SAND2019-15128C

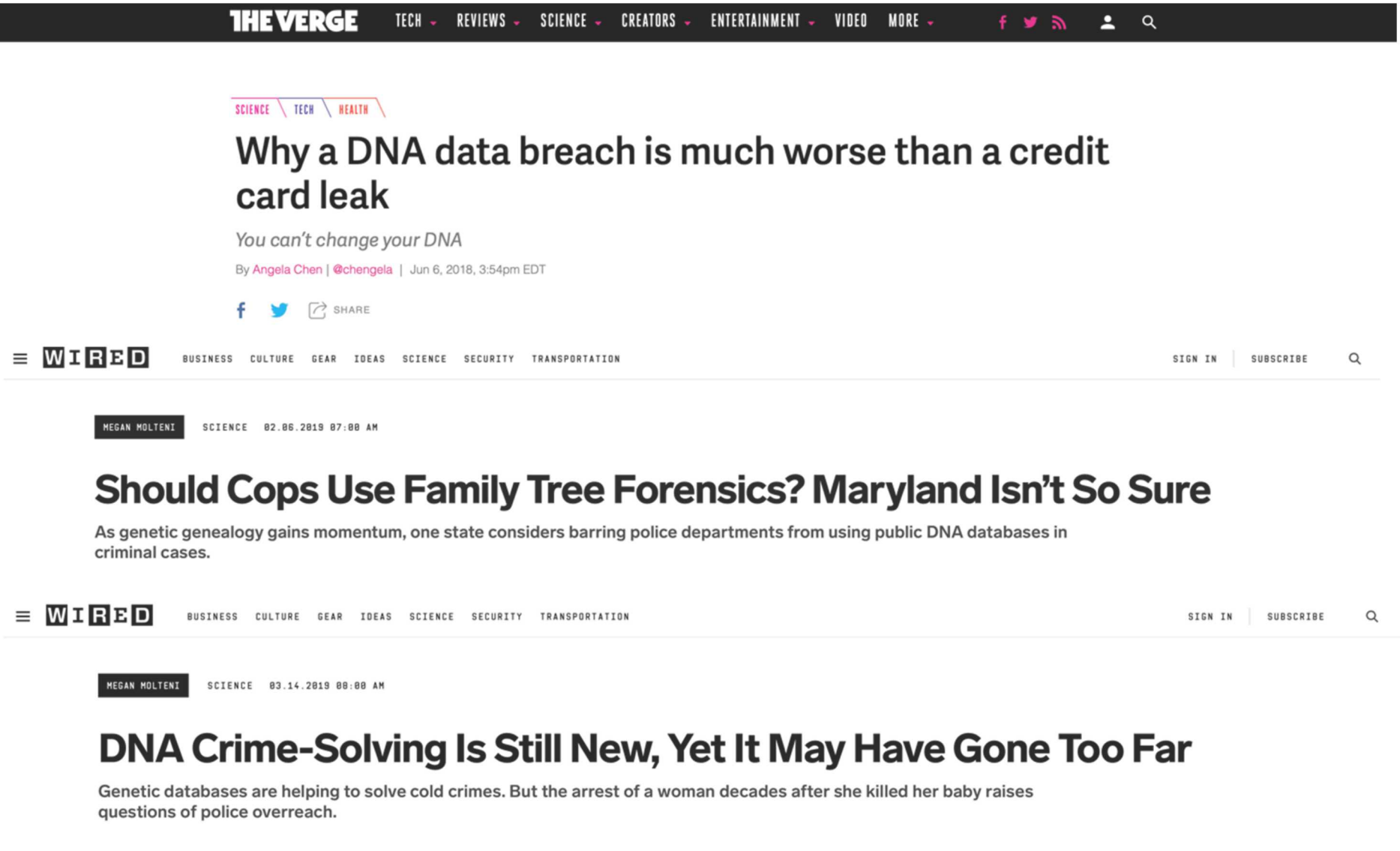
Reordering Genomic Sequences for Enhanced Classification via Compression Analytics

Christina Ting, Jacob Caswell, and Richard Field

PRESENTED BY

Renee Gooding

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.



THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

SCIENCE TECH HEALTH

Why a DNA data breach is much worse than a credit card leak

You can't change your DNA

By Angela Chen | @chengela | Jun 6, 2018, 3:54pm EDT

SHARE

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION SIGN IN SUBSCRIBE

MEGAN MOLTENI SCIENCE 02.06.2019 07:00 AM

Should Cops Use Family Tree Forensics? Maryland Isn't So Sure

As genetic genealogy gains momentum, one state considers barring police departments from using public DNA databases in criminal cases.

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION SIGN IN SUBSCRIBE

MEGAN MOLTENI SCIENCE 03.14.2019 08:00 AM

DNA Crime-Solving Is Still New, Yet It May Have Gone Too Far

Genetic databases are helping to solve cold crimes. But the arrest of a woman decades after she killed her baby raises questions of police overreach.

Privacy is in our DNA

Everyone deserves a secure, private place to explore and understand their genetics. At 23andMe, we put you in control of deciding what information you want to learn and what information you want to share.

See our privacy statement for more info.

Five key ways we ensure your privacy

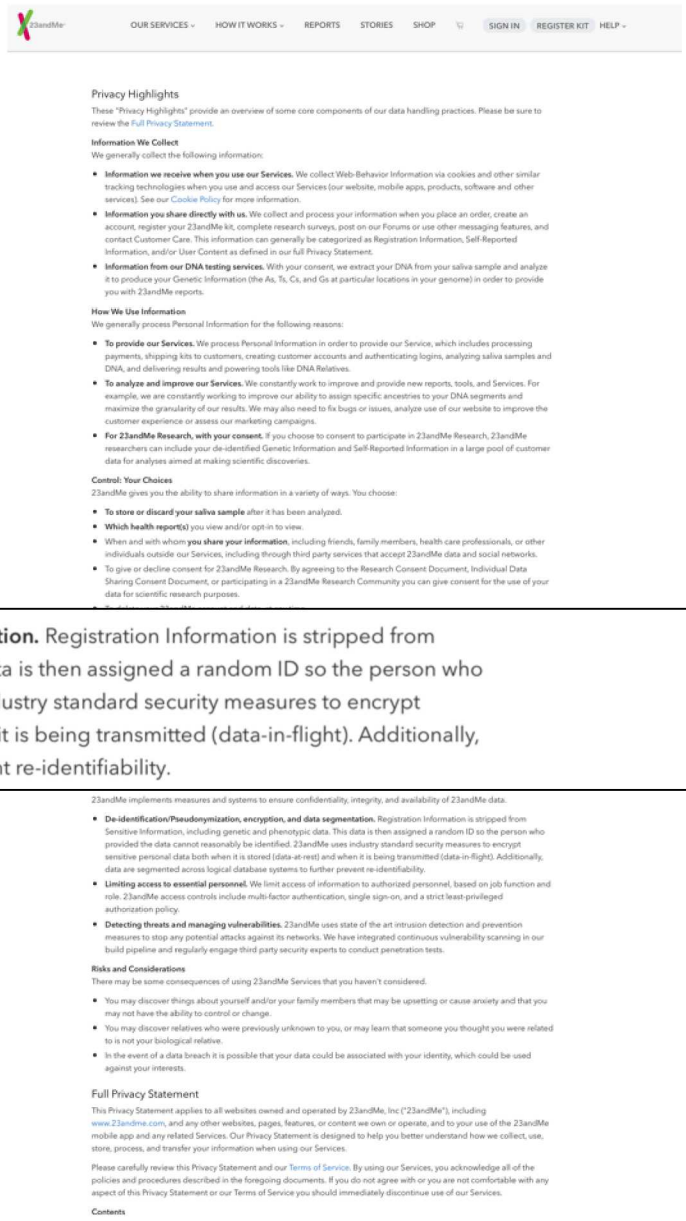
- Meaningful choice
- Privacy by design
- Third party sharing
- Security
- Research

The information presented here is meant to be a general guide to our privacy and security practices. For specific details about our practices, see our privacy statement, terms of service, research consent document, sample storage consent document and frequently asked questions.

We're committed to complying with the EU's new data protection law, referred to as the GDPR. Visit our GDPR page to learn about our data protection approach.

Please contact us at privacy@23andMe.com if you have questions.

23andMe's "Privacy Highlights"



OUR SERVICES · HOW IT WORKS · REPORTS · STORIES · SHOP · SIGN IN · REGISTER KIT · HELP

Privacy Highlights

These "Privacy Highlights" provide an overview of some core components of our data handling practices. Please be sure to review the Full Privacy Statement.

Information We Collect

We generally collect the following information:

- **Information we receive when you use our Services.** We collect Web-Behavior Information via cookies and other similar tracking technologies when you use and access our Services (our website, mobile apps, products, software and other services). See our Cookie Policy for more information.
- **Information you share directly with us.** We collect and process your information when you place an order, create an account, register your 23andMe kit, complete research surveys, post on our Forums or use other messaging features, and contact Customer Care. This information can generally be categorized as Registration Information, Self-Reported Information, and/or User Content as defined in our full Privacy Statement.
- **Information from our DNA testing services.** With your consent, we extract your DNA from your saliva sample and analyze it to produce your Genetic Information (the As, Ts, Cs, and Gs at particular locations in your genome) in order to provide you with 23andMe reports.

How We Use Information

We generally process Personal Information for the following reasons:

- **To provide our Services.** We process Personal Information in order to provide our Service, which includes processing payments, shipping kits to customers, creating customer accounts and authenticating logins, analyzing saliva samples and DNA, and delivering results and powering tools like DNA Relatives.
- **To analyze and improve our Services.** We constantly work to improve and provide new reports, tools, and Services. For example, we are constantly working to improve our ability to assign specific ancestries to your DNA segments and maximize the granularity of our results. We may also need to fix bugs or issues, analyze use of our website to improve the customer experience or assess our marketing campaigns.
- **For 23andMe Research, with your consent.** If you choose to consent to participate in 23andMe Research, 23andMe researchers can include your de-identified Genetic Information and Self-Reported Information in a large pool of customer data for analyses aimed at making scientific discoveries.

Control: Your Choices

23andMe gives you the ability to share information in a variety of ways. You choose:

- **To store or discard your saliva sample** after it has been analyzed.
- **Which health report(s)** you view and/or opt-in to view.
- When and with whom **you share your information**, including friends, family members, health care professionals, or other individuals outside our Services, including through third party services that accept 23andMe data and social networks.
- To give or decline consent for 23andMe Research. By agreeing to the Research Consent Document, Individual Data Sharing Consent Document, or participating in a 23andMe Research Community you can give consent for the use of your data for scientific research purposes.

- **De-identification/Pseudonymization, encryption, and data segmentation.** Registration Information is stripped from Sensitive Information, including genetic and phenotypic data. This data is then assigned a random ID so the person who provided the data cannot reasonably be identified. 23andMe uses industry standard security measures to encrypt sensitive personal data both when it is stored (data-at-rest) and when it is being transmitted (data-in-flight). Additionally, data are segmented across logical database systems to further prevent re-identifiability.

23andMe implements measures and systems to ensure confidentiality, integrity, and availability of 23andMe data.

- **De-identification/Pseudonymization, encryption, and data segmentation.** Registration Information is stripped from Sensitive Information, including genetic and phenotypic data. This data is then assigned a random ID so the person who provided the data cannot reasonably be identified. 23andMe uses industry standard security measures to encrypt sensitive personal data both when it is stored (data-at-rest) and when it is being transmitted (data-in-flight). Additionally, data are segmented across logical database systems to further prevent re-identifiability.
- **Limiting access to essential personnel.** We limit access of information to authorized personnel, based on job function and role. 23andMe access controls include multi-factor authentication, single sign-on, and a strict least-privileged authorization policy.
- **Detecting threats and managing vulnerabilities.** 23andMe uses state of the art intrusion detection and prevention measures to stop any potential attacks against its networks. We have integrated continuous vulnerability scanning in our build pipeline and regularly engage third party security experts to conduct penetration tests.

Risks and Considerations

There may be some consequences of using 23andMe Services that you haven't considered.

- You may discover things about yourself and/or your family members that may be upsetting or cause anxiety and that you may not have the ability to control or change.
- You may discover relatives who were previously unknown to you, or may learn that someone you thought you were related to is not your biological relative.
- In the event of a data breach it is possible that your data could be associated with your identity, which could be used against your interests.

Full Privacy Statement

This Privacy Statement applies to all websites owned and operated by 23andMe, Inc ("23andMe"), including www.23andme.com, and any other websites, pages, features, or content we own or operate, and to your use of the 23andMe mobile app and any related Services. Our Privacy Statement is designed to help you better understand how we collect, use, store, process, and transfer your information when using our Services.

Please carefully review this Privacy Statement and our Terms of Service. By using our Services, you acknowledge all of the policies and procedures described in the foregoing documents. If you do not agree with or you are not comfortable with any aspect of this Privacy Statement or our Terms of Service you should immediately discontinue use of our Services.

Contents

There exist both real and demonstrated examples of successful attacks that have revealed weaknesses in existing privacy protection methods

Genomic Privacy

Privacy risks may include re-identification, inference of sensitive attributes, and revelation of familial relationships

Understanding what attributes can be inferred from available information is a critical part of genomic privacy and security

However, the full implications of sharing genomic information are still largely unknown due to:

- Advances in the study of genomics
- Advances in genomic inference methods

Foundations

Genomics Foundations

DNA is a molecular sequence composed of base units called nucleotides

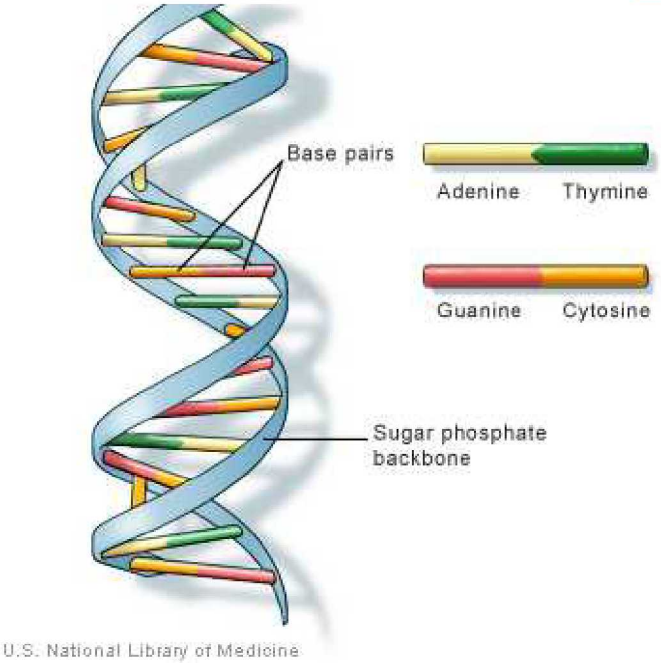
Four nucleotides:

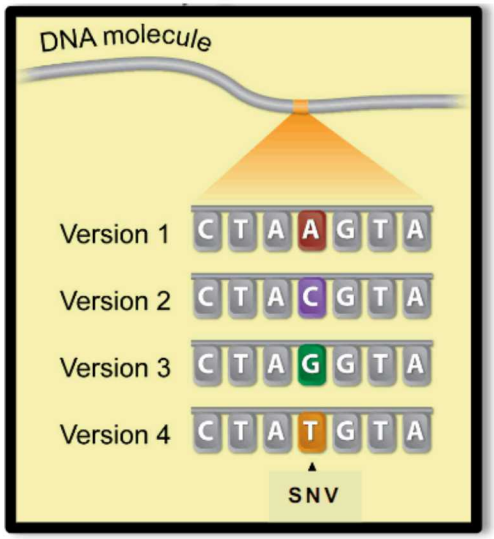
- cytosine (C), guanine (G), adenine (A), or thymine (T)

Human DNA is made up of **3.2 billion** base units; most are identical across the population

Single nucleotide variations (SNVs) occur at specific positions in the genome

- Carry privacy-sensitive information
- 0.3% of the genome
- Our analysis will focus on the sequence of SNVs



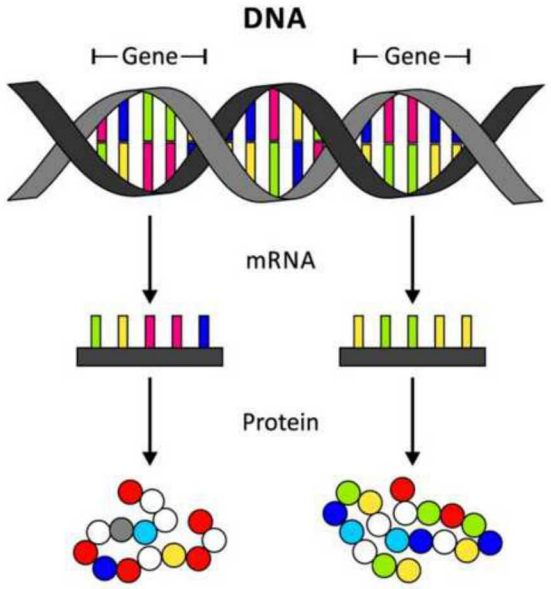


Background on genomic inference methods

- Pairwise correlation, or *linkage disequilibrium*, between genomic variants
- *k*-th order Markov chains
  - Probability of genomic variant occurring in the genome is conditioned on a contagious set of preceding *k* variants
- Standard Machine Learning algorithms
  - Rely on predetermined set of features requiring *a-priori* knowledge of best features
  - For genomic data, a standard feature set is the *k*-mer distribution

These methods rely on expert knowledge and contiguous sequences of *k* variants containing the predictive elements

Long range dependencies exist in genomic data



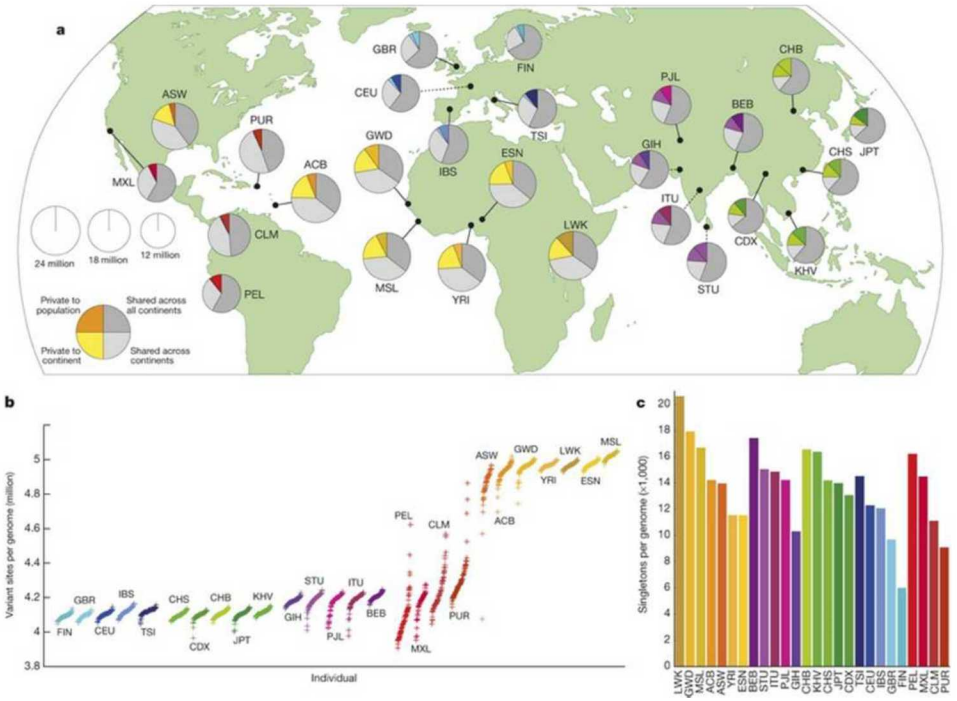
Genomic inference: our approach

1. Problem set up and data
2. An information-based machine learning model via compression analytics
3. Reordering genomic information for improved inference
4. Results

Population inference using the1000 genomes project

- Dataset: contains genetic variants with frequencies of at least 1% in the population
  - 2,504 individuals
  - 26 population groups

- Problem set up:
  - Focus on 3 populations (93, 99, 64 individuals)
  - Select a subset of 1000 variants and try to infer the population attribute



Fundamentals of compression analytics

Consider two sequences of 10K nucleotides:

x = 'AAA … AAA'
c(x) = 86 Bits

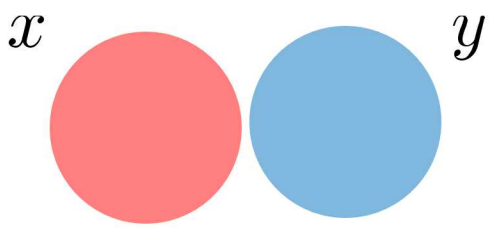
A lot of data; almost no information

y = 'ATGCC … CTCG'
c(y) = 86,418 Bits

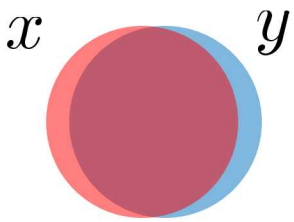
A lot of data; a lot of information

Fundamentals of compression analytics

Now consider **comparisons** between two sequences:



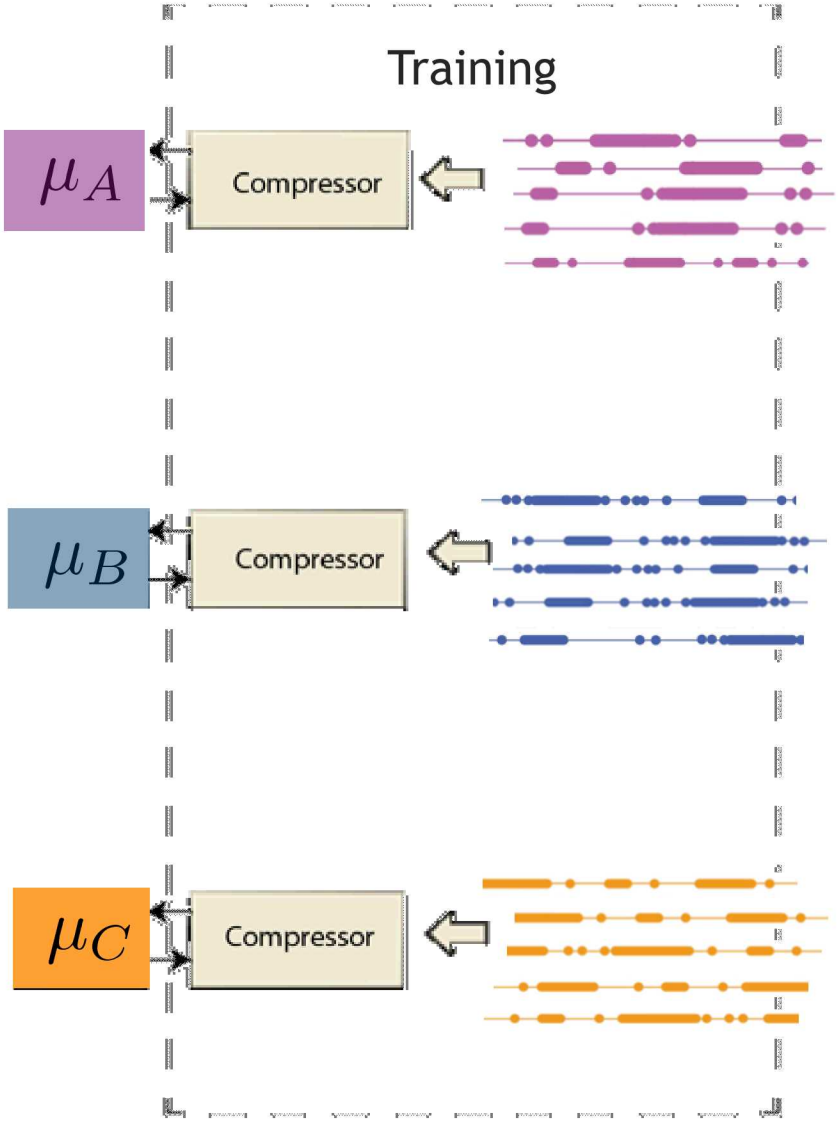
x and y completely different



x and y overlap

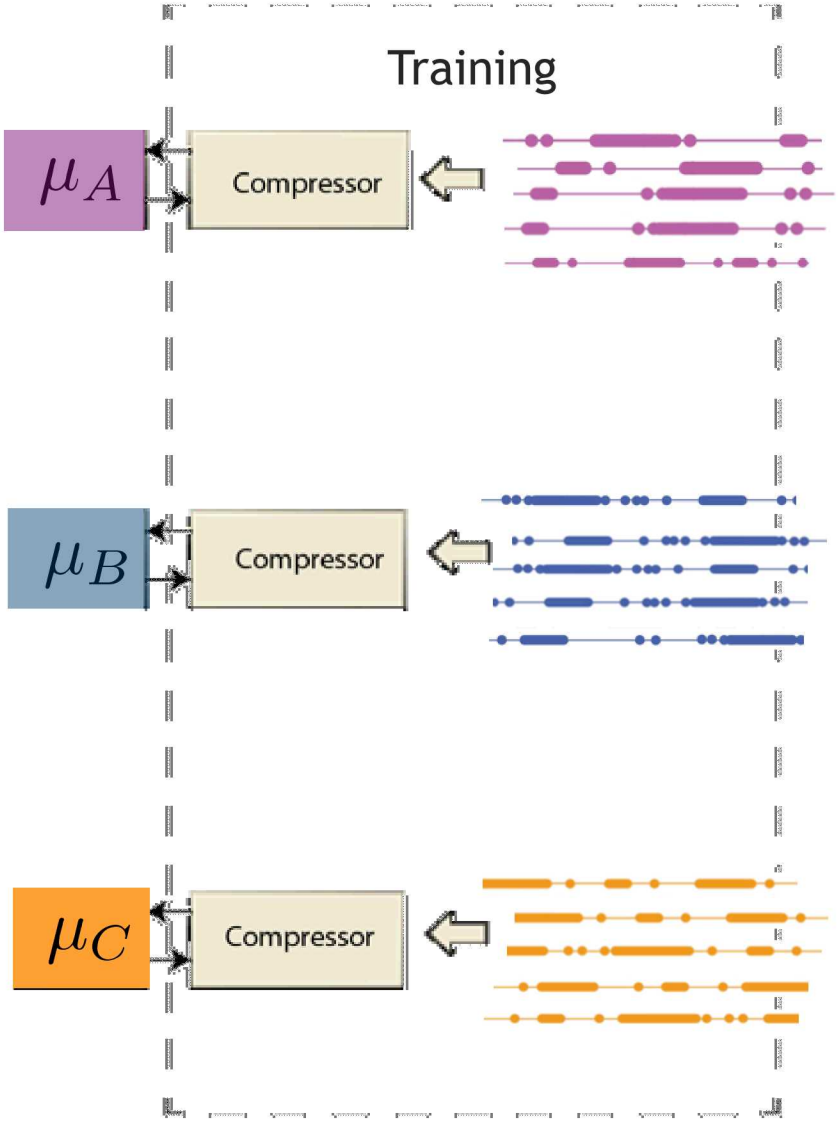
$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Compression analytics: training

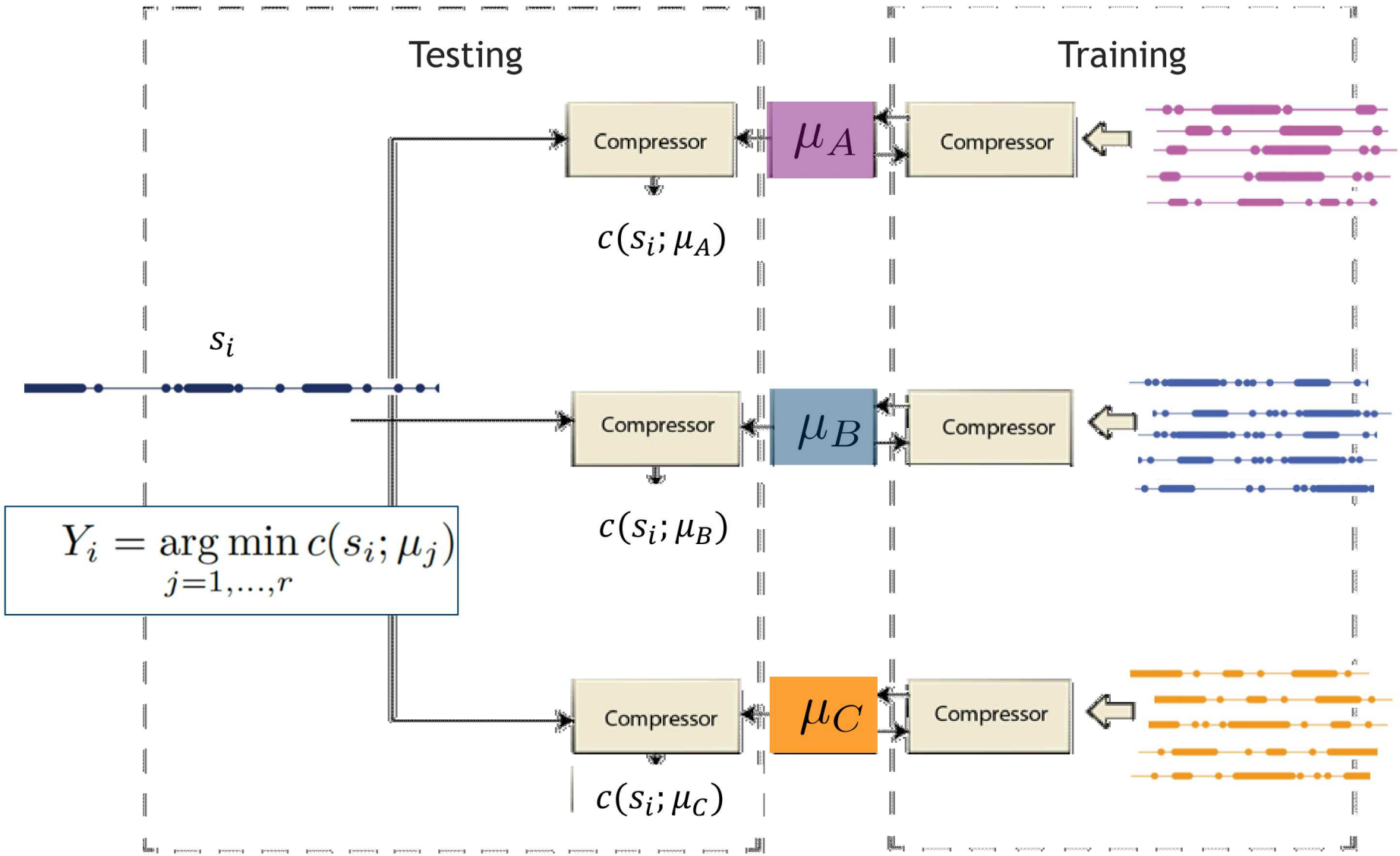


Compression analytics: testing

s_i

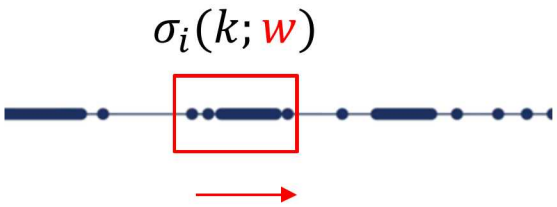


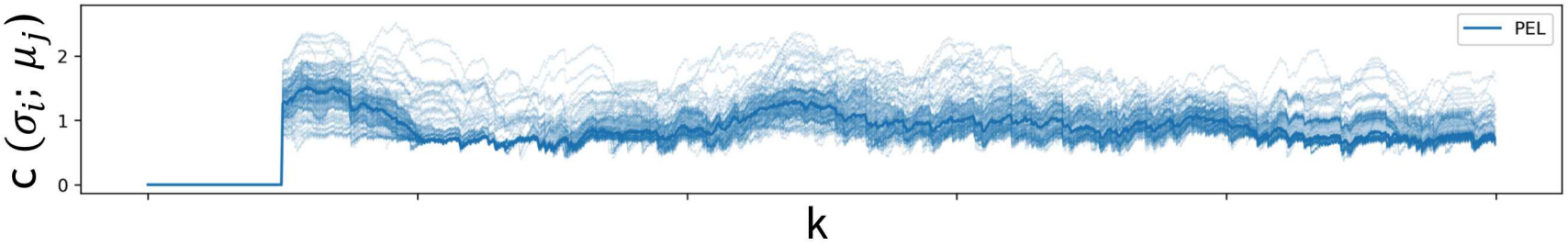
Compression analytics: testing



Slice compression analytics

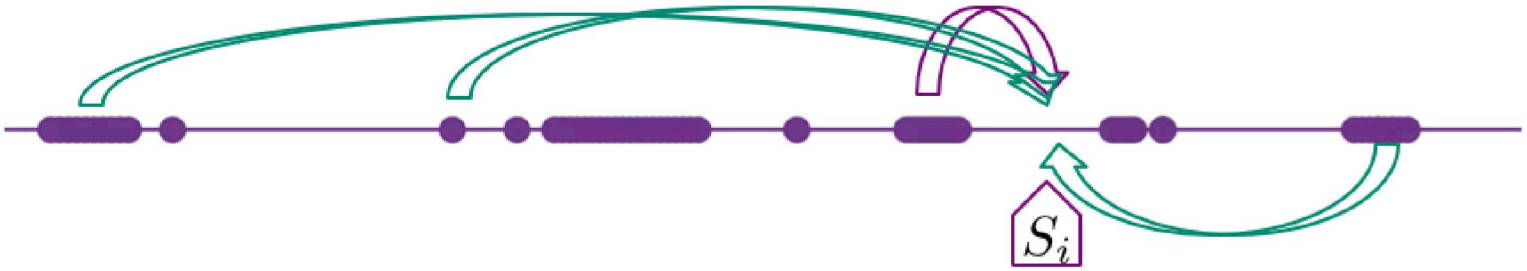
Slice of width w:

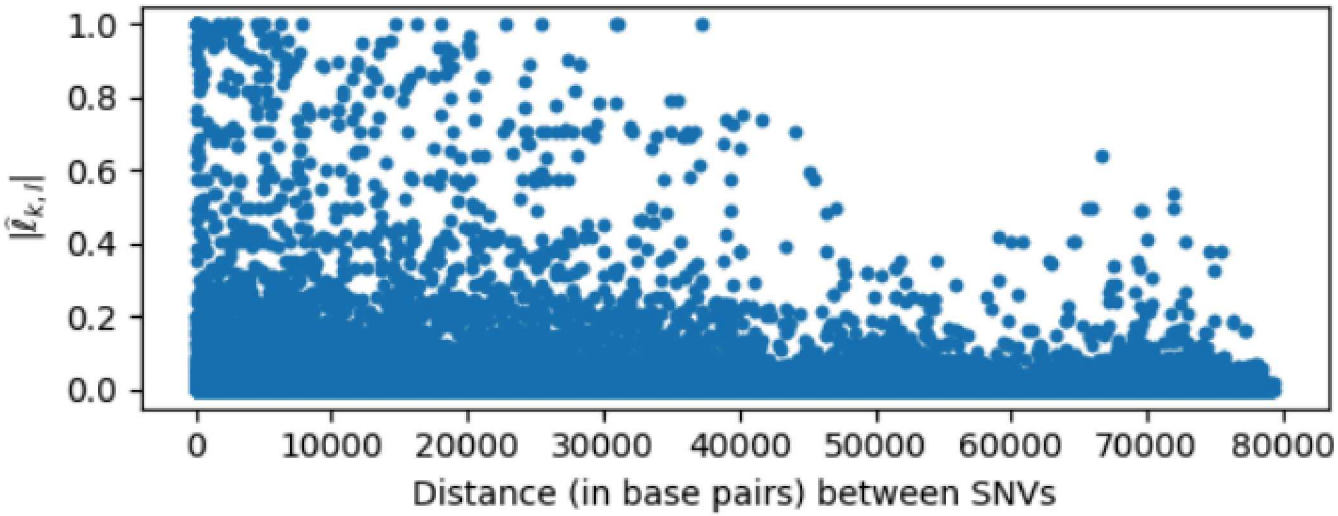




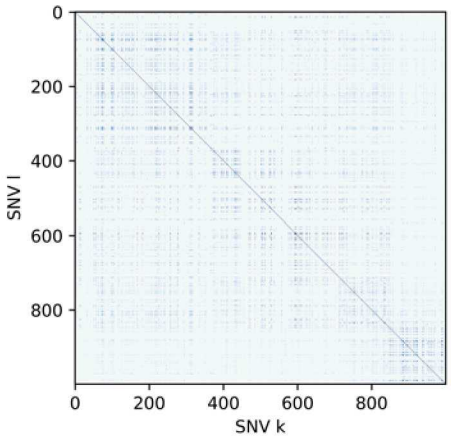
$$Y_i = \arg\min_{j=1,\ldots,r} \bar{c}_i(\sigma_i; \mu_j)$$

Additional challenge: genomic information is long ranged

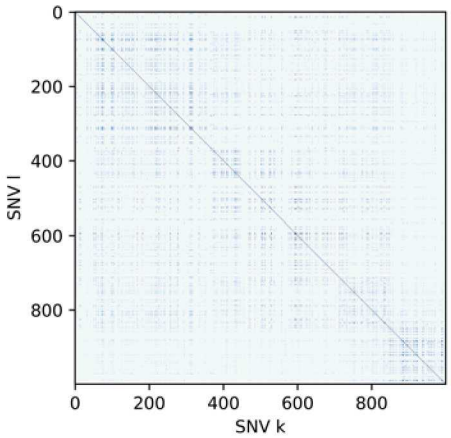




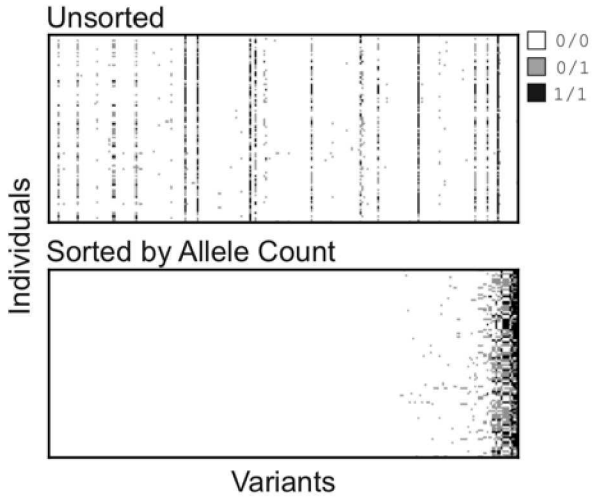
Reordering schemes to localize genomic information



Reordering schemes to localize genomic information

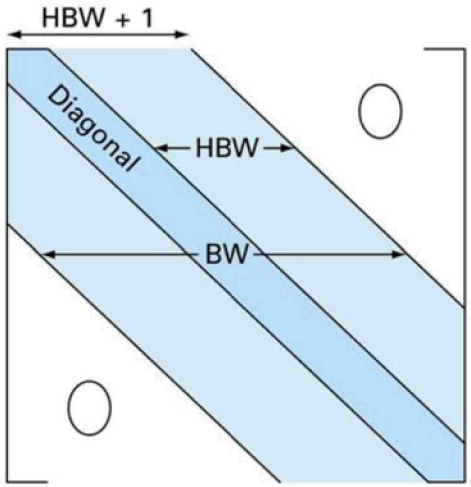


Data compression



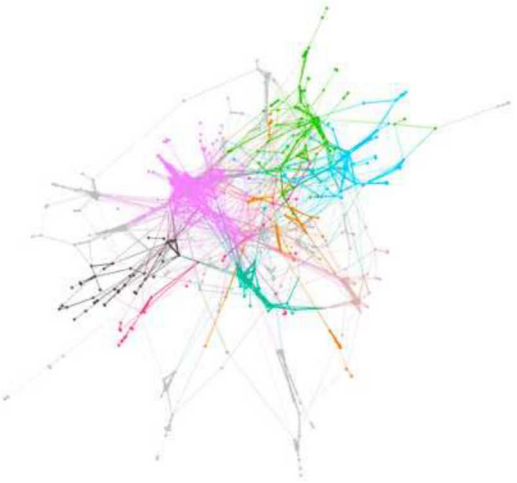
R. M. Layer, et al., "Efficient genotype compression and analysis of large genetic-variation data sets". Nature Methods (2016)

Linear algebra



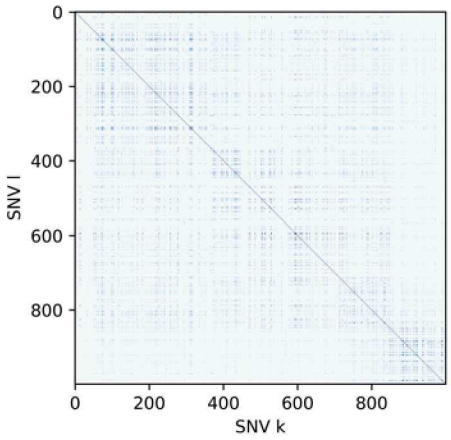
E. Cuthill and J. McKee. "Reducing the bandwidth of sparse symmetric matrices" In Proc. 24th Nat. Conf. ACM, pages 157–172, 1969

Graph theory

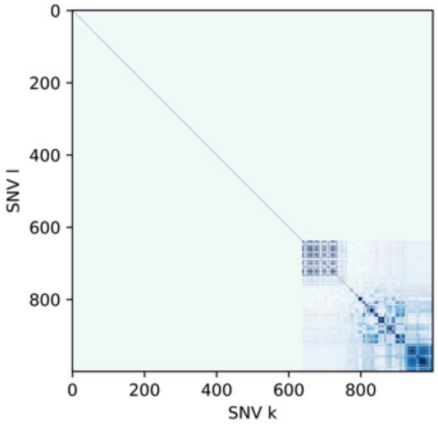


V. D. Blondel, et al., "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment.* **2008** (10): P10008

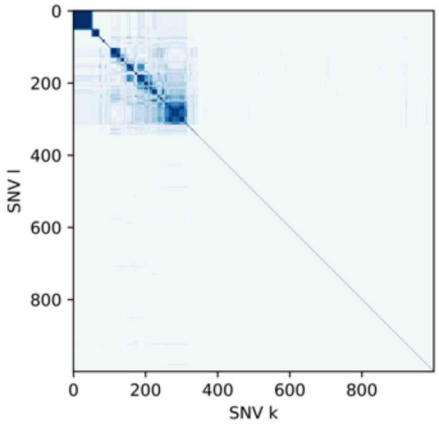
Reordering schemes to localize genomic information



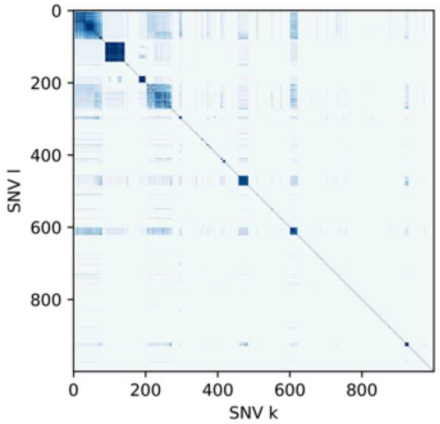
Data compression



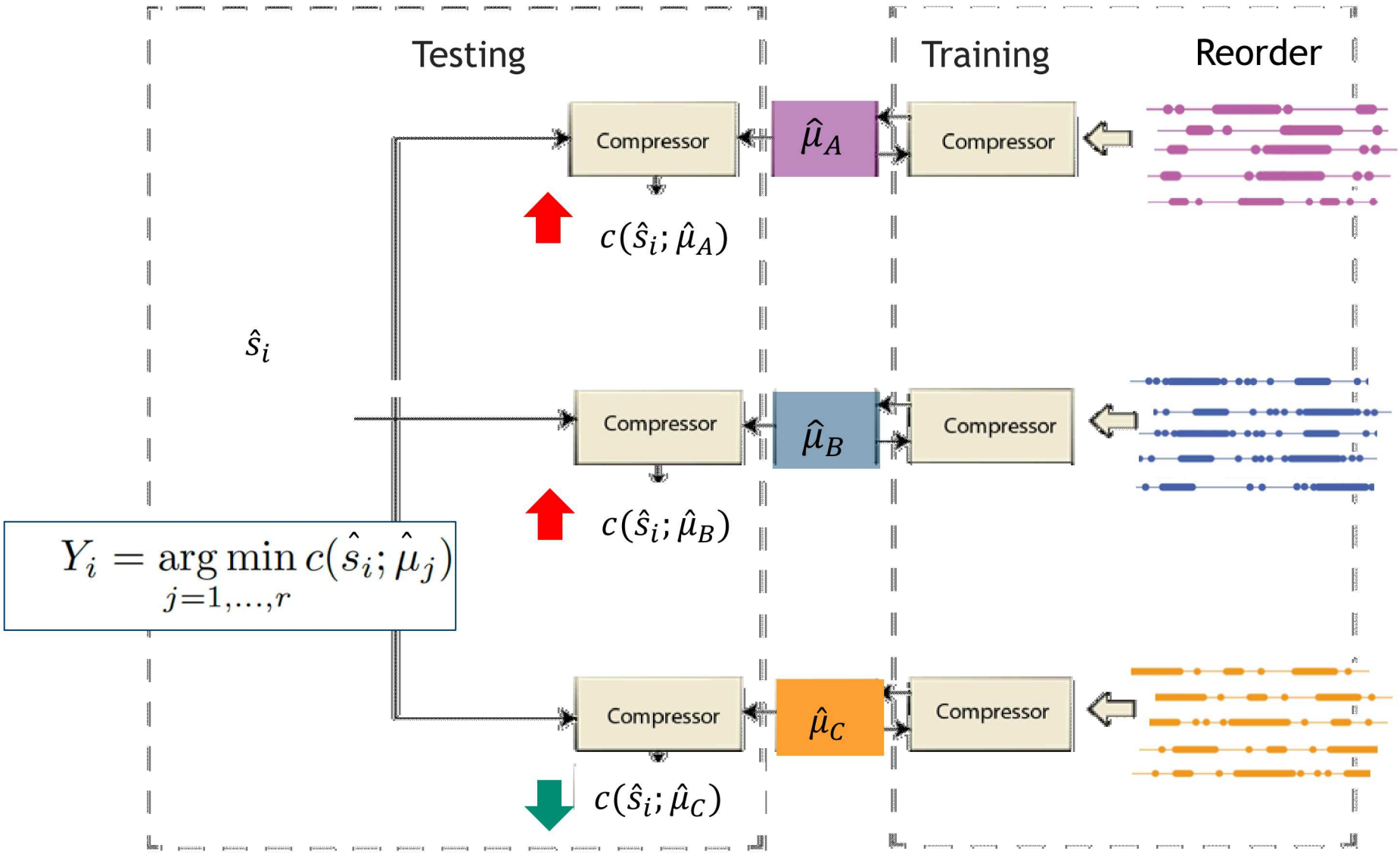
Linear algebra



Graph theory

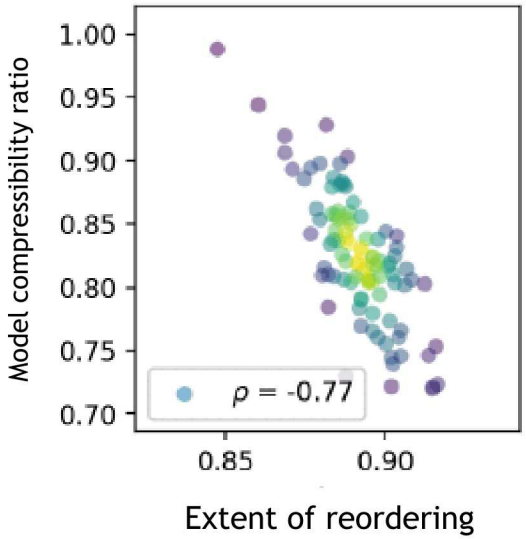


Compression analytics with reordering



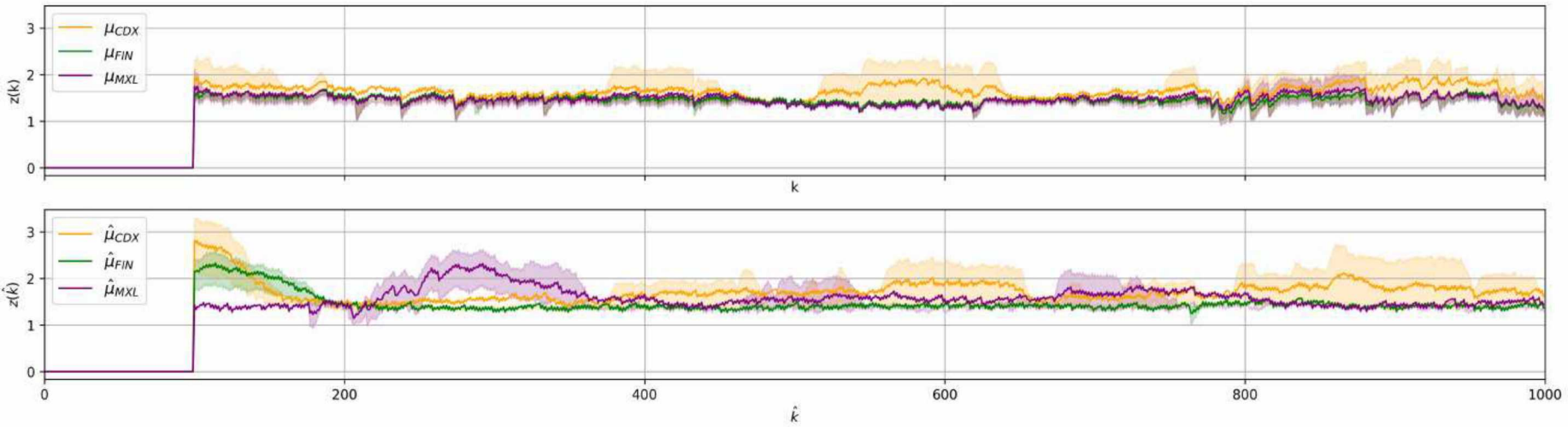
Impact of reordering on compressibility

$$\frac{c(\hat{\mu})}{c(\mu)}$$



The different reordering schemes have improved the ability of the model to **identify patterns** that are useful for a machine learning task.

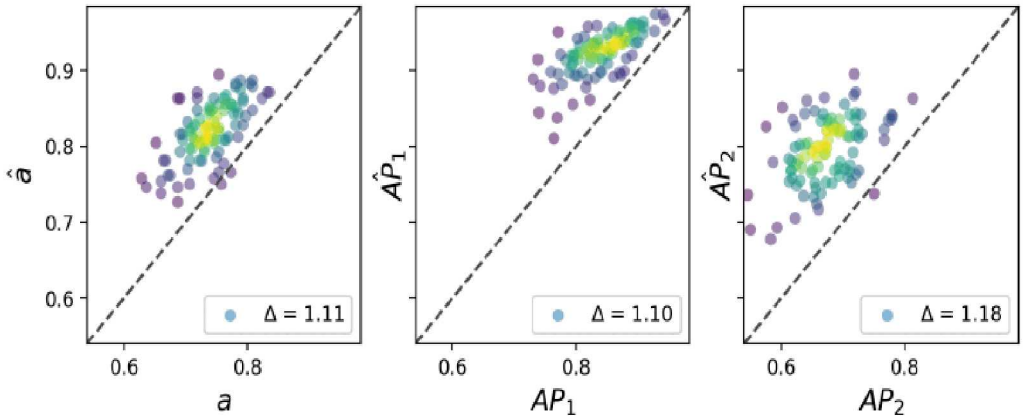
Impact of reordering on slice compression



Before reordering (upper), the SC scores show no notable structure and are not distinguishable from each other.

After reordering (lower), the SC scores are separated into distinguishable regions of high and low compressibility.

Impact of reordering on classification



Each measure of classification performance indicates an improvement after reordering

Summary

Method	Accuracy	Δ
RF (Native)	0.67	–
RF (k-mer histogram, $k = 1$)	0.55	–
RF (k-mer histogram, $k = 10$)	0.68	–
Compression (Native)	0.74	–
Compression (Louvain)	0.81	1.10
Compression (RCM)	**0.82**	**1.11**
Compression (Allele Frequency)	0.79	1.06
SC (Native)	0.70	–
SC (Louvain)	0.75	1.08
SC (RCM)	**0.77**	**1.10**
SC (Allele Frequency)	0.73	1.04

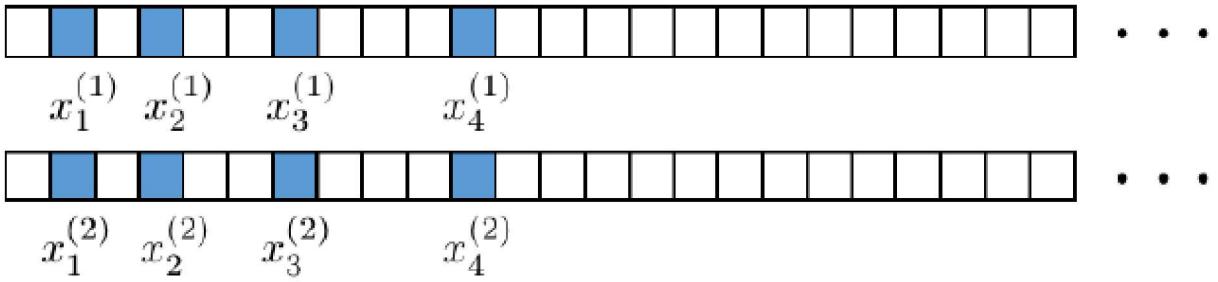
- All reordering methods considered improve classification accuracy.
- Information theoretic ML methods based on compression analytics outperform standard feature-based ML methods.
- Analogy between genomic code and computer code. These methods and others that deal with long-ranged dependencies will be useful beyond the application considered here.

Contact

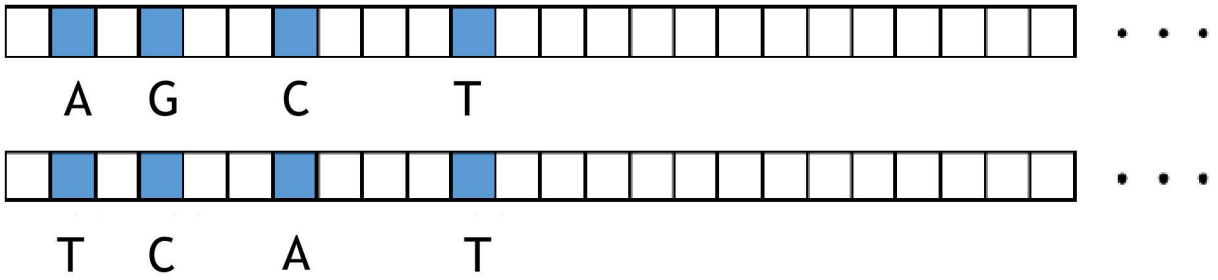
Renee Gooding

rlgoodn@sandia.gov

Compression analytics



Part I: Compression for classification



10 possible combinations for the pair $x_{i,k}^{(1)} x_{i,k}^{(2)}$ is encoded by a single token, $\chi_{i,k}$, where each is one of AA, AC/ CA, . . ., TT.

We represent a sequence of SNVs for an individual i as,

$$s_i = \chi_{i,1} \chi_{i,2} \cdots \chi_{i,m}$$

Linkage Disequilibrium

- We assume bi-allelic SNVs where the allele occurring most (least) often is referred to as the *major* (*minor*) allele

- $\Pr(X_k = 0)$ and $\Pr(X_k = 1)$ denote the major and minor *allele frequencies* of SNV *k*

- Because each SNV is assumed bi-allelic, we have

$$\Pr(X_k = 0, X_l = 0) = (1 - p_k)(1 - p_l) + l_{k,l}$$
$$\Pr(X_k = 0, X_l = 1) = (1 - p_k)p_l + l_{k,l}$$
$$\Pr(X_k = 1, X_l = 0) = p_k(1 - p_l) + l_{k,l}$$
$$\Pr(X_k = 1 X_l = 1) = p_k p_l + l_{k,l}$$

- The term $l_{k,l}$ is the *linkage disequilibrium* (LD)
  - Quantifies the degree of statistical dependence between SNVS *k* and *l*
  - $l_{k,l} = 0$ if and only if X_k and X_l are independent
  - Normalized LD, independent of allele frequency, satisfies $-1 \leq l_{k,l} \leq 1$