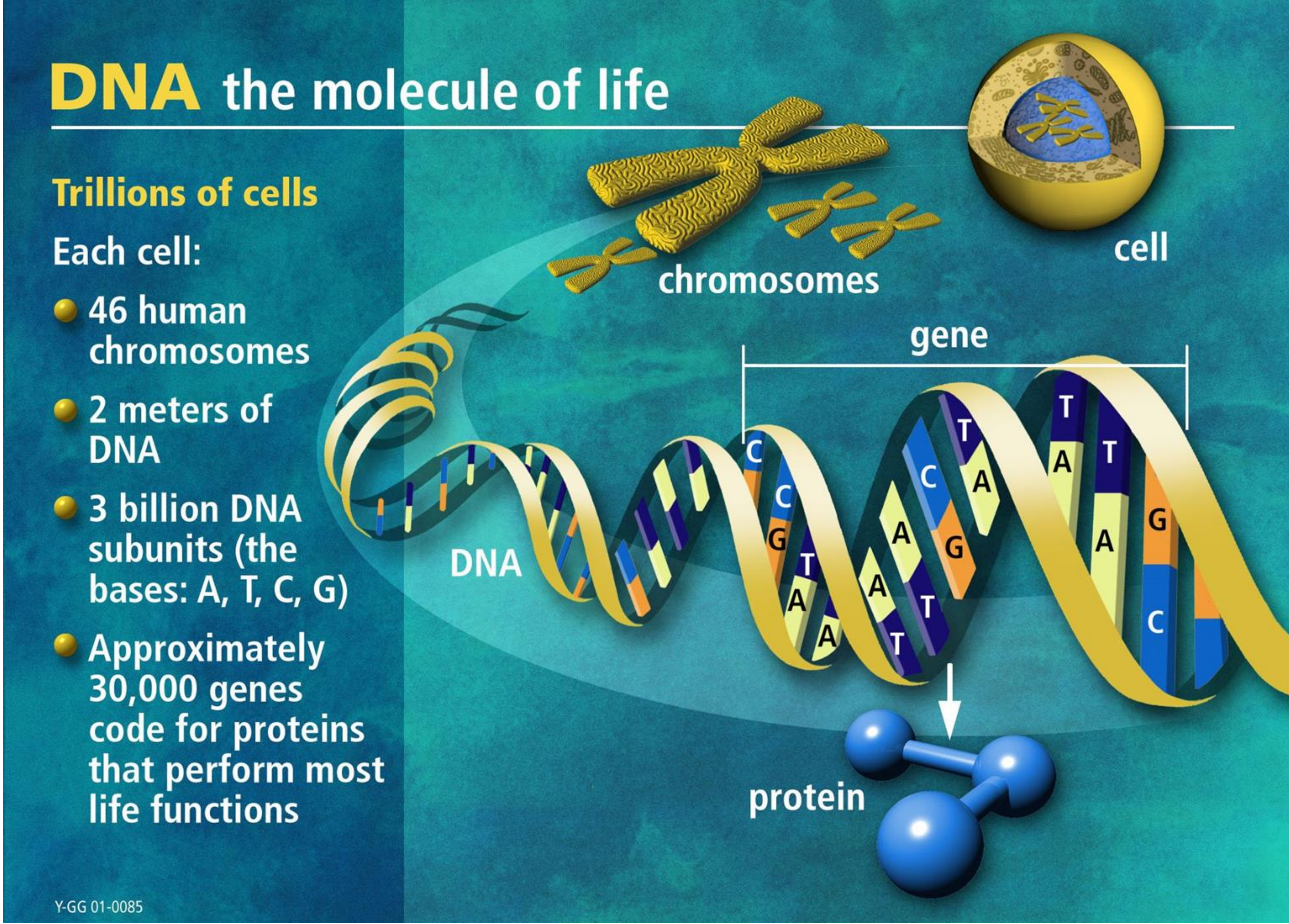
Поиск эпистаза в экспериментальных данных, полученных случайным мутагенезом

