Similarity of genomic intervals

Table of Contents

Introduction

When performing genomic analyses, data are often presented in the form of genomic intervals. These are nuleotide positions that are calculated with respect to some reference genome. They can correspond to the positions of particular genes on a given chromosome, deletions present in the genome of a person with a particular cancer, or a genomic loci where a particular protein binds.

In a study, a single sample can have many different genomic loci of interest, and it is often in the purview of the study to compare these intervals across a set of samples. For example, when identifying mutations in patients with a particular cancer, it is of interest to identify mutations that are recurrent across a set of patients. This type of analysis helped identify the gene *TP53* as a commonly mutated gene in cancer, leading to the identification of its role in the DNA repair mechanism [REF].

For genomic intervals that correspond to a single nucleotide, determining recurrence across a set of samples is intuitively straightforward if all samples are using the same reference genome; one must simply check whether the nucleotide of interest is present in the set of genomic intervals for each patient. For larger genomic intervals that are multiple nucleotides long, determining whether an interval is recurrent is not as straightfoward. Does overlap of a single nucleotide count as recurrence of the interval as a whole? Should there be a minimum overlap required between the intervals being compared? Should this overlap be the same between the two intervals, or should it be based on the percentage of the interval that is overlapped by the other? How does the binary operation of interval intersection extend to $n > 2$ samples? An example of the scenario can be visualized in 1.

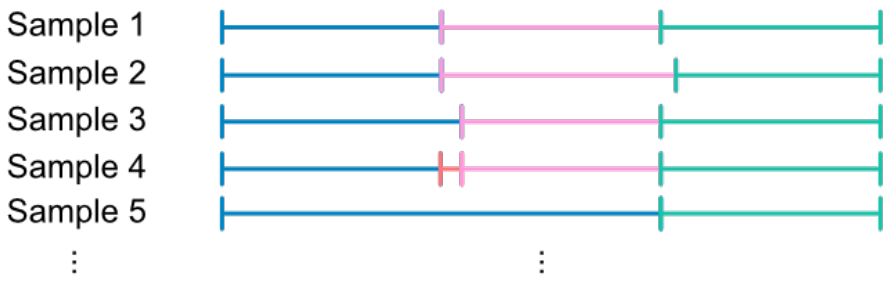

Figure 1: Genome segmentation in multiple samples. Segments are called independently in each sample. The problem consists of determining which segment matches with other segments in other samples, and whether the variation in endpoints is due to a true biological process or error associated with calling the segment boundaries.

In this paper, we address these questions by extending the notion of "equivalence" to a more general mathematical relation, and use this notion to define a measure of similarity that accounts for $n > 2$ samples. We use this as a guiding principle to explore how large genomic intervals should be considered in the context of recurrence.

In Section 1 we describe an examplary problem that inspires the methodology that is the subject of this work. In Section 2 we define a binary relation that extends the canonical equivalence relation of intervals and how to define a measure of similarity. In Section 3 we develop a method to identify "differential" intervals by extending the notion of similarity defined in Section 2. In Section 4 we use real chromosome conformation capture (Hi-C) data to demonstrate the utility of this method.

Problem inspiration

Consider the context of chromosome conformation capture sequencing, where DNA is sequenced in a certain way as to infer the three-dimensional structure of the genome. Briefly, certain genomic elements (genes, promoters, enhancers) come into spatial proximity with each other more frequently than certain distal genomic elements, leading to a hub of local interactions. These hubs, termed topologically associated domains (TADs), are linear intervals in the genome, and identifying these TADs can lead to understanding the regulation of the genes within them [REF].

Determining the boundaries of these TADs, and how similar they are between samples, can help identify novel modes of genetic regulation between samples. An example of this idea can be seen in 2. Our goal for the remainder of this paper is to develop a method to identify intervals that are different for a subset of samples that may be of biological interest.

