A Graph Auto-Encoder for Haplotype Assembly and Viral Quasispecies Reconstruction

Ziqi Ke, Haris Vikalo
Department of Electrical and Computer Engineering
The University of Texas at Austin
ziqike@utexas.edu, hvikalo@ece.utexas.edu

Abstract

Reconstructing components of a genomic mixture from data obtained by means of DNA sequencing is a challenging problem encountered in a variety of applications including single individual haplotyping and studies of viral communities. High-throughput DNA sequencing platforms oversample mixture components to provide massive amounts of reads whose relative positions can be determined by mapping the reads to a known reference genome; assembly of the components, however, requires discovery of the reads' origin – an NP-hard problem that the existing methods struggle to solve with the required level of accuracy. In this paper, we present a learning framework based on a graph auto-encoder designed to exploit structural properties of sequencing data. The algorithm is a neural network which essentially trains to ignore sequencing errors and infers the posterior probabilities of the origin of sequencing reads. Mixture components are then reconstructed by finding consensus of the reads determined to originate from the same genomic component. Results on realistic synthetic as well as experimental data demonstrate that the proposed framework reliably assembles haplotypes and reconstructs viral communities, often significantly outperforming state-of-the-art techniques. Source codes, datasets and supplementary document are available at https://github.com/WuLoli/GAEseq.

1 Introduction

Genetic makeup of a biological sample, inferred by means of DNA sequencing, will help determine an individual's susceptibility to a broad range of chronic and acute diseases, support the discovery of new pharmaceutical products, and personalize and improve the delivery of health care. However, before the promises of personalized medicine come to fruition, efficient methods for accurate inference of genetic variations from massive DNA sequencing data must be devised.

Information about variations in an individual genome is provided by haplotypes, ordered lists of single nucleotide polymorphisms (SNPs) on the individual's chromosomes (Schwartz 2010). High-throughput DNA sequencing technologies generate massive amounts of reads that sample an individual genome and thus enable studies of genetic variations (Schwartz 2010; Clark 2004; Sabeti et al. 2002). Haplotype reconstruction, however, remains challenging due to limited lengths of reads and presence of sequencing errors (Hashemi, Zhu, and Vikalo 2018). Particularly difficult is the assembly of haplotypes in polyploids, organisms with chromosomes organized in k-tuples with $k > 2$, where deep coverage is typically required to achieve desired accuracy. This implies high cost and often renders existing haplotype assembly techniques practically infeasible (Motazedi et al. 2018).

A closely related problem to haplotype assembly is that of reconstructing viral communities. RNA viruses such as HCV, HIV, and Ebola, are characterized by high mutation rates which give rise to communities of viral genomes, the so-called viral quasispecies. Determining genetic diversity of a virus is essential for the understanding of its origin and mutation patterns, and the development of effective drug treatments. Reconstructing viral quasispecies (i.e., *viral haplotypes*, as we refer to them for convenience) is even more challenging than haplotype assembly (Ahn, Ke, and Vikalo 2018) since the number of constituent strains in a community is typically unknown, and its spectra (i.e., strain frequencies) non-uniform.

Existing methods often approach haplotype assembly as the task of grouping sequencing reads according to their chromosomal origin into as many clusters as there are chromosomes. Separation of reads into clusters is rendered challenging by their limited lengths and the presence of sequencing errors (Hashemi, Zhu, and Vikalo 2018); such artifacts create ambiguities regarding the origin of the reads. The vast majority of existing haplotype assembly methods attempt to remove the aforementioned ambiguity by altering or even discarding the data, leading to minimum SNP removal (Lancia et al. 2001), maximum fragments cut (Duitama et al. 2010), and minimum error correction (MEC) score (Lippert et al. 2002) optimization criteria. Majority of haplotype assembly methods developed in recent years are focused on optimizing the MEC score, i.e., determining the smallest possible number of nucleotides in sequencing reads that should be altered such that the resulting dataset is consistent with having originated from k haplotypes (k denotes the ploidy of an organism) (Xie et al. 2016; Pisanti et al. 2015; Kuleshov 2014; Patterson et al. 2015). These include the branch-and-bound scheme (Wang et al. 2005), an integer

[1]This work was funded in part by the NSF grant CCF 1618427.

linear programming formulation in (Chen, Deng, and Wang 2013), and a dynamic programming framework in (Kuleshov 2014). All these techniques attempt to find exact solution to the MEC score minimization problem; the resulting high complexity has motivated search for computationally efficient heuristics. They include the greedy algorithm in (Levy et al. 2007) and methods that compute posterior joint probability of the alleles in a haplotype sequence via MCMC (Bansal et al. 2008) and Gibbs sampling (Kim, Waterman, and Li 2007). A max-cut algorithm for haplotype assembly in (Bansal et al. 2008) is motivated by the clustering interpretation of the problem. The efficient algorithm proposed there, HapCUT, has recently been upgraded as HapCUT2 (Edge, Bafna, and Bansal 2017). In (Aguiar and Istrail 2012), a novel flow-graph approach to haplotype assembly was proposed, demonstrating performance superior to state-of-the-art methods. More recent methods include a greedy max-cut approach in (Duitama et al. 2011), convex optimization framework in (Das and Vikalo 2015), and a communication-theoretic motivated algorithm in (Puljiz and Vikalo 2015).

Haplotype assembly for polyploids ($k > 2$) is more challenging than that for diploids ($k = 2$) due to a much larger space of possible solutions to be searched. Among the aforementioned methods, only HapCompass (Aguiar and Istrail 2012), SDhaP (Das and Vikalo 2015) and BP (Puljiz and Vikalo 2015) are capable of solving the haplotype assembly problem for $k > 2$. Other techniques that can handle reconstruction of haplotypes for both diploid and polyploid genomes include a Bayesian method HapTree (Berger et al. 2014), a dynamic programming method H-PoP (Xie et al. 2016) shown to be more accurate than the techniques in (Aguiar and Istrail 2012; Berger et al. 2014; Das and Vikalo 2015), and the matrix factorization schemes in (Cai, Sanghavi, and Vikalo 2016; Hashemi, Zhu, and Vikalo 2018).

On another note, a number of viral quasispecies reconstruction methods were proposed in recent years. Examples include ShoRAH (Zagordi et al. 2011) and ViSpA (Astrovskaya et al. 2011) that perform read clustering and read-graph path search, respectively, to identify distinct viral components. QuasiRecomb (Töpfer et al. 2013) casts the problem as the decoding in a hidden Markov model while QuRe (Prosperi and Salemi 2012) formulates it as a combinatorial optimization. PredictHaplo (Prabhakaran et al. 2014) employs non-parametric Bayesian techniques to automatically discover the number of viral strains in a quasispecies. More recently, aBayesQR (Ahn and Vikalo 2017) approached viral quasispecies reconstruction with a combination of hierarchical clustering and Bayesian inference while (Ahn, Ke, and Vikalo 2018) relies on tensor factorization.

In this paper, we propose a first ever neural network-based learning framework, named GAEseq, to both haplotype assembly and viral quasispecies reconstruction problems. The framework aims to estimate the posterior probabilities of the origins of sequencing reads using an auto-encoder whose design incorporates salient characteristics of the sequencing data. Auto-encoders are neural networks that in an unsupervised manner learn a low-dimensional representation of data; more specifically, they attempt to perform a dimensionality reduction while robustly capturing essential content of high-dimensional data (Goodfellow, Bengio, and Courville 2016). Auto-encoders have shown outstanding performance in a variety of applications across different fields including natural language processing (Socher et al. 2011), collaborative filtering (Berg, Kipf, and Welling 2017), and information retrieval (Kipf and Welling 2016), to name a few. Typically, auto-encoders consist of two blocks: an encoder and a decoder. The encoder converts input data into the so-called codes while the decoder reconstructs the input from the codes. The act of copying the input data to the output would be of little interest without an important additional constraint – namely, the constraint that the dimension of codes is smaller than the dimension of the input. This enables auto-encoders to extract salient features of the input data. For both the single individual and viral haplotype reconstruction problems, the salient features of data are the origins of sequencing reads. In our work, we propose a graph auto-encoder architecture with an encoder featuring a softmax function placed after the dense layer that follows graph convolutional layers (Masci et al. 2011; Berg, Kipf, and Welling 2017); the softmax function acts as an estimator of the posterior probabilities of the origins of sequencing reads. The decoder assembles haplotypes by finding the consensus sequence for each component of the mixture, thus enabling end-to-end solution to the reconstruction problems.

2 Methods

2.1 Problem Formulation

Let H denote a $k \times n$ haplotype matrix where k is the number of (single individual or viral) haplotypes and n is the haplotype length. Furthermore, let R denote an $m \times n$ SNP fragment matrix whose rows correspond to sequencing reads and columns correspond to SNP positions. Matrix R is formed by first aligning reads to a reference genome and then identifying and retaining only the information that the reads provide about heterozygous genomic sites. One can interpret R as being obtained by sparsely sampling an underlying ground truth matrix M, where the i^{th} row of M is the haplotype sampled by the i^{th} read. The sampling is sparse because the reads are much shorter than the haplotypes; moreover, the reads may be erroneous due to sequencing errors. Following (Ahn, Ke, and Vikalo 2018) , we formalize the sampling operation as

$$[\mathcal{P}_\Omega(M)]_{ij} = \begin{cases} M_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where Ω denotes the set of informative entries in R, i.e., the set of (i, j) such that the j^{th} SNP is covered by the i^{th} read, and $\mathcal{P}_\Omega$ is the projection operator denoting the sampling of haplotypes by reads. Sequencing is erroneous and thus $[\mathcal{P}_\Omega(R)]_{ij}$ may differ from $[\mathcal{P}_\Omega(M)]_{ij}$; in particular, given sequencing error rate p, $[\mathcal{P}_\Omega(R)]_{ij} = [\mathcal{P}_\Omega(M)]_{ij}$ with probability $1 - p$.

Since each read samples one of the haplotypes, $R = \mathcal{P}_\Omega(UH)$ where U denotes the $m \times k$ matrix indicating origins of the reads in R. In particular, each row of matrix U is one of the k-dimensional standard unit vectors $e_i^{(k)}, 1 \le i \le k$, with 1 in the i^{th} position and the remaining

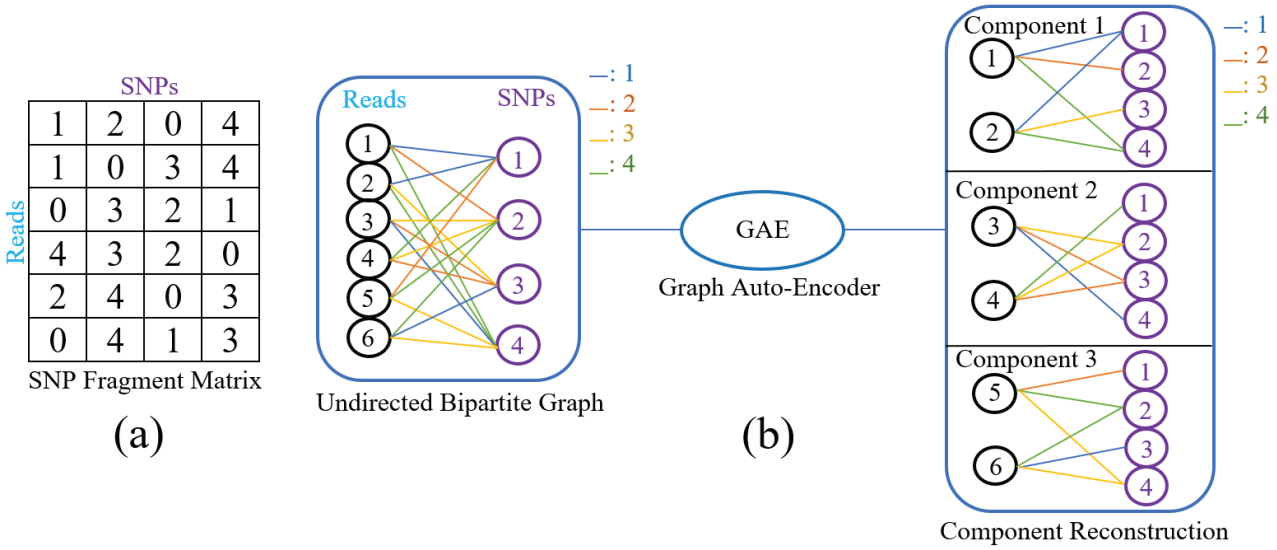

Figure 1: (a) Segment of the SNP fragment matrix. Non-zero entries represent SNP information provided by sequencing reads; labels 1-4 indicate the four nucleotides. Zero entries in a row indicate that the read does not cover corresponding SNP. In this illustration, the first two rows represent reads originating from the same haplotype; the third and fourth reads both originated from another haplotype; and so on. (b) The pipeline from the SNP fragment matrix to haplotypes via a graph auto-encoder.

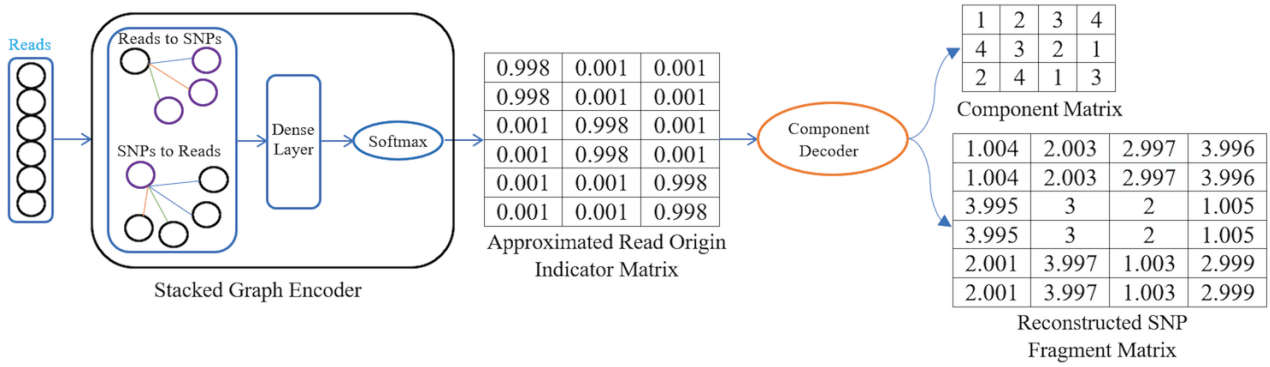

Figure 2: A forward pass through the graph auto-encoder consisting of a stacked graph encoder that passes messages between read and SNP nodes and constructs approximate read origin indicator matrix via the softmax function. Decoder reconstructs haplotypes and SNP fragment matrix.

entries 0. If i^{th} read samples j^{th} haplotype, the i^{th} row of U is $e_j^{(k)}$. If the origins of reads were known, each haplotype could be reconstructed by finding consensus of reads which sample that particular haplotype. We think of the assembly as a two-step procedure: given the SNP fragment matrix R we first identify the read origin indicator matrix U and then use U to reconstruct the haplotype matrix H.

To characterize the performance of haplotype assembly methods we rely on two metrics: the minimum error correction (MEC) score, which can be traced back to (Lippert et al. 2002), and the correct phasing rate, also referred to as reconstruction rate. The MEC score is defined as the smallest number of observed entries in R that need to be altered (i.e., corrected) such that the resulting data is consistent with having originated from k distinct haplotypes, i.e.,

$$\text{MEC} = \sum_{i=1}^{m} \min_{j=1,2,\ldots,k} \text{HD}(R_{i:}, H_{j:}), \qquad (2)$$

where $\text{HD}(\cdot,\cdot)$ denotes the Hamming distance between its arguments (sequences, evaluated only over informative entries), $R_{i:}$ denotes the i^{th} row of R and $H_{j:}$ denotes the j^{th} row of H. The correct phasing rate (CPR) is defined as

$$\text{CPR} = 1 - \frac{1}{kn}(\min \sum_{i=1}^{k} \text{HD}(H_{i:}, \mathcal{M}(H_{i:}))), \qquad (3)$$

where $\mathcal{M}$ is the one-to-one mapping from the set of reconstructed haplotype to the set of true haplotype (Hashemi, Zhu, and Vikalo 2018), i.e., mapping that determines the best

possible match between the two sets of haplotypes. To characterize performance of methods for reconstruction of viral quasispecies with generally a priori unknown number of components, in addition to correct phasing rate we also quantify *recall rate*, defined as the fraction of perfectly reconstructed components in a population (i.e., recall rate = $\frac{TP}{TP+FN}$), and *predicted proportion*, defined as the ratio of the estimated and the true number of components in a genomic mixture (Ahn, Ke, and Vikalo 2018).

To assemble haplotypes from a set of reads we design and employ a graph auto-encoder. Fig. 1 (b) shows the entire end-to-end pipeline that takes the collection of erroneous reads and generates reconstructed haplotypes. First, the SNP fragment matrix R is processed by the graph encoder to infer the read origin indicator matrix U; then, a haplotype decoder reconstructs matrix H. The graph auto-encoder is formalized in the next section.

2.2 Graph Auto-Encoders

Graph auto-encoders are a family of auto-encoders specifically designed for learning on graph-structured data (Berg, Kipf, and Welling 2017; Kipf and Welling 2016). In this paper, we design graph auto-encoders for the assembly of the components of a genomic mixture. As in conventional auto-encoder structures, the developed architecture consists of two parts: the graph encoder and the decoder. The graph encoder $Z = f(R, A)$ takes the SNP fragment matrix R and the $m \times n$ graph adjacency matrix A as inputs, and outputs the $m \times k$ node embedding matrix Z. Note that we impose constraints on the node embedding matrix so that the salient features extracted by a graph auto-encoder approximate the read origin indicator matrix U. Such a constraint does not prevent efficient training of the auto-encoders via backpropagation. The decoder $\hat{R} = g(Z)$ is utilized to reconstruct the SNP fragment matrix R and the haplotype matrix H from the node embedding matrix Z; this implies that the decoder is essentially capable of imputing the unobserved entries in the SNP fragment matrix.

To numerically represent information in the SNP fragment matrix R, we encode its entries R_{ij} using a set of 4 discrete values – one for each of 4 possible nucleotides – where the mapping between nucleotides and the discrete values can be decided arbitrarily. To this end, we may simply represent the nucleotides A, C, G and T by 1, 2, 3 and 4, respectively; non-informative entries in each row of R, i.e., SNP positions not covered by a read, are represented by 0. Note that the SNP fragment matrix can be represented by an undirected bipartite graph $G = (V, E, \mathcal{W})$ where the set of read nodes $r_i \in \mathcal{A}$ with $i \in \{1, ..., m\}$ and the set of SNP nodes $s_j \in \mathcal{B}$ with $j \in \{1, ..., n\}$ together form the set of vertices V, i.e., $\mathcal{A} \cup \mathcal{B} = V$. The weights $w \in \{1, 2, 3, 4\} = \mathcal{W}$ assigned to edges $(r_i, w, s_j) \in E$ are the discrete values used to represent nucleotides. With this model in place, we can rephrase the graph encoder as $Z = f(R, A_1, A_2, A_3, A_4)$, where $A_w \in \{0, 1\}^{m \times n}$ represents the graph adjacency matrix for a nucleotide encoded by w. Equivalently, A_w has 1's for the entries whose corresponding positions in R are encoded by w. Since we are interested in imputing the unobserved entries based on the observed entries in R instead of simply copying the observed entries to $\hat{R}$, it is beneficial to reformulate the decoder as $\hat{R} = g(Z, R)$. In other words, the auto-encoder is trained to learn from the observed entries in order to determine origin of reads, impute unobserved entries of R, and reconstruct haplotypes in the genomic mixture.

2.3 Read Origin Detection via Graph Encoder

Recall the interpretation that the SNP fragment matrix R is obtained by erroneously sampling an underlying ground truth matrix M. This motivates development of a specific graph encoder architecture, motivated by the ideas of the design in (Berg, Kipf, and Welling 2017), that is capable of detecting origin of sequencing reads in R via estimating the posterior probabilities of the origin of each read.

Let D_r denote an $m \times m$ diagonal read degree matrix whose entries indicate the number of SNPs covered by each read, and let D_s denote an $n \times n$ diagonal SNP degree matrix whose entries indicate the number of reads covering each SNP. We facilitate exchange of messages between read nodes and SNP nodes in the graph, initiating it from the set of read nodes $\mathcal{A}$; doing so helps reduce the dimensions of weights and biases since the number of reads m is far greater than the haplotype length n. Note that the dimension of messages keeps reducing during the message passing procedure.

The messages from read nodes to SNP nodes are

$$M_{(1)} = \sigma(\sum_{w=1}^{4} D_s^{-1} A_w^T R W_w^{(1)} + B_w^{(1)}), \tag{4}$$

where $W_w^{(1)}$ and $B_w^{(1)}$ denote the weights and biases of the first convolutional layer for the nucleotide encoded with w, respectively, σ denotes an element-wise activation function such as $\text{ReLU}(\cdot) = \max(\cdot, 0)$, and $(\cdot)^T$ denotes the transpose of a matrix. The dimension of both $W_w^{(1)}$ and $B_w^{(1)}$ is $n \times c_{(1)}$, where $c_{(1)}$ denotes the message length after the first message passing step.

The messages from SNP nodes to read nodes are

$$M_{(2)} = \sigma(\sum_{w=1}^{4} D_r^{-1} A_w M_{(1)} W_w^{(2)} + B_w^{(2)}), \tag{5}$$

where $W_w^{(2)}$ and $B_w^{(2)}$ denote the weights and biases of the second convolutional layer for the nucleotide encoded with w, respectively. The dimension of both $W_w^{(2)}$ and $B_w^{(2)}$ is $c_{(1)} \times c_{(2)}$, where $c_{(2)}$ denotes the message length after the second message passing step.

Repeating message passing and stacking the convolutional layers leads to formation of a deep model. The read nodes to SNP nodes layer is readily generalized as

$$M_{(2i+1)} = \sigma(\sum_{w=1}^{4} D_s^{-1} A_w^T M_{(2i)} W_w^{(2i+1)} + B_w^{(2i+1)}), \tag{6}$$

where $i \in \{0, 1, 2, ...\}$ and $M_{(2i)} = R$ for $i = 0$. The dimension of $M_{(2i)}$ is $m \times c_{(2i)}$. Furthermore, the SNP nodes to read nodes layer is generalized as

$$M_{(2i)} = \sigma(\sum_{w=1}^{4} D_r^{-1} A_w M_{(2i-1)} W_w^{(2i)} + B_w^{(2i)}), \tag{7}$$

where $i \geq 1$. The dimension of $M_{(2i-1)}$ is $n \times c_{(2i-1)}$. Note that the messages are passed from read nodes to SNP nodes when the subscript of M is odd, and otherwise traverse in the opposite direction.

Equation (6) and (7) specify the graph convolutional layer while the dense layer is defined as

$$O = \sigma(M_{(l)}W_d + B_d), \tag{8}$$

where O denotes the output of the dense layer, W_d and B_d are the weights and biases of the dense layer, respectively, $M_{(l)}$ is the output of the last graph convolutional layer, and l represents the number of graph convolutional layers. The dimension of W_d is $c_{(l)} \times k$ and the dimension of O and B_d is $m \times k$, where k denotes the ploidy (i.e., the number of components in a genomic mixture).

To find Z which approximates the read origin indicator matrix U (i.e., Z with each row close in the l_2-norm sense to a k-dimensional standard basis vector), we employ the softmax function

$$Z_{ij} = \frac{e^{\beta O_{ij}}}{\sum_{j=1}^{k} e^{\beta O_{ij}}}, \tag{9}$$

where in our experiments we set β to 200. Having estimated read origins by the node embedding matrix Z, the reads can be organized into k clusters. This enables straightforward reconstruction of haplotypes by determining the consensus sequence for each cluster.

2.4 Haplotype Decoder

Thus far, we have conveniently been representing alleles as the numbers in $\{1, 2, 3, 4\}$. It is desirable, however, that in the definition of a loss function the distance between numerical values representing any two alleles is identical, no matter which pair of alleles is considered; this ensures the loss function relates to the MEC score – the metric of interest in haplotype assembly problems. Following (Ahn, Ke, and Vikalo 2018), we define the loss function of the auto-encoder as the squared Frobenius norm of the difference between a one-hot SNP fragment matrix $\mathcal{R}$ and the reconstructed matrix $\hat{\mathcal{R}} = Z\mathcal{H}$ at the informative positions, i.e., $\mathcal{L} = \frac{1}{2}||\mathcal{P}_\Omega(\mathcal{R} - Z\mathcal{H})||_F^2$, where $\mathcal{R} \in \{0, 1\}^{m \times 4n}$ and $\mathcal{H} \in \{0, 1\}^{k \times 4n}$ are formed by substituting discrete values $w \in \{1, 2, 3, 4\}$ by the set of four dimensional standard basis vectors $e_i^{(4)}, 1 \leq i \leq 4$. With such a notational convention, the proposed loss function approximates the MEC score; it only approximates the score, rather than coincides with it, because Z is an approximation of the read-origin matrix U. Therefore, the graph auto-encoder is trained to approximately minimize the MEC score. Fig. 2 illustrates the data processing pipeline that takes as inputs reads in the SNP fragment matrix and produces the matrix of haplotypes as well as imputes missing entries in the SNP fragment matrix. The proposed graph auto-encoders for haplotype assembly and viral quasispecies reconstruction are formalized as Algorithm 1 and Algorithm 2, respectively. For the viral quasispecies reconstruction problem, the number of clusters k is typically unknown; detailed strategy based on (Ahn, Ke, and Vikalo 2018) for the automated inference of k can be found in Supplementary Document B.

Algorithm 1 Graph auto-encoder for haplotype assembly

Input: SNP fragment matrix R, the number of experiments n_{exp} and the number of haplotpyes k
Output: Reconstructed haplotypes H
while $n_{exp} \neq 0$ **do**
  Initialize $W_w^{(i)}$, $B_w^{(i)}$, W_d and B_d using Xavier initialization where $w \in \{1, 2, 3, 4\}$ and $i \in \{1, 2\}$
  for $n_{epoch} = 1$ to 100 **do**
    $M_{(1)} \leftarrow \sigma(\sum_w D_s^{-1} A_w^T R W_w^{(1)} + B_w^{(1)})$
    $M_{(2)} \leftarrow \sigma(\sum_w D_r^{-1} A_w M_{(1)} W_w^{(2)} + B_w^{(2)})$
    $O \leftarrow \sigma(M_{(2)} W_d + B_d)$
    $Z_{ij} \leftarrow \frac{e^{\beta O_{ij}}}{\sum_{j=1}^{k} e^{\beta O_{ij}}}$ with $\beta = 200$
    Calculate $\mathcal{H}$ by majority voting
    $\mathcal{L} \leftarrow \frac{1}{2}||\mathcal{P}_\Omega(\mathcal{R} - Z\mathcal{H})||_F^2$
    Record reconstructed haplotypes and the MEC score
    Update $W_w^{(i)}$, $B_w^{(i)}$, W_d and B_d using Adam Optimizer where $w \in \{1, 2, 3, 4\}$ and $i \in \{1, 2\}$
  end for
  $n_{exp} \leftarrow n_{exp} - 1$
end while
Output the reconstructed haplotypes H corresponding to the lowest MEC score

Algorithm 2 Graph auto-encoder for viral quasispecies reconstruction

Input: SNP fragment matrix R, the number of experiments n_{exp}, the MEC improvement rate threshold η and the estimated initial number of components k_0
Output: Reconstructed viral haplotypes H and the inferred frequencies
Initial $\tau \leftarrow 0$, MECflag $\leftarrow 0$ and $k_\tau \leftarrow k_0$
while $\tau = 0$ or $k_\tau = k_\tau - 1$ **do**
  for $k \in \{k_\tau, k_\tau + 1\}$ **do**
    Run Algorithm 1 with k
  end for
  if MECimpr$(k_\tau) \leq \eta$ **then**
    $k_{\tau+1} \leftarrow \lfloor (k_\tau + \max\{1, k_i\})/2 \rfloor, \{i \in \{1, \cdots, \tau\text{-}1\} : k_i \leq k_\tau\}$; MECflag $\leftarrow 1$
  else
    if MECflag $= 0$ **then**
      $k_{\tau+1} \leftarrow 2k_\tau$
    else
      $k_{\tau+1} \leftarrow \lfloor (k_\tau + \min k_i)/2 \rfloor, \{i \in \{1, \cdots, \tau\text{-}1\} : k_i > k_\tau\}$
    end if
  end if
  $\tau \leftarrow \tau + 1$
end while
Output the viral quasispecies H with $k = k_\tau + 1$ and the inferred frequencies

3 Results

The hyper-parameters of GAEseq are determined by training on 5 synthetic triploid datasets with coverage 30× and

Table 1: Performance comparison on biallelic Solanum Tuberosum semi-experimental data.

