bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ALIGNMENT BY THE NUMBERS: SEQUENCE ASSEMBLY USING REDUCED DIMENSIONALITY NUMERICAL REPRESENTATIONS

Avraam Tapinos, Bede Constantinides, David L Robertson*

Computational and Evolutionary Biology, Faculty of Life Sciences, The University of Manchester, Manchester, M13 9PT, UK.

*Correspondence to: david.robertson@manchester.ac.uk

ABSTRACT

DNA sequencing instruments are enabling genomic analyses of unprecedented scope and scale, widening the gap between our abilities to generate and interpret sequence data. Established methods for computational sequence analysis generally consider the nucleotide-level resolution of sequences, and while these approaches are sufficiently accurate, increasingly ambitious and data-intensive analyses are rendering them impractical for demanding applications such as genome and metagenome assembly. Comparable analytical challenges are encountered in other data-intensive fields involving sequential data such as signal processing and time series analysis. By representing nucleic acid composition numerically it is possible to apply dimensionality reduction methods from these fields to sequences of nucleotides, enabling their approximate representation. To explore the applicability of signal decomposition methods in sequence assembly, we implemented a short read aligner and evaluated its performance against simulated high diversity viral sequences alongside four existing aligners. Using our prototype implementation, approximate sequence representations reduced overall alignment time by up to 14-fold compared to that of uncompressed sequences, and without any reduction in alignment accuracy. Despite using heavily approximated sequence representations, our implementation yielded alignments of similar overall accuracy to existing aligners, outperforming all other tools tested at high levels of sequence variation. Our approach was also applied to the *de novo* assembly of a simulated diverse viral population. We have demonstrated that full sequence resolution is not a prerequisite of accurate sequence alignment and that analytical performance may be retained or even enhanced through appropriate dimensionality reduction of sequences.

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Contemporary sequencing technologies have massively parallelised the determination of nucleotide order within genetic material, making it possible to rapidly sequence the genomes of individuals, populations and interspecies samples (Margulies *et al.* 2005; Bentley *et al.* 2008; Eid *et al.* 2009; Rothberg *et al.* 2011). However, the sequences generated by these instruments are usually considerably shorter in length than the genomic regions they are used to study. Genomic analyses accordingly begin with the process of sequence assembly, wherein sequence fragments (reads) are reconstructed into the larger sequences from which they originated. Computational methods play a vital role in the assembly of short read datasets, and a wide variety of assemblers and related tools have been developed in tandem with emerging sequencing platforms (Schatz *et al.* 2010).

Where the objective of a nucleotide sequencing experiment is to derive a single consensus sequence representing the genome of an individual, various computational methods are applicable. Seed-and-extend alignment methods using suffix array derivatives such as the Burrows-Wheeler Transform have emerged as the preferred approach for assembling short reads informed by a supplied reference sequence (Li and Durbin 2009; Shrestha *et al.* 2014), while graph-based methods employing Overlap Layout Consensus (OLC) (Kececioglu and Myers 1995; Myers 1995) and de Bruijn graphs of *k*-mers (Pevzner *et al.* 2001) have become established for reference-free *de novo* sequence assembly. However, the effectiveness of these approaches varies considerably for characterising genetic variation within populations ('deep' sequencing), or interspecies biodiversity within a metagenomics sample.

Within populations comprised of divergent variants, such as those established by a virus within their host, bias associated with use of a reference sequence can lead to valuable read information being discarded during assembly (Archer *et al.* 2010). While this can be overcome by constructing a data-specific reference sequence after initial reference alignment, this still necessitates use of a reference sequence for the initial alignment, a limiting factor for sequencing unknown species, which, for example, are often abundant in metagenomic samples. On the other hand, while *de novo* approaches require little *a priori* knowledge of target sequence composition, they are very computationally intensive and their performance scales poorly with datasets of increasing size. Indeed, the underlying problem of *de novo* assembly is NP-hard (Myers 1995), with the presence of genetic diversity serving to further increase the complexity of computational solutions. As such, aggressive heuristics are typically employed in order to reduce the running time of *de novo* assemblers, which in turn can compromise assembly quality.

The challenge of diverse population assembly is exacerbated by the scale of sequencing datasets. Contemporary sequencing platforms may generate billions of reads, and their ongoing development is giving rise to increasing read lengths, posing growing challenges for efficient sequence analysis. A key challenge to be overcome when analysing any sufficiently large dataset arises when its size exceeds the capacity of a computer's working memory. If, for example, stored short read data exceeds the capacity of a computer's random access memory (RAM), this data must be exhaustively 'swapped' between RAM and disk storage that is orders

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of magnitude slower to access, forming a major analysis bottleneck (Yang and Wu 2006). *De novo* assembly in particular requires manyfold more memory than is needed to store the read information itself. Additionally, the use of all-to-all pairwise comparison for read alignment (and numerous data mining tasks) scales poorly (quadratic time complexity) with increasing data size. While indexing structures such as suffix arrays are often used reduce the burden of pairwise sequence comparison, their performance generally deteriorates with increasing sequence length in accordance with the phenomenon known as the 'curse of dimensionality' (Verleysen and François 2005).