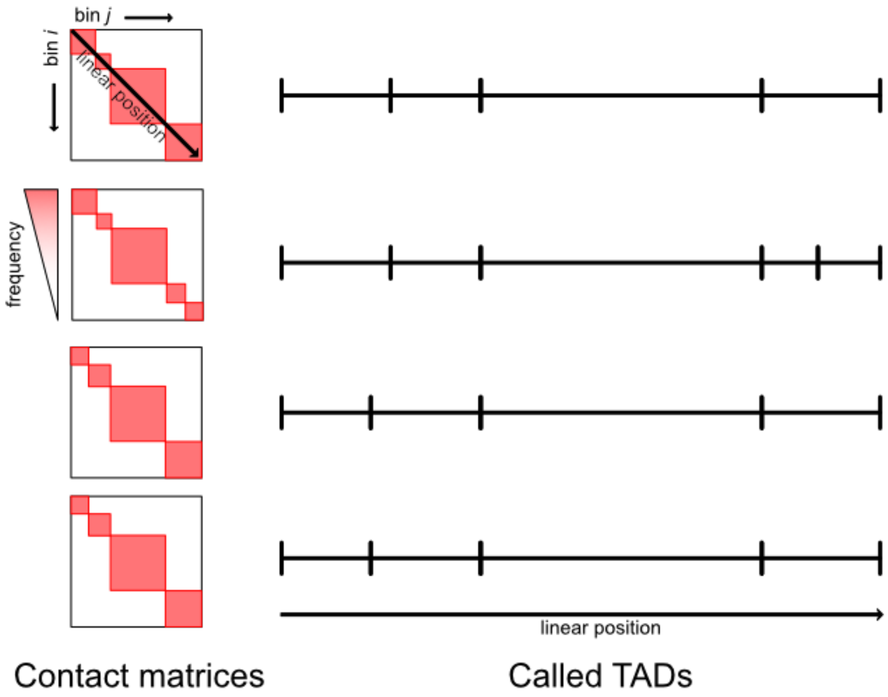

Figure 2: Schematic showing the problem of interest, where one calls topologically associated domains (TADs) from Hi-C contact matrices and compares the called TADs between samples.

Section 2

Definitions

Let $\mathcal{I} = \{[s, e) : s, e \in \mathbb{W}, e \geq s\}$ be the set of half-open discrete intervals on the whole numbers. For two intervals, $a = [a_0, a_1), b = [b_0, b_1) \in \mathcal{I}$, the canonical equivalence relation is $a = b \iff a_0 = b_0, a_1 = b_1$. For intervals themselves, the size of an interval is denoted $|a| = a_1 - a_0$. Let $\mathcal{B} = \{True, False\}$ be the Boolean set.

Relation on intervals

Consider the relation

$$
\begin{aligned}
e_f &: \mathcal{I} \times \mathcal{I} \rightarrow \mathcal{B} \\
e_f(a,b) &= \left\{ \frac{|a \cap b|}{|a|} \geq f \right\} \bigwedge \left\{ \frac{|a \cap b|}{|b|} \geq f \right\}
\end{aligned}
$$

where $f \in [0,1], a, b \in \mathcal{I}$.

Proposition: e_f is an equivalence relation $\iff f \in \{0,1\}$

Clearly, e_f is reflexive ($e_f(a,a)$ is true $\forall f \in [0,1], a \in \mathcal{I}$) and symmetric ($e_f(a,b) = e_f(b,a) \forall a, b \in \mathcal{I}$). It remains to be shown that e_f is transitive.

In the boundary cases for f, $f = 1$ reduces to the canonical equivalence relation on intervals, and $f = 0$ reduces to the trivial equivalence relation ($e_0(a,b) = True \forall a, b \in \mathcal{I}$). Thus $f \in \{0,1\} \implies e_f$ is an equivalence relation.

Let $F = \lceil \frac{1}{1-f} \rceil$, and consider the intervals $a_0 = [0, F), a_1 = [1, F+1), ..., a_F = [F, 2F+1)$ (3).

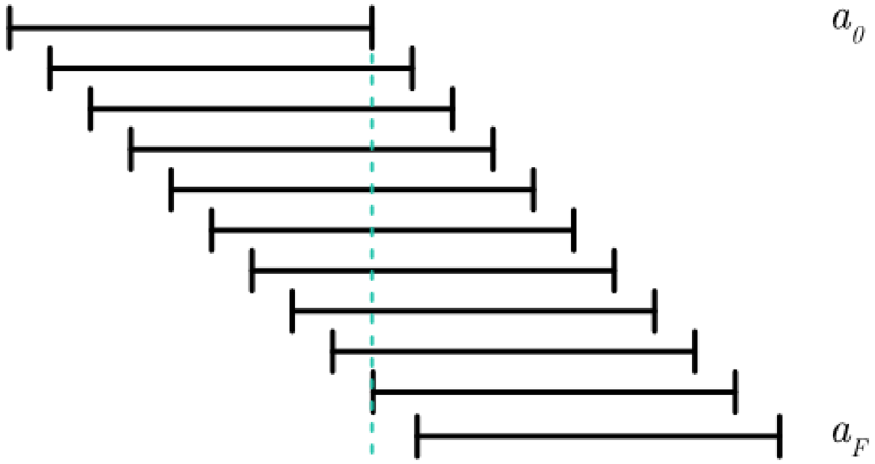

Figure 3: Shifting set of intervals, each with an offset of 1 from the previous interval. a_i and a_{i+1} are all equivalent for a given f, but a_0 and a_F have no intersection.

$$\begin{aligned} e_f(a_0, a_1) &= \left\{ \frac{|[1, F)|}{|[0, F)|} \geq f \right\} \bigwedge \left\{ \frac{|[1, F)|}{|[1, F+1)|} \geq f \right\} \\ &= \left\{ \frac{F-1}{F} \geq f \right\} = \left\{ 1 - \frac{1}{F} \geq f \right\} \\ &= \{1 - (1-f) \geq f\} \\ &= True \end{aligned}$$

Similarly, $e_f(a_i, a_{i+1}) = True \forall i \in 0, ..F-1$. However, $a_0 \cap a_F =$, thus $e_f(a_0, a_F) = False$. Since $e_f(a_0, a_1) = ... = e_f(a_{F-1}, a_F) = True$ and $e_f(a_0, a_F) = False$, e_f is not transitive for $f > 0$.