Coverage		MEC Mean	SD	CPR Mean	SD
15	GAEseq	**8.200**	**4.686**	**0.822**	0.048
	HapCompass	100.700	66.150	0.763	**0.046**
	H-PoP	28.700	32.667	0.783	0.066
	AltHap	59.100	28.125	0.709	0.054
25	GAEseq	**8.400**	**4.719**	**0.831**	0.081
	HapCompass	124.800	132.156	0.810	0.063
	H-PoP	33.800	47.434	0.798	**0.046**
	AltHap	92.600	83.649	0.756	0.068
35	GAEseq	**10.700**	**3.234**	**0.857**	0.087
	HapCompass	217.400	174.135	0.775	**0.072**
	H-PoP	41.700	53.971	0.823	0.094
	AltHap	164.000	101.583	0.754	0.093

Table 2: Performance comparison of GAEseq, PredictHap, TenSQR and aBayesQR on a real HIV-1 5-virus-mix data. Genes where all the strains are perfectly reconstructed are denoted as boldface.

		p17	p24	p2-p6	PR	RT	RNase	int	vif	vpr	vpu	gp120	gp41	nef
GAEseq	PredProp	**1**	1	**1**	**1**	**1.2**	**1**	**1**	**1**	**1**	1.2	1	1	1
	CPR_{HXB2}	**100**	99.4	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	96.2	96.7	100
	$CPR_{89.6}$	**100**	99.4	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.2	99.4	100	98.2
	CPR_{JR-SCF}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	99.9	100	99.3
	CPR_{NL4-3}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	100	100	99.8
	CPR_{YU2}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	99.6	100	98.1
PredictHap	PredProp	**1**	0.6	**1**	**1**	**1**	0.8	0.8	0.8	**1**	0.8	0.8	0.8	0.8
	CPR_{HXB2}	**100**	0	**100**	**100**	**100**	98.9	100	100	**100**	93.2	0	0	0
	$CPR_{89.6}$	**100**	100	**100**	**100**	**100**	100	99.8	100	**100**	0	97.8	100	98.8
	CPR_{JR-SCF}	**100**	100	**100**	**100**	**100**	100	100	100	**100**	100	99.7	100	100
	CPR_{NL4-3}	**100**	99.1	**100**	**100**	**100**	100	100	100	**100**	100	100	100	100
	CPR_{YU2}	**100**	0	**100**	**100**	**100**	0	0	0	**100**	100	98.6	100	100
TenSQR	PredProp	**1**	1.6	**1**	1	1.4	**1**	**1**	**1**	**1**	1.6	2.2	1.2	0.8
	CPR_{HXB2}	**100**	98.9	**100**	100	99.2	**100**	**100**	**100**	**100**	92.8	96.0	99.0	0
	$CPR_{89.6}$	**100**	100	**100**	100	98.0	**100**	**100**	**100**	**100**	94.0	97.2	100	95.7
	CPR_{JR-SCF}	**100**	100	**100**	100	100	**100**	**100**	**100**	**100**	100	98.3	97.7	99.8
	CPR_{NL4-3}	**100**	99.3	**100**	100	99.5	**100**	**100**	**100**	**100**	100	99.8	99.5	99.7
	CPR_{YU2}	**100**	99.3	**100**	99.7	99.7	**100**	**100**	**100**	**100**	100	94.9	100	98.6
aBayesQR	PredProp	1	1	**1**	**1**	1	**1**	1	1	**1.2**	1	0.8	0.8	1.2
	CPR_{HXB2}	**100**	99.4	**100**	**100**	98.5	**100**	99.9	100	**100**	99.6	98	0	95.8
	$CPR_{89.6}$	**100**	98.7	**100**	**100**	98.6	**100**	100	100	**100**	92	96.5	98.9	95.5
	CPR_{JR-SCF}	**100**	99.6	**100**	**100**	99	**100**	100	100	**100**	98.8	97.7	99.1	98.2
	CPR_{NL4-3}	**100**	100	**100**	**100**	98.9	**100**	100	99.8	**100**	100	96.3	98.8	100
	CPR_{YU2}	**100**	99.7	**100**	**100**	99.2	**100**	99.5	99.7	**100**	100	0	98.6	99.2

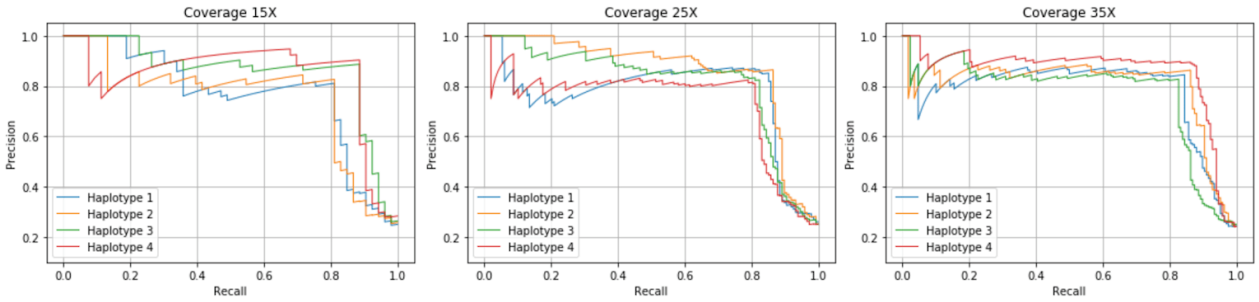

Figure 3: The precision-recall curves for Solanum Tuberosum semi-experimental data with coverage 15×, 25× and 35×

validated on different 5 synthetic triploid datasets with the same coverage. The results reported in this section are obtained on test data. Detailed description of the computational platform and the choice of hyper-parameters can be found in Supplementary Document A.

