DOI 10.18699/MM-HPC-BBB-2018-03

Finding epistasis in high-throughput experimental data

L. Aviño Esteban[1], N.S. Bogatyreva[1, 2, 3], F.A. Kondrashov[4], D.N. Ivankov[3, 4]*
[1] *Universitat Pompeu Fabra* (*UPF*)*, Barcelona, Spain*
[2] *Bioinformatics and Genomics Programme, Centre for Genomic Regulation* (*CRG*)*, Barcelona, Spain*
[3] *Laboratory of Protein Physics, Institute of Protein Research of the RAS, Pushchino, Moscow region, Russia*
[4] *Institute of Science and Technology, Klosterneuburg, Austria*
** e-mail: ivankov13@gmail.com*

Key words: epistasis, fitness, higher-order epistasis, multi-dimensional epistasis

Motivation and Aim: Epistasis is one of the most important factors of molecular evolution. Epistasis in its simplest form stands for a phenomenon when the fitness of double mutant differs from the fitness expected from the two single mutants [1]. For higher-order epistasis, we look for the deviation between the fitness of multiple mutant and the fitness expected from all the mutants of lower order [2]. Another concept in protein fitness landscapes is multi-dimensional epistasis. This is the type of epistasis when experimental data cannot be fitted by a monotonic function of fitness potential, the linear combination of contributions from single amino acid substitutions [3]. To analyze epistasis, we have to find hypercubes either in two-dimensional space or in a higher-dimensional space. Different designs of experiments can produce combinatorially complete datasets of genotypes [2] or much bigger datasets where nucleotide variants are generated randomly [1].
Methods and Algorithms: Three algorithms were designed and implemented to obtain the results.
Results: First, in the presented work we find all hypercubes in the random mutagenesis dataset of yeast protein HIS3 [4]. For more than 700 thousand measured phenotypes we found more than 170 millions hypercubes, the biggest dataset available so far. Next, we realize here an idea that genotypes can be searched at any distance. Thus, we can investigate epistasis in hyperrectangles, not only in hypercubes. Using this approach, we found much more rectangles in genotype space than squares. And last, we present here a completely new type of multi-dimensional epistasis when two groups of four genotypes fit unidimensional picture individually but not simultaneously. In the presented work we elucidated all >20000 cases when the multi-dimensional epistasis of that kind can occur in the experimental data of GFP [1].
Conclusion: Overall, the methods presented here have practical importance for the analysis of fitness landscapes.
Acknowledgements: Supported by the 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].
References
1. Sarkisyan K.S. et al. (2016) Local fitness landscape of the green fluorescent protein. Nature 533:397- 401.
2. Poelwijk F.J. et al. (2016) The context-dependence of mutations: a linkage of formalisms. Plos Comp. Biol. 12:e1004771.
3. Kondrashov F.A., Kondrashov A.S. (2001) Multidimensional epistasis and disadvantage of sex. PNAS 98:12089-12092.
4. Pokusaeva V. et al. (2017) Experimental assay of a fitness landscape on a macroevolutionary scale. BioRxiv: