Curriculum representation learning of genomes in cancer and healthy populations

Jacob Krol

1 Abstract

Rare cancers have few similar cases for comparative analysis to aid in diagnosis, treatment, and outcome prediction. And, genefusions are a subset of DNA mutations which commonly drive cancer progression forward. Therefore, if the genome of a patient with a rare cancer could be searched against a population database of other patient's genefusion mutations, then comparative analysis becomes feasible. However, no public databases currently exists which encode patient-wise genefusions for large populations. Furthermore, human curated genefusion databases mostly and are unlikely to house rare genefusion driver mutations. To address this gap, a population-scale genefusion search method is explored by embedding genefusion structural variants (GSVs) of a mixed population of cancer and healthy individuals with a siamese neural network (SNN) trained by curriculum learning. The SNN embeddings support fast, online vector databases for applications for practical use in hospital settings.

2 Introduction

Rare cancers are poorly understood and difficult to treat since few similar cases exist for comparative analysis. Finding the genetic mutations driving cancer progression can help clinicians choose the correct therapy for patients. Searching for genetically similar patients should ideally be fast and a practical database must readily support adding new individuals. However, genome sequencing data is exceptionally large which motivates computational methods to effectively encode human genomes to support fast similarity searches. Intel has recently developed vector databases to support fast similarity searches ([1]). However, human genome sequencing data is not directly translatable to a vector representation for use vector databases. Since genefusions are known to drive cancer progression, human genomes will be encoded by large genetic mutations, called structural variants (SVs) which support genefusion evidence, GSVs. However, the total number of possible GSVs is roughly $\binom{27000}{2}$ which is not a reasonably sized encoding vector. Therefore, extracting low dimensional vector representations of GSV encodings is necessary for practical use. However, the sparsity and nonlinearity of GSV encodings violates assumptions made by many common dimensionality reduction techniques, such as principal component analysis (PCA). Furthermore, it is crucial to preserve pairwise distances of genome encodings in the low dimensional space to ensure the similarity search of individual genomes is accurate. To ensure pairwise distances are preserved, a siamese neural network will learn representations from GSV encodings. Siamese neural networks have been successfully used for representation learning for signature [2] and face verification [10].

No public datasets exist with individual GSV encodings, however whole genome sequencing data exists for individuals from the 1000 genomes project [4] and the the Pan-Cancer Analysis of Whole Genomes [5] (PCAWG). The individuals from 1000 genomes do not have cancer while individuals from PCAWG do. Mixing both populations can simulate a real hospital setting and evaluate the efficacy of GSV encoding similarity searches.

GSV encodings of each human genomes must be transformed into low dimensional vectors to support vector databases, and SNNs have proven to be effective for learning low dimensional embeddings from paired training examples. However, a similarity search requires ranking individuals and SNNs are usually applied to verification tasks. To address the similarity search application, a curriculum learning approach will be used to train an SNN capable of maintaining pairwise GSV encoding distances into a low dimensional space.

3 Related work

Population scale analysis of SVs required development of tools to effectively handle the large scale of datasets. GIGGLE [8] is a search tool which optimally searches for SV data using B+ trees, and STIX [3] provides an interface to GIGGLE with greater resolution searches, statistical tests, and improved metadata support. GIGGLE and STIX can capture patterns of SVs across populations, and there is an opportunity to explore whether personalized representations from GSVs are feasible for personalized medicine applications. However, the number of possible GSV is roughly $\binom{27000}{2}$ which is an impractical vector size.

Many methods exist for dimensionality reduction, yet GSV encodings are sparse and non-linear. And for practical use, the GSV embedding method must be "online" to rapidly embed new individuals. Neural networks are powerful tools for learning embeddings, and they can quickly embed new samples without re-training. However, neural network embeddings are typically a byproduct of a model trained for a downstream task, such as regression or classification. Compared to classification and regression, representation learning with neural networks has niche, yet powerful applications. Siamese neural networks have been used for representation learning in signature verification [2] and face verification [10] to capture (dis)similarity geometrically in an embedding space. The efficacy of SNNs for genomic similarity searches has been shown by GenoSiS. GenoSiS showed embeddings derived from haplotype encoding of human genomes can accurately recall the population membership of individuals [9]. Here, an SNN will learn embeddings from GSV encodings of individuals. However, learning a shared embedding space for GSVs across both cancer and healthy populations presents the challenge of de-trivializing the model's optimization to separate the two populations without much consideration of within population genomic variation. For complex learning goals, curriculum learning is a technique which trains a model on different learning objectives, often sorted ascending by difficulty, to achieve better performance on complex tasks [6]. Here, curriculum learning will enforce two curricula for learning SV representations: 1) a within population curriculum and 2) an across population curriculum.

