Basodi et al. BMC Genomics 2020, 21(Suppl 6):405
https://doi.org/10.1186/s12864-020-6661-6

BMC Genomics

METHODOLOGY Open Access

Analysis of heterogeneous genomic samples using image normalization and machine learning

Sunitha Basodi[1*], Pelin Icer Baykal[1], Alex Zelikovsky[1,2], Pavel Skums[1†] and Yi Pan[1†]

From 8th IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2018)
Las Vegas, NV, USA. 18-20 October 2018

Abstract

Background: Analysis of heterogeneous populations such as viral quasispecies is one of the most challenging bioinformatics problems. Although machine learning models are becoming to be widely employed for analysis of sequence data from such populations, their straightforward application is impeded by multiple challenges associated with technological limitations and biases, difficulty of selection of relevant features and need to compare genomic datasets of different sizes and structures.

Results: We propose a novel preprocessing approach to transform irregular genomic data into normalized image data. Such representation allows to restate the problems of classification and comparison of heterogeneous populations as image classification problems which can be solved using variety of available machine learning tools. We then apply the proposed approach to two important problems in molecular epidemiology: inference of viral infection stage and detection of viral transmission clusters using next-generation sequencing data. The infection staging method has been applied to HCV HVR1 samples collected from 108 recently and 257 chronically infected individuals. The SVM-based image classification approach achieved more than 95% accuracy for both recently and chronically HCV-infected individuals. Clustering has been performed on the data collected from 33 epidemiologically curated outbreaks, yielding more than 97% accuracy.

Conclusions: Sequence image normalization method allows for a robust conversion of genomic data into numerical data and overcomes several issues associated with employing machine learning methods to viral populations. Image data also help in the visualization of genomic data. Experimental results demonstrate that the proposed method can be successfully applied to different problems in molecular epidemiology and surveillance of viral diseases. Simple binary classifiers and clustering techniques applied to the image data are equally or more accurate than other models.

Keywords: Next-generation sequencing data, Image normalization, Staging HCV infections, Outbreaks investigations, Clustering

* Correspondence: sbasodi1@student.gsu.edu
†Pavel Skums and Yi Pan contributed equally to this work.
[1]Department of Computer Science, Georgia State University, 25 Park Place NE, Atlanta, GA 30303, USA
Full list of author information is available at the end of the article

Background

Currently, viral epidemics continue to be critical public health issues. Many emerging and long-standing epidemics are associated with small (~ 10 kilobases long) positive-sense single stranded RNA virus, such Human Immunodeficiency Virus (HIV), Hepatitis C virus (HCV), Zika virus (ZIKV) and dengue virus (DENV). The paramount feature of these viruses is their extremely high mutation rate caused by error-prone replication, which can be as high as 10^{-4} mutations per site per replication cycle [1], thus resulting in generation of all possible single point mutations in each infected individual every day. As a result, RNA viruses exist in infected hosts as highly heterogeneous populations of genomic variants usually referred to as *viral quasispecies*. Intra-host and inter-host evolution of viral quasispecies is a complex phenomenon de defined by numerous factors such as virulence, infectivity, drug resistance, immune escape, transmission rates, behavorial patterns and otherphenotypic and epidemiological numerous factors such as virulence, infectivity, drug resistance, immune escape, transmission rates, behavorial patterns and other phenotypic and epidemiological features, which plays crucial role in disease progression and outcome of infection [2–6]. Challenges associated with understanding complex quasispecies evolution attracted many researchers in different domains, including virology, epidemiology, population genetics and systems biology.

Analysis of heterogeneous viral populations is one of the most challenging bioinformatics tasks owing both to the complexity of the underlying algorithmic problems and features and sheer amount of data [7, 8]. These challenges became especially complicated in the recent decade with the advent of high-throughput sequencing (HTS), which has now become a major tool for viral research, allowing to sample viral populations at unprecedented depth [9–15]. Modern computational virology continues mostly to rely on classical approaches, which include sequence analysis, phylogenetics/phylodynamics and structural bioinformatics [8, 16]. In the recent years, these approaches started to be complemented with the network analysis [17–19].

Significant number of computational molecular epidemiology problems could be defined using phylogenetics or clustering-based objective. These problems include inference of transmission clusters, detection of co-infections, therapy outcome prediction, infection staging and other research and medical questions. Such problems could be tackled by powerful methods of machine learning and deep learning. It should be expected that in the near future, in accordance with the general trend in AI and Computer Science research, machine learning and deep learning techniques will be utilized in viral research on a much wider scale.

In order to employ machine learning for viral studies, quasispecies populations should be transformed into feature vectors from a multidimensional euclidean space. Several encoding schemes have been used in the literature for transforming biomedical data into numerical data for machine learning [20]. However, the existing methods face significant challenges when applied to viral genomic data. These challenges are associated with extremely high heterogeneity of intra-host viral populations, sequencing errors and sampling biases, robustness to noise and difficulty of selection of relevant sets of features.

Contribution

In this work, we propose a novel method converting genomic data into images, which are then used for classification and clustering. The new approach allows to utilize a well-developed machine learning methodology from the domain of image processing in genomic analysis. The proposed scheme provides the data structure for the representation of intra-host population structure which is compact, easily adjustable, robust to technological noise and sampling bias, preserve structural properties of populations and can be used for a variety of classification problems, where machine learning is applicable.

We validated our approach by applying image processing techniques to two important molecular epidemiology problems. The first problem is the HCV infection staging, i.e. distinguishing between recent and chronic infections using viral sequences sampled by next-generation sequencing (NGS). It is known that in 80% of untreated cases HCV infection turns into a chronic infection leading to severe health problems such as liver cirrhosis and hepatocellular carcinoma (a form of liver cancer). HCV infection often does not manifest any clinical symptoms in its early stages, which impedes the timely diagnosis of disease. Furthermore, currently there are no diagnostic assays to determine the stage of HCV infection. Therefore, distinguishing recently infected patients from chronically infected patients using non-invasive methods such as analysis of genomic data would be highly important both for personalized therapeutic purposes and for epidemiological surveillance; e.g., for detection of incident HCV cases.