3.1 Performance Comparison on Solanum Tuberosum Semi-Experimental Data

We first evaluate performance of GAEseq on realistic simulations which, for convenience and to distinguish from perhaps more rich synthetic and experimental data discussed in supplementary documents, we refer to as "semi-experimental data". The semi-experimental data is obtained by simulating mutations, shotgun sequencing procedure, read alignment and SNP calling steps in a fictitious experiment on a single individual *Solanum Tuberosum* (polyploid with $k = 4$). Details on how exactly the semi-experimental data is generated and processed can be found in Supplementary Document C. We compare the performance of GAEseq on this data with publicly available software HapCompass (Aguiar and Istrail 2012), an algorithm that relies on graph-theoretic models to perform haplotype assembly, H-PoP (Xie et al. 2016), a dynamic programming method, and AltHap (Hashemi, Zhu, and Vikalo 2018), a method based on tensor factorization. The performance of different methods is evaluated in terms of the MEC score and CPR. All the considered softwares were executed with their default settings, i.e. we follow instructions in the papers they were originally proposed; there are no parameter tuning steps required for these methods. We report the MEC scores and CPR achieved by the considered algorithms in Table 1. For each sequencing coverage, the mean and standard deviation (SD) of the adopted metrics are evaluated over 10 samples. As shown in the table, GAEseq achieves the lowest average MEC score as well as the lowest standard deviation of the MEC score at all sequencing coverage settings. Moreover, GAEseq achieves the highest average CPR at all coverage settings. Note that the MEC score increases with sequencing coverage since higher coverage implies more reads. The results demonstrate that the adopted graph abstraction enables GAEseq to achieve high accuracy of the reconstruction task by learning posterior probabilities of the origins of reads. Fig. 3 shows the precision-recall curves for data with coverage $15\times$, $25\times$ and $35\times$. Note that GAEseq performs very accuratly at high sequencing coverage while its performance deteriorates at low coverage. An extended version of Table 1 with additional coverage settings is in Supplementary Document C.

We further test the performance of GAEseq on simulated biallelic diploid, polyallelic triploid and tetraploid data, and on real Solanum Tuberosum data; in addition to H-Pop, AltHap and HapCompass, comparisons on diploid data also include performance of HapCUT2 (Edge, Bafna, and Bansal 2017). GAEseq outperforms all the considered algorithms by achieving lower MEC score and higher CPR. Further details can be found in Supplementary Document D and E. Note that GAEseq handles deletions automatically while insertions can be handled by employing a downstream pipeline proposed in (Ahn, Ke, and Vikalo 2018).

3.2 Performance Comparison on Gene-Wise Reconstruction of Real HIV-1 Data

The real HIV-1 data with pairwise distances between $2.61\%-8.45\%$ and relative frequencies between 10% and 30% is an *in vitro* viral population of 5 known HIV-1 strains generated by Illumina's MiSeq Benchtop Sequencer (Di Giallonardo et al. 2014). These reads are then aligned to the HIV-1_{HXB2} reference genome. According to (Di Giallonardo et al. 2014), we remove reads of length lower than 150bp and mapping quality scores lower than 60 for better results. We compare the performance of GAEseq on gene-wise reconstruction of the HIV population to that of other state-of-the-art methods such as PredictHaplo (Prabhakaran et al. 2014), TenSQR (Ahn, Ke, and Vikalo 2018) and aBayesQR (Ahn and Vikalo 2017), following their default settings. For fair benchmarking, we use the same dataset as (Ahn, Ke, and Vikalo 2018) which is why the results of our benchmarking tests match those in (Ahn, Ke, and Vikalo 2018). The correct phasing rate and the inferred strain frequencies are evaluated for all reconstructed strains because the ground truth for the 5 HIV-1 strains is available at https://bmda.dmi.unibas.ch/software.html. Following (Ahn, Ke, and Vikalo 2018), we evaluate predicted proportion by setting the parameter η needed to detect the number of HIV-1 strains to 0.09. The results in Table 2 show that GAEseq perfectly reconstructs all 5 HIV-1 strains in 8 genes while other methods correctly reconstruct components in 5 or 6 genes. This demonstrates that GAEseq's inference of read origins based on posterior probabilities enables high accuracy of the reconstruction tasks. Regarding the 5 genes where GAEseq and other methods do not achieve perfect reconstruction (p24, vpu, gp120, gp41, nef): closer examination of viral strains reconstructed by various methods suggests possible rearrangements of short segments within those 5 genes in the "gold standard" dataset created by (Di Giallonardo et al. 2014), which may be the reason for mismatch between reconstructed strains and the ground truth. Further results on reconstruction of HIV viral communities can be found in Supplement Document F.

4 Conclusions

In this article, we introduce auto-encoders to the problem of reconstructing components of a genomic mixture from high-throughput sequencing data that is encountered in haplotype assembly and analysis of viral communities. In particular, a graph auto-encoder is trained to group together reads that originate from the same component of a genomic mixture and impute missing information in the SNP fragment matrix by learning from the available data. The graph convolutional encoder attempts to discover origin of the reads while the decoder aims to reconstruct haplotypes and impute missing information, effectively correcting sequencing errors. Studies on semi-experimental data show that GAEseq can achieve significantly lower MEC scores and higher CPR than the competing methods. Benchmarking tests on simulated and experimental data demonstrate that GAEseq maintains good performance even at low sequencing coverage. Studies on real HIV-1 data illustrate that GAEseq outperforms existing state-of-the-art methods in viral quasispecies reconstruction.

References

Aguiar, D., and Istrail, S. 2012. Hapcompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. *J Comput Biol.* 19(6):577–90.

Ahn, S., and Vikalo, H. 2017. abayesqr: A bayesian method for reconstruction of viral populations characterized by low diversity. *International Conference on Research in Computational Molecular Biology* 353–369.

Ahn, S.; Ke, Z.; and Vikalo, H. 2018. Viral quasispecies reconstruction via tensor factorization with successive read removal. *Bioinformatics (Oxford, England)* 34(13):i23–i31.