Семинар «Суперкомпьютерное моделирование в науке и инженерии», МИЭМ НИУ ВШЭ

20.11.2019

Дмитрий Иванков

Genome



DNA the molecule of life

Trillions of cells

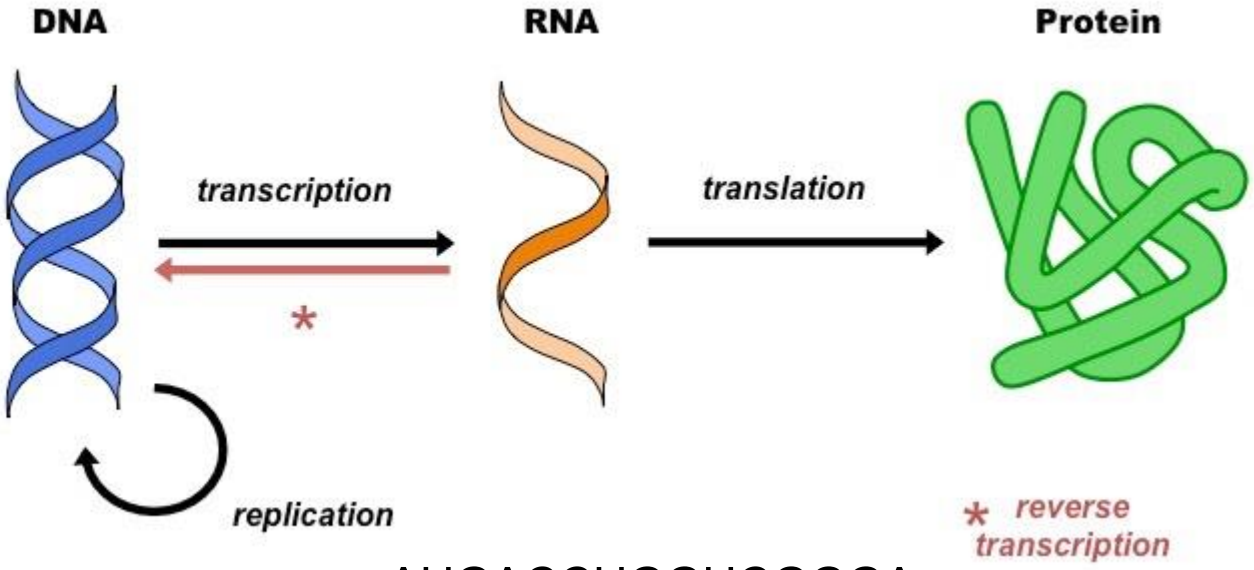
Each cell:

- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions

Y-GG 01-0085

Central dogma of molecular biology

- DNA and RNA alphabet: 4-letter (base pairs)
- Protein alphabet: 20-letter (amino acid residues)



