The arithmetic topology of genetic alignments

Christopher Barrett\textsuperscript{1,2} · Andrei Bura\textsuperscript{1} · Qijun He\textsuperscript{1} · Fenix Huang\textsuperscript{1} · Christian Reidys\textsuperscript{1,3}

Abstract
We propose a novel mathematical paradigm for the study of genetic variation in sequence alignments. This framework originates from extending the notion of pairwise relations, upon which current analysis is based on, to \(k\)-ary dissimilarity. This dissimilarity naturally leads to a generalization of simplicial complexes by endowing simplices with weights, compatible with the boundary operator. We introduce the notion of \(k\)-stances and dissimilarity complex, the former encapsulating arithmetic as well as topological structure expressing these \(k\)-ary relations. We study basic mathematical properties of dissimilarity complexes and show how this approach captures watershed moments of viral dynamics in the context of SARS-CoV-2 and H1N1 flu genomic data.

Christopher Barrett, Andrei Bura, Qijun He, Fenix Huang and Christian Reidys have contributed equally to this work

\begin{itemize}
\item Christian Reidys
  cmr3hk@virginia.edu
\item Christopher Barrett
  clb5xe@virginia.edu
\item Andrei Bura
  cb8wn@virginia.edu
\item Qijun He
  qh4nj@virginia.edu
\item Fenix Huang
  fwh3zc@virginia.edu
\end{itemize}

\textsuperscript{1} Biocomplexity Institute, University of Virginia, 994 Research Park Boulevard, Charlottesville, VA 22911, USA
\textsuperscript{2} Department of Computer Science, University of Virginia, 351 McCormick Road, Charlottesville, VA 22904, USA
\textsuperscript{3} Department of Mathematics, University of Virginia, 141 Cabell Drive, Charlottesville, VA 22904, USA
Introduction

Genetic variation is the observed difference at the genetic sequence level between individuals in a population and is the key contributor to phenotypic diversity. It affects population dynamics and ultimately the evolution of the entire system.

One of the key tasks of Biology is understanding the development of genetic variation within a given population. Namely, the extraction of evolutionary relationships and histories among the sequences present (i.e. their phylogenetics). These relationships are construed in the guise of the phylogenetic tree; a graph topological structure that underpins our understanding of the major evolutionary transitions appearing in the system. This structure is central to inferring everything from the emergence of new body plans, novel metabolism and the origin of new genes to detecting molecular adaptation, understanding morphological character evolution and reconstructing demographic changes in recently diverged species (Kapli et al. 2020).

Such a tree is generally constructed from metric (pairwise) information present at the sequence level. For two sequences of equal length, one naive approach is to employ the Hamming distance, which counts the number of positions in the sequence pair with different entries at those positions. For sequences of different length, alignment methods are applied to obtain equal length via the insertion of “gap” symbols (Needleman and Wunsch 1970; Smith and Waterman 1981). Other metrics can be used, such as the edit distance, which can still be viewed as weighted versions of the Hamming metric Berger et al. (2021).

As a result, metric-based phylogenetic tree constructions integrate only pairwise dissimilarity information. This information can be viewed as a complete graph where each node represents a sequence and the length of an edge represents the dissimilarity between the two corresponding sequences the edge links. The integration process in metric-based phylogeny can be summarized as finding a spanning tree whose inherited metric information “best fits” the metric in this complete graph (Felsenstein and Felsenstein 2004; Saitou and Nei 1987).

Graphs are natural choices but can only encode pairwise relations. However in a population of genetic sequences, there exist higher order interactions that cannot be expressed from pairwise relations alone. In a population of sequences, different positions (sites) exhibit different nucleotide diversities. Some sites are conserved allowing almost no polymorphism while others, the key contributors to genetic variation, are less conserved. Polymorphic sites that allow three or more nucleotide realizations (tri-allelic or higher) are of interest as they are particularly indicative of the quasi-species exploring the fitness landscape by stochastic mutation and post-selection (Westen et al. 2009; Huebner et al. 2007). For instance, in the SARS-CoV-2 genome, site 23012, located in the spike protein region, exhibits such polymorphisms, see Fig. 1.
Fig. 1 Site 23012 on the SARS-CoV-2 genome

Wild type  |  Beta variant  |  Kappa variant
--- | --- | ---
\[23012\] | \[\text{CAA}\] | \[\text{CAA}\]
\[\text{CAA}\] | \[\text{AAA}\] | \[\text{CAA}\]
\[\text{AAA}\] | \[\text{AAA}\] | \[\text{CAA}\]

Fig. 2 Segments of SARS-CoV-2 genomes and their GSAID IDs. \(X\): EPI-ISL-428678, \(Y\): EPI-ISL-509797, \(Z\): EPI-ISL-656231, \(X'\): EPI-ISL-467431, \(Y'\): EPI-ISL-428678, \(Z'\): EPI-ISL-415706. We cannot distinguish the triplet \(\{X, Y, Z\}\) from that of \(\{X', Y', Z'\}\) if we restrict ourselves to pairwise Hamming distance comparisons. That is because for hamming distance \(d_H\) we have
\[
d_H(X, Y) = d_H(X', Y'),
\]
\[
d_H(X, Z) = d_H(X', Z'),
\]
\[
d_H(Y, Z) = d_H(Y', Z').
\]
However, the triple \(\{X, Y, Z\}\) contains two tri-allelic sites, while \(\{X', Y', Z'\}\) does not.

This site has the nucleotide \(G\) in the wild type sequence, while in the Beta variant of the virus, it mutates to \(A\) and induces an amino acid change from \(E\) to \(K\) at position 484. \(E484K\) is one of the characteristic mutations of the Beta variant and improves the virus’s ability to evade the host’s immune system (Wise 2021). One the other hand, in the Kappa variant, site 23012 mutates to \(C\) and results in the amino acid change \(E484Q\), which is thought to enhance ACE2 receptor binding (Cherian et al. 2021), and may reduce the vaccine-stimulated antibodies’ ability to attach to this altered spike protein (Wilhelm et al. 2021).

Conservation at a site can be measured by various indices which are usually based on alphabet frequency distributions, information entropy or stereo-chemical matrix considerations. In Valdar (2002), Valdar provides a retrospective on such indices and presents the criteria a successful score should obey. Although the scores become sophisticated in terms of components they employ, such as means of accounting for sequence redundancy in the MSA and stereo-chemical variability or gap presence, the scores that yield the best results, are still based on sum of pairs constructions. When considering only symbol frequency distributions, key features of highly polymorphic sites in genetic alignments, like the site 23012 on the Spike protein of the SARS-CoV-2 genome, cannot be expressed employing measures derived from pairwise comparisons alone, see Fig. 2.

To capture multi-fold sequence interactions like the ones in Fig. 2, we will introduce a novel notion of dissimilarity (the complement of conservation), that augments the sums of pairs approaches in Valdar (2002) to \(k\)-tuples where \(k\) can be greater than two. We propose passing from weighted graphs to weighted simplicial complexes (that account for \(k\)-fold relations) embodied in a novel arithmetico-topological structure we call the dissimilarity complex which encodes these sequence hyper-relations.

The notion of dissimilarity entails more data than just a collection of certain scores. This is already apparent in the analysis of sequences based on Hamming distances.
Genetic data are in fact more deeply understood via tree or tree-like concepts, all of which reside in the weighted graph of Hamming distances. A tree is a maximally acyclic spanning graph of the complete graph over the set of its vertices. Tree edges form a maximally independent set, namely including any edge that connects two vertices we enclose a unique cycle. The distance between two such vertices in the complete graph can be approximated by the sum of involved tree edge lengths. In conclusion, the graph contextualizes the notions of trees instrumental in classical phylogeny.

The goal of this paper is to introduce the dissimilarity complex, which is the higher dimensional counterpart of the weighted graph as it appears in classical phylogeny. In fact this graph can be derived as the 1-skeleton of the dissimilarity complex. The dissimilarity complex is a weighted simplicial complex.

An (unweighted) simplicial complex is classically regarded as just a set composed of points, edges, triangles, tetrahedra and their higher dimensional counterparts. In addition to its combinatorics, a simplicial complex can also be studied from a global (topological) perspective as it naturally gives rise to a topological space. In particular, homological (algebraic) invariants of simplicial complexes often already capture important information about global structural features of the sequence phylogeny. For instance, $k$-dimensional holes in such spaces can be shown to correspond to multi-fold recombination (Chan et al. 2013). Recently, weighted homology, an augmentation of simplicial homology that further encodes arithmetical information for each simplex has been developed (Bura et al. 2021; Dawson 1990; Ren et al. 2018). These weights and the information they carry are central to the new theory as they enrich the algebraic invariants of classical simplicial homology with arithmetical torsion that encodes additional combinatorial information about the sequence phylogeny.

The dissimilarity complex, similar to the graph of Hamming distances, allows to derive the notion of a “phylogenetic complex”, the higher dimensional analogue of a tree. Distance based phylogeny shows the way: here one constructs a tree of sequences whose tree metric “best approximates” the complete graph on sequences weighted by their pairwise distances. To mimic this process, one would construct the phylogenetic complex, whose induced dissimilarity “best approximates” that of the dissimilarity complex of the given alignment of sequences. Such a derived object should retain “tree like” qualities and it is here where the choice of a discrete valuation rings is key, see (Li and Reidys 2022).

As a higher dimensional analogue of the phylogenetic tree, the phylogenetic complex should behave similarly: It should contain no (homological) cycles at any simplicial dimension and should be able to approximate the weight of any simplex that appears in the dissimilarity complex but not in the phylogenetic one. These notions can only be derived within the framework of the dissimilarity complex introduced here.

Our paper is organized as follows:

Section 1.1 provides context by reviewing the current phylogenetic study of a sequence alignment via Hamming distance optimizations and phylogenetic trees.

In Sect. 2 we introduce $k$-stances, a higher dimensional relation between $k$ aligned sequences representing a dissimilarity measurement that naturally generalizes the Hamming distance. Section 2.1 deals with the mathematical properties of this this
k-ary relation, both related to metric properties it inherits from the Hamming distance (pairwise case) and to properties that are intrinsic to its higher dimensionality. Section 2.2 discusses k-stances in the context of genetic recombination.