4 Methods

4.1 Problem setup

The SSN's objective is to convert GSV encodings into low dimensional embedding representations via a symbolic curriculum learning approach.

4.2 Data

GSVs are gathered from two different populations and studies: PCAWG (cancer) and 1000 genomes project (healthy). GSV evidence is directly gathered from the short-read sequences of each study. Briefly, short-reads are small segments of the sample genome aligned to a reference, and the resulting alignments are stored in a binary alignment map (BAMs) file. For each individual, GSV evidence is computed by the excord program which reports SV evidence in the form of genomic intervals; excord results are stored in a generic genomic interval file format called browser extensible data (BED) files. Next, GIGGLE [8] builds B+ tree to index the intervals from the full corpus of BED files to support scalable searching of many genomic interval files given a query interval. This study focuses on GSVs which are a subset of all SVs where both genomic intervals must intersect with a protein coding region in the reference genome. Each individual's genome receives a final encoding based on the count of GSVs recovered from sequencing data.

4.3 Siamese neural network (distributed component)

The SNN receives GSV encodings of individuals as input and outputs a low dimensional embedding. The SNN architecture is shallow with two hidden layers and one output layer. The size of each hidden layer is the same as the output layer, 512. For clarity, the output layer size is the embedding size.

4.4 Curriculum learning (symbolic component)

In curriculum one, the SNN learns the pairwise distances by penalizing the difference between GSV encoding distance, $||x_i - x_j||_2$, and the embedding distance $||x'_i - x'_j||_2$. Pairwise distance error is aggregated for each batch into mean squared error.

$$\text{MSE} = \frac{1}{N} \sum_{(i,j)} \left(||x'_i - x'_j||_2^2 - ||x_i - x_j||_2^2 \right)^2$$

The goal is to learn within population distances in curriculum one. Therefore, the training

set is downsampled to only include paired GSV encodings from individuals in the same population: healthy-healthy or cancer-cancer, not healthy-cancer (Figure 3).

Curriculum two trains the SNN to contrast between embeddings of different populations using triplet loss.

$$L(A,P,N) = max(||x'_A - x'_P||_2^2 - ||x'_A - x'_N||_2^2 + \alpha, 0)$$

where x'_A, x'_P, and x'_N are the embeddings of an anchor, positive, and negative such that anchor and positive are sample from the population, while the negative sample is from a different population.

GSV encodings for cancer are significantly different from healthy populations which suggests curriculum two will be easier to train than curriculum one. The SNN will be evaluated with all possible curricula permutations: C1, C2, C1+C2, and C2+C1.

5 Evaluation

Root mean squared error (RMSE) of the ground truth pairwise distances and the pairwise distances in the embedding space measures the conservation of relative distances in each model's embedding space. RMSE is computed solely on a test dataset consisting of pairs not included during training. The original l2 distance of GSV encodings serves as the ground truth. Loss is calculated by subtracting the squared l2 distance of pairs in the embedding space from the original squared l2 distance. The error is aggregated across all test pairs by computing RMSE. Since the number of pairs is very large, the log10(RMSE) is reported. Unfortunately, a mistake was made when computing RMSE. Mistakenly, the l2 distances of paired individuals were squared before taking the difference between ground truth and predicted. This error can increase the magnitude of error significantly for certain pairs and may effect the overall model rankings.

5.1 Baseline

A johnson-lindenstrauss (JL) transformation is chosen as a baseline method since JL is ideal for large dimensional datasets with relatively small sample sizes. The JL lemma [7] provides an upper bound on pairwise distance error parameterized by the number of samples, n, and an error threshold, ϵ.

$$(1-\epsilon)||x-y||_2^2 \leq ||f(x)-f(y)||_2^2 \leq (1+\epsilon)||x-y||_2^2.$$

By defining an error threshold of $\epsilon = 0.1$, the formula $k = \lceil (4 \cdot \ln(n) \frac{1}{\epsilon^2/2 - \epsilon^3/3} \rceil$ was used to compute k, the size of the JL transformed embedding space.

5.2 Data curation

The data curation steps are identical for the 1000 genomes and PCAWG studies.

