Sandia National Laboratories

SAND2019-1501C

Reordering Genomic Data for Improved Compression-Based Inference

Renee Gooding, Jacob Caswell, Richard Field, Christina Ting

Abstract

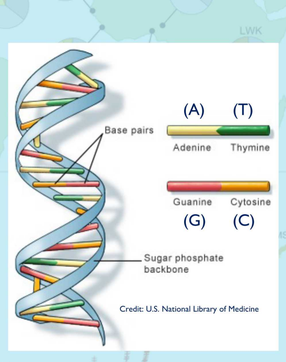
We have previously shown that compression algorithms can be extended in a variety of ways for useful application in machine learning and data analytics, including deception detection in text, boundary detection in audio, and anomaly detection in network traffic. Compression-based analytics rely on the data to occur locally and sequentially in order to identify patterns, which can be applied towards effective decision making. Although genomic data is nominally read as a sequence of nucleotides, the information content is neither local nor sequential; long-range interactions and regulations of genes with similar biochemical functions exist.

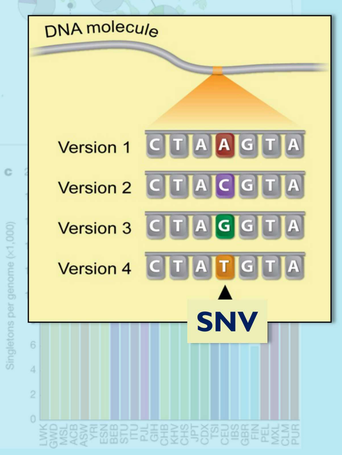
We study how the dependencies among single nucleotide variants (SNVs) revealed from pairwise linkage disequilibrium calculations can be used to re-order the genomic sequence and improve the ability of a compression-based analytic to identify patterns and make inferences. In particular, we apply Louvain community detection, a graph-based algorithm, to reorder the SNVs into sections of highly dependent SNVs. We use prediction by partial matching (PPM), an adaptive statistical data compression technique, to train local and global models on the re-ordered sequences. We demonstrate that the re-ordering by Louvain can improve a compression-based classifier's ability to infer a population attribute. Our results are compared to standard machine learning classifiers such as Random Forests. Ultimately, understanding how to improve inference can be used to understand how to improve genomic privacy.

Dataset: 1000 Genomes Project

Human DNA is made up of 3.2 billion bps; most are identical across the population. Variations occur at specific positions in the genome known as single nucleotide variations (SNVs), or 0.3% of the genome (10 million base pairs (bps)).

Within this dataset, our analysis will focus on
- subsets of 1000 SNVs
- 3 population groups





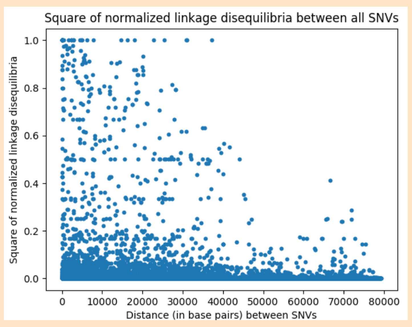
Identifying long-ranged and non-random association of SNVs

I. Linkage disequilibrium (LD)

The non-random association of SNVs can be quantified, to first order, by LD:

$$l_{ij} \sim P_{ij} - P_i P_j$$
$$-1 \leq l_{ij} \leq 1$$

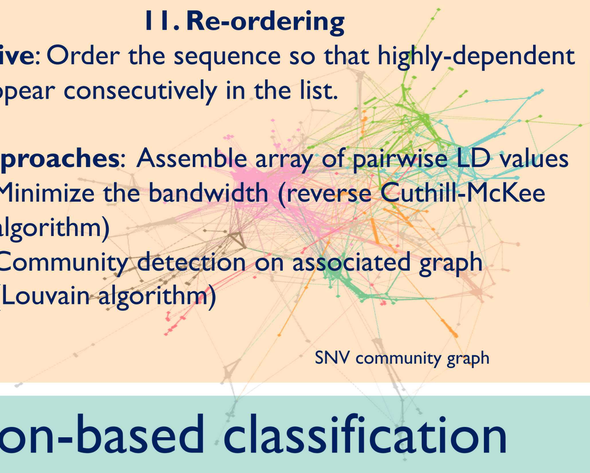
SNVs are not 'optimally' ordered, in the sense that highly correlated SNVs may be far apart in the genome.