Section 3 introduces the dissimilarity complex: an arithmetico-geometrical space whose topological structure encodes the k-stance relations among the sequences of an alignment by integration across all k-dimensions. We shall briefly review the notion of simplicial complexes and their build principles, usage and means of study. In Sect. 3.1 we provide the construction of the dissimilarity complex of a sequence alignment and establish some basic properties. In Sect. 3.2 we then introduce weighted homology and show how it is connected to the dissimilarity complex of an alignment.

Section 4 analyzes the “induction basis” of alignments of length one. We shall show here that all k-stances are in fact connected.

Section 5 considers the multiple column case. In Sect. 5.1 we present two case studies on the statistics of k-stances for alignments of SARS-CoV-2 and H1N1 flu genomes, while in Sect. 5.2 we hint at the relationships between the arithmetic torsion that arises in the weighted homology of a multi-column alignment and its connections to genetic recombination present in the population of the alignment.

Section 6 we integrate our results and discuss future directions of work. Finally, Sect. 7 contains all proofs.

1.1 Hamming distance and phylogeny

Let us begin by revisiting the underlying ideas of Hamming distance, its basic metric properties and current metric-based phylogeny.

Given an alphabet \( \mathcal{A} \), let \( \mathcal{A}^l \) denote the set of all sequences of length \( l \) over \( \mathcal{A} \). In the case of DNA sequences, \( \mathcal{A} = \{ A, C, G, T \} \). The Hamming distance between two sequences \( w_0, w_1 \in \mathcal{A}^l \), denoted by \( h(w_0, w_1) \), is the number of positions in which the two sequences differ.

It is easy to check that \( h \) satisfies the following axioms making it a metric. Namely, for any \( w_0, w_1, w_2 \in \mathcal{A}^l \)

1. (identity of indiscernibles) \( h(w_0, w_1) = 0 \iff w_0 = w_1 \).
2. (symmetry) \( h(w_0, w_1) = h(w_1, w_0) \).
3. (triangle inequality) \( h(w_0, w_2) \leq h(w_0, w_1) + h(w_1, w_2) \).

Given a set of sequences in \( \mathcal{A}^l \), all pairwise distance information can be encoded in a symmetric matrix, or equivalently, a complete weighted graph, where each node represents a sequence and the weight of the edge represents the Hamming distance between the two corresponding sequences.

This metric structure over \( \mathcal{A}^l \) allows to infer the phylogenetic relations of a given set of sequences via recursive clustering (i.e. Neighbor-joining, UPGMA, WPGMA etc. Saitou and Nei 1987; Sokal 1958) or via optimization (minimum evolution, least squares inference etc. Fitch and Margoliash 1967; Hendy and Penny 1982). These relations can then be represented in the phylogenetic tree where the input sequences become the tree’s leaves and the internal nodes can then be interpreted as common ancestors of their descendants. The key idea in all such algorithms is that no matter the metric information on the sequences, our target (the phylogenetic tree) is acyclic.
Fig. 3 LHS: a complete graph that encodes all pairwise dissimilarities among sequences labelled A, B, C and D. The edge label represents the distance between the corresponding sequences. RHS: the corresponding neighbor-joining tree: X and Y were added as interior nodes, and the distance between two original nodes is approximated by the sum of the length of the edges on the unique path the connects these nodes in the tree, e.g. \(d'(A, B) = 7 + 4.5 + 5.5 = 17\) approximates the original \(d(A, B) = 13\).

and spanning. As a result there exists a unique path between any two leaves, i.e. an unique distance which approximates the original distance, see Fig. 3.

In what follows we shall mimic this construction but move beyond pairwise metrics to \(k\)-wise comparisons (\(k\)-stances). As a canonical analogue of the weighted graph, encoding the metric structure, the dissimilarity complex will encapsulate the \(k\)-stances.

2 \(k\)-Stances

As previously noted, one encounters \(k\)-ary interactions that have no pairwise (Hamming) analogue. In this section we will derive a \(k\)-spectrum of measurements that capture such higher order dissimilarity relations among multiple sequences. The following Subsections deal with the properties of this measurement and its connection to a particular type of genetic recombination.

We begin by reformulating the Hamming distance between two sequences. Consider the following projections: for any \(j \in \{1, \ldots, l\}\), let \(f_j : A^l \rightarrow A\) where for a sequence \(w \in A^l\), \(f_j(w)\) is the letter at position \(j\) in \(w\). Then for two sequences \(w_0, w_1 \in A^l\), each position \(j\) contributes one unit to the Hamming distance between the two sequences if and only if \(f_j(w_0) \neq f_j(w_1)\).

Accordingly, a position \(j\) for the two sequences contributes one unit to the distance, if the number of distinct letters that appear at said position in \(w_0\) and \(w_1\) respectively, is equal to the number of sequences, i.e. two in this case. Stated this way, the definition of Hamming distance immediately hints at a generalization for any number \(k \geq 1\) of sequences as follows

**Definition 1** Let \(k \geq 1\). The \(k\)th order dissimilarity or \(k\)-stance, of \(k\) given sequences each of length \(l\), is given by

\[
d_k : (A^l)^k \rightarrow \mathbb{N}, \quad d_k(w_0, \ldots, w_{k-1}) := |\{j \in [1, l] : \bigcup_{i=0}^{k-1} \{f_j(w_i)\} = k\}|
\]

In other words, the \(k\)-stance of \(k\) given sequences is the number of positions in which the given sequences are all mutually distinct, see Fig. 4.
Fig. 4  $w_0 = AAGAGGCTT$, $w_1 = AAGAGACCT$ and $w_2 = GTGGGTCCC$. Note that at position 2 and position 6, all three sequences are mutually distinct. As such $d_3(w_0, w_1, w_2) = |\{j \in [1, 9] : \bigcup_{i=0}^{3-1} \{f_j(w_i)\} = 3\}| = ||2, 6|| = 2$

2.1 Basic properties

By definition, $d_1 = l$, i.e. the 1-stance reproduces the sequence length, while the 2-stance is exactly the Hamming distance, $d_2 = h$. In case of $k > 2$, the $k$-stances has the following properties.

**Proposition 1**  For any $k > 2$ and for any $w_0, \ldots, w_k \in A^l$, the $k$-stance satisfies

1. (implication of indiscernibles) if $w_i = w_{i'}$ for some $0 \leq i < i' \leq k - 1$ then $d_k(w_0, \ldots, w_{k-1}) = 0$.
2. (symmetry) $d_k(w_0, \ldots, w_{k-1}) = d_k(w_{\epsilon(0)}, \ldots, w_{\epsilon(k-1)})$ for any index permutation $\epsilon \in S_k$.
3. (polyhedron inequality) $d_k(w_0, \ldots, w_{k-1}) \leq \sum_{i=0}^{k-1} d_k(w_0, \ldots, \hat{w}_i, \ldots, w_k)$, where $\hat{w}_i$ denotes the omission of the $i$th sequence, see Fig. 5.

Accordingly, for $k > 2$, $d_k$ can be considered as a higher order pseudometric (Collatz 2014). Note that several generalization of metrics and their properties have been studied in the literature (Klein and Zhu 1998; Deza et al. 1997; Sommerville 2020). In particular, in Klein and Zhu (1998) such polyhedron type inequalities appear for certain graph invariants in the context of sums of powers of volumes of $k$-vertices in a graph. This higher dimensional “volume” is still derived from pairwise vertex quantities via Cayley-Menger type constructions (Sommerville 2020). In Deza et al. (1997) generalizations of the triangle inequality to hypermetrics are also presented in the context of the cut cone and integer programming, all of which are based on pairwise relations. Dissimilarity, in contrast, entails genuine $k$-interactions.

**Proposition 2**  For any $k > 2$ and for any $w_0, \ldots, w_{k-1} \in A^l$, the $k$-stance satisfies

1. (dimensional bounding) $d_k(w_0, \ldots, w_{k-1}) = 0$, for any $k > |A|$.
2. (dimensional monotonicity) $d_k(w_0, \ldots, w_{k-1}) \leq d_{k'}(w_{i(0)}, \ldots, w_{i(k'-1)})$, for any $1 \leq k' \leq k$ and any injection $i : \{0, \ldots, k'-1\} \rightarrow \{0, \ldots, k-1\}$.

2.2 $k$-Stances and genetic recombination

Genetic recombination can be defined as the exchange of genetic material among multiple sequences and is a key contributor to genetic variation (Rubio et al. 2001). In this section, we focus on a particular class of recombination and discuss its connections to $k$-stances.
Example 2 Let \( w_0, w_1, w_2 \) be two sets of length four sequences, exhibiting the same Hamming distance signature, and as such \( d_3(w_0, w_1, w_2) = 3 \). Note that for instance, \( d_3(w_0, w_1, w_2) = 4 \leq 8 = 2 + 3 + 3 = d_3(w_0, w_1) + d_3(w_0, w_1) + d_3(w_1, w_2), \) and this holds for any other permutation as well.

Definition 2 Let \( W = \{w_0, \ldots, w_{k-1}\} \subset A^l \) be \( k \) sequences and fix another sequence \( w \in A^l \). Then \( w \) is called a linear recombinant of \( W \), if for each \( j \in \{1, \ldots, l\} \) there exists an \( i \in \{0, \ldots, k-1\} \) such that \( f_j(w) = f_j(w_i) \).

Then we obtain

Proposition 3 Let \( W = \{w_0, \ldots, w_{k-1}\} \subset A^l \) be \( k \) fixed sequences. Then, \( d_k(w_0, \ldots, w_{k-1}) = 0 \) if there exists \( w \in W \) such that \( w \) is a linear recombinant of \( W \setminus \{w\} \).

Example 1 Consider \( W = \{w_0 = GCCT, w_1 = TTCA, w_2 = GCCA\} \). Firstly, \( d_1(w_0) = d_1(w_1) = d_1(w_2) = 4, d_2(w_0, w_1) = 4, d_2(w_0, w_2) = d_2(w_1, w_2) = 2 \). Since \( w_2 = f_1(w_0)f_2(w_0)f_3(w_1)f_4(w_1) \), \( w_2 \) is a linear recombinant of \( W \setminus \{w_2\} \), and as such \( d_3(w_0, w_1, w_2) = 0 \).

