Supplementary Note 2

Abstract

This is a supplementary note to the manuscript 'Deciphering causal genomic templates of complex molecular phenotypes'. We show how PSAs are a formal analogue of deciphering the amino-acid code to find protein coding sequences in the genome. We provide an abstract definition of alignment in Section 1. We show how this definition captures deciphering the amino-acid code in Section 2. We apply it to PSAs of CMPs in Section 3.

1 Formal encodings

We first provide a formal definition of alignment. The definition is a special case of the definition of structure in the sense of model theory (Grädel et al., 2007).

Definition 1 *Let $\mathcal{R}$ be a set of symbols. An* $\mathcal{R}$-structure *is a function $f : \mathcal{R} \to \mathcal{P}(A_f \times A_f)$, where A_f is the entities of f and $\mathcal{P}(A_f \times A_f)$ is the set of binary relations on A_f. The image $f(\mathcal{R})$ is the set of relations of f. Given $\mathcal{R}$-structures f and g, an* alignment of f with g *is a map $T : A_f \to A_g$ such that for any $R \in \mathcal{R}$ and all $x, y \in A_f$,*

$$(x, y) \in f(R) \Rightarrow (T(x), T(y)) \in g(R).$$

Under this definition, searching for an alignment means searching for a relation preserving function between sets of entities. We will show that this framework describes the procedure for finding protein coding sequences in the genome as well as for finding PSAs.

2 Protein sequences

We show how alignments of of protein sequences with the genome satisfy Definition 1. We first describe how protein sequences and genome sequences are $\mathcal{R}$-structures. We then show how finding an alignment, in the sense of Definition 1, captures the task of finding a coding sequence in the genome.

- Let $\mathcal{R} = \{\text{Precedes, Equals}\}$.
- Let $\mathcal{M} = \{\text{Met., Leu.}, \ldots\}$ be the set of amino acids. A protein, with amino-acid sequence $a_1, a_2, \ldots, a_n$ where $a_i \in \mathcal{M}$, is an $\mathcal{R}$ structure, as follows. The entities are $A = \{(a_1, 1), \ldots, (a_n, n)\}$ and $f : \mathcal{R} \to \mathcal{P}(A \times A)$ is

$$((a, i), (b, j)) \in f(\text{Precedes}) \Leftrightarrow j = i + 1 \qquad \text{and}$$
$$((a, i), (b, j)) \in f(\text{Equals}) \Leftrightarrow a = b$$

- Let $\mathcal{C} = \{AAA, AAT, ATT, \ldots\}$ be the set of three-nucleotide "codons". A genome, with nucleotide sequence $g_1, \ldots, g_k$, is an $\mathcal{R}$ structure, as follows. The entities are

$$B = \{(c, i) : \text{the three nucleotides starting at } i \text{ are given by } c\}$$

and $g : \mathcal{R} \to \mathcal{P}(B \times B)$ is given by

$$((c, i), (d, j)) \in g(\text{Precedes}) \Leftrightarrow j = i + 3 \qquad \text{and}$$
$$((c, i), (d, j)) \in g(\text{Equals}) \Leftrightarrow c = d$$

- An alignment is a map $T : A \to B$ such that

$$((a, i), (b, j)) \in f(\text{Precedes}) \Rightarrow (T((a, i)), T((b, j))) \in g(\text{Precedes})$$

and

$$((a, i), (b, j)) \in f(\text{Equals}) \Rightarrow (T((a, i)), T((b, j))) \in g(\text{Equals}).$$

That is, T assigns positions of amino acids in a protein sequence to the positions of codons in the genome sequence such that if the amino acids are equal at two positions, then so too are their assigned codons. This description does not account for synonymous codons and alignments with insertions or deletions.

3 Complex molecular phenotypes

We show that PSAs of CMPs with a genome satisfy Definition 1. We first describe how CMPs and genome sequences are $\mathcal{R}$-structures in the sense of Definition 1. We then show how finding an alignment in the sense of Definition 1 captures the task of finding PSAs as described in the manuscript.

- Let $\mathcal{R} = \mathbb{R}^+ \cup \{0\}$ be the nonnegative real numbers.
- Let $\mathcal{X} = (X, d)$ be phenotype space, for example, $\mathcal{X} = (\mathbb{R}^G, d)$ where G is the set of genes and d is the Euclidean distance between gene expression vectors. A complex phenotype, viewed as a finite subset $A \subset \mathcal{X}$, is an $\mathcal{R}$ structure as follows. The set of entities is A and $f : \mathcal{R} \to \mathcal{P}(A \times A)$ is given by

$$(x, y) \in f(r) \Leftrightarrow d(x, y) = r$$

- Let $\mathbf{G}$ be the genome, viewed as a metric space of k-mer functionals as described in the manuscript and Supplementary Note 1 . Then $\mathbf{G}$ is an $\mathcal{R}$-structure. The set of entities is $\mathbf{G}$ and $g : \mathcal{R} \to \mathcal{P}(\mathbf{G} \times \mathbf{G})$ is given by

$$(x, y) \in g(r) \Leftrightarrow d(x, y) = r$$

- An alignment is a map $T : A \to \mathbf{G}$ such that

$$(x, y) \in f(r) \Rightarrow (T(x), T(y)) \in g(r)$$

That is, T maps observations of the CMPs to the genome in a way that preserves distances.