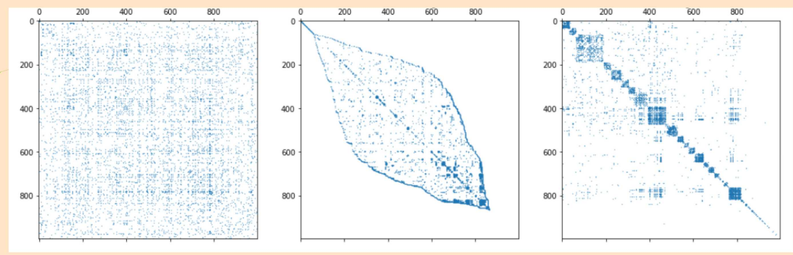


II. Re-ordering

Objective: Order the sequence so that highly-dependent SNVs appear consecutively in the list.

Two approaches: Assemble array of pairwise LD values
1. Minimize the bandwidth (reverse Cuthill-McKee algorithm)
2. Community detection on associated graph (Louvain algorithm)





Original matrix of LD values | Re-ordered using reverse Cuthill McKee | Re-ordered using Louvain community detection

Compression-based classification

Prediction by partial matching (PPM) with arithmetic coding (AC)

Statistical data compression techniques are based on Markov models of different contexts, i.e. a SNV (x) is predicted based on the previous *n* SNVs, referred to as a context ($c_1 c_2 \ldots c_n$):

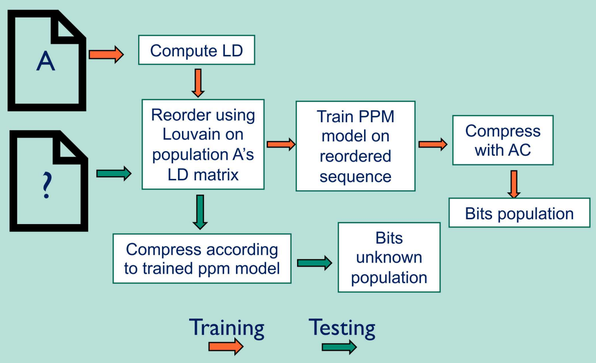
$$\Pr(x|c_1 c_2 \cdots c_n) = \frac{\Pr(c_1 c_2 \cdots c_n\ x)}{\Pr(c_1 c_2 \cdots c_n)}, \approx \frac{\mathrm{count}(c_1 c_2 \cdots c_n\ x)}{\mathrm{count}(c_1 c_2 \cdots c_n)}$$

Training → We build a *PPM model* for each population of observed contexts.

Testing → AC encodes an unknown sequence in such a way that higher probabilities are encoded in fewer bits. We exploit this property to use PPM-AC as a classifier.

I. Model-based compression

Classification is determined by comparing the compression score for an unknown population to models for known populations. The model that compresses the unknown population the best determines the classification.



II. Slice compression (SC)

With slice compression we aim to find local structure through a sequence of windowed compression scores.

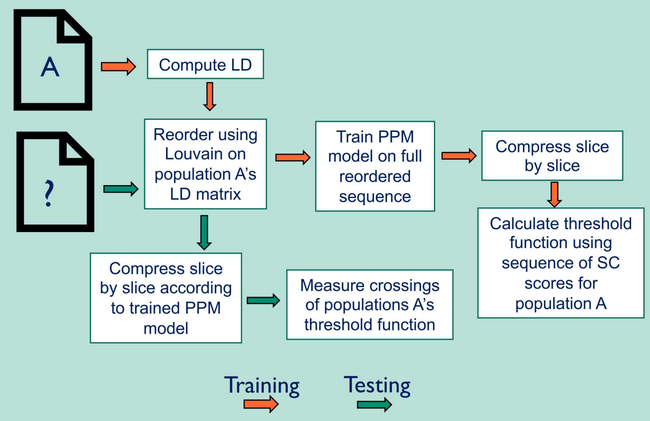
SC score:

$$z(k; w) := \frac{1}{w} C\big(S_k(w)\big),$$

w window width,
k data index

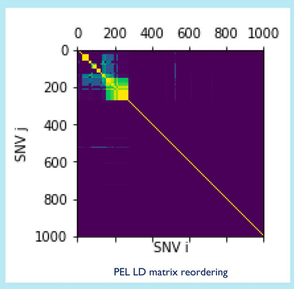
For classification, a threshold function is calculated by modeling the SC sequence of scores from the trained model as a random process. An unknown population is compressed slice by slice with respect to the trained model. Classification is determined by the threshold exceedances of the unknown population.