Astrovskaya, I.; Tork, B.; Mangul, S.; Westbrooks, K.; Mandoiu, I.; Balfe, P.; and Zelikovsky, A. 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC bioinformatics* 12(6):1.

Bansal, V.; Halpern, A.; Axelrod, N.; and Bafna, V. 2008. An mcmc algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.* 18(8):1336–46.

Berg, R.; Kipf, T.; and Welling, M. 2017. Graph convolutional matrix completion. *arXiv:1706.02263*.

Berger, E.; Yorukoglu, D.; Peng, J.; and Berger, B. 2014. Haptree: A novel bayesian framework for single individual polyplotyping using ngs data. *PLoS Comput Biol.* 10(3):e1003502.

Cai, C.; Sanghavi, S.; and Vikalo, H. 2016. Structured low-rank matrix factorization for haplotype assembly. *IEEE J Sel Top Sign Proc.* 10(4):647–657.

Chen, Z.-Z.; Deng, F.; and Wang, L. 2013. Exact algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics* 29(16):1938–1945.

Clark, A. 2004. The role of haplotypes in candidate gene studies. *Genet Epidemiol* 27(4):321–333.

Das, S., and Vikalo, H. 2015. Sdhap: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* 16(260).

Di Giallonardo, F.; Töpfer, A.; Rey, M.; Prabhakaran, S.; Duport, Y.; Leemann, C.; Schmutz, S.; Campbell, N.; Joos, B.; and Lecca, M. 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic acids research* 42(14):e115.

Duitama, J.; Huebsch, T.; Suk, E.-K.; and Hoehe, M. 2010. Refhap: a reliable and fast algorithm for single individual haplotyping. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational*, 160–169.

Duitama, J.; Huebsch, T.; Palczewski, S.; Schulz, S.; Verstrepen, K.; Suk, E.-K.; and Hoehe, M. 2011. Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.* 40(5):2041–2053.

Edge, P.; Bafna, V.; and Bansal, V. 2017. Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* 27(5):801–812.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press.

Hashemi, A.; Zhu, B.; and Vikalo, H. 2018. Sparse tensor decomposition for haplotype assembly of diploids and polyploids. *BMC Genomics* 19(191).

Kim, J.; Waterman, M.; and Li, L. 2007. Diploid genome reconstruction of ciona intestinalis and comparative analysis with ciona savignyi. *Genome Res.* 17(7):1101–10.

Kipf, T., and Welling, M. 2016. Variational graph auto-encoders. *arXiv:1611.07308*.

Kuleshov, V. 2014. Probabilistic single-individual haplotyping. *Bioinformatics* 30(17):379–385.

Lancia, G.; Bafna, V.; Istrail, S.; Lippert, R.; and Schwartz, R. 2001. Snps problems, complexity, and algorithms. *European symposium on algorithms* 2161:182–193.

Levy, S.; Sutton, G.; Ng, P.; Feuk, L.; Halpern, A.; Walenz, B.; Axelrod, N.; Huang, J.; Kirkness, E.; Denisov, G.; Lin, Y.; Macdonald, J.; Pang, A.; Shago, M.; Stockwell, T.; Tsiamouri, A.; Bafna, V.; Kravitz, S.; and Venter, J. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5(10):e254.

Lippert, R.; Schwartz, R.; Lancia, G.; and Istrail, S. 2002. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief Bioinformatics* 3(1):23–31.

Masci, J.; Meier, U.; Ciresan, D.; and Schmidhuber, J. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning - ICANN*, 52–59.

Motazedi, E.; Finkers, R.; Maliepaard, C.; and Ridder, D. 2018. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in bioinformatics* 19(3):387–403.

Patterson, M.; Marschall, T.; Pisanti, N.; Iersel, L.; Stougie, L.; Klau, G.; and Schönhuthstar, A. 2015. Whatshap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol.* 22(6):498–509.

Pisanti, N.; Bonizzoni, P.; Dondi, R.; Klau, G.; Pirola, Y.; and Zaccaria, S. 2015. Hapcol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics* 32(11):1610–1617.

Prabhakaran, S.; Rey, M.; Zagordi, O.; Beerenwinkel, N.; and Roth, V. 2014. Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM Trans. on Comput. Biol. Bioinform. (TCBB)* 11(1):182–191.

Prosperi, M., and Salemi, M. 2012. Qure: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 28(1):132–133.

Puljiz, Z., and Vikalo, H. 2015. Decoding genetic variations: communications-inspired haplotype assembly. *IEEE/ACM Trans Comput Biol Bioinform (TCBB)* 13(3):518–530.

Sabeti, P.; Reich, D.; Higgins, J.; Levine, H.; Richter, D.; Schaffner, S.; Gabriel, S.; Platko, J.; Patterson, N.; McDonald, G.; Ackerman, H.; Campbell, S.; Altshuler, D.; Cooper, R.; Kwiatkowski, D.; Ward, R.; and Lander, E. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Schwartz, R. 2010. Theory and algorithms for the haplotype assembly problem. *Communications in Info. and Sys.* 10(1):23–38.

Socher, R.; Pennington, J.; Huang, E.; Ng, A.; and Manning, C. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Conference on Empirical Methods in Natural Language Processing*, 151–161.

Töpfer, A.; Zagordi, O.; Prabhakaran, S.; Roth, V.; Halperin, E.; and Beerenwinkel, N. 2013. Probabilistic inference of viral quasispecies subject to recombination. *Journal of Computational Biology* 20(2):113–123.

Wang, R.-S.; Wu, L.-Y.; Li, Z.; and Zhang, X. 2005. Haplotype reconstruction from snp fragments by minimum error correction. *Bioinformatics* 21(10):2456–2462.

Xie, M.; Wu, Q.; Wang, J.; and Jiang, T. 2016. H-pop and h-popg: Heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics* 32(24):3735–3744.

Zagordi, O.; Bhattacharya, A.; Eriksson, N.; and Beerenwinkel, N. 2011. Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics* 1(119).