Supplementary Note 1

Abstract

This is a supplementary note to the manuscript 'Deciphering causal genomic templates of complex molecular phenotypes'. We describe the concepts and algorithms introduced in the main manuscript. We give our definition of phenotype sequence alignment (PSA) in Section 1. We define our Neural Alignment of CMPs and Reference genome (NACR) algorithm, used to identify PSAs, in Section 2. In Section 3, we explain the causal phenotype sequence alignment procedure (CPSA), which finds a locations within a PSA corresponding to phenotypic edits, and we define the experimental conditions for a PSA to be considered a template.

1 Phenotype Sequence Alignments

A PSA is a copy (in a precise sense) of a complex molecular phenotype in the genome. We formalize our notion of complex molecular phenotype (CMP) in Section 1.1. We formulate the genome as a metric space in Section 1.2. Finally, we define PSAs in Section 1.3.

1.1 Complex molecular phenotypes

We define a complex molecular phenotype (CMP) to be a measured metric space:

Definition 1 *A* complex molecular phenotype *(CMP) is a triple* (X, d, p)*, where* X *is a set,* d *is a metric on* X*, and* p *is a probability measure on* X*.*

For background on measured metric spaces relevant to our application, see (Mémoli, 2017). We consider two examples of CMPs:

1. $X = \mathbb{R}^{\#\text{genes}}$, d is the Euclidean metric, and p is the distribution of (log) gene expression measurements for single-cells in a biological experiment.
2. $X = \mathbb{R}^{\#\text{cell-types}}$, d is the Euclidean metric, and p is the distribution of (log) cell type counts for patches of cells extracted from a tissue via an imaging experiment.

In practice, we will work with empirical distributions p with finite support within X. Note that the support is a subset of X, not X itself. We refer to the pair (X, d) as the phenotype space of the CMP and this empirical distribution p as the points of the CMP.

1.2 The genome as a metric space

We formulate the genome in a way that enables searching for copies of CMPs within it. For this, we introduce the notion of *k-mer functional*. A k-mer s is a nucleotide string of length k; i.e., an element of the set $\mathcal{S}^k := \{A, T, C, G\}^k$. For example, the nucleotide string $AAG \in \mathcal{S}^3$ is a 3-mer.

Definition 2 *A* k-mer functional $\lambda : \mathcal{S}^k \to \mathbb{R}$ *assigns a real value to each element of* $\mathcal{S}^k$*.*

Remark 3 *k-mer functionals provide a natural way to restrict the complexity of real valued functions evaluated at each position in the genome, and will enable computationally finding PSAs, as discussed in the following sections.*

The support of a k-mer functional is the set of k-mers at which the functional takes non-zero value. The k-mer functionals are direct generalizations of k-mers: a k-mer $s \in \mathcal{S}^k$ corresponds to the functional that takes value 1 at s, and is 0 at all other k-mers.

We now represent a genome as a metric space of k-mer functionals.

Definition 4 *Let $\mathbb{G}$ be a genome sequence. The* metric space of k-mer functionals of $\mathbb{G}$, *denoted $\mathcal{M}^k(\mathbb{G})$, is the vector space of k-mer functionals supported on k-mers in $\mathbb{G}$, with distance between k-mer functionals λ and μ equal to*

$$d(\lambda, \mu) = \left(\frac{1}{N} \sum_s (\lambda(s) - \mu(s))^2 \right)^{\frac{1}{2}}, \tag{1}$$

where s runs over all k-mers in $\mathbb{G}$ and N is the number of k-mers in $\mathbb{G}$.

Remark 5 *The distance in* (1) *equals*

$$d(\lambda, \mu) = \left(\frac{1}{N} \sum_s n_s (\lambda(s) - \mu(s))^2 \right)^{\frac{1}{2}},$$

where s sums over the unique k-mers in $\mathbb{G}$ and n_s is the number of times s occurs in $\mathbb{G}$.

