# Finding multi-dimensional epistasis in high-throughput experimental data at the level of individual genotypes

Laura Aviñó Esteban

*Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain,* laura.avino@alum.esci.upf.edu

Natalya S. Bogatyreva

*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain; Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow region, 142290, Russia;* natali.bogatyreva@gmail.com

Fyodor A. Kondrashov

*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain; Laboratory; Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Pg. Lluís Companys, 08010 Barcelona, Spain;* fyodor.kondrashov@crg.es

Dmitry N. Ivankov

*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain; Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow region, 142290, Russia;* ivankov13@gmail.com

Epistasis is known as one of the major factors in the molecular evolution (1). In protein molecular evolution epistasis denotes a non-additive dependence of fitness on amino acid changes introduced into protein sequence (2). To find epistasis, we need to analyze experiments where fitness or some characteristics of the studied protein is measured for many different genotypes (2). Different designs of experiments can produce combinatorially complete dataset of genotypes (3) or a dataset where nucleotide variants are generated randomly (2).

If experimental data points can be fitted by a non-linear monotonic function of fitness potential, such epistasis is called unidimensional because only one explanatory variable, fitness potential, is sufficient for describing the measured phenotypes. Otherwise, epistasis is multi-dimensional because we need more than one dimensions to plot fitness as a many-variable monotonic function in a multi-dimensional genotype space (e.g., sign and reciprocal sign epistasis are cases of multi-dimensional epistasis). As for the interpretation, the

unidimensional epistasis can be ascribed to a non-linear relationship between the measured protein characteristics and fitness. Oppositely, the multidimensional epistasis seems to result from some physical interactions in the studied protein. So, for finding physical interactions in proteins it is important to develop methods for finding multi-dimensional epistasis in high-throughput data.

Here we present two methods for finding multi-dimensional epistasis in experimental high-throughput data at the level of individual genotypes.

The first approach is a modification of the analysis where cases of sign and reciprocal sign epistasis are found. Normally, this is done when four considered genotypes, $g_{00}$, $g_{01}$, $g_{10}$, and $g_{11}$ result from two individual amino acid substitutions, the first substitution corresponds to $g_{00}$ <–> $g_{01}$ and $g_{10}$ <–> $g_{11}$, while the second for $g_{00}$ <–> $g_{10}$ and $g_{01}$ <–> $g_{11}$. The four genotypes, thus, form quadrat with the side equal to one amino acid substitution. Our approach is to find and analyze all groups of four genotypes where substitutions do not need to be single, thus, forming the set of all rectangles in the measured part of protein sequence space.

The second method is a new type of multi-dimensional epistasis, having more fine structure than sign and reciprocal sign epistasis. Consider groups of four genotypes that fit the picture of unidimensional epistasis if taken individually (again, these groups of four do not need to contain only single substitutions). We found that at some conditions two groups of four genotypes cannot fit the unidimensional picture simultaneously. In the presented work we elucidated all cases when the multi-dimensional epistasis of that kind can occur. In words, this principle can be loosely formulated as follows: “if one amino acid change is better than some other amino acid change in one genetic context but is worse in another genetic context, this is a case of multi-dimensional epistasis”.

The methods presented here have practical importance for analysis of fitness landscapes.

The study was in part funded by HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2015-68723-P), Spanish Ministry of Economy and

Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME).

1. M.S.Breen, C.Kemena, P.K.Vlasov, C.Notredame, F.A.Kondrashov (2012) Epistasis as the primary factor in molecular evolution. *Nature* **490:**535-538.

2. K.S.Sarkisyan *et al*. (2016) Local fitness landscape of the green fluorescent protein. *Nature* **533:**397-401.

3. Z.R.Sailer, M.J.Harms (2017) Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205:**1079–1088.