP levels—i.e., class, class fold, superfamily, family (AUCs 0.62, 0.67, 0.78, 0.81)—by two word-based measures: normalized Google distance [59] (AUCs 0.63, 0.78, 0.80, 0.84) and



Consensus tree: ON

Scale: ON

Layout: Radial

Download

INFORMATION-BASED METHODS

○ d^{LZ}_{*1}

WORDS-BASED METHODS

○ d^{E}

○ d^{Eseq1}

○ $d^{Minkowski}$

○ d^{abs_mean}

○ $d^{Manhatan}$

○ $d^{Bray-Curtis}$

○ $d^{Canberra}$

○ d^{EVOL1}

○ d^{EVOL2}

○ d^{CV}

○ d^{Google}

○ d^{LCC}

○ d^{KL}

HYBRID METHODS

○ d^{W}

Fig. 3 Snapshot of the results returned by the alignment-free web tool (Alfree) for "example 1": HIV viral sequences obtained from dental patients in Florida [186]. Briefly, in the late 1980s some patients of an HIV-positive dentist in Florida were diagnosed as infected with HIV. An investigation by the Centers for Disease Control and Prevention did not uncover any hygiene lapses that could result in infection of patients. However, sequence comparison of the gene encoding gpg120 isolated from HIV strains from the dentist, his patients, and other individuals revealed that *PATIENT_A*, *PATIENT_B*, *PATIENT_C*, *PATIENT_E*, and *PATIENT_G* became infected while receiving dental care [183]. The phylogeny shown is based on the gp120 viral protein sequences from the dentist, the dentist's wife (*DENTIST WIFE*), eight patients (*PATIENT_A* to *PATIENT H*), and five individuals that never had contact with the accused (*CONTROL 1*, *2*, *3*, *4*, and *5*). The sphylogram was obtained as a majority-rule consensus tree that summarizes the agreement across 15 alignment-free methods (support values in scale from 0 to 1 are shown for every node of the tree). The web interface of the Alfree portal also provides an example case of phylogenetic reconstruction of mitochondrial genomes of 12 primates. Several additional options are available to explore and visualize the sequence comparison results, including selection of individual method, re-rooting trees, changing tree layouts, as well as collapsing or expanding different parts of the tree

Bray–Curtis distance [145] (AUCs 0.63, 0.77, 0.80, 0.84) (Additional file 2: Table S1). Three other word-based methods, including two variants of Squared Euclidean distance [53] as well as the Canberra distance [146], though less accurate in recognition of relationships within class, obtained higher overall scores (AUCs 0.744, 0.733, and 0.725, respectively) than the Smith–Waterman algorithm (AUC 0.72). These results support the assumption—very often taken for granted—that alignment-free methods can produce more accurate results than alignment-based solutions when applied to homologous sequences of low similarity. Interestingly, the Smith–Waterman algorithm was outperformed only by word-based methods with short *k*-mers of one to two residues, indicating that the conservation and order of longer sequence stretches are generally not preserved in the sequences, and the relationship between alignment similarity score and structural/evolutionary relationship breaks down. As alignment-free methods do not depend on where the words are found in the sequence, they are typically not confused by the complexities caused by mismatches, gaps, and sequence inversions that are often found in this type of distantly related homolog. It is also interesting to note that word-based methods achieved higher accuracy (AUC 0.67 ± 0.04) than information-theory based solutions (AUC 0.61 ± 0.06)

(Additional file 2: Table S1). Although explanation of this fact is not straightforward, it may indicate that compression procedures included in currently selected methods do not decipher the complexity of highly variable protein sequence, which would explain the broader application of the information-theory based methods in DNA sequence analyses. The full results of the benchmark can be interactively explored [145].

Remarkably, the duration time for the calculation of approximately 22 million pairwise protein comparisons by the Smith–Waterman algorithm took exactly 3 days, which was more than 1000-fold slower than the alignment-free methods (Additional file 3: Supplementary methods). On average, these methods need 4 minutes to complete the task, and the fastest approach (Hamming distance [146]) ran the analysis in only 19 seconds.