The second problem is the detection of outbreaks using NGS data. In molecular epidemiology, it is common to utilize the observation that viral populations from the same outbreak are genetically related. Thus, measures of genetic relatedness could be used as a predictor for epidemiological relatedness [21–23]. In other words, this problem could be considered as the problem of clustering of intra-host viral populations. Until recently, most available tools for outbreak investigations

analyzed only a single representative sequence per population (usually consensus sequence) [21, 23]. Although several recently published tools allow to take into account entire intra-host populations [18, 19, 22, 24], the problem of comparison and clustering of viral populations still remains challenging [25].

We demonstrate that classification and clustering techniques based on normalized image representations of intra-host viral populations allow to solve these two problems with high accuracy.

Methods

Data collection

Intra-host HCV populations sampled by sequencing of a highly heterogeneous genomic region (HVR1) are analyzed. The analyzed region of length 264 bp, which includes HVR1, has been sequenced using the GS FLX System and the GS FLX Titanium Sequencing Kit (454 Life Sciences, Roche, Branford, CT). Obtained sequences were processed using the error correction and haplotyping algorithm KEC [26], and the obtained haplotypes were aligned using Muscle [27]. The data [16, 28] used for classification of intra-host HCV populations as recent and chronic consists of 365 NGS samples, including 108 datasets corresponding to recently infected hosts and 257 datasets belonging to chronically infected hosts. Recent samples either belong to patients with the known times since seroconversion, or to the collection of HCV outbreaks, where epidemiological investigations revealed that secondary cases were infected within few months from the dates of sample collection, thus allowing to classify them as recently infected. Chronic samples are obtained from several molecular surveillance studies. For clustering and identication of outbreaks, we use the benchmark dataset [18, 19, 22] that consists of HCV intra-host populations collected from 335 infected individuals in 2008-2013. Of these, 142 HCV samples belong to epidemiologically curated outbreaks involving from 2 to 19 patients, while the remaining datasets are epidemiologically isolated cases.

Sequence image normalization

We transform sequence data into an image by the preprocessing method further referred to as Sequence Image Normalization. We assume that sequences are aligned and ordered by their counts, with sequences of the same counts being sorted lexicographically. Next, each symbol l 2 $f^0A^{0;0}$ $C^{0;0}$ $T^{0;0}$ $G^{0;0}$ ^0g is associated with a particular color thus transforming the sequence alignment into an image. Finally, the images corresponding to different infected hosts are normalized by transforming them into fixed size images. The colors to represent nucleotides are selected from the set of colors of higher variation in order to simplify identification of discriminative size images. The colors to represent nucleotides are selected from the set of colors of higher variation in order to simplify identification of discriminative features characterizing particular intra-host populations. Fig. 1 demonstrates an example of sequence image normalization output. Normalized images thus allow to captures entire viral population structure using a single data representation independent of the number of sequences and with minimum loss of existing data or introduction of artificial data.

Raw pixel data of generated images are used as features to train machine learning models for the