In the manuscript, where it is unambiguous, we referto $\mathcal{M}^k(\mathbb{G})$ as 'the genome'. Note that a single genome is a metric space of k-mer functionals, not an individual k-mer functional.

1.3 Phenotype sequence alignments

Now that we have represented both CMPs and the genome as metric spaces, we can define phenotype sequence alignments. Let $\text{supp}(p)$ denote the support of a distribution p.

Definition 6 *Let $\mathcal{P} = (X, d, p)$ be a complex phenotype and $\mathbb{G}$ be a genome sequence. A* phenotype sequence alignment (PSA) *of $\mathcal{P}$ with $\mathbb{G}$ is a function $T : X \to \mathcal{M}^k(\mathbb{G})$ such that*

$$\sup_{x,y \in supp(p)} (d(T(x), T(y)) - d(x, y)) = 0. \tag{2}$$

We make some remarks on Definition 6:

- With no restriction on the map T, PSAs for a CMP with finitely supported probability measure exist with any genome with sufficiently many unique k-mers. The empirical results of the main manuscript explore relevant restrictions on T for identification of biologically meaningful PSAs.
- Given restrictions on the complexity of T, we do not expect there to always exist exact PSAs; the term PSA is used in the main manuscript to refer to local minimizers of the magnitude of the quantity on the left hand side of (2).

- The notion of distance in (1) is the ∞-Gromov-Wasserstein distance (Mémoli et al., 2017). It can be relaxed to q-Gromov-Wasserstein distances by replacing (2) by

$$\int |d(T(x), T(y)) - d(x, y)|^q \, dp(x)dp(y) = 0.$$

Exploration of such relaxations is outside the scope of this work.

2 Neural Alignment of CMPs with a Reference genome

In the following, let (X, d, p) be a CMP and let $\mathbb{G}$ be a genome sequence. Our computational strategy to find approximate PSAs is called "Neural Alignment of CMPs with a Reference genome" (NACR). It is defined as follows:

1. Define a map $T_\theta : X \to \mathcal{M}^k(\mathbb{G})$ by:

$$T_\theta(x)(s) = \langle F_\theta(x), G_\theta(s) \rangle,$$

where $F_\theta : \mathbb{X} \to \mathbb{R}^m$ and $G_\theta : \{A, T, C, G\}^k \to \mathbb{R}^m$ are neural networks parameterized by θ. The map F_θ is called the *phenotype-embedding network*. Its input is a point $x \in X$ in phenotype space and its output is a point in $\mathbb{R}^m$. The map G_θ is called the *sequence-embedding network*. Its input is a k-mer $s \in \{A, T, C, G\}^k$ and its output is a point in $\mathbb{R}^m$. The space $\mathbb{R}^m$ is a latent inner product space.

2. Fix a batch B of samples from the phenotype distribution and a set S of k-mers from $\mathbb{G}$. An overall loss function is computed from the following three loss functions:

   - An isometry deviation loss:

$$\begin{aligned} \mathcal{L}_{\text{isom}} &= \max_{(x,y) \in B \times B} \left| \frac{1}{|S|} \sum_{s \in S} |T_\theta(x)(s) - T_\theta(y)(s)|^2 - d(x, y)^2 \right| \\ &\approx \max_{(x,y) \in B \times B} \left| d(T_\theta(x), T_\theta(y)) - d(x, y)^2 \right|. \end{aligned}$$

   The quantity on the second line is a Monte Carlo estimate, since S is a sample of k-mers from $\mathbb{G}$.

   - A sparsity regularization loss:

$$\mathcal{L}_{\text{sp}} = \sum_{x \in B} \sum_{s \in S} |T_\theta(x)(s)|.$$

   This encourages the model to be nonzero on a small number of genomic regions (which can be subsequently interpreted).

   - A complexity regularization loss $\mathcal{L}_{\text{c}}$. This is the usual ℓ_2 loss on neural network weights and is included to limit the complexity of the functionals in the trained model.