The implementations of all alignment-free methods used in this study are provided as a stand-alone Python application [147]. We also supplement this article with the benchmark dataset [148] for reference analysis that can be readily reproduced by enthusiasts or developers building new alignment-free solutions.

Conclusions

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. Will these new techniques spell doom for traditional alignments? Most likely not in the authors' lifetime. Alignment is still irreplaceable in many aspects of today's biology, such as the annotation of conserved protein domains and motifs, tracking phenotype-related sequence polymorphisms, reconstruction of ancestral DNA sequences, determining the rate of sequence evolution, and homology-based modeling of three-dimensional protein structures. In addition, the research on, and the development of, alignment-free methods is still relatively young, holding considerable potential for improvement, whereas alignment approaches are already mature and only a few alignment-free methods have really challenged the validity and reliability of alignment-based techniques.

Most published articles about alignment-free sequence comparison methods are still mainly technical, exploring their mathematical foundations and theoretical performance (versus alignment-based approaches), very often evaluated with individually selected, mostly simulated, data sets. Although many alignment-free programs exist (as shown in Tables 1 and 2), the majority of published alignment-free methods are still not supplemented with software implementations and thus cannot easily be compared on common sets of data. The absence of well-defined benchmarks covering various evolutionary scenarios of sequence divergence creates a major obstacle for researchers who simply need to know the current "best" tool. Consequently, it is still difficult to state which alignment-free method might be particularly suited for a certain task. The stage thus appears to be now set for application of alignment-free methods on real world data sets, which seems to be the only way for these methods to be widely accepted by scientists in biology and related fields [149].

Although alignment-free methods are computationally relatively easily scalable to multigenome data, they do have some "skeletons in their closet". For example, using long *k*-mers in word-based methods may impose a substantial memory overhead (the total number of possible DNA words of length 14 is 4^{14}, which is about 4 GB). Although information-theory methods that are based on the compression algorithms are more memory efficient and computationally inexpensive, they may fail to decipher complex organization levels in the sequences [39] (also shown in results obtained in this study; Additional file 2: Table S1). 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 [54, 150].

Nevertheless, alignment-free algorithms are rapidly extending the range of their applications [151–154] and answering previously intractable questions in phylogenomics and horizontal gene transfer (reviewed in [57]), population genetics (reviewed in [111]), evolution of regulatory sequences, and links between the genome and epigenome (reviewed in [155]). Disadvantages of next-generation sequencing data processing and analysis seem to be particularly well addressed by the alignment-free methods (reviewed in [156]). The currently dominant *k*-mer approaches are bound to novel measures for biological applications (e.g., Google distance [59]) and application of advanced information theory-based methods should improve the available alignment-free and alignment-based tool box. From this fair competition between alignment-based and alignment-free camps, scientists can get only the best. In this respect, the next years should be very exciting.

Additional files

Additional file 1: Figures S1. and **Figure S2.** Kraken algorithm for taxonomic labeling of metagenomic DNA sequences (based on Wood and Salzberg, 2014) [83]. (DOCX 864 kb)

Additional file 2: Table S1. Ranking list of alignment-free methods and the Smith-Waterman algorithm based on the area under the curve measures across four structural levels of the SCOP2 database. (DOCX 39 kb)

Additional file 3: Supplementary methods. (DOCX 37 kb)

Authors' contributions

AZ and WMK conceived and designed the study. AZ implemented the web application and performed analyses. SV and JA designed the benchmark framework. AZ, SV, JA, and WMK wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the KNOW RNA Research Centre in Poznan (No. 01/KNOW2/2014) and partially by FCT through projects UID/EMS/50022/2013 and IF/00653/2012.