…AUGACCUGGUCGGGA…

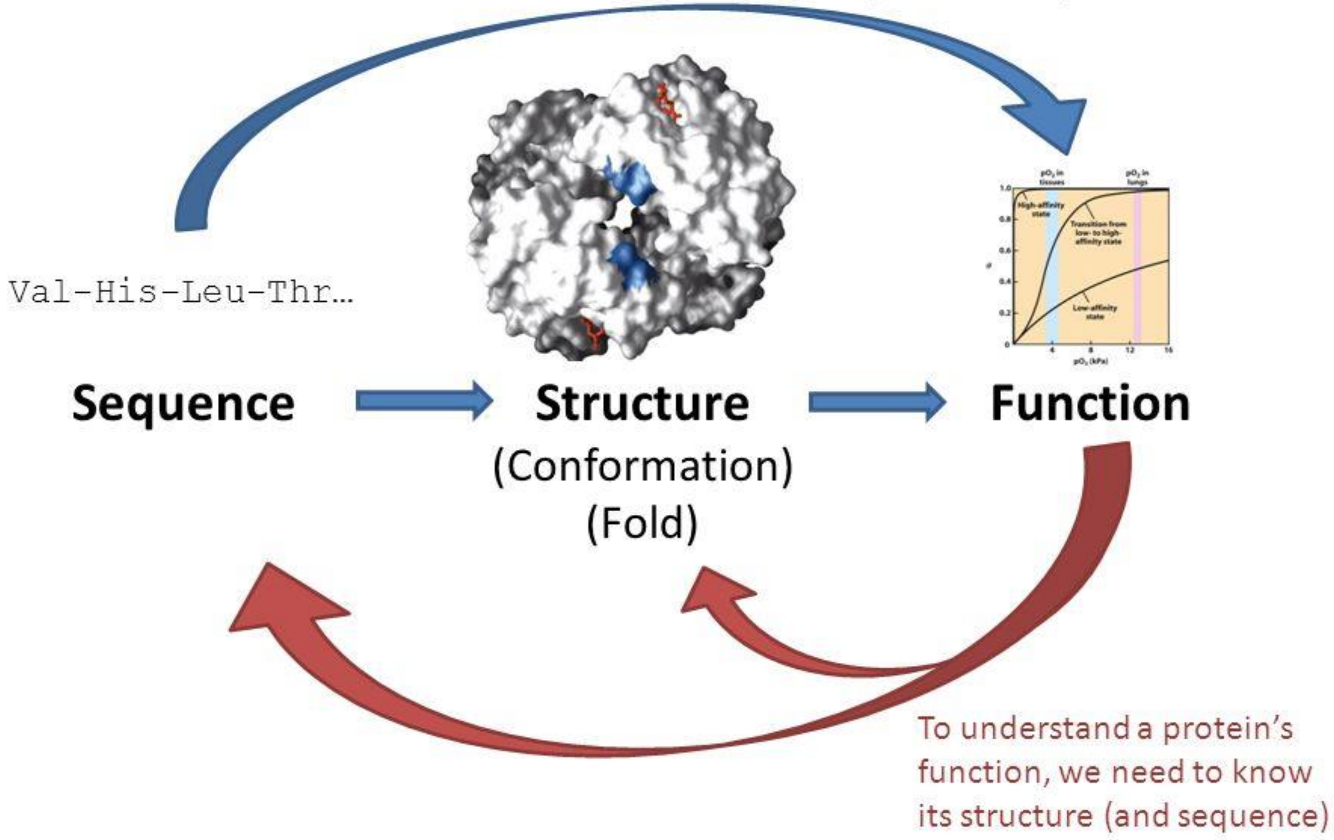
…-Met-Thr-Trp-Ser-Gly…

…ATGACCTGGTCGGGA…

… M T W S G …

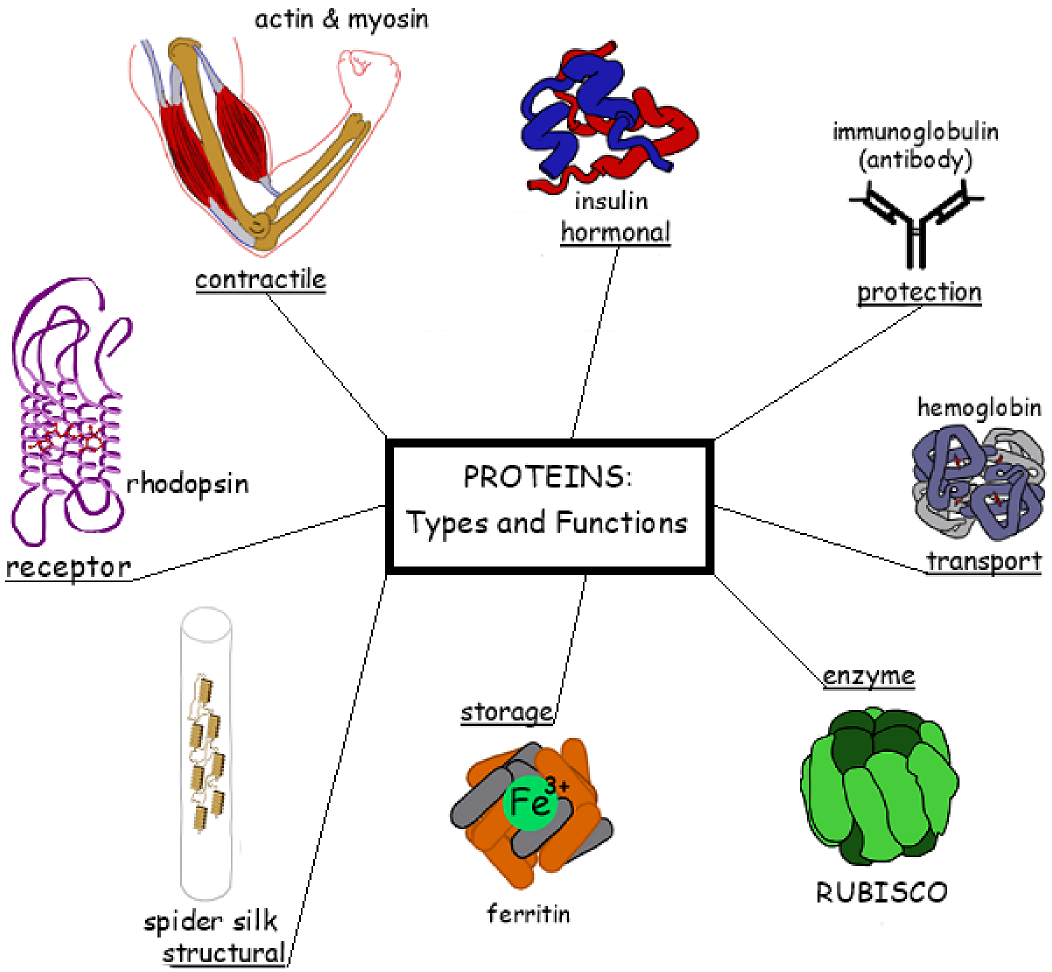
Proteins: sequence, structure, function

A protein’s function derives from its structure, and its structure is determined by its sequence.