- Download BAMs from the project's (1000 genomes or PCAWG) public databases.
- Compute SV evidence for each BAM with excord.
- Build a GIGGLE index for all excord BEDs in the population.
- Query the GIGGLE index for protein coding genes of GRCh37.
- Intersect the GIGGLE results back with the protein coding genes of GRCh37 to get a GSVs.
- Normalize the GSV count distribution by dividing by the sum.

5.3 Model training

SNN training was done in Python 3 using Pytorch. Custom classes were defined for 1) SNN architecture, 2) loss functions, and 3) datasets. The pairwise data of 1000 genomes and PCAWG individuals were partitioned into a 70/30 train/test split. To address the large difference in sample counts between populations ($\tilde{2}$.6k in 1000 genomes and 150 in PCAWG), random splits were stratified such that 70% of individuals from both cancer and healthy populations constituted the training set, similarly for testing. Furthermore, a custom dataset class ensured each batch had a 50/50 split of population pairs and triplets. In curriculum 1, a batch constituted 50% healthy-healthy pairs and 50% cancer-cancer pairs; no cross-population individuals were provided in curriculum 1. In curriculum 2, the batches constituted 50% healthy-healthy-cancer and 50% cancer-cancer-healthy triplets where the first sample is the anchor, second sample is the positive, and the third sample is the negative.

Curriculum learning was implemented by changing the loss function after one curriculum completed a full training cycle. For example, S1-C1+C2 trained on curriculum 1, then curriculum 2. Curriculum 1 was trained for 20 epochs using a batch size of 256. Curriculum 2 was trained for 1 epoch with a batch size of 4. The reduction in epochs for curriculum 2 is due to time complexity constraints due to unoptimized batch construction code. During experiments however, curriculum 2 rapidly converged in 1 epoch.

5.4 Code availability

Model training code and data curation code is available on GitHub.

6 Experimental

6.1 Pairwise embedding distance error

SNN embeddings better preserved pairwise distances compared to a JL transformation of GSV encodings, and permutes of curricula learning show that the triplet loss curricula (curriculum 2) produced the best embeddings (Table 1). The log10(RMSE) of distorted distance shows that the SNN trained on curriculum 2 is best (log10(RMSE)=4.88), followed closely by JL (log10(RMSE)=4.95). Curriculum 1 performed the worst (log10(RMSE)=8.39), and, interestingly, training both curricula 1 and 2 in either order produced the same result (log10(RMSE=7.30)). Also, the embedding size of all SNNs was set to 512 compared to JL's embedding size of 2104.

Model	RMSE	Embedding size
Siamese-C1	8.39	512
Siamese-C1+C2	7.3	512
Siamese-C2+C1	7.3	512
JL ($\epsilon = 0.1$)	4.95	2104
Siamese-C2	**4.88**	512

Table 1: Model performance and embedding size. C1 and C2 are curricula 1 and 2. RMSE is $\log_{10}$ transformed.

.

6.2 t-SNE

t-distributed stochastic neighbor embedding (t-SNE) plots show that curricula 1 and JL accurately group healthy and cancer individuals while curricula 2 segments the healthy individuals (Figure 1).

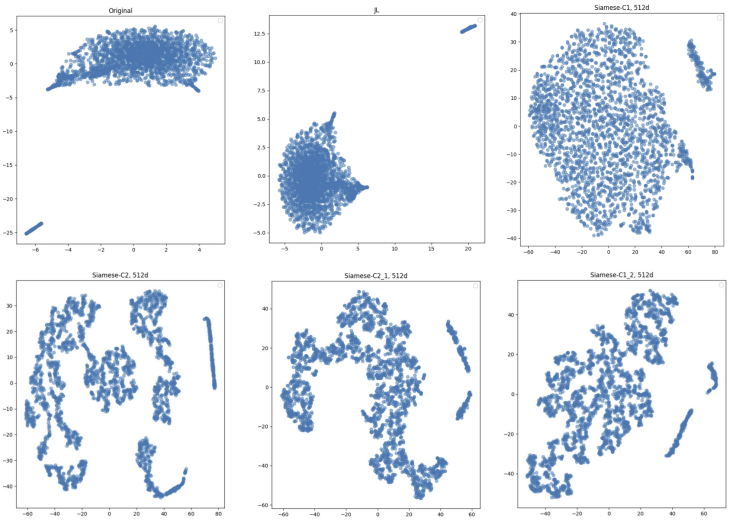

Figure 1: *t*-SNE plots of GSV encodings (original) and SNN embeddings with all curricula training permutations. Top left: GSV encodings. Top-middle: JL. Top-right: Siamese curriculum 1. Bottom-left: Siamese C2. Bottom-middle: Siamese C2+C1. Bottom-right: Siamese C1+C2.