Using higher order dissimilarity (as low in dimensionality as 3-stance) we obtain a more refined description of a sequence set. The following example illustrates this point in providing two sets of sequences, exhibiting the same Hamming distance signature, while differing on the level of 3-stances.

Example 2 Let \( W_0 = \{w_0^0 = TTCA, w_1^0 = CTCG, w_2^0 = TTTG\} \) and \( W_1 = \{w_0^1 = CTCG, w_1^1 = TTAG, w_2^1 = ATTG\} \) be two sets of length four sequences. We have \( d_1(w_0^j) = d_1(w_1^j) = d_1(w_2^j) = 4, d_2(w_0^j, w_1^j) = d_2(w_0^j, w_2^j) = d_2(w_1^j, w_2^j) = 2 \) for \( j = 0, 1 \). However, \( d_3(w_0^0, w_0^1, w_2^0) = 0 \neq 2 = d_3(w_0^1, w_1^1, w_2^1) \). Accordingly for \( d_1 \) and \( d_2 \), \( W_0 \) and \( W_1 \) exhibit identical dissimilarity while their \( d_3 \)-dissimilarities are distinct.

It is worth pointing out that, in the above example, none of the \( W_0 \)-sequences is a linear recombinant of the remaining two, while their 3-stance is still zero. This indeed shows that the implication of indiscernibles for \( k \)-stances with respect to linear recombinants is a sufficient but not a necessary condition.
3 The dissimilarity complex

To understand the extra arithmetic information encapsulated within the dissimilarity complex, we recall the notion of simplicial complexes. We adopt here a data-centric point of view and eschew abstract topological and category theoretical considerations.

Suppose we are given a discrete, finite set of data points $W$ as measurements of a system, and suppose that among the points in this data set there exists a multi-fold relation denoted by $\wedge$, i.e. a collection of $k$-ary relations for all $1 \leq k \leq |W|$. This relation might model an intrinsic dependency in the system that manifests in the measurements in $W$. Suppose that the relations in $\wedge$ are restriction compatible, i.e. they have the property that for any subsets $W'' \subseteq W' \subseteq W$ we have that $\wedge W' \implies \wedge W''$. Namely, if a subset of elements in $W$ are in a relation in $\wedge$, then any subset of those elements are in a lower arity relation $\wedge$ as well. For instance, suppose $W := \{w_0 = AC, w_1 = AT, w_2 = AG\}$ and the relation $\wedge$ would be “contains a nucleotide in common”. Then, clearly $\wedge\{w_0, w_1, w_2\}$ but also $\wedge\{w_0, w_1\}, \wedge\{w_0, w_2\}$ and $\wedge\{w_1, w_2\}$. This is since among the triplet and among any pairs, the nucleotide $A$ always appears in common. We shall further assume that for each individual measurement $w \in W$ we have $\wedge\{w\}$. Then, the combinatorial structure consisting of the totality of such $\wedge$-satisfying subsets is called the simplicial complex of the data set $W$ under the relation $\wedge$ and is denoted by $(W, \wedge)$. A single such element $W' \in (W, \wedge)$ is called a $|W'|-1$-dimensional simplex. This complex can be organised into a topological object by embedding each simplex $W' \in (W, \wedge)$ into a $(|W'|-1)$-dimensional $\mathbb{R}$-linear (Euclidean) polytope, and gluing these polytopes along their common faces, see Fig. 6.

We shall use this perspective to build the dissimilarity complex by letting $W$, the data set, be the sequences in a (genetic) alignment, with the relation $\wedge := \text{“mutual dissimilarity at at least one position”}$. Furthermore, we keep track of the degree of dissimilarity for each such $k - 1$-simplex via the $k$-stance among its constituent sequences.

3.1 Construction of the dissimilarity complex

We are now in position to formally introduce the natural mathematical structure encapsulating higher order dissimilarity information among a collection of sequences.
Definition 3 An alignment $W = [w_1, \ldots, w_n]$, $w_i \in \mathcal{A}_l$, is a finite ordered tuple of sequences of equal length $l$ over an alphabet $\mathcal{A}$. We can view $W$ as a matrix whose entries are letters in $\mathcal{A}$ with $w_i$ being the sequence (ordered tuple of letters) in the $i$th row of $W$. Furthermore, the $j$th column of $W$ is the ordered tuple $[f_j(w_1), \ldots, f_j(w_n)]$. The integer $l$ is called the length of the alignment (the number of columns) while the integer $n$ is called the size of the alignment (the number of sequences).

Given an alignment, the totality of dissimilarity it contains can be expressed via a weighted simplicial complex. Such a complex subsumes said dissimilarity information but also contains extra non-local information that is geometric in nature. This is similar to how a weighted tree and a weighted graph can be metrically equated via UPGMA in the context of phylogeny, but differ in their cycle structure as geometric objects.

Definition 4 Let $W = [w_1, \ldots, w_n]$, $w_i \in \mathcal{A}_l$, be a given alignment and let $k \geq 0$ be fixed. A simplex of dimension $k$ over $W$ is a $k+1$-subset of $W$, $\sigma = \{w_0, \ldots, w_k\}$ of sequences from $W$, such that $d_{k+1}(w_0, \ldots, w_k) > 0$. We denote by $K_k(W)$ the set of all $k$-simplices over $W$, and set $X(W) = \bigcup_{k \geq 0} K_k(W)$. Let $R \supset \mathbb{Z}$ be a discrete valuation ring with uniformizer $\pi$. Let

$$v_W : X(W) \to R, \quad v_W(\sigma = [w_0, \ldots, w_k]) = \pi^{d_{k+1}(w_0, \ldots, w_k)}.$$ 

Then, $v_W$ is called the weight function associated to $X(W)$ and we call the pair $(X(W), v_W)$ the dissimilarity complex of $W$.

For all intents and purposes, the discrete valuation ring can be taken to be the ring of formal power series with rational coefficients and a transcendental variable we denote by $\pi$, namely $R = \mathbb{Q}[[\pi]]$. We use the powers of the transcendental $\pi$ to encode the integer weights of simplices as exponents. This is done since in what follows, the new kind of boundary operator that will be introduced which yields a weighted version of homology as per (Dawson 1990; Ren et al. 2018; Bura et al. 2021), requires certain divisibility conditions as we shall see in Sect. 3.2.

Given a dissimilarity complex, we can construct its geometrical realization by constructing the geometrical realizations at each $k$ dimension and then integrating them via gluing along common faces, see Ex 3.

Example 3 Let $W = [w_0 = AAGAGGCTT, w_1 = AGGAGACCT, w_2 = GTGGGTCCC]$. We can construct the geometrical realization of $(X(W), v_W)$ by first constructing its 0-simplices (see Fig. 7a), 1-simplices (see Fig. 7b) and 2-simplices (see Fig. 7c). The geometrical realization of $(X(W), v_W)$ is then obtained by integrating all $k$-simplices via gluing (see Fig. 7d).

An immediate motivation for this construction is that more information can be encoded when we lift from weighted graphs to the weighted complex structure, see Ex 4. We are now in position to identify information not present in the graph— which is only the so called 1-skeleton of the complex.
Fig. 7 The geometric realizations at different dimensions and their integration of $X([w_0 = AAGAGGCTT, w_1 = AGGAGACTCT, w_2 = GTGGGTCCC])$ with $K_0$ in a, $K_1$ in b, $K_2$ in c and the integration into $(X(W), v_W)$ in d.

Fig. 8 Dissimilarity complexes corresponding to $W_0 = [w_0^0 = TTCA, w_1^0 = CTCG, w_2^0 = TTTG]$ and $W_1 = [w_0^1 = CTCG, w_1^1 = TTAG, w_2^1 = ATTG]$, with their respective weights. Note that $X(W_0)$ is an “empty” triangle while $X(W_1)$ is “filled.”

Example 4 Consider the dissimilarity complexes associated to the alignments in Ex 2. For $W_0 = [w_0^0 = TTCA, w_1^0 = CTCG, w_2^0 = TTTG]$, we have $d_1(w_0^0) = d_1(w_2^0) = 4 \neq 0$, hence $K_0(W_0) = \{[w_0^0], [w_1^0], [w_2^0]\}$ and since $d_2(w_0^0, w_1^0) = d_2(w_0^0, w_2^0) = d_2(w_1^0, w_2^0) = 2 \neq 0$, we also have $K_1(W_0) = \{[w_0^0, w_1^0], [w_0^0, w_2^0], [w_1^0, w_2^0]\}$. Finally $d_3(w_0^0, w_1^0, w_2^0) = 0$, thus $K_2(W_0) = \emptyset$, see Fig. 8 (LHS). For $W_1 = [w_0^1 = CTCG, w_1^1 = TTAG, w_2^1 = ATTG]$ we have $K_0(W_1) \cong K_0(W_0)$ and $K_1(W_1) \cong K_1(W_0)$ however $d_3(w_0^1, w_1^1, w_2^1) = 2 \neq 0$ yields $[w_0^1, w_1^1, w_2^1] \in K_2(W_1)$, see Fig. 8 (RHS).

$(X(W), v_W)$ represents an augmentation of classical simplicial complexes and in particular a generalization of weighted graphs. In addition, $(X(W), v_W)$ has some nice combinatorial properties that facilitate the study of a weighted version of homology as detailed in Sect. 3.2.
Proposition 4 $X(W)$ is a simplicial complex that is bounded in dimension, namely, for any $\sigma \in X(W)$, $\dim(\sigma) \leq |A| - 1$.

Proposition 5 Let $\sigma \in X(W)$ be a $k$-simplex and let $\tau \subseteq \sigma \in X(W)$ be a $k'$ face of $\sigma$. Then we have $v_W(\sigma) | v_W(\tau)$.

Definition 5 Let now $\epsilon \in S_n$ be a permutation on $\{1, \ldots, n\}$. The ordered tuple $W_\epsilon := [w_{\epsilon(1)}, \ldots, w_{\epsilon(n)}]$, is called a row shuffle of $W$.

Let $\omega \in S_l$ be a permutation on $\{1, \ldots, l\}$. The ordered tuple $W^{\omega} := [w_1^{\omega}, \ldots, w_l^{\omega}]$, with $w_i^{\omega} = [f_{\omega^{-1}(1)}(w_i), \ldots, f_{\omega^{-1}(l)}(w_i)]$, for all $1 \leq i \leq l$, is called a column shuffle of $W$.

Proposition 6 For any alignment $W$ and any pair of row and column shuffles $(\epsilon, \omega) \in S_n \times S_l$, we have $(W_\epsilon)^\omega = (W^{\omega})_\epsilon$. Furthermore denoting $W^{\omega}_\epsilon := (W_\epsilon)^\omega = (W^{\omega})_\epsilon$ we have that, for any pair of row and column shuffles $(\epsilon, \omega) \in S_n \times S_l$,

$$(X(W), v_W) \cong (X(W^{\omega}_\epsilon), v_W^{\omega}_\epsilon).$$

I.e. any row or column shuffle of $W$ produces a weighted dissimilarity complex that, up to vertex relabelling, is the same as the original dissimilarity complex of $W$.

Proposition 5 shows that $(X(W), v_W)$ is amenable to construct weighted homology by means of a novel boundary operator, compatible with the weight function, see Sect. 3.2. Proposition 6 shows that the dissimilarity complex has nice symmetry properties.

### 3.2 Weighted homology

Passing from graphs to simplicial complexes not only provides us with the degree of freedom to encode additional information, but it also enables us to study a multiple sequence alignment from a novel mathematical perspective. Any simplicial complex gives rise to a topological space. Studying topological properties of the dissimilarity complex enhances our conceptual understanding of the multiple sequence alignment itself.

In algebraic topology, simplicial homology is a useful tool for the study of features of a simplicial complex. It comes about as a sequence of abelian groups $H_1, \ldots, H_n, \ldots$, one for each dimension, whose structures yield surprising information (invariants with respect to continuous deformations) about the space in question, such as the structure
of its $k$-dimensional holes and its orientability (geometric torsion). This information is of key relevance and to date, dynamically tracking the birth and death of generators of these homology groups is an integral part of topological data analysis in the guise of Persistent Homology (Zomorodian and Carlsson 2005).

The dissimilarity complex constitutes a simplicial complex with additional weight information. By coherently integrating this weight information into a new homology theory that mimics the classical case, we can study the augmented arithmetic version of its topology. In this case, its torsion encodes $k$-stance level information among the sequences, and thus we can gain more insight about the structure of the alignment the dissimilarity complex is modeling.

Given an alignment $W$, let $(X, v) = (X(W), v_W)$ be its corresponding weighted dissimilarity complex. Let $C_{n,R}(X)$ denote the free $R$-module generated by all $n$-simplices in $X$, with $R$ being the co-domain ring of the simplex weight function. Setting a simplicial ordering (Hatcher 2005), namely a linear order on the 0-simplices, we can now consider a simplex $\sigma$ as an ordered tuple of sequences instead of a set. This allows one to define

$$\partial^n_v : C_{n,R}(X) \to C_{n-1,R}(X), \quad \partial^n_v(\sigma) = \sum_{i=0}^n v(\hat{\sigma}_i) \cdot v(\sigma) \cdot (-1)^i \hat{\sigma}_i,$$

where the face $\hat{\sigma}_i \subset \sigma$ is obtained by dropping the $i$th position in $\sigma$. We have $v(\sigma)$ divides $v(\hat{\sigma}_i)$ and as a result, $\partial^n_v$ is a well defined $R$-module homomorphism. Note here that it is crucial that $v$ takes values in the DVR as defined. This is since then, by Proposition 5, the divisibility condition required in the definition of $\partial^n_v$ is contingent only on $d_{n-1}(\hat{\sigma}_i) \geq d_n(\sigma)$ and not the latter dividing the former.

Indeed, in Dawson (1990); Ren et al. (2018) a similar weighted perspective is considered. However in those instances, the weighted complexes considered must have weights constrained by the divisibility condition, such as for instance integers that divide each other, as otherwise the fraction coefficients obtained exit the integer ring. In our formulation, although $d_n$ takes integer values, passing to the DVR eschews the issue by encoding the integers in the powers of the transcendental $\pi$. As such, only the order relation on the integers is required, and not divisibility. The DVR setup guarantees that $\frac{v(\hat{\sigma}_{i,j})}{v(\sigma)} = \frac{v(\hat{\sigma}_{i,j})}{v(\sigma)} \cdot \frac{v(\hat{\sigma}_i)}{v(\sigma)}$, hence we obtain $\partial^n_{n-1}(\partial^n_v(\sigma)) = 0$.

In view of this, $\partial^n_v$ is indeed a boundary map and we can define a homology theory accordingly, $H^n_v(X) = \text{Ker}(\partial^n_v) / \text{Im}(\partial^n_{n+1})$ denoting the weighted homology modules of $(X, v)$. Furthermore, it is a well known result that the weighted homology modules are in fact independent of our initial choice of simplicial order (Hatcher 2005).

**Proposition 7** Let $W$ be an alignment over an alphabet $A$ and let $(X(W), v_W)$ be its corresponding weighted dissimilarity complex. Then $H^n_v(X(W)) = 0$, for all $k \geq |A|$.

**Proposition 8** Let $W$ be an alignment consisting of $n$ sequences of length $l$. For any pair $(\epsilon, \omega) \in S_n \times S_l$, and any $k \in \mathbb{N}$, we have $H^v_{\omega\epsilon}(X(W)) \cong H^v_{\omega\epsilon}(X(W_{\omega\epsilon}))$. 

 Springer
Fig. 9 The partition of a single column alignment $W$ into bins

Fig. 10 A single column alignment with only two bins and its dissimilarity complex, the complete bipartite graph $K_{4,3}$

4 The single column case

In this section we present a relation between the $k$-stances for different $k$ for alignments of length one. Furthermore we compute their weighted homology. To this end, let $W = [w_0, \ldots, w_{n-1}]$ be an alignment of length one and size $n$.

We can organize $W$ via bins $W = \bigcup_{i=1}^{s} B_i = \bigcup_{i=1}^{s} \{ w \in W | f_1(w) = a_i \} \neq \emptyset$ where $s$ is the number of distinct letters that appear in $W$’s column and we let $b_i = |B_i|$ for all $1 \leq i \leq s$ be the size of $B_i$, i.e. the multiplicity in $W$’s column of the letter $a_i$, see Fig. 9.

Given the bin partitioning of $W$, it is easy to see that any simplex $\sigma \in X(W)$ has weight $v(\sigma) = \pi^1 = \pi$. Furthermore $X(W)$ is a pure simplicial complex as all of its maximal simplices are of dimension $s - 1$. By construction, each $s - 1$-simplex is obtained by picking one sequence (0-simplex) from each of the $s$ bins. Therefore, $X(W)$ is a complete $k$-partite simplicial complex, which is a natural generalization of the complete bipartite graph. Note that in the case of $s = 2$, $X(W)$ is precisely the classical complete bipartite graph $K_{b_1,b_2}$, see Fig. 10.

Definition 6 The total $k$-stance contribution in $W$ is defined to be

$$c_k = \sum_{Y \subseteq W, |Y| = k} d_k(y_0, \ldots, y_{k-1}).$$
where the sum is taken over all size $k$ subsets $Y = \{y_0, \ldots, y_{k-1}\} \subset W$. We integrate this information over all $k$ into a polynomial in indeterminate $x$ called the dissimilarity polynomial of $W$

$$D_W(x) = x^s + \sum_{k=1}^{s} (-1)^k c_k x^{s-k}.$$  

**Theorem 9** Let $W$ be a single column alignment. The size of each bin of $W$ is a root of $W$’s dissimilarity polynomial, and this polynomial has no other roots.

**Example 5** Let $W$ be the single column alignment shown in Fig. 9. We have $b_1 = 3$, $b_2 = b_3 = 2$ and $b_4 = 1$. Furthermore, we have $c_1 = 8$, $c_2 = 23$, $c_3 = 28$ and $c_4 = 12$. Then $D_W(x) = x^4 - 8x^3 + 23x^2 - 28x + 12 = (x - 3)(x - 2)^2(x - 1)$.

We note here that for each $k > 1$, the coefficient $c_k$, in the dissimilarity polynomial for the column $W$, is a straightforward augmentation of sum of pairs type scores for nucleotide conservation, as defined in Valdar (2002). This augmentation is achieved by passing to sums of $k$-tuples. A natural score that integrates all such information is $D_W(1)$ which is easily computable as per Theorem 9.

Note that, in certain cases when conservation within the column is high, we can observe a higher sensitivity for higher order $k$-stances than for that of lower order ones. Consider the following example:

**Example 6** Let $W$ be an alignment of size 100 with two bins $|b_1| = 98$ and $|b_2| = 2$. Then $d_2(W) = 2 \times 98$ and $d_3(W) = 0$. Suppose now that a single nucleotide in bin $b_2$ mutates, creating the alignment $W'$ with bins $b'_1 = b_1$, $b'_2 \subset b_2$ and $|b'_2| = 1$ and lastly $|b'_3| = 1$ with the last bin containing the mutated nucleotide. Then, $d_2(W') = 2 \times 98 + 1$ while now $d_3(W') = 98$. As such the relative increase in 3-stance is much higher when compared to the one in 2-stance.

**Theorem 10** Let $W$ be a single column alignment, let $(X, v) := (X(W), v_W)$ be its corresponding weighted simplicial complex and denote by $b = \prod_{i=1}^{s} |b_i - 1|$. Then, all homology modules $H^v_k(X)$ are free and furthermore

- (a) $H^v_0(X) = R$,
- (b) $H^v_{s-1}(X) = R^b$
- (c) $H^v_k(X) = 0$ for any $k > 0$, $k \neq s - 1$.

**Example 7** Let $W$ be the single column alignment shown in Fig. 10. We have $b_1 = 4$ and $b_2 = 3$. Then $H^v_0(X) = R$, $H^v_1(X) = R^{(4-1)(3-1)} = R^6$ and $H^v_{n \geq 2}(X) = 0$.

