Genetics and population analysis

**HypercubeME: two hundred million combinatorially complete datasets from a single experiment**

Laura A. Esteban¹, Lyubov R. Lonishin², Daniil M. Bobrovskiy³, Gregory Leleytner⁴, Natalya S. Bogatyreva¹,⁵,⁶, Fyodor A. Kondrashov⁷ and Dmitry N. Ivankov¹,⁸,*

¹Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain, ²Faculty of Medical Physics, Institute of Biomedical System and Technologies, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg 19251, Russia, ³Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow 119234, Russia, ⁴Department of Innovation and High Technology, Moscow Institute of Physics and Technology, Moscow 141701, Russia, ⁵Bioinformatics and Genomics Programme, Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain, ⁶Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, Moscow 142290, Russia, ⁷Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria and ⁸Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow 121205, Russia

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on August 29, 2019; revised on November 1, 2019; editorial decision on November 7, 2019; accepted on November 18, 2019

**Abstract**

**Motivation:** Epistasis, the context-dependence of the contribution of an amino acid substitution to fitness, is common in evolution. To detect epistasis, fitness must be measured for at least four genotypes: the reference genotype, two different single mutants and a double mutant with both of the single mutations. For higher-order epistasis of the order \(n\), fitness has to be measured for all \(2^n\) genotypes of an \(n\)-dimensional hypercube in genotype space forming a 'combinatorially complete dataset'. So far, only a handful of such datasets have been produced by manual curation. Concurrently, random mutagenesis experiments have produced measurements of fitness and other phenotypes in a high-throughput manner, potentially containing a number of combinatorially complete datasets.

**Results:** We present an effective recursive algorithm for finding all hypercube structures in random mutagenesis experimental data. To test the algorithm, we applied it to the data from a recent HIS3 protein dataset and found all 199,847,053 unique combinatorially complete genotype combinations of dimensionality ranging from 2 to 12. The algorithm may be useful for researchers looking for higher-order epistasis in their high-throughput experimental data.

**Availability and implementation:** https://github.com/ivankovlab/HypercubeME.git.

**Contact:** d.ivankov@skoltech.ru

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

Epistasis, the dependence of the impact of a mutation on the genetic context, is abundant and important phenomenon in molecular evolution (Breen *et al.*, 2012). Formally, epistasis is characterized by coefficients \(z\) having two or more indices in the following representation of fitness \(f\) as a function of a genotype \(g\) (assuming, for simplicity, that maximum of one mutation is allowed at any position):

\[
f(g) = \text{const} + \sum_{i=1}^{N} z_i g_i + \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ij} g_i g_j + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} z_{ijk} g_i g_j g_k + \cdots
\]

where sums are taken over \(N\) considered positions, \(g_i = 1\) if \(i\)-th position is mutated in genotype \(g\); otherwise \(g_i = 0\) or \(g_i = -1\), depending on the formalism of epistasis description (Poelwijk *et al.*, 2016). Coefficients \(z\) correspond to a single effect of the mutation in the \(i\)-th position. Coefficients \(z_{ij}\), having two indices, represent the pairwise epistasis between positions \(i\) and \(j\), while coefficients \(z\) having three or more indices correspond to higher-order epistasis (de Araujo and Guimaraes, 2016; Orwinowski *et al.*, 2018; Poelwijk *et al.*, 2016; Sailer and Harms, 2017a, b, c; Tuo, 2018; Weinreich *et al.*, 2013, 2018).

To detect epistatic terms of the order \(n\) by means of Walsh transform (Weinreich *et al.*, 2013), one has to measure phenotypes of all \(2^n\) genotypes forming an \(n\)-dimensional hypercube in genotype space. Such experimental datasets are called 'combinatorially complete datasets' (Weinreich *et al.*, 2013). Up to now, higher-order
epistasis was studied using only a handful of examples carefully designed to have all \(2^n\) combinations (Weinreich et al., 2013). On the other hand, high-throughput experiments using (quasi-)random mutagenesis have produced vast amounts of data: 51 715 measured genotypes for GEP (Sarkisyan et al., 2016), over 65 000 for arginine tRNA (Li et al., 2016), 956 648 for HIS3 protein (Pokusaeva et al., 2019), to name a few. These experiments may contain a number of combinatorially complete datasets as subsets of a general dataset. However, the extraction of such hypercubes from a large dataset is not straightforward, and may not be feasible to do in a brute-force manner.

2 Algorithm

The algorithm uses the fact that any \(n\)-dimensional hypercube contains two opposite hyperfacets, which are parallel to each other. Those hyperfacets are hypercubes of dimensionality \((n−1)\), which, in turn, are built from parallel hypercubes of dimensionality \((n−2)\), etc. down to hypercubes of 0-th dimensionality (which are simply points in genotype space, i.e. genotypes).

The algorithm consists of repeating steps. At each step, all possible \(n\)-dimensional hypercubes are generated from the set of \((n−1)\)-dimensional hypercubes. Informally, at each step, the algorithm takes all pairs of existing parallel hypercubes and if the distance between the hypercubes in the pair is one, the pair composes the hypercube of a higher dimensionality (Fig. 1). We need to define the diagonal of a combinatorially complete dataset as a list of mutations transforming a genotype of the dataset to the most distant genotype of the same dataset. An \(n\)-dimensional combinatorially complete dataset therefore contains \(2^{n−1}\) diagonals, each of which (if not empty) can be written in the forward and reverse direction.

Formally, the algorithm consists of the following steps:

Each step of the algorithm can be easily parallelized. The multi-processor version can be found at https://github.com/ivankovlab/HypercubeME.git.

3 Results

We have applied the algorithm to the recently published fitness landscape for HIS3 protein (Pokusaeva et al., 2019), the biggest fitness map...
landscape published so far. The HIS3 protein was divided into 12 segments, and quasi-random mutagenesis has been done in each segment separately. We had to exclude indels and mutations outside the segment, so the number of considered experimentally measured genotypes ranged from 16 182 for segment S7 to 82 081 for segment S2, overall summing up to 721 791 genotypes (Supplementary Table S1).

We have found all 199 847 053 hypercubes having dimensionality from 2 to 12. The single-processor working time ranged from 2 h for S7 to almost 10 days for S5. Among the found hypercubes, the percentage of squares was 12%, while the remaining 88% had dimensionality 3 and higher and, thus, can be used for exploring higher-order epistasis. The number of found hypercubes throughout segments is given in Supplementary Table S2.

Acknowledgement
The authors are grateful to Cathy Shufro for help with editing the text.

Funding
This work was supported by the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME) and Startup package to the Ivankov laboratory at Skolkovo Institute of Science and Technology. The work was started at the School of Molecular and Theoretical Biology 2017 supported by the Zimin Foundation. N.S.B. was supported by the Woman Scientists Support Grant in Centre for Genomic Regulation (CRG).

Conflict of Interest: none declared.

References