6.3 Training loss

Both curricula converged during training, yet curriculum 1 begins with significantly higher loss and requires many training samples to converge compared to curriculum 2 (Figure 2).

7 Discussion of findings and limitations

Overall, SNNs effectively learn low dimensional representations of GSVs across healthy and cancer individuals. A JL baseline which preserves pairwise distances with an upper bound of $\epsilon = 0.1$ fractional error was outperformed by an SNN model. Furthermore, the SNN embedding size was roughly four times smaller than the JL embedding size. Two training curricula were evaluated, and the most performant was triplet loss which trained to minimize the GSV embedding distance of individuals in the same population while maximizing the distance of individuals across populations. The curriculum designed to penalize pairwise distances produced qualitatively good t-SNE plots, yet the pairwise distance error compared to ground truth was worse compared to other approaches. This qualitative result seems contradictory and can be explored by error analysis. An intuitive guess is that curriculum 1 solely learned to across population distances with little regard to within population distances; however, this does not explain the consistent convergence in the training curves. Curricula 2's low

pairwise distance error and distorted t-SNE plot, with respect to the t-SNE of GSV encodings, may indicate a simpler geometric structure exists to organize individual GSVs while preserving pairwise distances. Adding more metrics and visualizations can explain the RMSE of each method, such as computing the pairwise distance error separately for within and across population pairs.

8 Conclusions and future work

8.1 Conclusions

A siamese neural network trained on triplet loss (curriculum 2) of human genomes effectively preserved pairwise distances of gene fusion encodings. Compared to the JL transformation baseline, the SSN embeddings are only 512 dimensions compared to 2104 used for JL transformation, and the SNN RMSE of pairwise distances on testing data is 4.88 compared to 4.95 for JL. Furthermore, SNN is an online model supporting rapid embedding of new genomes without re-training, unlike JL. The efficacy of SNN representation learning suggests a practical use case for embedding human genomes by gene fusion evidence to support similarity searches in fast vector databases. These findings lay the groundwork to assist treatment of rare genetic diseases, such as rare cancers, by finding similar patients. Once similar patients are found, the genotypic and phenotypic characteristics of similar patients can undergo comparative analysis to hopefully improve understanding, diagnosis, and treatment of rare genetic diseases.

8.2 Future work

Embedding other cancer types, besides prostate, would directly test whether embeddings cluster by cancer type. Evaluation can be improved by directly performing similarity searches and computing precision and recall at top k. Where, the ground truth is the top k nearest neighbors (KNN) in the original encoding space and predicted top k are neighbors in the embedding space. In addition, error analysis can be improved by comparing pairwise distance loss within and across populations. More parameter tuning could minimize the error of pairwise distances in the embeddings space: embedding size, hidden layer count and size, activation functions, etc. Other improvements could include early stopping, weight freezing for certain curricula, and increasing the periodicity of curricula swapping, such as swapping after k batches. The simplest improvement of performance and evaluation is to include all available data into training, validation, and testing. Due to time constraints, only a small fraction of the total amount of paired data was used for training and testing.

References

[1] Cecilia Aguerrebere et al. “Similarity search in the blink of an eye with compressed indices”. In: *Proceedings of the VLDB Endowment* 16.11 (2023), pp. 3433–3446.

[2] Jane Bromley et al. “Signature verification using a "Siamese" time delay neural network”. In: *Proceedings of the 6th International Conference on Neural Information Processing Systems*. NIPS’93. Denver, Colorado: Morgan Kaufmann Publishers Inc., 1993, pp. 737–744.

[3] Murad Chowdhury et al. “Searching thousands of genomes to classify somatic and novel structural variants using STIX”. In: *Nature Methods* 19.4 (2022), pp. 445–448. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01423-4. URL: https://doi.org/10.1038/s41592-022-01423-4.

[4] Susan Fairley et al. “The International Genome Sample Resource (IGSR) collection of open human genomic variation resources”. In: *Nucleic Acids Research* 48.D1 (Oct. 2019), pp. D941–D947. ISSN: 0305-1048. DOI: 10.1093/nar/gkz836. eprint: https://academic.oup.com/nar/article-pdf/48/D1/D941/31697918/gkz836.pdf. URL: https://doi.org/10.1093/nar/gkz836.