## 5 Dissimilarity and $k$-stances of multi-column sequence alignments

For general alignments we have at present no analytical (closed form) expression connecting its $k$-stances and weighted homology modules in terms of the bins of its various columns. A way of piecing together column information inductively is
currently under investigation and the idea here would be to employ some version of Mayer-Vietoris sequences for weighted complexes. However, $k$-stance statistics as well as the modules of weighted homology can be computed, effectively. We have developed a framework for computing weighted homology and can provide a link to a free underlying software module created for this purpose (software module). In the following, we shall illustrate that both $k$-stances and weighted homology provide new insights into aligned genetic data and reveal biologically relevant features of said alignments.

In Sect. 5.1 we present case studies for SARS-CoV-2 and for H1N1, respectively, where $k$-stance signatures are seen to reflect distinct phases in the evolution of these pathogens in the human population. In Sect. 5.2 we illustrate connections between the structure of the weighted homology modules and the $k$-stances present in the alignments.

### 5.1 $k$-Stance statistics

In this Subsection we present two case studies that illuminate the usage of higher order $k$-stance statistics to infer biologically relevant information on viral population dynamics. In these case studies, $k$-stance statistics are computed over the alphabet \{A, C, G, T, −\}, where the gap symbol “−” was not accounted as a distinguished symbol for any of the contributions computed. Therefore by construction, $k = 4$ is the maximum possible $k$ with nontrivial $k$-stance. Here we focus on $k = 2, 3$ as the vast majority of the sites exhibit at most three different nucleotide types. In general, $k$-stances for larger $k$ values can still be of interest when the alphabet is sufficiently large, such as when considering amino acid sequences.

**Case study 1: SARS-CoV-2**

The multiple sequence alignment considered here is comprised of all SARS-CoV-2 genomes submitted to GISAID (Shu and McCauley 2017) prior to 2021-01-11. This amounts to 254,148 sequences, each exhibiting 29,903 aligned sites. For each site, we computed its total 2-stance and 3-stance contribution respectively (i.e. the total number of pairs and the total number of triplets that are mutually distinct). We now partition the set of logarithms of these numbers (shifted by 1 for technical reasons) into 100 bins of the same width and plot their corresponding histograms (bin vs. frequency), see Fig. 11.

The 2-stance and the 3-stance exhibit distinctive distributions. Firstly, the 2-stance contains approx. 6000 sites in the zero-th bin while the 3-stance contains approx. 20,000 in the zero-th order bin. This is due to the fact that approx. 14,000 sites exhibit bi-allelic but not tri-allelic SNPs. This should be expected as it is an instance of the observation preceding Theorem 10, regarding sensitivity of higher order $k$-stances in highly conserved columns. In any other bin the two distributions also differ distinctively. The 2-stance exhibits a sharper absolute decay in frequency when compared to the 3-stance frequencies, as the latter are approximately an order of magnitude lower. The Pearson’s coefficients of dispersion (Pearson 1920) for the two distributions are 0.49 for 2-stance and 1.24 for 3-stance. Having a closer look at the polymorphic site 23012 mentioned in the Introduction, corresponding to the
E484Q and E484K mutations, we find rank 1028 for 2-stance and rank 327 for 3-stance. Note that in the aggregated population, more than 99.88% of the sequences have nucleotide G at site 23012. Such a high degree of conservation results in a relative low 2-stance rank. On the other hand, the remaining 0.12% sequences exhibit two types of nucleotides: A and C. The existence of tri-allelic SNPs results in a more prominent 3-stance rank in comparison to its 2-stance rank.

We proceed by investigating sites with high $k$-stance rank for $k = 2, 3$. For each $k$, we collect the top 100 sites in $k$-stance. The 2-stance collected sites have 40 sites in common with their 3-stance counterparts, see Fig. 12. When we restrict focus to the Spike protein region of the SARS-CoV-2 genome, we observe that 13 of the top 2-stance collected sites are located on the spike protein region with 6 of which being among the top 3-stance sites. In addition, there are 10 Spike protein sites that are among the top 3-stance collected sites but are not found in the 2-stance collected cohort. Cross referencing these 10 sites with characteristic mutations of VoCs, we found that 4 of them actually correspond to characteristic mutations of VoCs, see Table 1.

Case study 2: H1N1

We study 2-stances and 3-stances within a sliding window of 100 sequences across a temporally ordered alignment of GISAID H1N1 flu data from 2009 to 2018. The
**Table 1** Highlighted spike protein sites that are captured by 3-stance but not 2-stance and their correspondence to characteristic mutations of certain VoCs

| Site  | Amino acid mutation | VoC   |
|-------|---------------------|-------|
| 21974 | D138Y               | Gamma |
| 22205 | D215G               | Beta  |
| 23593 | Q677H               | Eta   |
| 23604 | P681H               | Alpha |

**Fig. 13** Time evolution of the ensembles of 2-stance and 3-stance in a sliding window of 100 sequences across a temporally ordered alignment of H1N1 flu data from 2009 to 2018

The **y-axis** represents the sum over all possible \( k \)-stances \((k = 2, 3)\) which we refer to as the ensemble of 2-stance and 3-stance, respectively, for each window as time evolves in the **x-axis**, see Fig. 13.

2- and 3-stances capture the two outbreaks (Jan 2009 and Nov 2013) of the virus and we speculate that the peaks in this dissimilarity signal appear due to the virus’ genetic variation being elevated as it explores its fitness landscape. However, note that the Apr 2016 flu season that exhibited a change of the dominant strain is not captured by 2- but 3-stances. We stipulate that this is the case because 3-stance exhibits higher conservation decreases in the column when compared to 2-stance as discussed in Ex 6.

### 5.2 Multi column weighted homology

Li and Reidys (2022) provides structure theorems for the weighted homologies of arbitrary weighted complexes (not necessarily arising from dissimilarity) relating simplicial homology with coefficients in certain valuation rings to weighted simplicial homology. The idea being here is to create a homomorphic image of the “known” homology into the “unknown” homology and then to study the quotient via homological algebra. The concepts developed in the process suggest employing a version of Nakayama’s Lemma (Nakayama 1951) to reduce the coefficients controlling this
quotient down to rational numbers. This enables very fast computation of all weighted homology modules (software module).

The weighted homology modules of the dissimilarity complex exhibit non-trivial torsion, which genuinely stems from \( k \)-stances and reflects interesting features about the structure of the alignment itself. We present two pertinent examples that allude to this fact:

**Example 8** Consider the alignment \( W = [w_0 = AGCTTT, w_1 = ATTCAA, w_2 = AGCCAA] \). Firstly, we have \( d_1(w_0) = d_1(w_1) = d_1(w_2) = 6 \). Then we have \( d_2(w_0, w_1) = 5 \), \( d_2(w_0, w_2) = 3 \) and \( d_2(w_1, w_2) = 2 \). Finally, we have \( d_3(w_0, w_1, w_2) = 0 \). Since the maximum dimension of \( X(W) \) is one, we have two nontrivial weighted homology modules, namely, \( H^1_1(X(W)) = R \) and \( H^0_0(X(W)) = R \oplus R/\pi \oplus R/\pi^3 \). Note that \( H^1_1(X(W)) \) has free rank one, and this is due to the fact that \( w_2 \) is a linear recombinant of \( w_0 \) and \( w_1 \), namely \( w_2 = f_1(w_0)f_2(w_0)f_3(w_0)f_4(w_1)f_5(w_1)f_6(w_1) \). Furthermore, \( H^0_0(X(W)) \) has free rank one and two torsion components. The first torsion component \( R/\pi = R/\pi^{(6−5)} \) corresponds to the largest 2-stance among the three sequences and the second torsion component \( R/\pi = R/\pi^{(6−3)} \) corresponds to the second largest 2-stance among the three sequences.

**Example 9** Let \( W = [w_0 = AAGAGGCTT, w_1 = AGGAGACCT, w_2 = GTGGGTCCC] \). Firstly, we have \( d_1(w_0) = d_1(w_1) = d_1(w_2) = 9 \). Then we have \( d_2(w_0, w_1) = 3 \), \( d_2(w_0, w_2) = 6 \) and \( d_2(w_1, w_2) = 5 \). Finally, we have \( d_3(w_0, w_1, w_2) = 2 \). Since the maximum dimension of \( X(W) \) is two, we have \( H^1_1(X(W)) = 0 \), for all \( k \geq 3 \). In fact, we have \( H^1_2(X(W)) = 0 \), \( H^1_3(X(W)) = R/\pi \) and \( H^0_0(X(W)) = R \oplus R/\pi^3 \oplus R/\pi^4 \). Since \( H^1_0(X(W)) = R/\pi \) are full torsion, none of \( w_0 \), \( w_1 \) or \( w_2 \) is a linear recombinant of the remaining two. Furthermore, the torsion \( R/\pi = R/\pi^{(3−2)} \) corresponds to the difference between the 3-stance and the minimum pairwise 2-stance among the three sequences.

In the following, we present a theoretical application that motivates the homological framework approach: the structure of certain degenerate primers.

**Definition 7** Let \( A = \{a_1, \ldots, a_s\} \) be a finite alphabet, for instance \( A = \{a, c, g, t\} \). A primer \( P \) of length \( k \) is a string of \( k \) nonempty subsets of \( A \), namely, \( P = p_1 \ldots p_k \), where \( p_i \subseteq A \). Each \( p_i \) represents the potential nucleotides that could appear at position \( i \) in the primer \( P \). The degeneracy of \( P \) is defined to be

\[
d(P) = \prod_{i=1}^{k} |p_i|.
\]

One key procedure in PCR is to design efficient and specific primer sequences. A primer sequence can form a double-stranded initiation site with the target sequence at which the polymerase can bind. In real applications, there might be more than one target sequence. In a typical situation, one has a collection of related target sequences, for instance, DNA sequences of homologous genes, and the goal is to design primers...
that will match as many of them as possible. In this case, a single unique primer is typically not sufficient. Instead, one might consider designing a degenerate primer. On the one hand, a primer with high degeneracy is likely to match more target sequences, while on the other hand, high degeneracy may lead to reduced specificity, namely the primer is more likely to bind to wrong binding sites. Therefore, a key task of the degenerate primer design problem is to find a balance between degeneracy and the number of matched target sequences. Despite the already established hardness for the general problem, a topological classification can often shed light on the computational complexity of the class and reveal sub-classes with low topological complexity corresponding to tractable instances. Here we provide a structure theorem that can hint at said complexity.

A primer $P$ can be interpreted as an alignment consisting of the $d(P)$ distinct sequences $P$ can produce. As we are only interested in $X(P)$, by Proposition 6, w.l.o.g. we can assume the columns of the alignment $P$ are sorted by their respective multiplicities, namely, $|p_i| \geq |p_j|$, for any $i < j$. In other words, $P$ can be organised as a concatenation of blocks of columns, where each block consists of columns with the same multiplicity and the blocks are arranged from left to right in decreasing order of their respective multiplicities.

**Definition 8** For a given primer $P$, let $X^{\geq m}(P) \subseteq X(P)$ denote the subcomplex generated by all the columns with multiplicity at least $m$

$$X^{\geq m}(P) := \bigcup_{|p_j| \geq m} X(P_j),$$

where $P_j$ is the $j$th column of the primer alignment $P$.

Since, columns with $|p_j| = 1$ will not generate non-zero simplices in $X(P)$, $X^{\geq 2}(P) = X(P)$, for any primer $P$ with $|p_1| \geq 2$. As such we can w.l.o.g. consider only primers $P$, such that $|p_j| \geq 2$, for all $1 \leq j \leq l$.

**Theorem 11** Let $P$ be a length $l > 1$ primer that contains a single column with maximum multiplicity $s$. Namely, $|p_1| = s$ and $|p_1| > |p_j|$, for all $2 \leq j \leq l$. Then for the unweighted dissimilarity complex of $P$ (i.e. $v(\sigma) := 1_R$) $H_k(X(P)) = 0$, for all $k \geq s$, and $H_{s-1}(X(P)) = R^b$, where

$$b = \left( -1 + \prod_{2 \leq j \leq l} |p_j| \right)^s.$$

### 6 Discussion

In this paper we introduce the notion of higher order dissimilarities, naturally generalizing the concept of Hamming distance. We have shown that such dissimilarities emerge within alignments of viral sequences and that these are not independent of each other. In fact we give explicit formulae for these dependencies in specific instances.
We can thus conclude that, in case of genetic sequences and the underlying four letter alphabets, there is more information than is reflected by Hamming distance alone by considering 3- and 4-stances. It is therefore noteworthy that all the information we currently obtain is based on or derived from Hamming distance.

We then provide a mathematical context for these hyper-distances by means of the dissimilarity complex. Here $k$-stances manifest as weights of certain simplices. To be concrete, simplices are comprised of $k$-sequences that exhibit in at least one site a $k$th order polymorphism and the weight of this simplex is the actual number of the sites exhibiting such $k$th order polymorphisms. The weighted complex homology can be readily computed via weighted homology (Li and Reidys 2022) and in case of sequence alignments, an inductive computation by means of patching the complex column by column that is based on the single column case—which we compute here—is currently under investigation.

As for future work, along the lines of constructing the phylogenetic tree within a complete weighted graph of an alignment, we work on constructing a “tree-analogue” within the Dissimilarity Complex of the given alignment. This “phylogenetic complex” should generalize the well known phylogenetic tree. The central idea of distance based phylogeny is to construct a tree of sequences whose tree metric “best approximates” the complete graph on sequences weighted by their pairwise distances. To mimic this process, construct the phylogenetic complex, whose induced dissimilarity “best approximates” that of the dissimilarity complex of the given alignment of sequences. It is natural then to ask what sort of properties such a derived object should possess:

A tree is a maximally acyclic spanning graph of the complete graph over the set of its vertices. Tree edges form a maximally independent set, namely including any edge that connects two vertices we enclose a unique cycle. The distance between two such vertices can be approximated by the sum of involved tree edge lengths. As a higher dimensional analogue of the phylogenetic tree, the phylogenetic complex should also exhibit such “tree-like” properties. Firstly, the phylogenetic complex should contain no (homological) cycles at any simplicial dimension. Secondly, the phylogenetic complex should be able to approximate the weight of any simplex that appears in the dissimilarity complex but not in the phylogenetic one, just as the phylogenetic tree distance between two vertices, approximates the weight of the complete graph edge that closes the cycle between them in the tree.

The phylogenetic complex will inevitably include higher order, pseudometric information that arises as a result of an optimization process that is fundamentally different from clustering. This is clear since the very notion of clustering is based on pairwise relations. There are two mechanisms that we are currently investigating as construction principles for this object:

Higher order “clustering”: one approach to construct a phylogenetic tree based on a pairwise distance matrix is by recursive clustering. At each step, small clusters are combined into a larger one with the distance matrix being updated to reflect the new inter-cluster distances. The selection of which clusters to merge is typically based on a local criterion. To adapt this clustering approach to the construction of the phylogenetic complex, one key challenge is the design of a higher dimensional analogue of clustering. One potential way to mitigate this challenge is to generalize the split decomposition presented in Bandelt and Dress (1992). That is since, a split can be viewed as
a minimal clustering result and any phylogenetic tree can thus be decomposed into a set of compatible splits.

Combinatorial optimization: an alternate approach to constructing a phylogenetic tree based on pairwise distance information is by combinatorial optimization. A tree is selected based on certain global criteria such as least square error or minimum evolution. For example, one can select the tree that minimizes the total tree length, as in the case of the “Steiner tree”. In this case the estimation of the length of each edge can be generally dealt with within the least-squares framework. To adapt this approach to the construction of the phylogenetic complex, one key step is the definition of an objective function that incorporates higher dimensional dissimilarity information, beyond pairwise distance errors.

7 Proofs

Proposition 1.

Proof Implication of indiscernibles: if \( w_i = w_{i'} \) for some \( 0 \leq i < i' \leq k - 1 \), then \( f_j(w_i) = f_j(w_{i'}) \) for any \( j \in \{1, \ldots, l\} \) and the claim follows by definition of \( d_k \).

Symmetry: since \( \bigcup_{i=0}^{k-1} \{f_j(w_i)\} = \bigcup_{i=0}^{k-1} \{f_j(w_{\epsilon(i)})\} \) for any \( \epsilon \in S_k \), the claim follows by definition of \( d_k \).

Polyhedron inequality: Let \( I : A^k \to \{0, 1\} \) be an indicator function for which \( I(f_j(w_0), \ldots, f_j(w_{k-1})) = 1 \) if \( |\bigcup_{i=0}^{k-1} \{f_j(w_i)\}| = k \) and \( I(f_j(w_0), \ldots, f_j(w_{k-1})) = 0 \) otherwise. We note then that

\[
d_k(w_0, \ldots, w_{k-1}) = \sum_{j=1}^{l} I(f_j(w_0), \ldots, f_j(w_{k-1})).
\]

It suffices then to show that \( I \) satisfies the polyhedron inequality. Furthermore, since \( I \) is always non-negative, it suffices to only consider the case \( I(f_j(w_0), \ldots, f_j(w_{k-1})) = 1 \). If \( f_j(w_k) \neq f_j(w_i) \) for all \( 0 \leq i \leq k - 1 \), then \( I(f_j(w_0), \ldots, f_j(w_i), \ldots, f_j(w_{k-1})) = 1 \) for any \( 0 \leq i \leq k - 1 \). In this case the polyhedron inequality holds for \( I \). The other possibility is that \( f_j(w_k) = f_j(w_{i^*}) \) for some distinguished \( 0 \leq i^* \leq k - 1 \). But then, \( I(f_j(w_0), \ldots, f_j(w_{i^*}), \ldots, f_j(w_{k})) = 1 \) which still implies the claim for \( I \), completing the proof. \( \square \)

Proposition 2.

Proof Dimensional bounding: if \( k > |A| \) then for any \( j \in \{1, \ldots, l\} \) we must have \( |\bigcup_{i=0}^{k-1} \{f_j(w_i)\}| \leq |A| < k \) and the claim follows by definition of \( d_k \).

Dimensional monotonicity: Fixing \( 1 \leq k' \leq k \), by Proposition 1 (symmetry), it suffices to prove the claim for \( \iota = \text{id}|_{\{0, \ldots, k'-1\}} \). Namely, we want to show

\[
d_k(w_0, w_1, \ldots, w_{k-1}) \leq d_{k'}(w_0, w_1, \ldots, w_{k'-1}).
\]
This however follows immediately from the definition of $d_k$ by observing that for any $j \in \{1, \ldots, l\}$ for which $|\bigcup_{i=0}^{k-1} f_j(w_i)| = k$ we must in turn have $|\bigcup_{i=0}^{k'-1} f_j(w_i)| = k'$.

Proposition 3.

Proof It suffices to note that if $w \in W$ is a linear recombinant of $W \setminus \{w\}$ then by definition, for each $j \in \{1, \ldots, l\}$ there exists an $i \in \{0, \ldots, k-1\}$ such that $f_j(w) = f_j(w_i)$. This means that for each $j \in \{1, \ldots, l\}$ we must have $|\bigcup_{i=0}^{k-1} f_j(w_i)| \leq k - 1 < k$ and as such $d_k(w_0, \ldots, w_{k-1}) = 0$ as claimed.

Proposition 4.

Proof This is an immediate consequence of Proposition 2 (dimensional bounding), for any $k > |A|$.

Proposition 5.

Proof Again, by Proposition 2 (dimensional monotonicity), we have $d_{k+1}(\sigma) \leq d_{k'+1}(\tau)$ and so immediately $v_W(\sigma) = \pi^{d_{k+1}(\sigma)}|\pi^{d_{k'+1}(\tau)} = v_W(\tau)$.

Proposition 6.

Proof The first claim, $(W_\epsilon)_{\omega} = (W_{\omega})_\epsilon$, follows immediately by observing the commutative identity for each entry in the alignment matrix of $W$

\[
\begin{align*}
  f_j(w_i) &\xrightarrow{\omega} f_{\omega(j)}(w_i) \\
  \downarrow &\quad \quad \downarrow \\
  f_j(w_{\epsilon(i)}) &\xrightarrow{\omega} f_{\omega(j)}(w_{\epsilon(i)})
\end{align*}
\]

For the second claim it suffices to investigate the action a row and column shuffle pair $(\epsilon, \omega)$ has on a fixed $k - 1$-simplex $\sigma = \{w_0, \ldots, w_{k-1}\} \in X(W)$.