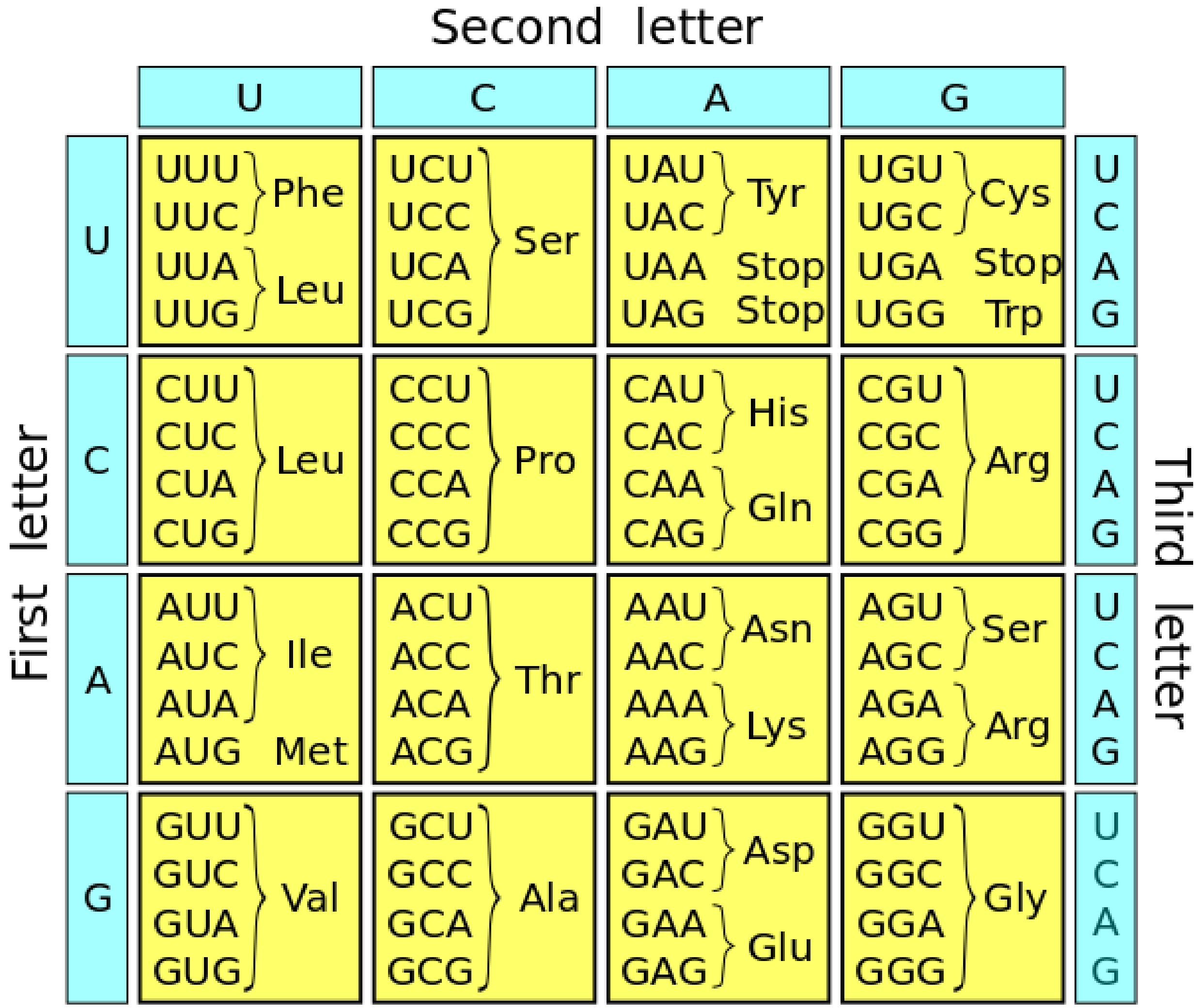


To understand a protein’s function, we need to know its structure (and sequence)