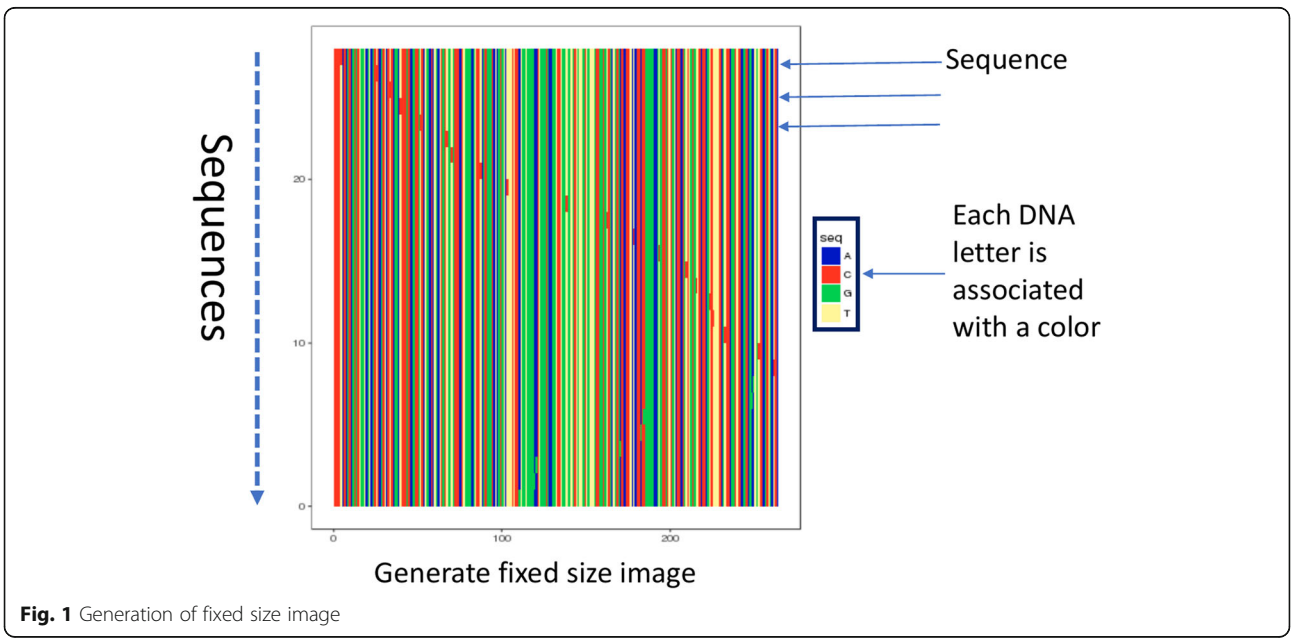

Fig. 1 Generation of fixed size image

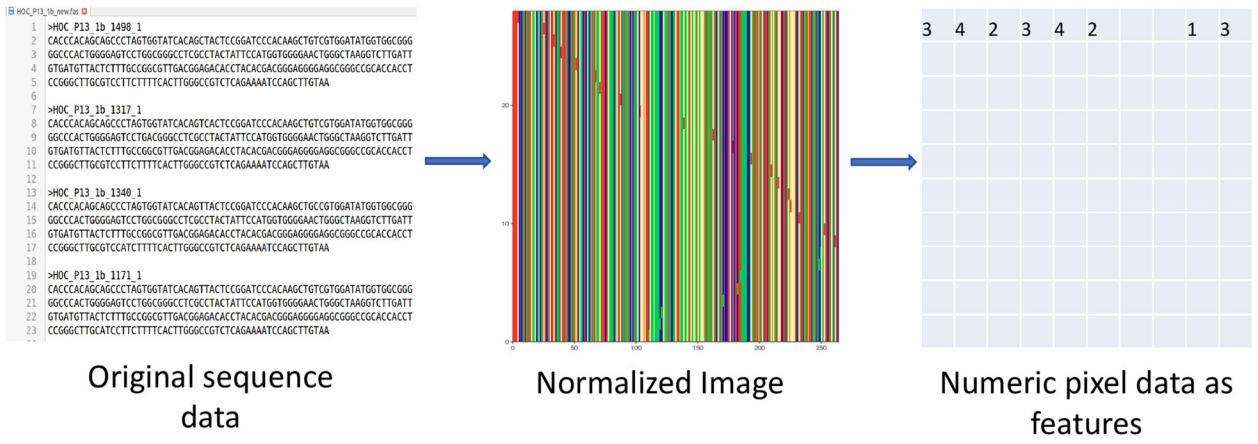

Fig. 2 Pipeline of sequence image normalization of a fasta file

consecutive analysis, as demonstrated in Fig. 2. The number of features depends on the image resolution: each image of the resolution x× y corresponds to x× y × 3 feature vector, with each pixel having 3 RGB components. In our experiments, sequence datasets have been analyzed for different resolutions ranging from 50 × 50 to 550 × 550 with the step size of 50 in each dimension. Results were generated using resolution 480 × 480 at which both models performed most accurately.

Classification of recent and chronic HCV infections

Identification of HCV infection stages is considered as a binary classification problem. Fig. 3 shows typical normalized images of HCV populations from recent and chronic infections. Visual inspection of images allows for identification of typical patterns associated with both classes - images of recent infection have pronounced diagonal lines while chronic images are choppy.

Images corresponding to intra-host viral populations have been labeled based on the stage of infection as recent or chronic and used to train the following machine learning classification models: Stochastic Gradient Descent (SGD), Decision Tree, Gaussian Naive Bayes (Gaussian NB), Linear Support Vector Machine (Linear SVM), Random Forest and *k*-Nearest Neighbours (*k*NN). We used models' implementations from python *scikit-learn* library [29]. Different SVM kernels have been explored of which SVM with linear kernel produced the best results. In linear SVM model, there is a regularization parameter *c* which helps in generalizing the model by controlling testing and training errors. In this model, grid search is performed on *c* values in the range [-2, 20]. For kNN models, we selected the best model among the models with euclidean and manhattan metrics and with *k* from the range [3, 20]. For random forest, the best model has been chosen by performing grid search on the number of trees in the range [10, 100].

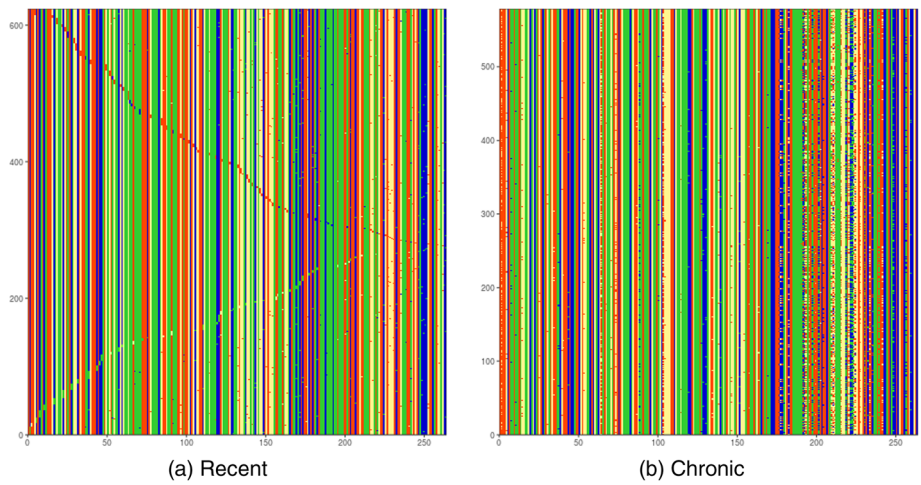

Fig. 3 Examples of normalized images of intra-host populations from (a) recent HCV infection and (b) chronic HCV infection

Trained classifiers have been validated based on their accuracy, area under the curve (AUC), precision, and recall. Accuracy (Acc) is defined as the proportion of test cases correctly classified, as either recent or chronic. Precision (Prec) measures the fraction of the correctly classified populations within each predicted infection class, while recall (Rec) measures the fraction of the true recent or chronic populations that are correctly predicted. Validation has been performed via stratified 10-fold cross-validation. Specifically, in addition to the standard 10-fold cross-validation, we employ "leave-one-outbreak-out" cross-validation and random undersampling methods to balance the datasets. In our current data, some of the samples come from the same HCV outbreak. Such samples are close to each other by their nucleotide composition, thus their presence may lead to over-fitting of any particular method. In "leave-one-outbreak-out" cross-validation, data from each of these outbreaks was used in the validation set, while other samples are used in the training sets. Random undersampling has been performed to balance the difference in sizes of datasets of recent and chronic hosts. In this method, chronic dataset size is reduced by random subsampling to match the recent dataset size.