Availability of data and materials

Alfree web application is available at http://www.combio.pl/alfree.
The protein sequences used for the benchmark analysis in this manuscript were obtained from the SCOPe/STRAL database.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computational Biology, Faculty of Biology, Adam Mickiewicz University in Poznan, Umultowska 89, 61-614 Poznan, Poland. [2]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal. [3]Stony Brook University (SUNY), 101 Nicolls Road, Stony Brook, NY 11794, USA.

Published online: 03 October 2017

References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
2. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8.
3. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
4. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
5. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
6. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
7. Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5:e11147.
8. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse alignments with BLASTZ. Genome Res. 2003;13:103–7.
9. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004;14:708–15.
10. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9:267–76.
11. Song N, Joseph JM, Davis GB, Durand D. Sequence similarity network reveals common ancestry of multidomain proteins. PLoS Comput Biol. 2008; 4, e1000063.
12. Terrapon N, Weiner J, Grath S, Moore AD, Bornberg-Bauer E. Rapid similarity search of proteins using alignments of domain arrangements. Bioinformatics. 2014;30:274–81.
13. Xiong J. Essential bioinformatics. 1st ed. Cambridge: Cambridge University Press; 2006.
14. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12:85–94.
15. Chattopadhyay AK, Nasiev D, Flower DR. A statistical physics perspective on alignment-independent protein sequence comparison. Bioinformatics. 2015; 31:2469–74.
16. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22:1035–6.
17. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res. 2005;33:2433–9.
18. Capriotti E, Marti-Renom MA. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. BMC Bioinformatics. 2010;11:322.
19. Lange K. Mathematical and statistical methods for genetic analysis. 2nd ed. New York, NY: Springer New York; 2002.
20. Eddy SR. What is dynamic programming? Nat Biotechnol. 2004;22:909–10.
21. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
22. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27:334–42.
23. Pevzner P, Tesler G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res. 2003;13:37–45.
24. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003;100:11484–9.
25. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, et al. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res. 2000;10:577–86.
26. Earl D, Nguyen N, Hickey G, Harris RS, Fitzgerald S, Beal K, et al. Alignathon: a competitive assessment of whole-genome alignment methods. Genome Res. 2014;24:2077–89.
27. Prakash A, Tompa M. Measuring the accuracy of genome-size multiple alignments. Genome Biol. 2007;8:R124.
28. Chatzou M, Magis C, Chang J-M, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Brief Bioinform. 2015;17:1–15.
29. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, et al. Phylo: a citizen science approach for improving multiple sequence alignment. PLoS One. 2012;7:e31362.
30. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008;319:473–6.
31. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. Nat Biotechnol. 2008;26:274–5.
32. 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.
33. Bernard G, Chan CX, Ragan MA. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Sci Rep. 2016;6:28970.
34. Jun S-R, Sims GE, Wu GA, Kim S-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci U S A. 2010;107:133–8.
35. Sims GE, Jun S-R, Wu GA, Kim S-H. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. Proc Natl Acad Sci U S A. 2009;106:17077–82.
36. Leimeister C-A, Sohrabi-jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. Bioinformatics. 2017;33:1–9.
37. Vinga S. Editorial: Alignment-free methods in computational biology. Brief Bioinform. 2014;15:341–2.
38. Vinga S, Almeida J. Alignment-free sequence comparison—a review. Bioinformatics. 2003;19:513–23.
39. Vinga S. Information theory applications for biological sequence analysis. Brief Bioinform. 2014;15:376–89.
40. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. J Comput Biol. 2006;13:336–50.
41. Haubold B, Pierstorff N, Möller F, Wiehe T. Genome comparison without alignment using shortest unique substrings. BMC Bioinformatics. 2005;6:123.
42. Pinho AJ, Ferreira PJSG, Garcia SP, Rodrigues JMOS. On finding minimal absent words. BMC Bioinformatics. 2009;10:137.
43. Yang L, Zhang X, Wang T, Zhu H. Large local analysis of the unaligned genome and its application. J Comput Biol. 2013;20:19–29.
44. Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990;18:2163–70.
45. Almeida JS. Sequence analysis by iterated maps, a review. Brief Bioinform. 2014;15:369–75.
46. Wen J, Zhang Y. A 2D graphical representation of protein sequence and its numerical characterization. Chem Phys Lett. 2009;476:281–6.

47. Randić M, Zupan J, Balaban AT. Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem Phys Lett. 2004;397:247–52.
48. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington; 2005.
49. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4.
50. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol. 1999;16:1391–9.
51. Vinga S. Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification. In: Pham TD, Yan H, Crane DI, editors. Advanced computational methods for biocomputing and bioimaging. New York : Nova Science; 2007. p. 70–105.
52. Höhl M, Rigoutsos I, Ragan MA. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online. 2006;2:359–75.
53. Höhl M, Ragan MA. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol. 2007;56:206–21.
54. Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep. 2014;4:6504.
55. Sims GE, Jun S, Wu GA, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A. 2009;106:2677–82.
56. Wang Y, Liu L, Chen L, Chen T, Sun F. Comparison of metatranscriptomic samples based on k-tuple frequencies. PLoS One. 2014;9, e84348.
57. Bernard G, Chan CX, Chan Y, Chua X-Y, Cong Y, Hogan JM, et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Brief Bioinform. 2017;286:1443a.
58. 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. Bioinformatics. 2005;21:4125–32.
59. Lee JC, Rashid NA. Adapting normalized google similarity in protein sequence comparison. International Symposium on Information Technolnology. September 2008. p. 1–5.
60. Li M, Vitányi P. An introduction to Kolmogorov complexity and its applications. New York, NY: Springer New York; 2008.
61. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. Bioinformatics. 2003;19:2122–30.
62. Li M, Chen X, Li X, Ma B, Vitanyi PMB. The similarity metric. IEEE Trans Inf Theory. 2004;50:3250–64.
63. Giancarlo R, Rombo SE, Utro F. Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. Brief Bioinform. 2014;15:390–406.
64. Tribus M, McIrvine EC. Energy and information. Sci Am. 1971;225:179–88.
65. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22:79–86.
66. Microbiology by numbers. Nat Rev Microbiol. 2011;9:628.
67. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462–4.
68. Zhang Z, Wang W. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. Bioinformatics. 2014;30:i283–92.
69. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
70. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
71. Shajii A, Yorukoglu D, William Yu Y, Berger B. Fast genotyping of known SNPs through approximate k-mer matching. Bioinformatics. 2016;32:i538–44.
72. Rudewicz J, Soueidan H, Uricaru R, Bonnefoi H, Iggo R, Bergh J, et al. MICADo – looking for mutations in targeted PacBio cancer data: an alignment-free method. Front Genet. 2016;7:214.
73. Pajuste F-D, Kaplinski L, Möls M, Puurand T, Lepamets M, Remm M. FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. Sci Rep. 2017;7:2537.
74. Li Y, Heavican TB, Vellichirammal NN, Iqbal J, Guda C. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. Nucleic Acids Res. 2017;45:1–18.
75. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, et al. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience. 2015;4:35.
76. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623–30.
77. Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103–10.
78. Fan H, Ives AR, Surget-Groba Y, Cannon CH. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 2015;16:522.
79. Ren J, Song K, Deng M, Reinert G, Cannon CH, Sun F. Inference of Markovian properties of molecular sequences from NGS data and applications to comparative genomics. Bioinformatics. 2016;32:993–1000.
80. Gardner SN, Hall BG. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One. 2013;8:e81760.
81. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015;31:2877–8.
82. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013;29:2253–60.
83. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
84. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32:3823–5.
85. Roosaare M, Vaher M, Kaplinski L, Möls M, Andreson R, Lepamets M, et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017;5:e3353.
86. Gupta A, Jordan IK, Rishishwar L. stringMLST: a fast k-mer based tool for multilocus sequence typing. Bioinformatics. 2017;33:119–21.
87. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep. 2017;7:1559.
88. Jin H, Wan Y-W, Liu Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. BMC Bioinformatics. 2017;18:117.
89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
90. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
91. Almeida JS, Grüneberg A, Maass W, Vinga S. Fractal MapReduce decomposition of sequence alignment. Algorithms Mol Biol. 2012;7:12.
92. Wilkinson SR, Almeida JS. QMachine: commodity supercomputing in web browsers. BMC Bioinformatics. 2014;15:176.
93. Marçais G, Yorke JA, Zimin A. QuorUM: An error corrector for Illumina reads. PLoS One. 2015;10:1–13.
94. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol. 2014;15:509.
95. Lim EC, Müller J, Hagmann J, Henz SR, Kim ST, Weigel D. Trowel: A fast and accurate error correction module for Illumina sequencing reads. Bioinformatics. 2014;30:3264–5.
96. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11:25–37.
97. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:19233.
98. 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.
99. Dai Q, Yang Y, Wang T. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics. 2008;24:2296–302.
100. van Helden J. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics. 2004;20:399–406.
101. Göke J, Schulz MH, Lasserre J, Vingron M. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. Bioinformatics. 2012;28:656–63.
102. Kantorovitz MR, Robinson GE, Sinha S. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics. 2007;23:i249–55.
103. Ivan A, Halfon MS, Sinha S. Computational discovery of *cis*-regulatory modules in *Drosophila* without prior knowledge of motifs. Genome Biol. 2008;9:R22.

104. Domazet-Lošo M, Haubold B. Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics. 2011;27:1466–72.
105. Haubold B, Krause L, Horn T, Pfaffelhuber P. An alignment-free test for recombination. Bioinformatics. 2013;29:3121–7.
106. Maetschke SR, Kassahn KS, Dunn JA, Han S-P, Curley EZ, Stacey KJ, et al. A visual framework for sequence analysis using n-grams and spectral rearrangement. Bioinformatics. 2010;26:737–44.
107. Martin J, Anamika K, Srinivasan N. Classification of protein kinases on the basis of both kinase and non-kinase regions. PLoS One. 2010;5, e12460.
108. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 2005;33:e6.
109. Domazet-Lošo M, Haubold B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. Mob Genet Elements. 2011;1:230–5.
110. Cong Y, Chan Y-B, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. Sci Rep. 2016;6:30308.
111. Haubold B. Alignment-free phylogenetics and population genetics. Brief Bioinform. 2014;15:407–18.
112. Sims GE, Kim S-H. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci U S A. 2011;108:8329–34.
113. Cheung M, Li L, Nong W, Kwan H. 2011 German *Escherichia coli* O104: H4 outbreak: whole-genome phylogeny without alignment. BMC Res Notes. 2011;4:533.
114. Li Q, Xu Z, Hao Bailin B. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. J Biotechnol. 2010;149:115–9.
115. Joseph J, Sasikumar R. Chaos game representation for comparison of whole genomes. BMC Bioinformatics. 2006;7:243.
116. Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. 2009;37:174–8.
117. Qi J, Wang B, Hao BI. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol. 2004;58:1–11.
118. Bromberg R, Grishin NV, Otwinowski Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. PLoS Comput Biol. 2016;12:e1004985.
119. Li Y, He L, He RL, Yau SS-T. Zika and flaviviruses phylogeny based on the alignment-free natural vector method. DNA Cell Biol. 2017;36:109–16.
120. Wang H, Xu Z, Gao L, Hao B. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol. 2009;9:195.
121. Hatje K, Kollmar M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Front Plant Sci. 2012;3:192.
122. Ng K-H, Ho C-K, Phon-Amnuaisuk S. A hybrid distance measure for clustering expressed sequence tags originating from the same gene family. PLoS One. 2012;7:e47216.
123. Burke J. d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res. 1999;9:1135–42.
124. Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics. 2012;13:174.
125. Hauser M, Mayer CE, Söding J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics. 2013;14:248.
126. Albayrak A, Otu HH, Sezerman UO. Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets. BMC Bioinformatics. 2010;11:428.
127. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics. 2007;8:4.
128. Doytchinova IA, Flower DR. Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. Vaccine. 2007;25:856–66.
129. Kuksa P, Pavlovic V. Efficient alignment-free DNA barcode analytics. BMC Bioinformatics. 2009;10:S9.
130. Little DP. DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. PLoS One. 2011;6:e20552.
131. Göker M, Grimm GW, Auch AF, Aurahs R, Kučera M. A clustering optimization strategy for molecular taxonomy applied to planktonic foraminifera SSU rDNA. Evol Bioinform Online. 2010;6:97–112.
132. Liu C, Liang D, Gao T, Pang X, Song J, Yao H, et al. PTIGS-IdIt, a system for species identification by DNA sequences of the psbA-trnH intergenic spacer region. BMC Bioinformatics. 2011;12:S4.