Comparable analytical challenges involving high dimensional sequential data are encountered in other data-intensive fields such as signal processing and time series analysis, in which a number of effective dimensionality reduction methods have been proposed, including the discrete wavelet transform (DWT) (Chan and Fu 1999), the discrete Fourier transform (DFT) (Agrawal *et al.* 1993) and piecewise aggregate approximation (PAA) (Geurts 2001; Keogh *et al.* 2001). Through selective preservation of data features, the use of such approaches can allow accurate analyses to be performed with reduced computational overhead. Crucially, dimensionality reduction may be performed to approximate data to lower dimensional space without the loss of useful information, since only minor data characteristics are discarded (Ye 2003). Due to the successive nature of base observations within nucleic acids, many of these approximation approaches can also be applied to sequences of nucleotides (Silverman and Linsker 1986) and amino acids (Cheever *et al.* 1989). Since most of these representation approaches are suitable only for numerical sequences, an appropriate numerical sequence transformation must be performed prior to using these methods.

Many methods for transforming symbolic nucleotide sequences into numerical sequences have been proposed (Kwan and Arniker 2009), permitting the application of various digital signal processing techniques to appropriately transformed nucleotide sequences. They include the Voss (1992) method (Voss 1992), the DNA walk (Berger *et al.* 2004), (Lobry 1996), the real number method (Chakravarthy *et al.* 2004), the complex number (Anastassiou 2001), and the tetrahedron method (Silverman and Linsker 1986). Fundamentally, all the nucleotide to numerical sequence transformation approaches convert a nucleotide sequence into a numerical sequence by assigning a unique number or vector to each nucleotide base. Some of the methods, including the complex number and the real number methods, introduce biasing mathematical properties to the transformed sequence, yet are appropriate for use in certain applications such as the detection of AT or GC biases in sequence composition (Arneodo *et al.* 1998). The Voss and the tetrahedron methods, among others, are notable as they do not introduce biases in internucleotide distance, and thus represent good starting points for numeric analysis of nucleotide sequences.

Although various methods from the field of signal processing have previously been applied to nucleotide sequences (Silverman and Linsker 1986; Cheever *et al.* 1989; Katoh *et al.* 2002), they have yet to be applied to the problem of short read assembly. We tested the use of established signal processing techniques for assembling short DNA sequencing reads both with and without use of a reference sequence, and present a prototype approach for reducing the

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

computational expense of read alignment using signal processing techniques from the field of time series data mining. We benchmarked the alignment accuracy of a proof-of-concept short read alignment implementation against existing tools, before demonstrating the applicability of our approach to *de novo* assembly.

RESULTS

The four stages of our approach are: *i)* transforming symbolic nucleotide sequences to numerical sequences, *ii)* creating approximate representations of reads and, where appropriate, of a reference sequence, *iii)* performing accelerated comparison of the reduced dimensionality sequence representations to identify candidate alignments, and i*v)* verifying and finishing candidate alignments using original, full resolution sequences.

The symbolic to numerical sequence transformation can be performed using one of several methods, including the real number (Chakravarthy *et al.* 2004) and three dimensional DNA walk methods (Lobry 1996) methods illustrated in figures 1 and 3D respectively. The subsequent approximation step can be performed using a representation method such as the DFT, the DWT or PAA. Each of these can be used to approximate transformed sequences to different levels of resolution, permitting reduced complexity sequence analysis (compared in figure 1). Similarity between pairs of reduced dimensionality sequence representations can then be efficiently established using an appropriate distance measure, enabling candidate alignments to be identified. Finally, these alignments are rigorously assessed using original, full resolution data and conventional pairwise alignment algorithms.

Sequence approximation

Effective methods for approximating sequential data should: *i)* accurately represent data without loss of useful information, *ii)* have low computational overheads, *iii)* facilitate rapid comparison of data, and *iv)* provide lower bounding — where the distance between data representations is always smaller or equal to that of the original data — guaranteeing no false negative results (Faloutsos *et al.* 1994).

The DFT, the DWT and PAA are three widely used sequential data representation methods which satisfy these requirements (Mitsa 2010). The DFT decomposes a numerically transformed nucleotide sequence with *n* positions (dimensions) into a series of *n* frequency components ordered by their frequency. A subset of the resulting Fourier frequencies can be used to represent the original sequence at reduced resolution (Agrawal *et al.* 1993), and the tradeoff between analytical speed and accuracy can be varied according to the number of frequencies considered (Mörchen 2003).

The DWT is a set of averaging and differencing functions that may be used recursively to represent sequential data at different resolutions (Chan and Fu 1999). Unlike the DFT, the DWT provides time-frequency localisation, and so better accommodates changes in signal frequency over time (i.e. non-stationary signals) compared with the DFT and related methods (Wu *et al.* 2000). A drawback of the DWT is its requirement of input with length of an integer exponent of

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

two (i.e. 2^n). Where sequences have a length other than 2^n, they are padded with zeros up to the next integer exponent of two prior to application of the DWT. The corresponding DWT representations are then truncated in order to remove the bias associated with artificial padding (Percival and Walden 2006). For example, in order to approximate a time series with 500 data points to a resolution of three (i.e. 2^3), artificial padding must be added to increase its length to 512 (2^9) – the next integer exponent of two. In this case, the final, eighth wavelet should be truncated so as to avoid introducing bias.

In PAA (Geurts 2001; Keogh *et al.* 2001), a numerical sequence is divided into *n* equally sized windows, the mean values of which together form a compressed sequence representation. The selection of *n* determines the resolution of the representation. While PAA is faster and easier to implement than the DFT and the DWT, unlike these two methods it is irreversible meaning that the original sequence cannot be recovered from the representations. Figure 1 depicts DFT, DWT and PAA sequence representations of a numerically transformed viral sequence (HIV-1).

Reference-based alignment

To demonstrate the applicability of sequential data representations in the process of read alignment, we implemented a naive sequential scanning *k*-Nearest Neighbours (*k*NN) read alignment algorithm in the Matlab environment. In our proof-of-concept implementation (table 1), the prerequisite numerical sequence transformation is performed using a four dimensional binary mapping approach proposed by Voss (1992); chosen since it introduces no biases in internucleotide distance, and since its binary nature removes the need for normalisation prior to analysis. Approximate representations of transformed sequence *k*-mers are constructed using one of three implemented methods: the DFT, the DWT and PAA. Euclidean distance was used as a similarity measure for read and reference sequence comparison as: *i)* it is fast and easily implemented, *ii)* it is applicable to many different representation methods, and *iii)* its performance compares with more sophisticated '*elastic*' similarity methods for *k*NN search in medium to large datasets (Wang *et al.* 2013). After generating candidate alignments between reads and reference *k*-mers through pairwise comparison, these candidates alignments are verified through Needleman-Wunsch (NW) global alignment of their corresponding original sequences. Finally NW alignment scores are used to identify best alignments for each read and to reject false positives, and the algorithm's gapped output is used to construct alignments in the widely used Sequence Alignment/Map (SAM) file format.

Benchmarking

The accuracy our proof-of-concept aligner implementation (three variants) and four existing tools was evaluated against simulated HIV-1 reads generated with CuReSim (Caboche *et al.* 2014). Sixteen viral population datasets with combined rates of base insertions, deletions and substitutions of 0-10% were simulated, and CuReSim's companion tool CuReSimEval was used to quantify alignment accuracy in terms of *F*-score, a balanced measure of precision and recall.

The relative performance of three numerical sequence approximation methods was assessed against simulated reads with 6% and 10% variation from the simulated reference sequence. These two simulated datasets were aligned using an otherwise identical implementation using:

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

i) DWT representations of *k*-mers, *ii)* DFT representations of *k*-mers, *iii)* PAA representations of *k*-mers, and finally *iv)* numerically transformed but uncompressed *k*-mers (thus, removing the overhead of building approximate sequence representations).

We observed that in spite of their associated overheads, the use of approximate sequence representations dramatically reduced execution time and yielded alignments of equal or greater accuracy than could be obtained using uncompressed sequences. For simulated reads with 6% variation, alignment of uncompressed (baseline) sequences was performed in 734 seconds (s) with an *F*-score of 0.697. For the same dataset, PAA approximated sequences were aligned fastest and most accurately with an execution time of 54s (~14-fold faster than baseline) and an *F*-score of 0.704, while the DWT approximated sequences were aligned in 445s (~2 fold faster) with an *F*-score of 0.698, and the DFT approximated sequences were aligned in 120s (~6-fold faster) with an *F*-score of 0.692.

Alignment of reads with 10% overall variation was performed in 853s using uncompressed sequences giving rise to an *F*-score of 0.567. The PAA approximated sequences were again aligned fastest in 60s with an *F*-score of 0.572, while the DWT approximated sequences were aligned most accurately (*F*-score 0.577) but most slowly with an execution time of 516s, and the DFT approximated sequences were aligned in 138s with an *F*-score of 0.575.

The alignment accuracy of our signal decomposition approach was also evaluated alongside the existing read aligners Bowtie2 (Langmead and Salzberg 2012), BWA-MEM (Li and Durbin 2009), Mosaik (Lee *et al.* 2014), and Segemehl (Otto *et al.* 2014). Three variants of our approach using the DFT, the DWT and PAA dimensionality reduction methods were tested, using otherwise identical parameters and *k*-mers of length 100-300 nucleotides. As existing tools were all configured with default parameters and run once per dataset, so as to provide a conservative comparison of our tool's performance, the results we present for our approach correspond to the worst-performing *k*-mer for each dataset.

Using a strict definition of mapping correctness where reads are considered correctly mapped only if their exact start position is identified, the performance of the tested aligners was relatively similar. Segemehl generally produced the most accurate alignments in terms of *F*-score (figure 2A-C), its accuracy falling behind those of other tools only at the highest rate of sequence variation tested. Our implementation was outperformed by several existing tools in terms of strict start position accuracy. This can be mostly attributed to the use of a global alignment algorithm for the alignment finishing step, which, in the presence of insertion variation near the beginning of reads, tended to insert gaps rather than truncate the aligned region. Consequently, in some cases our approach identified a starting position one or two bases prior to that deemed correct by CuReSimEval. The challenges associated with assessing alignment correctness for benchmarking purposes are discussed by Holtgrewe *et al*. (2011).

Accordingly, a relaxed correctness definition (correct start position ±10 bases) was also considered, negating the impact of: *i)* multiple possible alignments associated with simulated variation near the start of reads, *ii)* use of different gapped alignment algorithms, and *iii)* the use

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of different algorithm parameters including scoring matrices and match, mismatch and gap extension penalties. Under this relaxed definition, our signal decomposition approach yielded the most accurate overall alignments for reads containing both insertion/deletion and substitution variation (figure 2F), while our DFT-based implementation offered joint best performance for reads containing only substitutions (figure 2E). Alignment accuracy for reads containing only insertion/deletion variation was comparable but slightly below the average of existing tools tested (figure 2D). Notably, the relative performance of our approach improved considerably with increasing rates of sequence variation, and existing tools were outperformed at the highest rates tested.

De novo assembly

To demonstrate the applicability of our approach to *de novo* assembly of short reads, we implemented a naive algorithm for all-against-all *k*-mer comparison using wavelet representations. Reads are first transformed using the Voss method. Every *k*-mer of each transformed read is subsequently identified and approximated to a reduced dimensionality wavelet representation using the DWT. These approximated *k*-mers are then compared with one another to establish their pairwise similarities in terms of Euclidean distance, and to construct a weighted graph such as shown in figure 3A. Finally, a breadth-first search (BFS) algorithm identifies the shortest path through the graph (figure 3B), and after attribution of *k*-mers to their corresponding reads, yields an assembly of short reads (figure 3C). Additionally, a numerical transformation such as the DNA walk may subsequently be used to aid visualisation of the assembly (figure 3D).

We implemented this algorithm and assembled simulated HIV populations. The dimensionality of the numerically transformed sequences was reduced by 16-fold using the DWT prior to alignment, and the deviation of the resulting assemblies from the reference sequence used for read simulation was quantified using CuReSimEval. Figure 4A illustrates the three dimensional 'walk' of the HIV-1 reference sequence HXB2 (accession number: K03455.1), while figures 4B, 4C and 4D depict the three dimensional surface plots, of *de novo* alignments for reads with 2%, 4% and 6% variation. The alignments of the 0%, 2%, 4% and 6% variation datasets had *F*-scores of 1, 0.9979, 0.9255 and 0.7772 respectively.

DISCUSSION

We have demonstrated application of a flexible sequence alignment heuristic leveraging established signal decomposition methods. Our prototype implementation aligned simulated viral reads with comparable overall precision and recall to existing tools, and excelled in the alignment of reads with the high levels of sequence diversity often observed in RNA virus populations (Archer *et al.* 2010). Our results show that base-level sequence resolution is not a prerequisite of accurate sequence alignment and that analytical performance can be preserved or even enhanced through appropriate dimensionality reduction of sequences. For example, respective six and fourteen fold reductions in execution time were observed during our tests of DFT and PAA-represented sequences, yet in both cases alignment accuracy was marginally better than that obtained using full resolution sequences. The approach's applicability to *de*

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

novo assembly of divergent sequences was also demonstrated. While our implementation makes use of *k*-mers, the approximate nature of the representation approaches means that optimal *k*-mer selection is considerably less important than it is with conventional exact *k*-mer matching methods. The inherent error tolerance of the approach also permits use of higher *k* values than would be suitable for use with conventional seed-and-extend algorithms, reducing the burden of pairwise comparison, and, in *de novo* assembly, the complexity of building and searching an assembly graph.

As demonstrated, nucleotide sequences may be represented as series of numeric vectors, enabling the application of existing analytical methods from a variety of mathematical and engineering fields. Indeed, to represent the composition of nucleic acid molecules as symbolic sequences is entirely arbitrary. By using proven signal decomposition methods, it is possible to create approximate representations of nucleotide sequences, permitting substantial reductions in the spatiotemporal complexity of their analysis without necessarily compromising analytical accuracy. In the context of sequence alignment, pairwise comparison of conservative, lower bounding representations of sequences may be performed to quickly reduce the search space of a more rigorous and computationally demanding final alignment stage, all without giving rise to false negative alignments. While we have applied this approach heuristically, an end-to-end alignment strategy using sequence representations of iteratively increasing resolution is conceivable, and an optimised implementation could allow time-efficient analyses of large, high complexity sequencing datasets without sacrificing analytical accuracy.

Efficient mining of terabase-scale biological sequences requires us to look beyond substring indexing algorithms towards more versatile methods of compression for both data storage and analysis. The use of probabilistic data structures can considerably reduce the computer memory required for in-memory sequence lookups at the expense of a few false positives, and Bloom filters and related data structures have seen broad application in *k*-mer centric tasks such as error correction (Shi *et al.* 2010), *in silico* read normalisation (Zhang *et al.* 2014) and *de novo* assembly (Salikhov *et al.* 2013; Berlin *et al.* 2014). However, while these hash-based approaches perform very well on datasets with high sequence redundancy, for large datasets with many distinct *k*-mers, large amounts of memory are still necessary (Zhang *et al*. 2014). Lower bounding approximation (such as DFT, DWT and PAA) exhibits the same attractive one-sided error offered by these probabilistic data structures, yet – rather than hashes – constructs intrinsically useful sequence representations, permitting their comparison with one another. Furthermore, approximation allows compression of standalone sequence composition, enabling flexible reduction of sequence resolution according to analytical requirements, so that redundant sequence precision need not hinder analysis. In large datasets, the associated reductions in resource usage can be significant. While the problem of read alignment to a known reference sequence is largely solved, the assembly of large and/or poorly characterised sequenced genomes remains limited by computational methods. Moreover, consideration of the metagenomic composition of mixed biological samples further extends the scope and scale of the assembly problem beyond what is tractable using conventional sequence comparison approaches. Through implementing prototypes of a reference-based and a *de novo* aligner, we

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

have demonstrated that numerical sequence approximations represent a tractable and versatile approach to short read analysis.

METHODS

Symbolic to numeric sequence transformation

The Voss transformation is a fixed mapping approach, which turns a nucleotide sequence with *n* dimensions into a 4*n*-dimensional binary matrix. Each row vector in the matrix represents a nucleotide base (symbols G, C, A and T), while each column vector represents a sequence position. Binary values are assigned to each cell, indicating the presence or absence of each nucleotide base at each sequence position (equation 1). Where V_{4i} *i*s the binary indicator for presence of a nucleotide in the ith position of the sequence *S* with *n* bases.

$$V_{4i} = \begin{bmatrix} 1i\begin{cases}1 \; if \;\; i = C \\ 0 \; otherwise\end{cases} \\ 2i\begin{cases}1 \; if \;\; i = G \\ 0 \; otherwise\end{cases} \\ 3i\begin{cases}1 \; if \;\; i = A \\ 0 \; otherwise\end{cases} \\ 4i\begin{cases}1 \; if \;\; i = T \\ 0 \; otherwise\end{cases} \end{bmatrix}, \; \forall \; i \in S_n \qquad (1)$$

Since the Voss transformation does not introduce inter-base mathematical bias, the pairwise distances between different transformed bases is the same.

DNA walks are graphical portrayals of the trajectories of nucleotide sequences in Hilbert space. In two dimensional DNA walks (Berger *et al*. 2004), the pyrimidines cytosine and thymine (symbols C and T) have an upward trajectory while the purines adenine and guanine (A and G) have a downward trajectory. The sequence's graph expands in a cumulative manner with each consecutive nucleotide in sequence (equation 2). In three dimensional *DNA walks (Lobry 1996)*, the cardinal directions north, south, east or west are assigned to each nucleotide base. In this way, a nucleotide sequence is represented as a three dimensional graph. DNA walks allow intuitive visualisation of sequence composition and variation (such as the viral genome in figure 4*A*).

Similarity search approaches for sequential data

Suitable methods for measuring the similarity of sequential data include the Lp-norms (Yi and Faloutsos 2000), Dynamic Time Warping (DTW) (Keogh and Ratanamahatana 2005), Longest Common SubSequence (LCSS) (Vlachos *et al.* 2002) and alignment algorithms such as the Needleman-Wunsch and Smith-Waterman algorithms. Euclidean distance is arguably the most widely used Lp-norm method for sequential data comparison. Lp-norms are straightforward and fast to compute, but require input data of the same dimensionality (sequence length). For

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comparing sequences of different length, a workaround is to truncate the longer sequence to the same length as the shorter sequence. Furthermore, Lp-norm methods do not accommodate for shifting in the *x*-axis (time or position) and are thus limited for identifying similar features between data that are offset. Elastic similarity/dissimilarity methods such as LCSS, DTW and various alignment algorithms permit comparison of data with different dimensions and tolerate shifts in the *x*-axis. These properties of elastic similarity methods can be very useful in the analysis of datasets such as speech signals, but are computationally expensive (Kotsakos *et al.* 2013) in comparison with measures of Euclidean distance.

Similarity search strategies can be broadly classified into whole matching and subsequence matching methods. Whole matching is appropriate for comparing sequences with similar overall characteristics—including length—with the queried sequence. Subsequence matching, by contrast, is more suited to identifying similarity between a short query sequence and limited, subsequence regions of longer sequences. Subsequence matching can however be effectively adapted for whole matching purposes through copying to a new dataset each subsequence falling within a sliding window of the longer sequence. The newly created dataset is then used for whole matching similarity search (Das *et al.* 1998). In spite of the storage redundancies associated with this approach, it is both fast and easily implemented, and so was used for our prototype alignment algorithm.

Read simulation ….after transformations

Using CuReSim, sixteen pyrosequencing runs of an HIV-1 HXB2 reference sequence (GenBank accession number: K03455.1) were simulated with a mean single read length of 400 nucleotides and mean coverage depth of 100 reads: *i)* in the absence of variation or sequencing error, *ii)* with base insertion/deletion rates of 1-5%, *iii)* with base substitution rates of 1-5%, and *iv)* with matching insertion/deletion and substitution rates of 1-5% (2%, 4%, 6%, 8%, and 10% overall variation) so as to simulate the heterogeneity present in diverse viral populations. Read quality was simulated at a fixed value of Sanger Q30. CuReSim recorded the exact origin of each simulated read with respect to the reference sequence, enabling critical evaluation of alignments using CuReSimEval in terms of precision, recall and *F*-score.

Benchmarking

Alignment accuracy was evaluated according to strict and relaxed definitions of mapping correctness. Correct read mapping according to the strict definition required exact identification of a read's starting position on the reference sequence, while the relaxed definition extended correctness to include ten positions to either side (±10 bases) of the known start position. These criteria were assessed using CureSimEval.

The relative performance of three dimensionality reduction methods was evaluated using simulated reads with 6% and 10% variation. These two read datasets were aligned using an otherwise identical implementation using: *i)* DWT representations of Voss-transformed *k*-mers, *ii)* DFT representations of Voss-transformed *k*-mers, *iii)* PAA representations of Voss-transformed *k*-mers, and finally *iv)* Voss-transformed but uncompressed *k*-mers (and thus removing the overhead of building approximate sequence representations).

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The three variants of our alignment implementation were evaluated in terms of mapping accuracy alongside the four existing read aligners Bowtie2 2.2.3, BWA-MEM 0.7.10, MOSAIK 2.2.3 and Segemehl 0.1.9 under Mac OS 10.9.4 using default parameters. The accuracy of these tools together with our proof-of-concept implementation was evaluated against the simulated reads using CuReSimEval.

In our prototype reference aligner implementation, DWT *k*-mer representations had a resolution of eight wavelets and were truncated where necessary to remove bias associated with artificially padding sequence length to the next highest integer exponent of two, while PAA representations had a length of eight, and the DFT representations used the first six frequencies from each sequence's Fourier series. Different *k*-mer lengths of 100-300 nucleotides (in increments of ten) were evaluated. In all cases, the Euclidean distance was used as a measure of sequence similarity. In the final, full resolution sequence alignment stage, default Matlab NW scoring parameters and the NUC44 [ftp://ftp.ncbi.nih.gov/blast/matrices/] were used. In order to provide a conservative comparison of our tool's performance with the existing tools which were tested with default parameters, the results presented correspond to the worst-performing *k*-mer length in terms of *F*-score for each read dataset.

Application to *de novo* assembly

All-to-all pairwise comparison of DWT *k*-mer representations was performed using Euclidean distance to inform the construction of a weighted assembly graph. The shortest path through the graph was identified using a BFS algorithm implemented in Matlab R2013a. Simulated HIV-1 HXB2 reads with sequence variation of 0%, 2%, 4% and 6% were assembled. Forward reads were represented by their initial *k*-mers of length 256, and these 256-mers were approximated to a resolution of 16 wavelets prior to graph construction.

ACKNOWLEDGEMENTS

We thank Mattia Prosperi and Douglas B. Kell for helpful discussion. AT has been supported by the Welcome Trust [097820/Z/11/B] and BBSRC [BB/H012419/1 & BB/M001121/1]. BC has been supported by a BBSRC DTP studentship.

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. A numerically transformed DNA sequence represented at different levels of spatial resolution using (A) discrete Fourier transforms (DFT), (B) discrete wavelet transforms (DWT) and (C) piecewise aggregate approximation (PAA). A 30 nucleotide sequence (top) is shown transformed into a numerical sequence (black lines) using the real number transformation method (T=1.5, C=0.5, G=-0.5 and A=-1.5). (A) DFT representations of the sequence with 5 (red), 3 (blue) and 1 (green) respective Fourier frequencies. (B) DWT representations of the same sequence with 8 wavelets (red), 4 wavelets (blue), and 2 wavelets (green). (C) PAA representations of the same sequence with 8 (red), 5 (blue) and 3 (green) respective coefficients.

Figure 2. Accuracy of our prototype aligners and four established tools in aligning simulated 400 nucleotide (mean) reads with varying levels of sequence variation. A-C depict alignment accuracies according to a strict criteria, requiring identification of a read's exact starting position determined by the read simulator, while in D-F a relaxed (±10 base) criteria is used. A and D show results for reads with 0-5% insertion/deletion variation, while B and E correspond to reads with 0-5% substitution variation. C and F show obtained accuracies for reads with combined, equally contributing insertion/deletion and substitution rates of 0-10%.

Figure 3. A *de novo* read assembly methodology for numerically represented nucleotide sequences. All-against-all sequence comparison (A) enables construction of a read graph with weighted edges. The weight assigned to each edge is the smallest pairwise distance between every possible *k*-mer representation of the two reads. The shortest path in the graph is identified with a breadth-first search algorithm (B), thereby enabling read alignment (C). A DNA walk of aligned reads (D) may subsequently be used to illustrate alignment characteristics.

Figure 4. A three-dimensional DNA walk of the HIV-1 HXB2 genome (A) also plotted with *de novo* alignments of three simulated HXB2 sequencing datasets of 2%, 4%, and 6% rates of combined insertion/deletion and substitution variation (B-D respectively). Plotting DNA walks of aligned short reads enables intuitive visualisation of the nature and extent of sequence diversity across a genomic region, with sequence variants each represented by a distinct trajectory through space.

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLES

Table 1: Implementation of Sequential Scanning alignment

1)	Transform short reads (of forward and reverse strand origin) and reference genome into numerical sequences
2)	Select appropriate *k*-mer length
3)	Create approximate representations (**ref_kmer_reps**) of each *k*-mer component of the numerically transformed reference sequence.
4)	Create approximate representations (**read_reps**) of the initial *k*-mer of each numerically transformed read.

5)	Identify potential candidate positions using DWT representations.
	
	`for each read i`
	`  best_dist = null`
	`  candidate_positions = []`
	`  for each ref_kmer j`
	`  if dist(read i,ref_kmer j) < best_dist`
	`  best_dist = dist(read i,ref_kmer j)`
	`  candidate_positions_i[1] = j`
	`  elseif dist(read i,ref_kmer j) == best_dist`
	`  candidate_positions_i[+1] = j`
	`  end`
	`  end`
	`end`

6)	Align approximate results on original data with Needleman-Wunsch algorithm (NWA).
	