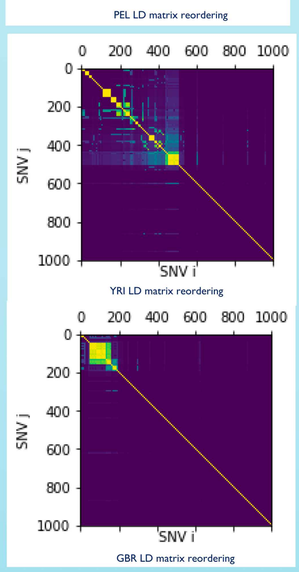
Threshold function:
$t(k; w) = \mu(k; w) + \sigma(k; w) F^{-1}(p)$



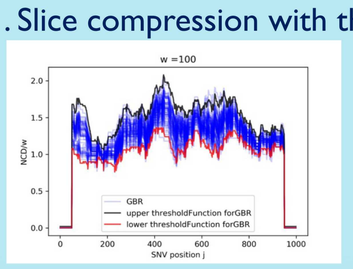
Results

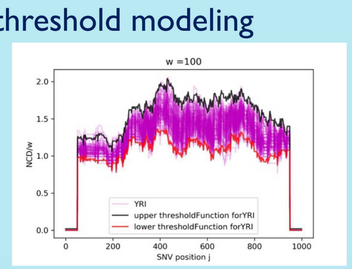
I. Slice compression with threshold modeling

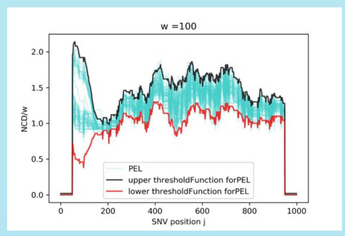




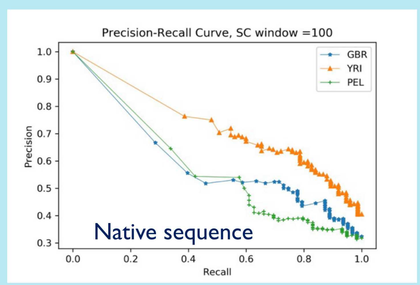
Extent of reordering by population attribute (Reverse Cuthill McKee)

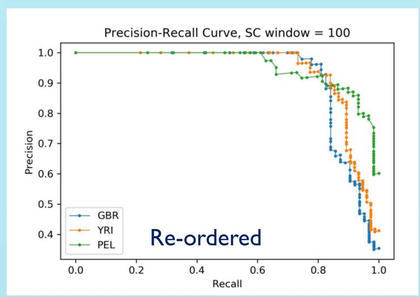




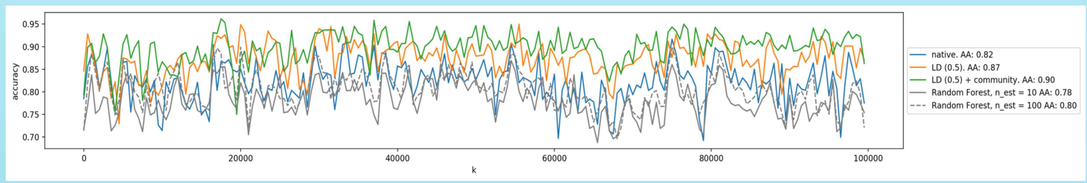


SC scores on reordered sequences for each respective population, together with upper and lower threshold functions,





II. Model-Based compression



Accuracy comparison between different reordering methods for Model-based Compression and Random Forest classification.

Conclusions

We apply different reordering schemes to the genomic sequence to localize predictable structure. This reordering improves a compression classifier's ability to identify population attributes of an unknown sequence.

1. Using model-based compression, we find that the accuracy improves as we move from native ordering to reordering by Louvain communities, to finally reordering by Louvain communities followed by selecting significant communities.

2. When compared to a Random Forest classifier, it can be seen that model-based compression classifier, even on the native sequence, out performs Random Forest.

3. Slice compression and threshold modeling on the reordered sequence further identifies deviations from expected *local* structure to classify an unknown population with high precision and recall rates.

4. Future work will involve understanding when transformations of the SNV sequence yield optimal improvements in accuracy.

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____

Sandia National Laboratories

Title up to two lines

- Color transparent boxes below are for resizing and using to highlight content. Delete if not needed.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____