133. La Rosa M, Fiannaca A, Rizzo R, Urso A. Alignment-free analysis of barcode sequences by means of compression-based methods. BMC Bioinformatics. 2013;14:S4.
134. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 1986;83:5155–9.
135. Reinert G, Chew D, Sun F, Waterman MS. Alignment-free sequence comparison (I): statistics and power. J Comput Biol. 2009;16:1615–34.
136. Liu X, Wan L, Li J, Reinert G, Waterman MS, Sun F. New powerful statistics for alignment-free sequence comparison under a pattern transfer model. J Theor Biol. 2011;284:106–16.
137. Huang G, Zhou H, Li Y, Xu L. Alignment-free comparison of genome sequences by a new numerical characterization. J Theor Biol. 2011;281:107–12.
138. Pizzi C. MissMax: alignment-free sequence comparison with mismatches through filtering and heuristics. Algorithms Mol Biol. 2016;11:6.
139. Comin M, Verzotto D. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms Mol Biol. 2012;7:34.
140. Leimeister C-A, Morgenstern B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. Bioinformatics. 2014;30:2000–8.
141. Wu TJ, Burke JP, Davison DB. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics. 1997;53:1431–9.
142. Hide W, Burke J, Davison DB. Biological evaluation of d2, an algorithm for high-performance sequence comparison. J Comput Biol. 1994;1:199–215.
143. Vinga S, Gouveia-Oliveira R, Almeida JS. Comparative evaluation of word composition distances for the recognition of SCOP relationships. Bioinformatics. 2004;20:206–15.
144. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42:D304–9.
145. Alfree: Benchmark. http://www.combio.pl/alfree/benchmark. Accessed 23 Aug 2017.
146. Jones E, Oliphant T, Peterson P, et al. SciPy: Open source scientific tools for Python. 2001. http://www.scipy.org/. Accessed 23 Aug 2017.
147. alfpy. https://github.com/aziele/alfpy. Accessed 23 Aug 2017.
148. Alfree: Benchmark dataset. http://www.combio.pl/alfree/download/data/. Accessed 23 Aug 2017.
149. Schwende I, Pham TD. Pattern recognition and probabilistic measures in alignment-free sequence analysis. Brief Bioinform. 2014;15:354–68.
150. Bernard G, Ragan MA, Chan CX. Recapitulating phylogenies using k-mers: from trees to networks. F1000Research. 2016;5:2789.
151. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics. 2016;17:754.
152. Zhao J, Song X, Wang K. lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. Sci Rep. 2016;6:34838.
153. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017;45:39–53.
154. Glouzon J-PS, Perreault J-P, Wang S. The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures. Bioinformatics. 2017;33(8):1169–78. doi:10.1093/bioinformatics/btw773.
155. Pinello L, Lo Bosco G, Yuan G-C. Applications of alignment-free methods in epigenomics. Brief Bioinform. 2014;15:419–30.
156. Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. Brief Bioinform. 2014;15:343–53.
157. Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen APM, Wallace DC, et al. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. Bioinformatics. 2015;31:1310–2.
158. Solovyov A, Lipkin W. Centroid based clustering of high throughput sequencing reads based on n-mer counts. BMC Bioinformatics. 2013;14:268.
159. Comin M, Leoni A, Schimd M. Clustering of reads with alignment-free measures and quality values. Algorithms Mol Biol. 2015;10:4.
160. Song K, Ren J, Zhai Z, Liu X, Deng M, Sun F. Alignment-free sequence comparison based on next-generation sequencing reads. J Comput Biol. 2013;20:64–79.

161. Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ. 2014;2, e425.
162. Pham D-T, Gao S, Phan V. An accurate and fast alignment-free method for profiling microbial communities. J Bioinform Comput Biol. 2017;15:1740001.
163. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
164. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol. 2016;17:111.
165. Jiang B, Song K, Ren J, Deng M, Sun F, Zhang X. Comparison of metagenomic samples using sequence signatures. BMC Genomics. 2012;13:730.
166. Ulyantsev VI, Kazakov SV, Dubinkina VB, Tyakht AV, Alexeev DG. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. Bioinformatics. 2016;32:2760–7.
167. Ren J, Song K, Sun F, Deng M, Reinert G. Multiple alignment-free sequence comparison. Bioinformatics. 2013;29:2690–8.
168. Thankachan SV, Chockalingam SP, Liu Y, Apostolico A, Aluru S. ALFRED: A practical method for alignment-free distance computation. J Comput Biol. 2016;23:452–60.
169. Thankachan SV, Chockalingam SP, Liu Y, Krishnan A, Aluru S. A greedy alignment-free distance estimator for phylogenetic inference. BMC Bioinformatics. 2017;18:238.
170. Haubold B, Klötzl F, Pfaffelhuber P. andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics. 2015;31:1169–75.
171. Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F. CAFE: aCcelerated Alignment-FrEe sequence analysis. Nucleic Acids Res. 2017;45:2015–7.
172. Qi J, Luo H, Hao B. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 2004;32:45–7.
173. Zuo G, Hao B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. genomics, proteomics bioinforma. Genomics Proteomics Bioinforma. 2015;13:321–31.
174. Wu Q, Yu Z-G, Yang J. DLTree: efficient and accurate phylogeny reconstruction using the dynamical language method. Bioinformatics. 2017. doi:10.1093/bioinformatics/btx158.
175. Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. Estimating mutation distances from unaligned genomes. J Comput Biol. 2009;16:1487–500.
176. Horwege S, Lindner S, Boden M, Hatje K, Kollmar M, Leimeister C-A, et al. Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches. Nucleic Acids Res. 2014;42:W7–11.
177. Vialle RA, Pedrosa FO, Weiss VA, Guizelini D, Tibaes JH, Marchaukoski JN, et al. RAFTS3: rapid alignment-free tool for sequence similarity search. bioRxiv. 2016;55269.
178. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017;45:e57.
179. Pratas D, Silva RM, Pinho AJ, Ferreira PJSG. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Sci Rep. 2015;5:10203.
180. Cong Y, Chan Y, Phillips CA, Langston MA, Ragan MA. Robust inference of genetic exchange communities from microbial genomes using TF-IDF. Front Microbiol. 2017;8:21.
181. Ward LD, Bussemaker HJ. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics. 2008;24:i165–71.
182. Koohy H, Dyer NP, Reid JE, Koentges G, Ott S. An alignment-free model for comparison of regulatory sequences. Bioinformatics. 2010;26:2391–7.
183. Liao W, Ren J, Wang K, Wang S, Zeng F, Wang Y, et al. Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length markov chains. Sci Rep. 2016;6:37243.
184. Struck D, Lawyer G, Ternes A-M, Schmit J-C, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 2014;42:e144.
185. Yin C, Yau SS-T. A coevolution analysis for identifying protein-protein interactions by Fourier transform. PLoS One. 2017;12, e0174862.
186. Centers for Disease Control (CDC). Update: transmission of HIV infection during invasive dental procedures—Florida. MMWR Morb Mortal Wkly Rep. 1991;40:377–81.