Note first that $d_k(\epsilon.\sigma) = d_k(\sigma)$ by construction, where $\epsilon.\sigma \in W_\epsilon$ is the simplex in $W_\epsilon$ corresponding to $\sigma$. On the other hand, for $I$ the indicator function in the proof of Proposition 1, we have that

\[
d_k(\sigma) = \sum_{j=1}^{l} I(f_j(w_0), \ldots, f_j(w_{k-1}))
= \sum_{j=1}^{l} I(f_{\omega(j)}(w_0), \ldots, f_{\omega(j)}(w_{k-1})) = d_k(\omega.\sigma).
\]

By the previous claim, the order in which we apply the two actions does not matter, and as such $W_{\sigma_\epsilon}^{\omega} \supseteq \sigma_\epsilon^{\omega} = \omega.\epsilon.\sigma$ with $d_k(\sigma_\epsilon^{\omega}) = d_k(\sigma)$ and the proposition follows.

Proposition 7.

Proof This is an immediate consequence of Proposition 4 which bounds the dimensionality of the complex.
Proposition 8.

**Proof** This follows immediately from Proposition 6 which homeomorphically equates the dissimilarity complex of an alignment with its row column shuffle and shows that the arithmetic weight information is preserved under such a transformation. □

Theorem 9.

**Proof** It suffices to examine the polynomial \( P_W(x) = \prod_{i=1}^{s} (x - b_i) \) where \( b_i = |B_i| \) and \( s \) is the number of distinct letters that appear in \( W \)'s column. Vieta’s formulae for \( P_W(x) \) yield, for each \( 0 \leq k \leq s \),

\[
\sum_{1 \leq i_0 < i_1 < \cdots < i_k \leq s} \left( \prod_{j=0}^{k} b_{i_j} \right) = (-1)^k \frac{p_{s-k}}{p_s}
\]

where \( p_q \) is the coefficient of the term \( x^q \) in \( P_W(x) \) for \( 0 \leq q \leq s \). In this particular case \( p_s = 1 \). If we showed that \( p_{s-k} = c_k \) for any \( 0 \leq k \leq s \) then \( P_W(x) = D_W(x) \) and the theorem would follow. To this end, for \( I \) the indicator function in the proof of Proposition 1, we can write

\[
c_k = \sum_{\sigma \in K_k(W)} I(\sigma) = |K_k(W)|.
\]

To construct a simplex in \( K_k(W) \) it suffices to select \( k + 1 \) bins and select one sequence from each bin. As such, the theorem follows from

\[
|K_k(W)| = \sum_{1 \leq i_0 < i_1 < \cdots < i_k \leq s} \left( \prod_{j=0}^{k} b_{i_j} \right) = p_{s-k}.
\]

□

Theorem 10.

Note that this Theorem is equivalent to Theorem 4 in Bolker (1976). The original proof was based on simplicial joins and a Mayer-Vietoris type sequence. Here we present an alternate, more combinatorial proof of the claims.

We note that our alternate proof consists in developing a novel, explicit combinatorial construction of the relevant generators for homology at the \( s-1 \) dimension which could prove useful in tackling the multi-column case. Furthermore the argument for item c), contains an imbricated induction technique on bins and dimensions, that we believe provides hints as to similar results on the multi-column closed form case.

Before we present our proof, let us fix a simplicial ordering. Without the loss of generality, we can pick a simplicial order that is compatible with the bin ordering, see Fig. 14a. Then for any maximal simplex \( \sigma = [x_1, \ldots, x_s] \), we have \( x_i \in B_i \), and a simplex is now considered an ordered tuple.

Springer
Proof Item (a) Since there exist two different sequences in \( W, X \) is connected, and \( H_0(X) = \mathbb{Z} \).

Item (b) Since all simplices in \( X(W) \) have weight \( \pi \), we have \( \partial^i_k(\sigma) = \sum_{i=0}^{k} (-1)^i \frac{\pi}{k!} \sigma_i = \sum_{i=0}^{k} (-1)^i \sigma_i \).

It suffices to find a set of \( R \)-linearly independent set of generators for \( \text{Ker}(\partial^s_0) = H^s_0(X) \) of size \( b \). By construction, \( C_{s-1,R}(X) = \langle M \rangle_R \neq 0 \). Fix \( \sigma_0 = [x_1, \ldots, x_s] \in M \) and consider

Case 1 there exists no other simplex \( \sigma' = [y_1, \ldots, y_s] \in M \) such that \( y_i \neq x_i \) for all \( 1 \leq i \leq s \). In this case, at least one bin has size 1 and thus we have \( b = 0 \).

Claim \( H^s_{s-1}(X) = 0 \).

To prove this, consider \( c \in \text{Ker}(\partial^s_0) \) and grade it by \( |\sigma_0 \cap \sigma| \), denoting the number of vertices the two maximal simplices share. In this case, the grading starts at \( |\sigma_0 \cap \sigma| = 1 \)

\[
c = \sum_{k=1}^{s} \sum_{|\sigma_0 \cap \sigma| = k} a_{\sigma} \sigma.
\]

Let \( \sigma^* \in c \) satisfy \( |\sigma_0 \cap \sigma^*| = 1 \). Then \( \sigma^* = [y_1, \ldots, x_i, \ldots, y_s] \), for some \( 1 \leq i \leq s \), while \( y_j \neq x_j \) for all \( j \neq i \). Let \( \hat{\sigma}_i^* = [y_1, \ldots, \hat{x}_i, \ldots, y_s] \). Consider all possible \( \sigma^{**} \in M \) with \( |\sigma^{**} \cap \sigma_0| \geq 1 \), such that \( \hat{\sigma}_i^* = \sigma^{**} \subset \sigma^{**} \). Since \( |\sigma^{**} \cap \sigma_0| \geq 1 \) and \( |\sigma^*_i \cap \sigma_0| = 0 \), we must have \( \sigma^{**} = \sigma^* \). Namely, \( \sigma^* \) is the only simplex in \( M \) that contains \( \sigma^*_i \) as a face. Then \( \partial^s_1(c) = 0 \implies a_{\sigma^*} = 0 \). This holds independently for all \( \sigma^* \) with \( |\sigma_0 \cap \sigma^*| = 1 \). Therefore we have

\[
c = \sum_{k=2}^{s} \sum_{|\sigma_0 \cap \sigma| = k} a_{\sigma} \sigma.
\]

We proceed similarly for each \( k \geq 2 \) in order, which eventually leads to \( c = 0 \).

Case 2 there exist at least one simplex \( \sigma' = [y_1, \ldots, y_s] \in M \) such that \( y_i \neq x_i \) for all \( 1 \leq i \leq s \). In this case, each bin must contain at least 2 vertices, see Fig. 14b. Let

\[
L(\sigma') := \{[z_1, \ldots, z_s] \in M | z_j = x_j \text{ or } z_j = y_j, \text{ for all } 1 \leq j \leq s \},
\]

with \( x_i \) or \( y_i \) appearing at the same coordinate since they are chosen from the same bin and the 0-simplices follow an order that is compatible with the bin order, see Fig. 14c. We make the Ansatz

\[
\beta = \|L(\sigma')\| := \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} |\sigma^*| \in M, |\sigma_0 \cap \sigma'| = 0),
\]

noting that \( |\beta| = \prod_{i=1}^{s} ||B_i| - 1| = b \).
Claim $\beta$ is a $R$-basis for $\text{Ker}(\partial_v^{s-1})$.

A fixed $\sigma'$ only ever appears in $l(\sigma')$, therefore $\beta$ is a $R$-linearly independent set.

We next show $\beta \subset \text{Ker}(\partial_v^{s-1})$. Fix $l(\sigma') \in \beta$ and consider

$$\partial_v^{s-1}(l(\sigma')) = \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} \partial_v^{s-1}(\sigma) = \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} \sum_i (-1)^i \hat{\sigma}_i.$$  

Note that $\hat{\sigma}_i = [z_1, \ldots, \hat{z}_i, \ldots, z_s]$ appears in exactly two images,

$$(-1)^{|\sigma_0 \cap \sigma|} \partial_v^{s-1}(\sigma^*) \text{ and } (-1)^{|\sigma_0 \cap \sigma^{**}|} \partial_v^{s-1}(\sigma^{**}),$$

where $\sigma^* = [z_1, \ldots, x_i, \ldots, z_s]$ and $\sigma^{**} = [z_1, \ldots, y_i, \ldots, z_s]$, and it does so with opposite signs, hence $l(\sigma') \in \text{Ker}(\partial_v^{s-1})$, see Fig. 14d.

Finally, we show $\langle \beta \rangle_R = \text{Ker}(\partial_v^{s-1})$. Consider

$$\text{Ker}(\partial_v^{s-1}) \ni c = \sum_{\sigma \in K_{s-1}(W)} a_{\sigma} \sigma$$

$$= \sum_{k=0}^{s} \sum_{|\sigma_0 \cap \sigma| = k} a_{\sigma} \sigma = \sum_{|\sigma_0 \cap \sigma'| = 0} a_{\sigma'} \sigma' + \sum_{k=1}^{s} \sum_{|\sigma_0 \cap \sigma| = k} a_{\sigma} \sigma.$$  

Claim: $c = \sum_{|\sigma_0 \cap \sigma'| = 0} a_{\sigma'} l(\sigma').$

Let

$$\text{Ker}(\partial_v^{s-1}) \ni c^* = c - \sum_{|\sigma_0 \cap \sigma'| = 0} a_{\sigma'} l(\sigma').$$

By construction, the coefficient of $\sigma' \in c^*$, with $|\sigma' \cap \sigma_0| = 0$, is 0 hence

$$c^* = \sum_{k=1}^{s} \sum_{|\sigma_0 \cap \sigma| = k} a_{\sigma}^* \sigma.$$  

Iterating Case 1, yields $c^* = 0$.