Clustering of intra-host viral populations from outbreaks

We cluster images representing intra-host viral populations into transmission clusters using standard clustering algorithms {agglomerative hierarchical clustering, *k*-means clustering and mini-batch *k*-means clustering. As before, we used models' implementations from python *scikit-learn* library [29]. Several distance measures have been employed, including euclidean, manhattan and cosine metrics. Hierarchical clustering has been executed using complete, average and ward linkage approaches.

Normalized Mutual Information (NMI) [30], homogeneity [31] and completeness [31] scores as used as metrics to analyze the clustering performance. These measures evaluate the assigned cluster labels after clustering compared to the actual cluster class label of each intra-host viral population. Homogeneity score measures if the all members of a cluster actually belong to one cluster class label, while the completeness scores measure if all the members of an actual cluster class label are grouped into the same cluster. NMI measures the mutual information shared between the individuals in the clusters. All these measures range from 0 to 1 and the values closer to 1 refer to better clustering efficiency. To evaluate the effectiveness of the normalization method in detecting relatedness between any pair of samples, we compute AUROC (Area under ROC curve) is computed (as done in [18]). Viral populations taken from the same outbreak are considered as genetically related, otherwise as unrelated. There are 55,945 pairs of samples, and 479 of them are related. After computing the distances between each pair of samples, all the pairs crossing a threshold value are considered as related. To compute AUROC curve, false-postive rate (FPR) and true-postive rate (TPR) are measured by modifying the threshold starting from the best threshold value where there are no false positives.

Results

Classification of infection stages

Stratified 10-fold cross-validation has been initially performed to analyze the performance of several classification methods trained using the normalized image data. Fig. 4 shows accuracy and AUC of the best models for each of the methods using box plots, with the average metrics being indicated by red line. Linear SVM demonstrated superior performance compared to all other models, with an average accuracy of 97.545% and low accuracy variance. Other models with the exception of Gaussian NB have accuracy greater than 85%, thus exceeding accuracy of existing methods, which are primarily based on feature extraction methods (see Comparison with previous methods subsection). Accuracy metric alone cannot define performance of the

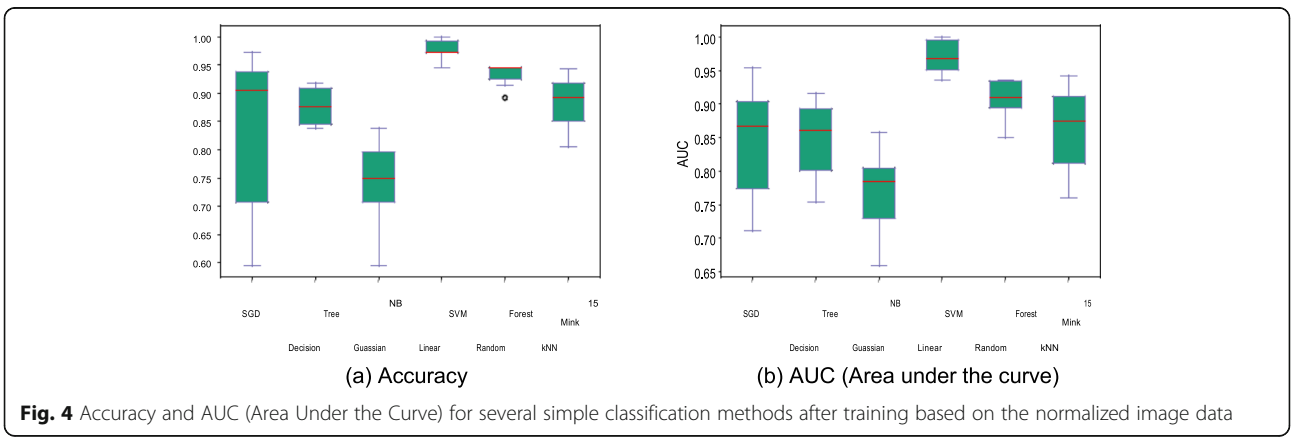

Fig. 4 Accuracy and AUC (Area Under the Curve) for several simple classification methods after training based on the normalized image data

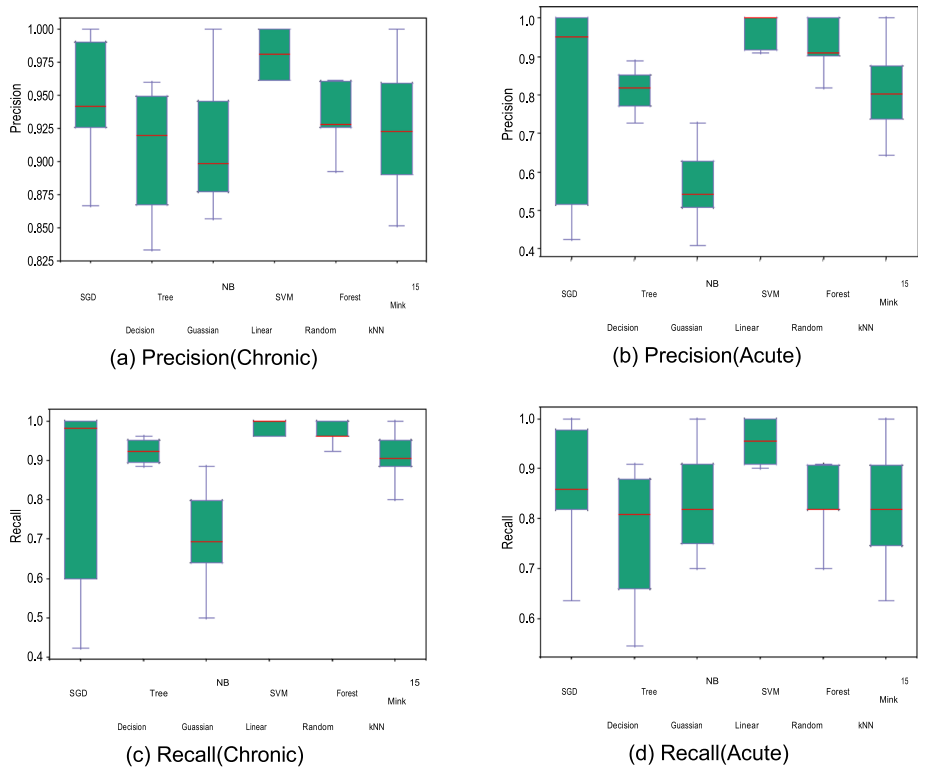

Fig. 5 Precision and Recall comparisons of several simple classification methods after training based on the image data generated using sequence image preprocessing method

