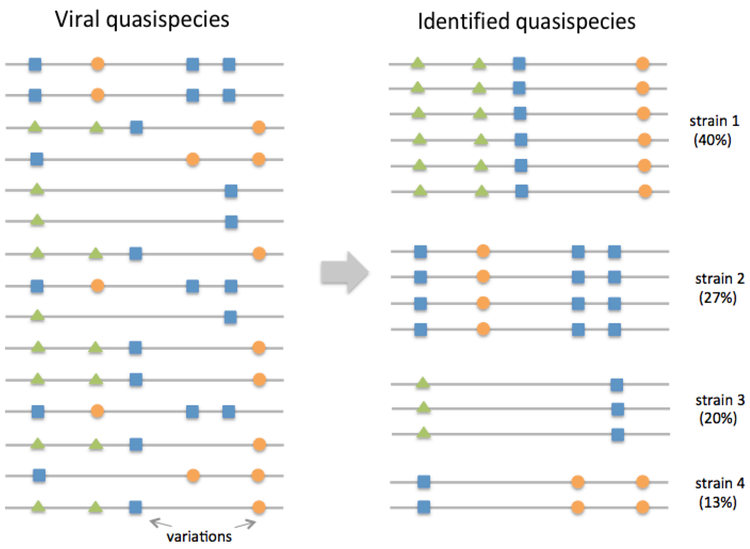
A Graph Auto-Encoder for Haplotype Assembly and Viral Quasispecies Reconstruction

Ziqi Ke and Haris Vikalo

Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA

BACKGROUND

- Haplotype assembly
  - Reconstructing haplotypes, ordered lists of single nucleotide polymorphisms on an individual's chromosomes, from high-throughput DNA sequencing data
  - Applications: diagnosis of genetic diseases, personalized medicine
- Viral quasispecies reconstruction
  - Reconstructing a priori unknown number of viral sequences in a population and estimating their relative frequencies
  - Applications: antiviral vaccine designs, discovery of new pharmaceutical products



- Most prior works are based on branch-and-bound schemes, integer linear programming, dynamic programming, matrix factorization, Bayesian inference and so on
- Our approach: the first neural network-based learning framework, GAEseq, to both haplotype assembly and viral quasispecies reconstruction problems

PROBLEM FORMULATION

- Notation

H : the haplotype matrix

R : the SNP fragment matrix

Z : an approximation of the read-origin matrix

Ω : the set of informative entries in the SNP fragment matrix

$\mathcal{P}_\Omega$: the projection operator denoting the sampling of haplotypes by reads

- Solve the *NP-hard* problem with required level of accuracy

$$\min_{Z,\mathcal{H}} \frac{1}{2}||\mathcal{P}_\Omega(\mathcal{R} - Z\mathcal{H})||_F^2$$

- Evaluation metrics
  - Minimum error correction (MEC) score:

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

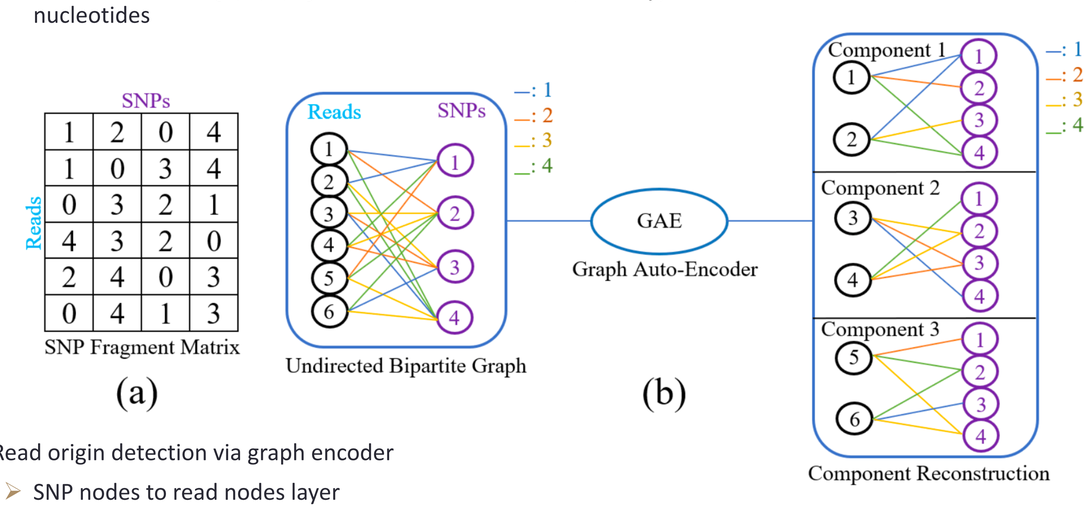
  - Correct phasing rate:

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

where *k* is the number of haplotypes and *n* is the haplotype length.

A GRAPH AUTO-ENCODER

- An undirected bipartite graph $G = (V, E, \mathcal{W})$
  - The set of read nodes $r_i \in \mathcal{A}$ and the set of SNP nodes $s_j \in \mathcal{B}$ form the set of vertices
  - 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



- Read origin detection via graph encoder
  - SNP nodes to read nodes layer

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

  - Read nodes to SNP nodes layer

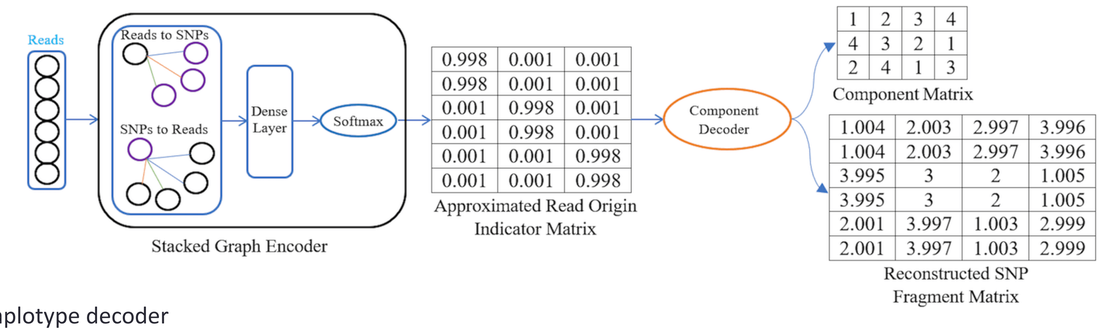
$$M_{(2i)} = \sigma(\sum_{w=1}^{4} D_r^{-1} A_w M_{(2i-1)} W_w^{(2i)} + B_w^{(2i)})$$

  - Dense layer

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

  - Softmax layer

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



- Haplotype decoder
  - Majority voting based on the approximation of the read-origin matrix Z

ALGORITHMS

Algorithm 1 Graph auto-encoder for haplotype assembly

Input: SNP fragment matrix R, the number of experiments n_{exp} and the number of haplotypes 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

RESULTS

- Performance comparison on Solanum Tuberosum semi-experimental data

Coverage		MEC Mean	MEC SD	CPR Mean	CPR 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

- Performance comparison on gene-wise reconstruction of real HIV-1 data

		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(20.5)**	99.4(17.1)	**100(21)**	**100(30.9)**	**100(12.1)**	**100(9.6)**	**100(13.6)**	**100(10.4)**	**100(6.6)**	100(34.3)	96.2(8.7)	96.7(2.8)	100(6.6)
	$CPR_{89.6}$	**100(18.8)**	99.4(21.8)	**100(20)**	**100(18)**	**100(18.2)**	**100(20.9)**	**100(18.2)**	**100(20.4)**	**100(20.3)**	99.2(10.5)	99.4(24.1)	100(25.6)	98.2(22.9)
	CPR_{JR-SCF}	**100(30.9)**	100(31.5)	**100(27)**	**100(21.6)**	**100(23.5)**	**100(20.2)**	**100(21.5)**	**100(29.8)**	**100(34.6)**	100(36.4)	99.9(33.0)	100(27)	99.3(19.7)
	CPR_{NL4-3}	**100(17.4)**	100(18.3)	**100(14.1)**	**100(19.7)**	**100(30.6)**	**100(33.1)**	**100(37.1)**	**100(32.8)**	**100(30.6)**	100(7.9)	100(29.6)	100(34.5)	99.8(39.1)
	CPR_{YU2}	**100(12.3)**	100(11.2)	**100(18)**	**100(9.9)**	**100(11.9)**	**100(16.2)**	**100(9.6)**	**100(6.6)**	**100(7.9)**	100(10.2)	99.6(4.6)	100(10)	98.1(11.7)
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(17.8)**	0(0)	**100(18.7)**	**100(15.2)**	**100(12.2)**	98.9(25.4)	100(12.1)	100(17.7)	**100(10.2)**	93.2(10.8)	0(0)	0(0)	0(0)
	$CPR_{89.6}$	**100(19.9)**	100(46.4)	**100(21.7)**	**100(22.2)**	**100(19.4)**	100(18.2)	99.8(27.6)	100(20.9)	**100(22.1)**	0(0)	97.8(20.7)	100(26.7)	98.8(20.7)
	CPR_{JR-SCF}	**100(31.9)**	100(21.8)	**100(30.3)**	**100(26.9)**	**100(23.4)**	100(23.2)	100(22.3)	100(24.9)	**100(23.7)**	100(34.1)	99.7(42.7)	100(28.9)	100(23.2)
	CPR_{NL4-3}	**100(17)**	99.1(31.8)	**100(16.4)**	**100(20.9)**	**100(30.2)**	100(33.2)	100(38.1)	100(36.6)	**100(35.5)**	100(47.1)	100(28.6)	100(32.7)	100(39.3)
	CPR_{YU2}	**100(13.4)**	0(0)	**100(12.9)**	**100(14.8)**	**100(14.7)**	0(0)	0(0)	0(0)	**100(8.5)**	100(7.9)	98.6(7.9)	100(11.7)	100(16.9)
TenSQR	PredProp	**1**	1.6	**1**	1	1.4	**1**	**1**	**1**	**1**	1.6	2.2	1.2	0.8
	CPR_{HXB2}	**100(18.7)**	98.9(13.1)	**100()17.4**	100(9.9)	99.2(12.1)	**100(9.2)**	**100(8.1)**	**100(9.6)**	**100(7.2)**	92.8(5.9)	96.0(18)	99.0(11.5)	0(0)
	$CPR_{89.6}$	**100(18.4)**	100(19.6)	**100(20.1)**	100(17.2)	98.0(13.5)	**100(17.2)**	**100(16.7)**	**100(25)**	**100(19.3)**	94.0(15)	97.2(10.3)	100(27.8)	95.7(26)
	CPR_{JR-SCF}	**100(33.8)**	100(33)	**100(33.6)**	100(21.7)	100(20.7)	**100(24.6)**	**100(23.3)**	**100(20.5)**	**100(20.3)**	100(31.4)	98.3(33.5)	97.7(18.8)	99.8(19)
	CPR_{NL4-3}	**100(17)**	99.3(19.7)	**100(17.2)**	100(21.4)	99.5(26.7)	**100(37.7)**	**100(41.2)**	**100(38.4)**	**100(46.2)**	100(38.8)	99.8(9.2)	99.5(23.2)	99.7(42.7)
	CPR_{YU2}	**100(12.1)**	99.3(14.6)	**100(7.7)**	99.7(29.8)	99.7(14.5)	**100(11.4)**	**100(10.7)**	**100(6.5)**	**100(7.1)**	100(4.1)	94.9(10.5)	100(10.2)	98.6(12.3)
aBayesQR	PredProp	1	1	**1**	**1**	1	**1**	1	1	**1.2**	1	0.8	0.8	1.2
	CPR_{HXB2}	**100(16.3)**	99.4(21.1)	**100(22.2)**	**100(12.5)**	98.5(24.3)	**100(16.1)**	99.9(9.7)	100(9.2)	**100(16.4)**	99.6(17)	98(30.3)	0(0)	95.8(11.4)
	$CPR_{89.6}$	**100(27.1)**	98.7(17)	**100(17.3)**	**100(17.3)**	98.6(18.1)	**100(19.7)**	100(22.2)	100(20.6)	**100(16.3)**	92(10.4)	96.5(20.2)	98.9(23.7)	95.5(16.4)
	CPR_{JR-SCF}	**100(31.3)**	99.6(24.6)	**100(25.8)**	**100(29.9)**	99(21.5)	**100(22.1)**	100(20.8)	100(32.7)	**100(27)**	98.8(26.7)	97.7(21.4)	99.1(29.7)	98.2(21.1)
	CPR_{NL4-3}	**100(12.9)**	100(21.6)	**100(25.6)**	**100(20.1)**	98.9(17.7)	**100(30)**	100(39.5)	99.8(28.5)	**100(23.2)**	100(41.3)	96.3(28)	98.8(36.6)	100(31.8)
	CPR_{YU2}	**100(12.4)**	99.7(15.8)	**100(9.2)**	**100(20.3)**	99.2(18.5)	**100(12.2)**	99.5(7.9)	99.7(9)	**100(17.1)**	100(4.6)	0(0)	98.6(10.1)	99.2(14)

Predicted Proportion (PredProp) and Correct Phasing Rate (CPR (%)) for GAEseq, PredictHaplo, TenSQR and aBayesQR applied to reconstruction of HIV-1HXB2, HIV-189.6, HIV-1JR-CSF, HIV-1NL4-3 and HIV-1YU2 for all 13 genes of the HIV-1 dataset. Frequencies are reported in parenthesis.

CONCLUSIONS

- Designed a graph auto-encoder for both haplotype assembly and viral quasispecies
- Benchmarking tests on simulated and experimental data demonstrate that GAEseq achieves 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.