3. The overall loss $\mathcal{L}_{\text{isom}} + \alpha\mathcal{L}_{\text{sp}} + \beta\mathcal{L}_{\text{c}}$ is minimized by stochastic gradient descent, sampling sequence batches S and data batches B.

A *null alignment*, used for statistical hypothesis testing, is obtained by randomly initializing but not training the neural networks of NACR.

3 Causal phenotype sequence alignment

We defined PSAs in Section 1, and NACR, a computational strategy to identify approximate PSAs, in Section 2. We now describe our procedure for generating predictions whose verification establishes whether a PSA is a template. This procedure is termed "causal phenotype sequence alignment" (CPSA), because it generates predictions about which genomic loci are causally sufficient to achieve an edit to a CMP.

We define the discrepancy functional for a phenotypic edit in Section 3.1. The genomic loci where the discrepancy functional for an edit is high are those loci in the PSA predicted to achieve the edit. We discuss how to compute the discrepancy functional in Section 3.2.

3.1 The discrepancy functional

We define edits to a CMP in Section 3.1.1 , before defining the discrepancy functional in Section 3.1.2 and the CPSA procedure in Section 3.1.3.

3.1.1 Edits to a CMP

In the manuscript, we pictured edits as moving around the points of a CMP (see Figure 1B.2 in the main manuscript). We make this mathematically precise:

Definition 7 *Let $\mathcal{P} = (X, d, p)$ be a CMP. An* edit *to $\mathcal{P}$ is a probability distribution $\pi(x, y)$ on $X \times X$ that marginalizes over the second argument to p:*

$$p(x) = \int_X \pi(x, y)dy.$$

The edited *CMP is (X, d, q) where $q(x) = \int_X \pi(x, y)dx$.*

This definition of edit corresponds to moving points of the CMP in a probabilistic way, as follows. Suppose we have a CMP (X, d, p) where p is a probability distribution given by uniformly weighting points in the set $A := \{x_1, \ldots, x_n\} \subset X$. Then moving around the points of the CMP corresponds to selecting a function $m : A \to X$, which defines the probability distribution on $X \times X$ that assigns a uniform weight to points in $\{(x_1, m(x_1)), \ldots, ((x_n, m(x_n)))\} \subset X \times X$. Thus, given an edit $\pi(x, y)$ of p, $\pi(y|x)$ represents the probability that a point $x \in X$ is sent by the edit to y.

3.1.2 The discrepancy functional

We now define a k-mer functional corresponding to a given edit to a CMP.

Definition 8 *Let $\mathcal{P} = (X, d, p)$ be a CMP, $\mathbb{G}$ be a genome sequence, $T : X \to \mathcal{M}^k(\mathbb{G})$. The* discrepancy functional *$dT_\pi(s)$ of T for an edit π at the k-mer s is:*

$$dT_\pi(s) = \int_{X \times X} \pi(x, y)(T(x)(s) - T(y)(s))^2 dxdy.$$

If the edit corresponds to shifting a single cluster mean from $x \in X$ to $y \in X$, the k-mers that have a high discrepancy are those where $T(x)$ differs highly from $T(y)$.

We provide intuition for this definition of the discrepancy functional. Intuitively, the cost of an edit π to a CMP can be expressed in terms of how far each point moves, i.e.,

$$|\pi| := \int_{X\times X} d(x,y)^2 d\pi(x,y)$$

If T is a PSA, for all edits π of $\mathcal{P}$, we have

$$\begin{aligned}
|\pi| &= \int_{X\times X} d(x,y)^2 d\pi(x,y) \\
&\approx \int_{X\times X} \frac{1}{N} \sum_{k-\text{mers } s \text{ in } \mathbb{G}} \left(T(x)(s) - T(y)(s)\right)^2 d\pi(x,y) \\
&= \frac{1}{N} \sum_{k-\text{mers } s \text{ in } \mathbb{G}} dT_\pi(s),
\end{aligned}$$

where the $\approx$ is based on the assumption that T approximately preserves distances. Therefore, the value of the discrepancy functional at a k-mer can be interpreted as how much that k-mer contributes to the cost of the edit.

3.1.3 Selecting the edits

We want CPSA to make predictions whose experimental validation can be regarded as confirming that a given PSA is a template. We model an experiment as observing an original CMP $\mathcal{P} = (X, d, p)$, making some perturbation to the genome, and subsequently observing a new CMP $\mathcal{Q} = (X, d, q)$. Therefore CPSA takes as input a "target" CMP $\mathcal{Q} = (X, d, q)$ and predicts genomic loci whose manipulation is predicted to achieve a change from the original CMP to the target CMP.

Given two distributions p and q on X, there are multiple distributions on $X \times X$ that marginalize to p and q in each factor respectively. Therefore, CPSA has to make an additional assumption on which edit to choose.

Definition 9 *Causal phenotype sequence alignment (CPSA) is the procedure with input:*

- *An observed CMP $\mathcal{P} = (X, d, p)$ in an organism with genome $\mathbb{G}$,*
- *A PSA $T : X \to \mathcal{M}^k(\mathbb{G})$, and*
- *A target CMP $\mathcal{Q} = (X, d, q)$.*

The output is the functional dT_{π^*}, where π^* is the optimal transport coupling between p and q; i.e. the distribution on $X \times X$ that minimizes the cost

$$|\pi| := \int_{X\times X} d(x,y)^2 d\pi(x,y)$$

subject to

$$\int_X \pi(x,y)dy = p, \text{ and } \int_X \pi(x,y)dx = q$$

For background on optimal transport, see (Peyré and Cuturi, 2019).

3.2 CPSA in practice

We now discuss the practical considerations for computationally implementing CPSA. We discuss the implementation in Section 3.2.1 and its limitations in Section 3.2.2.

3.2.1 Implementation of CPSA

When editing CMPs in the manuscript, distributions considered are represented as the mean and frequency of different non-overlapping subsets of the data (i.e., clusters of cells in Figure 3, or the set of image patches from a patient sample in Figure 4J). Specifically, the CMP is an empirical distribution on phenotype space $(p_1, x_1), \ldots (p_m, x_m)$ where p_i denotes the weight assigned to x_i and the target CMP is an empirical distribution $(q_1, y_1), \ldots (q_n, y_n)$.

The output of CPSA is the discrepancy functional, as defined above. The general procedure is:

- Compute the optimal transport coupling between p and q

  $$\pi^* = \arg\min |\pi|$$

  using an empirical optimal transport solver (Flamary et al.).

- For each k-mer s, compute

  $$d_T(\pi^*)(s) = \sum_{i,j} \pi^*_{i,j} |T(x_i)(s) - T(y_i)(s)|^2.$$

  This expression can be computed using the trained genotype-embedding and phenotype-embedding networks of NACR, defined in Section 2.

The discrepancy functional for a phenotypic change is subsequently aggregated over annotated genomic regions (genes, promoters, codons) for biological interpretation.

3.2.2 Limitations of the implementation

We conclude this supplementary note with some remarks on the limitations of CPSA.

- Working with empirical distributions corresponding to cluster means and frequencies ignores the covariance structure of the distribution for each cluster.

- In the special case that the phenotypic change shifts one cluster mean x to mean $x + \delta x$, without changing the frequencies of clusters), we assume that the optimal transport coupling is given by translation $\pi : x \mapsto x + \delta x$. Thus, the discrepancy functional is assumed to be:

  $$dT_\pi(s) = (T(x)(s) - T(x + \delta x)(s))^2.$$

- The use of optimal transport couplings as defined here does not incorporate local sources and sinks for the measures in question (in the same way that searching for exact matchings between sequences does not allow for insertions or deletions). Other notions of coupling may provide a more effective and biologically relevant discrepancy computation.