[5] Mary J. Goldman et al. “A user guide for the online exploration and visualization of PCAWG data”. In: *Nature Communications* 11.1 (2020), p. 3400. ISSN: 2041-1723. DOI: 10.1038/s41467-020-16785-6. URL: https://doi.org/10.1038/s41467-020-16785-6.

[6] Guy Hacohen and Daphna Weinshall. “On The Power of Curriculum Learning in Training Deep Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 2535–2544. URL:

https://proceedings.mlr.press/v97/hacohen19a.html.

[7] William Johnson and J. Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Conference in Modern Analysis and Probability* 26 (Jan. 1982), pp. 189–206.

[8] Ryan M. Layer et al. "GIGGLE: a search engine for large-scale integrated genome analysis". In: *Nature Methods* 15.2 (2018), pp. 123–126. ISSN: 1548-7105. DOI: 10.1038/nmeth.4556. URL: https://doi.org/10.1038/nmeth.4556.

[9] Kristen Schneider et al. "GenoSiS: A Biobank-Scale Genotype Similarity Search Architecture for Creating Dynamic Patient-Match Cohorts". In: *bioRxiv* (2024). DOI: 10.1101/2024.11.02.621671. eprint: https://www.biorxiv.org/content/early/2024/11/03/2024.11.02.621671.full.pdf. URL: https://www.biorxiv.org/content/early/2024/11/03/2024.11.02.621671.

[10] Yaniv Taigman et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220.

9 Appendix

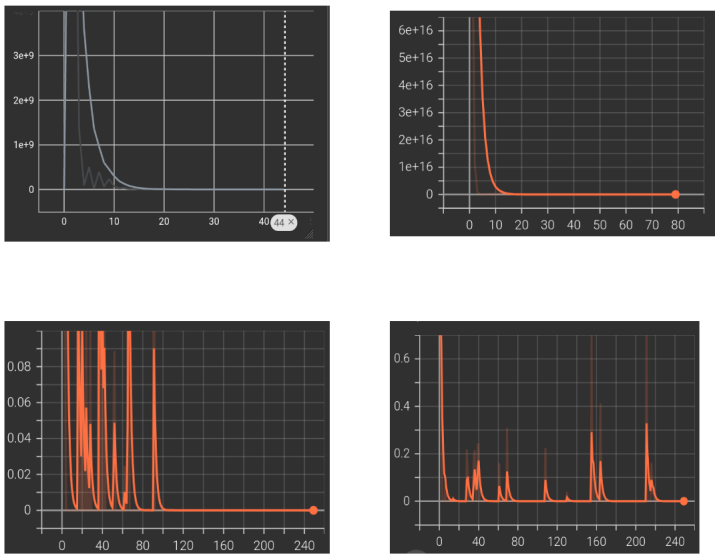

Figure 2: Screenshots of loss over epochs for training all curricula permutations. Top-left: curriculum 1 alone. Top right: curriculum 1 after training curriculum 2. Bottom left: curriculum 2 after curriculum 1. Bottom right: curriculum 2 alone.

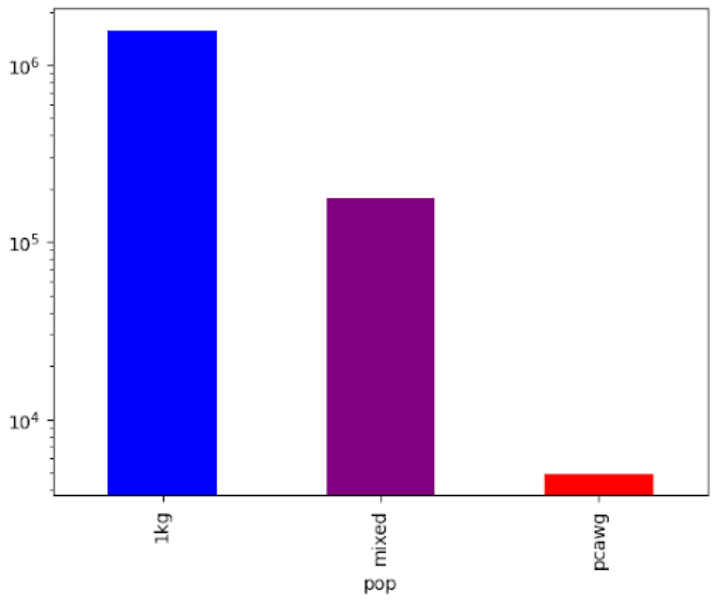

Figure 3: Number of pairs per population