Item (c)

Clearly $H_{k>s-1}(X) = 0$. It remains to show $H_{1 \leq k \leq s-2}(X) = 0$. We will prove this by induction (called the outer induction) on $s \geq 3$ the number of bins. To this end, first fix $s > 3$ and assume the outer induction hypothesis that, for any complex $\tilde{X}$ over $s-1$ bins we have already shown that $H_{1 \leq k \leq s-3}(\tilde{X}) = 0$. We wish to prove this is still the case for any complex $X$ over $s$ bins. To achieve this, we proceed by another layer of induction (called the inner induction) on the number of 0-simplices (sequences) of $X$-spaces over exactly $s$ bins. Note that, $X := X(W) \leq X' := X(W \cup \{x\})$, i.e. adding one more sequence $x$ to $W$ is tantamount to adding a vertex to the complex over $W$.
with its respective accompanying simplices and, in terms of bins, \( x \) will be added to an already existing bin which we will denote by \( B_x \). We have the following relative sequence

\[
\cdots \rightarrow H_{k+1}(X', X) \rightarrow H_k(X) \rightarrow H_k(X') \rightarrow H_k(X', X) \rightarrow \cdots.
\]

We assume the inner induction hypothesis that \( H_k(X) = 0 \) for any \( 1 \leq k \leq s - 2 \), and we want to show that \( H_{2 \leq k \leq s-2}(X', X) = 0 \), which will yield \( H_{2 \leq k \leq s-2}(X') \cong H_{2 \leq k \leq s-2}(X) = 0 \), and where \( H_1(X', X) = 0 \) will have to be proved separately. To this end, fix a \( 2 \leq k \leq s - 2 \), and let

\[
c = \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i [x_i, x] \in \text{Ker}(\partial_k)
\]

where \( x_i \in K_{k-1}(X) \), and \([x_i, x]\) is the tuple that contains the vertices in \( x_i \) with the vertex \( x \) appended at the end as per a simplicial ordering we fix on \( W \) where \( x \succ w \) for any \( w \in W \). Then

\[
0 = \partial_k(c) = \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i \partial_k([x_i, x]) = \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i (-1)^k [x_i] + \sum_{i=1}^{\left| K_{k-1}(W) \right|} \sum_{j=0}^{k-1} a_j^s [\hat{x}_{ij}, x]
\]

implies that the coefficient, in the second term above, of any \([\hat{x}_{ij}, x]\) must be zero. As such, since \([\hat{x}_{ij}, x]\) is tantamount to the face \([\hat{x}_{ij}]\) of \([x_i] \in K_{k-1}(W)\) with an extra \( x \)-label appended at the end, we have

\[
\partial_{k-1} \left( \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i [x_i] \right) = 0.
\]

**Claim** there exists \( c' = \sum_{i=1}^{\left| K_k(W) \right|} a_i [y_i] \in C_k(X) \) such that for any of the simplices \([y_i]\) we have the property that \( y_i \neq x' \) for any \( x' \in B_x \), and furthermore \( c' \) is such that \( \partial_k(c') = \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i [x_i] \). To show this, it suffices to first note that no \([x_i]\) simplex contains elements of \( B_x \) and as such, \( \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i [x_i] \in C_{k-1}(\hat{X}) \) for \( \hat{X} := X(W \setminus B_x) \) a space on \( s - 1 \) bins. By the outer induction hypothesis for the space \( \hat{X} \), we have \( H_{k-1}(\hat{X}) = 0 \) and so there exists \( c' \in C_k(\hat{X}) \) such that \( \partial_k(c') = \sum_{i=1}^{\left| K_{k-1}(W) \right|} a_i [x_i] \) and the Claim follows.

In view of the Claim, we can construct \( c^* = \sum_{i=1}^{\left| K_k(W) \right|} a_i [y_i, x] \in C_{k+1}(X) \), for which it is clear that \( \partial_{k+1}(c^*) = c \) and so \( H_{2 \leq k \leq s-2}(X', X) = 0 \) follows. It remains now to prove \( H_1(X', X) = 0 \).

Since \( x \) is the largest 0-simplex, any \( c = \sum_y \alpha_y [y, x] \in \text{Ker}(\partial_1) \) must have \( \sum_y \alpha_y = 0 \), where \( y \notin B_3 \). Thus, each edge with positive sign can be paired with an
edge with negative sign and so \( c = \sum \{ [y, x] - [y', x] \} \), with \( y \neq y' \) where w.l.o.g. we can assume \( y <_A y' \). For each \( \{ [y, x] - [y', x] \} \), if \( y \) and \( y' \) belong to different bins, then \([y, x] - [y', x] = \partial_2([y, y', x]) \). If however \( y \) and \( y' \) belong to the same bin, then since \( s \geq 3 \), there exists a bin \( Q \) such that \( x, y, y' \notin Q \). We can select \( z \in Q \) such that \([y, x] - [y', x] = [y, x] - [z, x] + [z, x] - [y', x] = \partial_2([y, z, x]) - \partial_2([y', z, x]). \) Thus, \( c \in \text{Im} (\partial_2) \), and hence \( H_1(X', X) = 0 \).

In view of the inner induction hypothesis and the long exact sequence, the above concludes the proof for the inner induction step. The inner induction base case is \( W = A \), which makes \( X(W) \) an \( s-1 \) simplex, for which it is trivial that \( H_1 \leq k \leq s-2(X) = 0 \). This concludes the proof for the outer induction step.

It remains to show the outer induction base case, \( s = 3 \)—i.e. for any space \( X \) over three bins, \( H_1(X) = 0 \). To this end, fix a simplicial ordering such that, for any \( x \in B_3 \), if \( x \in \sigma \), \( x \) is the last coordinate of \( \sigma \). We proceed by induction on \( m := |B_3| \). First, the base case \( m = 1 \). Let \( \tilde{X} := X(B_1 \cup B_2) \leq X := X(B_1 \cup B_2 \cup \{x\}) \).

Claim: \( i_* (H_1(\tilde{X})) = 0 \). There are two distinct cases:

(a) \( H_1(\tilde{X}) = 0 \). Then immediately \( i_* (H_1(\tilde{X})) = 0 \).

(b) \( H_1(\tilde{X}) \neq 0 \). By previous arguments \( H_1(\tilde{X}) = \langle \{l(\sigma')\} \rangle \), where \( l(\sigma') = \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} |\sigma, x] \in C_2(X), \)

obtained by appending \( x \) to each term in \( l(\sigma') \). Then \( \partial_2([l(\sigma'), x]) = -l(\sigma') + \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} \sum_{i=1}^2 (-1)^{i} [\hat{\sigma}_i, x]. \)

Note that, \( 0 = \partial_1 (l(\sigma')) = \sum_{\sigma \in L(\sigma')} (-1)^{|\sigma_0 \cap \sigma|} \sum_{i=1}^2 (-1)^{i} \hat{\sigma}_i, \)

which yields \( \partial_2([l(\sigma'), x]) = -l(\sigma') \). In other words, \( l(\sigma') \in \text{Im} (\partial_2(X)) \) and as such \( i_* (l(\sigma')) = 0 \), and so \( i_* (H_1(\tilde{X})) = 0 \). Since \( s = 3 \), by an identical argument to the distinguished case of the inner inductive step, we immediately conclude \( H_1(X, \tilde{X}) = 0 \). By exactness of the \((X, \tilde{X})\) relative homology sequence and the Claim, \( H_1(X) = 0 \), and the base case \( m = 1 \) follows.

Finally, for the inductive step, suppose that \( H_1(X(B_1 \cup B_2 \cup B_3)) = 0 \) for all \( 1 \leq |B_3| \leq \mu \) and consider the case \( m = \mu + 1 \). Let \( x \in B_3 \) and \( \tilde{X} := X(W \setminus \{x\}) \). Then \( \tilde{X} \) is a complex over three bins where \(|B_3| = \mu \). By inductive hypothesis, \( H_1(\tilde{X}) = 0 \), and hence \( i_* (H_1(\tilde{X})) = 0 \). Now, \( H_1(X, \tilde{X}) = 0 \) by the exact same argument as in the \( m = 1 \) case, and by exactness of the \((X, \tilde{X})\) relative homology sequence, \( H_1(X) = 0 \).

The proof of the outer induction base case thus follows, completing the proof of the theorem. \( \square \)
Fig. 14  a A single column alignment $W = [A, A, A, C, C, G]$. b A fixed $\sigma_0 = [w_0, w_3, w_5]$ and $\sigma' = [w_1, w_4, w_6]$, that do not share any vertex in common. c All 8 triangles $\sigma$ in $L(\sigma')$, the grading is given by $|\sigma_0 \cap \sigma|$. d Geometric illustration of $l(\sigma')$ as an element in Ker$(\partial_{s-1} = 2)$, the boundary of an octahedron. We have $H_2^a(X) = R^{(3-1)(2-1)(2-1)} = R^2$ and the other generator corresponds to $\sigma'' = [w_2, w_4, w_6]$.

**Theorem 11.**

**Proof** The theorem follows immediately from Theorem 10 and the following

**Claim:** let $P$ be a primer and let $X(P)$ be its corresponding dissimilarity complex. Then for all $k \geq m - 1$,

$$H_k(X(P)) = H_k(X_{\geq m}(P)).$$

To prove the claim, it suffices to note that column $j$ in $P$ with $|p_j| \leq m - 1$, can not generate a $k$-simplex in $X(P)$ if $k \geq m - 1$. $\Box$

**Acknowledgements** We want to thank Thomas Li for comments and discussions. We also thank the anonymous reviewers for their constructive feedback.

**Data availability** Data and code will be made available upon request.

**References**

Bandelt H-J, Dress AW (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol 1(3):242–252

Berger B, Waterman MS, Yu YW (2021) Levenshtein distance, sequence comparison and biological database search. IEEE Trans Inf Theory 67(6):3287–3294

Bolker ED (1976) Simplicial geometry and transportation polytopes. Trans Am Math Soc 217:121–142

Bura A, He Q, Reidys C (2021) Weighted homology of bi-structures over certain discrete valuation rings. Mathematics 9(7):744