	`for each read i`
	`  best_aln_score = null`
	`  best_aln = []`
	`  for each ref_kmer j in candidate_positions_i`
	`  if NWA_score(ref_kmer j,read i) > best_aln_score`
	`  best_aln_score = NWA_score(ref_kmer j,read i)`
	`  best_aln = NWA_aln(ref_kmer j,read i)`
	`  end`
	`  end`
	`end`

7)	Output alignment in Sequence Alignment/Map (SAM) format .

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Agrawal R, Faloutsos C, Swami A. 1993. *Efficient similarity search in sequence databases*. Springer.

Anastassiou D. 2001. Genomic signal processing. *Signal Processing Magazine, IEEE* **18**(4): 8-20.

Archer J, Rambaut A, Taillon BE, Harrigan PR, Lewis M, Robertson DL. 2010. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time—an ultra-deep approach. *PLoS Computational Biology* **6**(12): e1001022.

Arneodo A, D'Aubenton-Carafa Y, Audit B, Bacry E, Muzy J, Thermes C. 1998. What can we learn with wavelets about DNA sequences? *Physica A: Statistical Mechanics and its Applications* **249**(1): 439-448.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Berger JA, Mitra SK, Carli M, Neri A. 2004. Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute* **341**(1): 37-53.

Berlin K, Koren S, Chin C-S, Drake J, Landolin JM, Phillippy AM. 2014. Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing. *bioRxiv*: 008003.

Caboche S, Audebert C, Lemoine Y, Hot D. 2014. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC genomics* **15**(1): 264.

Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. 2004. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP Journal on Applied Signal Processing* **2004**: 13-28.

Chan K-P, Fu A-C. 1999. Efficient time series matching by wavelets. In *Data Engineering, 1999 Proceedings, 15th International Conference on*, pp. 126-133. IEEE.

Cheever E, Searls D, Karunaratne W, Overton G. 1989. Using signal processing techniques for DNA sequence comparison. In *Bioengineering Conference, 1989, Proceedings of the 1989 Fifteenth Annual Northeast*, pp. 173-174. IEEE.

Das G, Lin K-I, Mannila H, Renganathan G, Smyth P. 1998. Rule Discovery from Time Series. In *KDD*, Vol 98, pp. 16-22.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910): 133-138.

Faloutsos C, Ranganathan M, Manolopoulos Y. 1994. *Fast subsequence matching in time-series databases*. ACM.

Geurts P. 2001. Pattern extraction for time series classification. In *Principles of Data Mining and Knowledge Discovery*, pp. 115-127. Springer.

Holtgrewe M, Emde A-K, Weese D, Reinert K. 2011. A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics* **12**(1): 210.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**(14): 3059-3066.

Kececioglu JD, Myers EW. 1995. Combinatorial algorithms for DNA sequence assembly. *Algorithmica* **13**(1-2): 7-51.

Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. 2001. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM SIGMOD Record* **30**(2): 151-162.

Keogh E, Ratanamahatana CA. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems* **7**(3): 358-386.

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kotsakos D, Trajcevski G, Gunopulos D, Aggarwal CC. 2013. Time-Series Data Clustering.

Kwan HK, Arniker SB. 2009. Numerical representation of DNA sequences. In *Electro/Information Technology, 2009 eit'09 IEEE International Conference on*, pp. 307-310. IEEE.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.

Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS ONE* **9**(3): e90581.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Lobry J. 1996. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie* **78**(5): 323-326.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Mitsa T. 2010. *Temporal data mining*. CRC Press.

Mörchen F. 2003. Time series feature extraction for data mining using DWT and DFT. Univ.

Myers EW. 1995. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology* **2**(2): 275-290.

Otto C, Stadler PF, Hoffmann S. 2014. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*: btu146.

Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown CT. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences* **109**(33): 13272-13277.

Percival DB, Walden AT. 2006. *Wavelet methods for time series analysis*. Cambridge University Press.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98**(17): 9748-9753.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**(7356): 348-352.

Salikhov K, Sacomoto G, Kucherov G. 2013. Using cascading Bloom filters to improve the memory usage for de Brujin graphs. In *WABI*, pp. 364-376.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research* **20**(9): 1165-1173.

Shi H, Schmidt B, Liu W, Müller-Wittig W. 2010. A Parallel Algorithm for Error Correction in High-Throughput Short-Read Data on CUDA-Enabled Graphics Hardware. *Journal of Computational Biology* **17**(4): 603-615.

Shrestha AMS, Frith MC, Horton P. 2014. A bioinformatician's guide to the forefront of suffix array construction algorithms. *Briefings in bioinformatics* **15**(2): 138-154.

Silverman B, Linsker R. 1986. A measure of DNA periodicity. *Journal of Theoretical Biology* **118**(3): 295-300.

Verleysen M, François D. 2005. The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems*, pp. 758-770. Springer.

Vlachos M, Kollios G, Gunopulos D. 2002. Discovering similar multidimensional trajectories. In *Data Engineering, 2002 Proceedings 18th International Conference on*, pp. 673-684. IEEE.

Voss RF. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters* **68**(25): 3805.

bioRxiv preprint doi: https://doi.org/10.1101/011940; this version posted November 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E. 2013. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery* **26**(2): 275-309.