□

While e_f is not explicitly an equivalence relation, we can use it to define a notion of similarity between intervals.

Similarity of intervals

Intuitively, we can define the *similarity* of two intervals by the maximum fraction of symmetric overlap. Mathematically,

$$\begin{aligned} s &: \mathcal{I} \times \mathcal{I} \to [0, 1] \\ s(a, b) &= \arg\max_f \{e_f(a, b) = True\} \\ &= \frac{|a \cap b|}{\max\{|a|, |b|\}} \end{aligned}$$

This can easily be extended to an arbitrary number of intervals by considering all pairs of intervals in the set.

$$\begin{aligned} s &: 2^{\mathcal{I}} \to [0, 1] \\ s(A) &= \arg\max_f \{e_f(a, b) = True \forall a, b \in A\} \\ &= \inf_{a, b \in A} \left\{ \frac{|a \cap b|}{|a|} \right\} \end{aligned}$$

It is worth noting that if $|A| < \infty$, s is monotonically increasing on subsets of A, which we prove below.

Proposition: s is monotonically increasing on subsets of A

Let $A = \{a_1, ..., a_n\}$ be a finite subset of $\mathcal{I}$. Then

$$\begin{aligned} s(A) &= \min\left\{ \min_{a,b \in \{a_1,...,a_{n-1}\}} \left\{ \frac{|a \cap b|}{|a|} \right\}, \min_{a \in \{a_1,...,a_{n-1}\}} \left\{ \frac{|a \cap a_n|}{|a|}, \frac{|a_n \cap a|}{|a_n|} \right\}, \frac{|a_n \cap a_n|}{|a_n|} \right\} \\ &= \min\left\{ s(\{a_1, ..., a_{n-1}\}), \min_{a \in \{a_1,...,a_{n-1}\}} \left\{ \frac{|a \cap a_n|}{|a|}, \frac{|a_n \cap a|}{|a_n|} \right\}, 1 \right\} \\ &\leq s(\{a_1, ..., a_{n-1}\}) \end{aligned}$$

This is true for any $a \in A$, thus $s(A) \leq s(B) \forall B \subset A$. $\square$

This allows us to think of the similarity of a set of intervals as a value between 0 and 1. Given that s is monotonic, we should consider the value of similarity of 2 intervals differently than we should for 3 or 10. This is analogous to the problem of vanishing distance in high-dimensional data [REF].

Now that we have a quantitative notion[1] of similarity of intervals, we can use this in the context of finding differential intervals between a set of samples.

Section 3

Differential analysis in bioinformatics is the process of identifying statistically significant differences between conditions. For example, this could be differential expression of transcripts in mutant versus wild type conditions. Differential analysis typically comes in two varieties: *a priori* and *de novo*.

A priori differential analysis

A priori differential analysis results from known differences in condition, like the previous example. The response variable is modelled as a (often linear) function of the covariate(s) of interest. This model is fit to the data to estimate noise and account for technical artefacts[2], and null hypothesis significance testing is then carried out on the coefficient to the covariate(s) of interest [REF].

[1] We avoid using the term “measure” here, since we have not defined nor proved that s is a measure, in the strict sense.

[2] In DNA sequencing data, technical artefacts can include over-dispersion of variance, sequencing depth of each sample, and sequencing batch. Differential analysis also requires controlling for nuisance variables that may confound the results if not properly controlled for (e.g. the effect of gender when measuring drug response in patients with or without a mutation).

De novo differential analysis

While the goals of *de novo* and *a priori* differential analysis are the same they differ in detection and controlling for false discoveries. *De novo* differential analysis stems from the discovery of differences between subsets of samples that aren't previously stratified according to some known covariate. These differences must be detected in the first place (without knowledge of nuisance covariates), which itself is typically a statistical procedure. Since the detection step enriches for likely differential intervals, the resulting p-values from the hypothesis testing will not abide by the uniform distribution assumed by most multiple-testing correction procedures.

If we wish to perform some type of differential analysis on our intervals in n samples (or at the very least, identify locations where some subset of samples differ), this amounts to partitioning n samples into $k \geq 2$ groups.