model as it needs to achieve higher precision and recall metrics for each infection type as well. Fig. 5a-d demonstrate the precision and recall metrics for chronic and recent samples separately. As before, linear SVM achieves the best performance over all other models with an average precision and recall values of 98.11 and 98.45% for chronic populations and 96.52 and 95.36% for recent populations, respectively. This model also has low variance across the values obtained from all the folds. Noticeably, other models with the exception of Gaussian NB also achieve more than 80% values for these metrics.

Linear SVM model has been analyzed further with leave-one-outbreak-out and random undersampling validation combined with 10-fold cross-validation. Table 1 shows the results of these methods compared to the standard 10-fold cross-validation on the whole dataset. The classification accuracy remains stable under the additional sampling methods.

Detection of transmission clusters

The results of *k*-means, mini-batch *k*-means and hierarchical clustering models are shown in Table 2. In our experiments, agglomerative hierarchical clustering with ward linkage and euclidean distance between images demonstrated the best performance. Furthermore, we evaluated the accuracy of detection of epidemiologically related pairs. Two intra-host viral populations are considered to be related, if the distance between corresponding images is below a specified. threshold. ROC curves for the accuracy of detection of epidemiologically related pairs for different distance measures and thresholds are shown in Fig. 6. All distance measures produced

Table 1 Performance metrics of Linear SVM classi er assessed by standard 10-fold cross validation, leave-one-outbreak-out validation and random undersampling methods

Sampling Methods	Accuracy	Precision-Chronic	Precision-Recent	Recall-Chronic	Recall-Recent	AUC
Standard 10-fold cross-validation	97.545%	98.105%	96.515%	98.446%	95.364%	96.905%
Leave-one-outbreak-out	96.075%	97.004%	91.0%	98.446%	83.5%	90.973%
Random undersampling	95.164%	96.328%	94.661%	94.155%	96.173%	95.164%

Table 2 Performance metrics of various clustering methods

Clustering Method	NMI	homogeneity	completeness
k-means	0.986	0.994	0.978
Mini-batch *k*-means	0.985	0.992	0.978
Hierarchical	0.987	0.994	0.979

consistent results, with AUC exceeding 0.99 for all of them.

Effect of image resolution

All experimental results discussed above have been obtained using the default image resolution 480 × 480. We analyzed impact of image resolution on the classification and clustering performance. Resolution values varied from 50 × 50 to 550 × 550 with step size of 50. Fig. 7a shows the performance metrics of stratified. 10-fold cross-validation using LinearSVM model for detecting stage of HCV infections based on different image resolutions. Highest accuracy is achieved at the resolution 450 × 450, although the accuracy mostly saturates approximately after the resolution 300 × 300. Similar performance has been observed for agglomerative hierarchical clustering (Fig. 7b).

Comparison with previous methods

A previously published model [32] classifies stages of HCV infection using one of the following 3 parameters: variant frequencies entropy, average position-wise nucleotide entropy and the average distance from viral variants to the most frequent variant of the population. In our data, AUC for these parameters was equal to ~81, ~66 and ~78%, respectively, while the proposed classifier based on image normalization yielded ~96.9% AUC.

We also compared clustering sensitivity and AUROC (of the inference of genetic relatedness between a pair of HCV samples) for our method and consensus-based approach (see e.g. [9, 33]) for the two population based methods VOICE and ReD proposed in [18]. The consensus-based method compares intra-host viral population using one representative sequence per population, which is most often the consensus sequence, while VOICE and ReD methods analyze whole quasispecies populations. Consensus algorithm achieves clustering sensitivity of 93.94% and AUROC (genetic relatedness) of 98.7%. ReD method achieves clustering sensitivity of 96.3% and VOICE method achieves clustering sensitivity of 98.2% and AUROC (genetic re-latedness) of ~99%. Image clustering method achieves sensitivity of 98.181% and AUROC of 99.2% which are higher values than consensus and ReD methods and has similar performance to the VOICE algorithm.

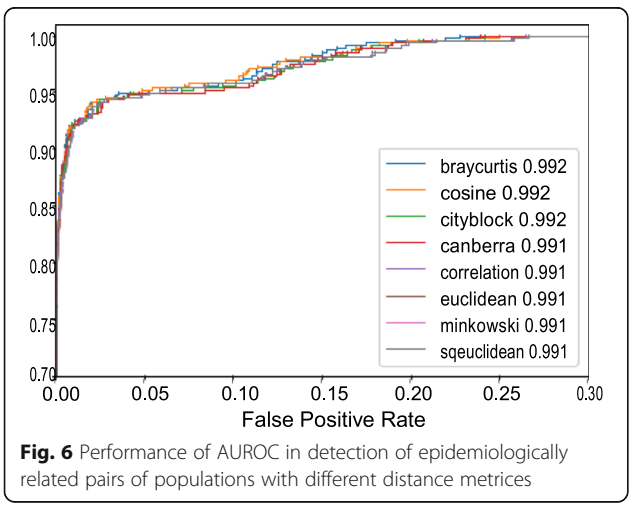

Fig. 6 Performance of AUROC in detection of epidemiologically related pairs of populations with different distance metrices

Discussion

The sequence-image normalization method described here provides a way to transform genomic data into image data which can be directly employed by machine learning methods. The proposed preprocessing method was specifically designed to addresses multiple challenges that currently impede applications of machine learning and deep learning methods to viral studies. These challenges could be thematically classified as follows:

Challenges associated with technological limitations

High-throughput sequencing technologies are prone to errors and biases, which may significantly affect viral data. Indeed, frequencies of minor viral variants are often comparable with the level of sequencing noise; however, such variants should not be simply discarded based on some frequency threshold, since often they are the ones responsible for transmission, immune escape or therapy failure [3, 5, 6, 34–36]. Presence of sequencing errors introduces noise to data and produces outlier viral variants, which negatively affect the quality and accuracy of machine learning classifiers.