Wu Y-L, Agrawal D, El Abbadi A. 2000. A comparison of DFT and DWT based similarity search in time-series databases. In *Proceedings of the ninth international conference on Information and knowledge management*, pp. 488-495. ACM.

Yang Q, Wu X. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making* **5**(04): 597-604.

Ye N. 2003. *The handbook of data mining*. Lawrence Erlbaum Associates Mahwah, NJ.

Yi B-K, Faloutsos C. 2000. Fast time sequence indexing for arbitrary Lp norms. VLDB.

Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. 2013. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *arXiv preprint arXiv:13092975*.

Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. 2014. These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure. *PLoS ONE* **9**(7): e101271.

Figure 1

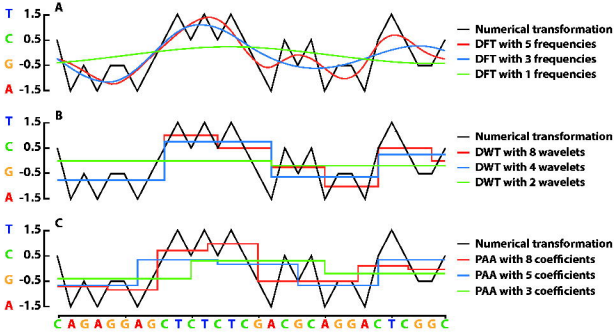

A
T 1.5
C 0.5
G -0.5
A -1.5
Numerical transformation
DFT with 5 frequencies
DFT with 3 frequencies
DFT with 1 frequencies
B
T 1.5
C 0.5
G -0.5
A -1.5
Numerical transformation
DWT with 8 wavelets
DWT with 4 wavelets
DWT with 2 wavelets
C
T 1.5
C 0.5
G -0.5
A -1.5
Numerical transformation
PAA with 8 coefficients
PAA with 5 coefficients
PAA with 3 coefficients
C A G A G G A G C T C T C T C G A C G C A G G A C T C G G C

Figure 2

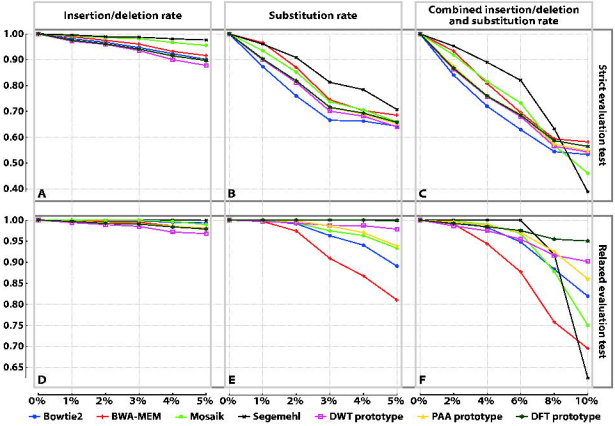

Insertion/deletion rate
Substitution rate
Combined insertion/deletion and substitution rate
Strict evaluation test
Relaxed evaluation test
A
B
C
D
E
F
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.95
0.85
0.75
0.65
0%
1%
2%
3%
4%
5%
6%
8%
10%
Bowtie2
BWA-MEM
Mosaik
Segemehl
DWT prototype
PAA prototype
DFT prototype

Figure 3

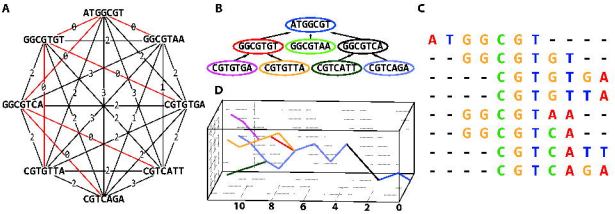
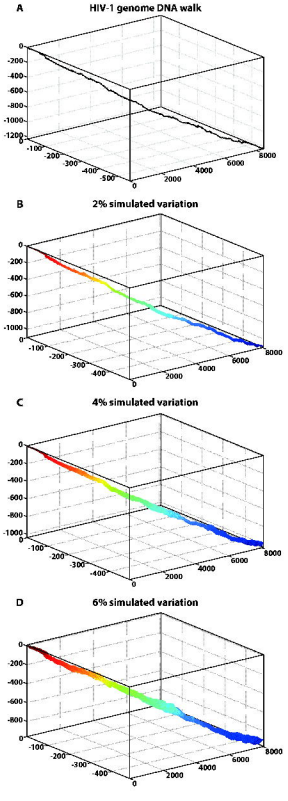

A
ATGGCGT
GGCGTGT
GGCGTAA
GGCGTCA
CGTGTGA
CGTGTTA
CGTCATT
CGTCAGA
B
ATGGCGT
GGCGTGT
GGCGTAA
GGCGTCA
CGTGTGA
CGTGTTA
CGTCATT
CGTCAGA
C
A T G G C G T - - - -
- - G G C G T G T - -
- - - - C G T G T G A
- - - - C G T G T T A
- - G G C G T A A - -
- - G G C G T C A - -
- - - - C G T C A T T
- - - - C G T C A G A
D
10 8 6 4 2 0

Figure 4


A
HIV-1 genome DNA walk
B
2% simulated variation
C
4% simulated variation
D
6% simulated variation