Another important problem is sampling and sequencing bias resulting in significant irregularities in the number and length of viral sequences from different infected individuals. If classifiers capture these artificial differences as significant associations, it may result in overfitting and decline of accuracy. Thus, application of machine learning to heterogeneous viral population data should be preceded by a preprocessing step to eliminate these irregularities via normalization procedure. However, selection of an appropriate normalization approach is challenging. For instance, if we use text classification techniques for preprocessing, difference in the number of sequences among different files needs to be controlled either by truncation or padding. This preprocessing, however, causes data loss (in case of truncation) or introduces irrelevant data (in case

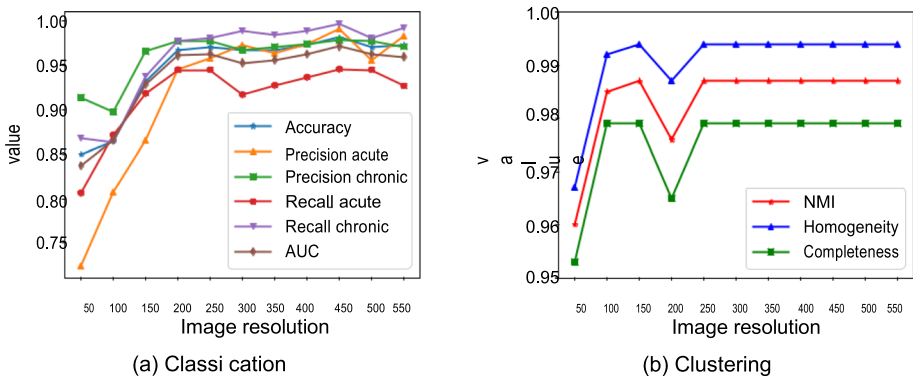

Fig. 7 Performance metrics (Y-axis) of classification and clustering methods based on different image resolutions(X-axis)

of padding). An optimal preprocessing method should not introduce such issues.

Challenges associated with feature selection and feature extraction

Before applying machine learning methods to classification of heterogeneous viral populations, genomic data should be mapped into the euclidian space R^n. It is usually achieved by identifying numerical features that are relevant to the problem under consideration. They can include various diversity measures [32], popula-tion genetic parameters [37], physico-chemical properties [16] and other parameters specifically tailored to particular problems. These features are generally identified in consultation with domain experts. Selection of the most relevant features is daunting and resource-consuming. A role of feature selection in determining classification performance is paramount. Selection of a limited number of features from certain domains inevitably results in loss of information, while increase of feature space dimensionality increases risk of overfitting and compromises the algorithm's scalability.

An optimal feature selection method should be able to capture the entire population structure using a relatively simple and easily contractible data representation. Furthermore, it should use a standard universal data format, which has a fixed number of features and is applicable to different problems. Since genomic data is essentially a textual information, it is tempting to utilize well-developed machinery from the text classification domain [38, 39] for the purpose of construction of such representation. Viral populations could be mapped to a euclidian spase using word2vec approaches [40], and classified using various available deep learning models [38, 39]. However, application of text processing approaches to viral research could be impeded by several factors. Since they are based on deep neural network models with large numbers of hyperparameters, it requires large annotated datasets to train these models. However, in molecular epidemiology, the amount of available training data is usually limited in comparison with the text processing domain. The datasets of several hundred intra-host viral populations analyzed in this paper are typical in this context. Although, word2vec or document embedding methods can be directly employed, it is challenging to train a model to get a higher classi classification performance. Furthermore, since viral haplotypes are unique, the trained model could overfit the data.

Challenges associated with data comparison

Clustering of intra-host viral populations requires an inter-population distance measure, which takes into account complex population structures. It has been shown that among simple alignment-based population distance measures, the minimal distance between population variants allows to achieve the highest clustering accuracy [41]. However, this measure is sensitive to noise and presence of outliers, and does not take into account the whole population structure. Recently, several simulation-based and network-based distance measures have been proposed [18, 19], which overcome above-mentioned limitations at the cost of lesser scalability. Thus, the universal, accurate and efficiently computable inter-population distance measure, which takes into account complex population structures still has to be developed.

Our proposed preprocessing method converts the viral population genomic data sampled by NGS into a scaled image. Irregularities in the data are thus handled by generating a fixed size image. The number of features in this case remains same. Therefore, it can be directly used for machine learning applications without any explicit feature selection methods. High accuracy of machine learning classification and clustering techniques based on image representation applied to several molecular epidemiology tasks signifies validity of our approach. The case of infection staging is particularly illustrative. Previous studies demonstrated that diversity of intra-host viral populations often increases with progression of

HCV infection [28, 32, 37]. In addition to immune escape, which is usually responsible for the diversity increase, complex adaptation mechanisms get engaged during intra-host HCV evolution, such as antigenic cooperation [6], which may result in increase of negative selection and selection of viral variants with particular properties, allowing HCV to survive in host environment for prolonged periods of time [17, 42–45]. The major features of such evolutionary processes include (but not limited to) low DN/DS ratio, skewed distributions of physico-chemical properties and presence of particular sequence motifs [16, 17, 37]. These and other features can be taken into account by inclusion of the features based on various genomic and biochemical parameters into machine learning classifiers. However, most of them are already implicitly included into the image representations, and thus are taken into account when the image-based classifiers are trained. It allowed us to achieve a high classification accuracy. In future work, sequence image normalization machinery can be applied to other challenging problems in viral genomics, such as detection of co-infections and prediction of drug resistance and therapy outcome.

Conclusions

Here, we propose a novel method for generation of a fixed set of features representing heterogeneous viral populations, which is widely applicable for various classification and clustering tasks addressed by machine learning. The method converts sequence data into fixed-size images, thus reducing several issues associated with comparison of viral populations by machine learning methods. Simplicity of the sequence image normalization method allows for a robust conversion of genomic data into numerical data. Image data also help in visualization of genomic data. Experimental results demonstrate that the proposed method can be successfully applied to different problems in molecular epidemiology and surveillance of viral diseases. Simple binary classifiers and clustering techniques applied to the image data are equally or more accurate than other models.

Acknowledgements

We sincerely thank Dr. Yury Khudyakov from Centers for Disease Control and Prevention (CDC) for providing data and helping to improve the presentation of the manuscript.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 6, 2020: Selected articles from the 8th IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2018): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-6.

Authors' contributions

SB designed and implemented algorithms, analyzed the data and wrote the paper; PIB designed algorithms and analyzed the data; AZ designed the algorithms; PS designed algorithms, analyzed the data, wrote the paper and supervised the project; YP designed algorithms, analyzed the data, wrote the paper and supervised the project. All authors read and approved the final manuscript.

Funding

PS was partially supported by NIH grant 1R01EB025022. A.Z. has been partially supported by NSF Grants DBI-1564899 and CCF-1619110 and NIH Grant 1R01EB025022. SB and PIB were supported by GSU Molecular Basis of Disease fellowship. The funding bodies have not played any roles in the design of the study and collection, analysis and interpretation of data in writing the manuscript. Publication costs are funded by NIH grant 1R01EB025022.

Availability of data and materials

The data used in this paper has been published in [13] and partially in [8]. It can be shared by reasonable request. The developed software is freely available at https://github.com/compbel/SequenceImageNormalization

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All the authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, Georgia State University, 25 Park Place NE, Atlanta, GA 30303, USA. [2]The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow 11991, Russia.

Received: 5 March 2020 Accepted: 9 March 2020
Published: 21 December 2020

References

1. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010;84(19):9733–48.
2. Apostolou A, Bartholomew ML, Greeley R, Guilfoyle SM, Gordon M, Genese C, Davis JP, Montana B, Borlaug G. Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011. MMWR Morb Mortal Wkly Rep. 2015;64(7):165–70.
3. Campo DS, Skums P, Dimitrova Z, Vaughan G, Forbi JC, Teo C-G, Khudyakov Y, Lau DTY. Drug resistance of a viral population and its individual intrahost variants during the rst 48 hours of therapy. Clin Pharmacol Ther. 2014;95(6):627–35.
4. Lengauer T, Rahnenfuehrer J, Roomp K, Beerenwinkel N, Sing T. Computational methods for the design of e ective therapies against drug resistant HIV strains. Bioinformatics. 2005;21:3943–50.
5. Rhee S-Y, Liu TF, Holmes SP, Shafer RW. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. PLoS Comput Biol. 2007;3(5):e87.
6. Skums P, Bunimovich L, Khudyakov Y. Antigenic cooperation among intrahost hcv variants organized into a complex network of cross-immunoreactivity. Proc Natl Acad Sci. 2015;112(21):6653–8.
7. Astrovskaya I, Mancuso N, Tork B, Mangul S, Artyomenko A, Skums P, Ganova-Raeva L, Mandoiu I, Zelikovsky A, Park MD. Inferring viral quasispecies spectra from shortgun and aplicon next-generation sequencing reads. Genome Anal Curr Proced Appl. 2014.
8. Marz M, Beerenwinkel N, Drosten C, Fricke M, Frishman D, Hofacker IL, Mann DH, Middendorf M, Rattei T, Stadler PF, et al. Challenges in rna virus bioinformatics. Bioinformatics. 2014;30(13):1793–9.
9. Bartlett SR, Wertheim JO, Bull RA, Matthews GV, Lamoury FMJ, Scheffler K, Hellard M, Maher L, Dore GJ, Lloyd AR, et al. A molecular transmission network of recent hepatitis c infection in people with and without hiv: Implications for targeted treatment strategies. J Viral Hepat. 2017;24(5):404–11.
10. Skums P, Mancuso N, Artyomenko A, Tork B, Mandoiu I, Khudyakov Y, Zelikovsky A. Reconstruction of viral population structure from next-generation sequencing data using multicommodity ows. BMC Bioinformatics. 2013;14(Suppl 9):S2. https://link.springer.com/article/10.1186/1471-2105-14-S9-S2#citeas.

11. Collier MG, Khudyakov YE, Selvage D, Adams-Cameron M, Epson E, Cronquist A, Jervis RH, Lamba K, Kimura AC, Sowadsky R. Outbreak of hepatitis a in the usa associated with frozen pomegranate arils imported from turkey: an epidemiological case study. Lancet Infect Dis. 2014;14(10): 976–81.
12. Grabowski MK, Redd AD. Molecular tools for studying hiv transmission in sexual networks. Curr Opin HIV AIDS. 2014;9(2):126–33.
13. Hellinger WC, Bacalis LP, Kay RS, Thompson ND, Xia G-L, Lin Y, Khudyakov YE, Perz JF. Health care associated hepatitis c virus infections attributed to narcotic diversion. Ann Intern Med. 2012;156(7):477–82.
14. Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y, othes. Characterization of quasispecies of pandemic 2009 in uenza a virus (a/h1n1/2009) by de novo sequencing using a next-generation dna sequencer. PLoS One. 2010;5(4):e10256.
15. Seña AC, Moorman A, Njord L, Williams RE, Colborn J, Khudyakov Y, Drobenuic J, Xia G-L, Wood H, Moore Z. Acute hepatitis b outbreaks in 2 skilled nursing facilities and possible sources of transmission north carolina, 2009, 2010. Infect Control. 2013;34(07):709–16.
16. Lara J, Teka M, Khudyakov Y. Identication of recent cases of hepatitis c virus infection using physical-chemical properties of hypervariable region 1 and a radial basis function neural network classi er. BMC Genomics. 2017;18(10): 880.
17. David S, Campo ZD, Yamasaki L, Skums P, Lau DTY, Vaughan G, Forbi JC, Teo C-G, Khudyakov Y. Next-generation sequencing reveals large connected networks of intra-host hcv variants. BMC Genomics. 2014;15(Suppl 5):S4.
18. Glebova O, Knyazev S, Melnyk A, Artyomenko A, Khudyakov Y, Zelikovsky A, Skums P. Inference of genetic relatedness between viral quasispecies from sequencing data. BMC Genomics. 2017;18(10):918.
19. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, Mandric I, Ramachandran S, Campo D, Jha D, et al. Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 2017;34(1):163–70.
20. Yu N, Li Z, Yu Z. Survey on encoding schemes for genomic data representation and feature learning|from signal processing to machine learning. Big Data Min Analytics. 2018;1(3):191–210.
21. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL. The global transmission network of hiv-1. J Infect Dis. 2014;209(2):304–13.
22. Campo DS, Xia G-L, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, Punkova L, Ramachandran S, Thai H, Skums P, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. J Infect Dis. 2015;213(6): 957–65.
23. Wertheim JO, Kosakovsky Pond SL, Forgione LA, Mehta SR, Murrell B, Shah S, Smith DM, Scheer K, Torian LV. Social and genetic networks of hiv-1 transmission in New York city. PLoS Pathog. 2017;13(1):e1006000.
24. Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, Gall A, Cornelissen M, Fraser C. The Maela Pneumococcal Collaboration STOP-HCV Consortium, and The BEEHIVE Collaboration. Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. Mol Biol Evol. 2017;35(3):719–33.
25. Gunthard HF, Kouyos R. Can directionality of hiv transmission be predicted by next generation sequencing data? J Infect Dis. 2018.
26. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y. E cient error correction for next-generation sequencing of viral amplicons. BMC Bioinformatics. 2012;13:S6. BioMed Central.
27. Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
28. Astrakhantseva IV, Campo DS, Araujo A, Teo C-G, Khudyakov Y, Kamili S. Di erences in variability of hypervariable region 1 of hepatitis c virus (hcv) between acute and chronic stages of hcv infection. In Silico Biol. 2011;11(5): 163–73.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
30. Strehl A, Ghosh J. Cluster ensembles|a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3(Dec):583–617.
31. Rosenberg A, Hirschberg J. V-measure: A conditional entropy-based external cluster evaluation measure. In: Proc 2007 Joint Conf Empir Methods Nat Lang Process Comput Nat Lang Learn (EMNLP-CoNLL); 2007. p. 410–20.
32. Montoya V, Olmstead AD, Janjua NZ, Tang P, Grebely J, Cook D, Harrigan PR, Krajden M. Di erentiation of acute from chronic hepatitis c virus infection by nonstructural 5b deep sequencing: A population-level tool for incidence estimation. Hepatology. 2015;61(6):1842–50.
33. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL. The global transmission network of hiv-1. J Infect Dis. 2013;209(2):304–13.
34. Nabel GJ, Douek DC, Kwong PD. The rational design of an AIDS vaccine. Cell. 2006;124:677–81.
35. Fischer GE, Schaefer MK, Labus BJ, Sands L, Rowley P, Azzam IA, Armour P, Khudyakov YE, Lin Y, Xia G. Hepatitis c virus infections from unsafe injection practices at an endoscopy clinic in las vegas, nevada, 2007, 2008. Clin Infect Dis. 2010;51(3):267–73.
36. Holland JJ, De La Torre JC, Steinhauer DA. RNA virus populations as quasispecies. Curr Top Microbiol Immunol. 1992;176:1–20.
37. Baykal PI, Artyomenko A, Ramachandran S, Khudyakov Y, Zelikovsky A, Skums P. Assessment of hcv infection stage as recent or chronic using multi-parameter analysis and machine learning. In 2017 IEEE 7th Int Conf Comput Adv Bio Med Sci (ICCABS). 2017. 1. IEEE.
38. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for e cient text classi cation. ArXiv Preprint ArXiv. 2016;1607:01759.
39. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classi cation. In AAAI. 2015;333:2267–73.
40. Mikolov T, Chen K, Corrado G, Dean J. E cient estimation of word representations in vector space. ArXiv Preprint ArXiv. 2013:1301–3781.
41. Campo DS, Xia G-L, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, Punkova L, Ramachandran S, Thai H, Skums P, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. J Infect Dis. 2016;213(6): 957–65.
42. Ramachandran S, Campo DS, Dimitrova ZE, Xia G-L, Purdy MA, Khudyakov YE. Temporal variations in the hepatitis c virus intrahost population during chronic infection. J Virol. 2011;85(13):6369–80.
43. Lu L, Tatsunori N, Li C, Waheed S, Gao F, Robertson BH. Hcv selection and hvr1 evolution in a chimpanzee chronically infected with hcv-1 over 12 years. Hepatol Res. 2008;38(7):704–16.
44. Palmer BA, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis c virus mapped by ultradeep pyrosequencing. J Virol. 2014;88(23):13709–21.
45. Gismondi MI, Carrasco JMD, Valva P, Becker PD, Guzman CA, Campos RH, Preciado MV. Dynamic changes in viral population structure and compartmentalization during chronic hepatitis c virus infection in children. Virology. 2013;447(1):187–96.

Publisher's Note