Proteins perform all functions

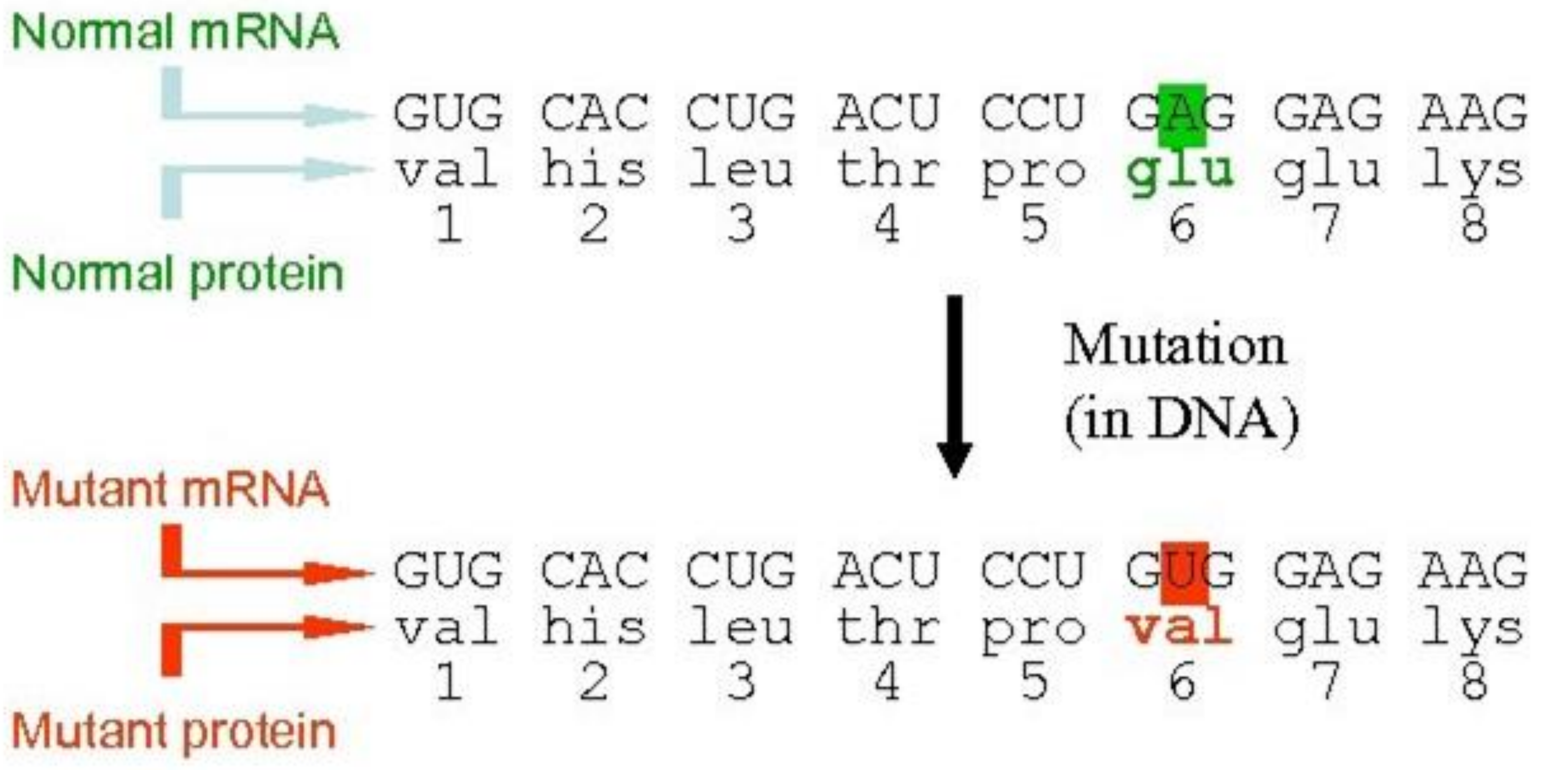


Genetic code



	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Mutations



Normal mRNA

GUG CAC CUG ACU CCU GAG GAG AAG

val his leu thr pro glu glu lys

1 2 3 4 5 6 7 8

Normal protein

Mutation (in DNA)

Mutant mRNA

GUG CAC CUG ACU CCU GUG GAG AAG

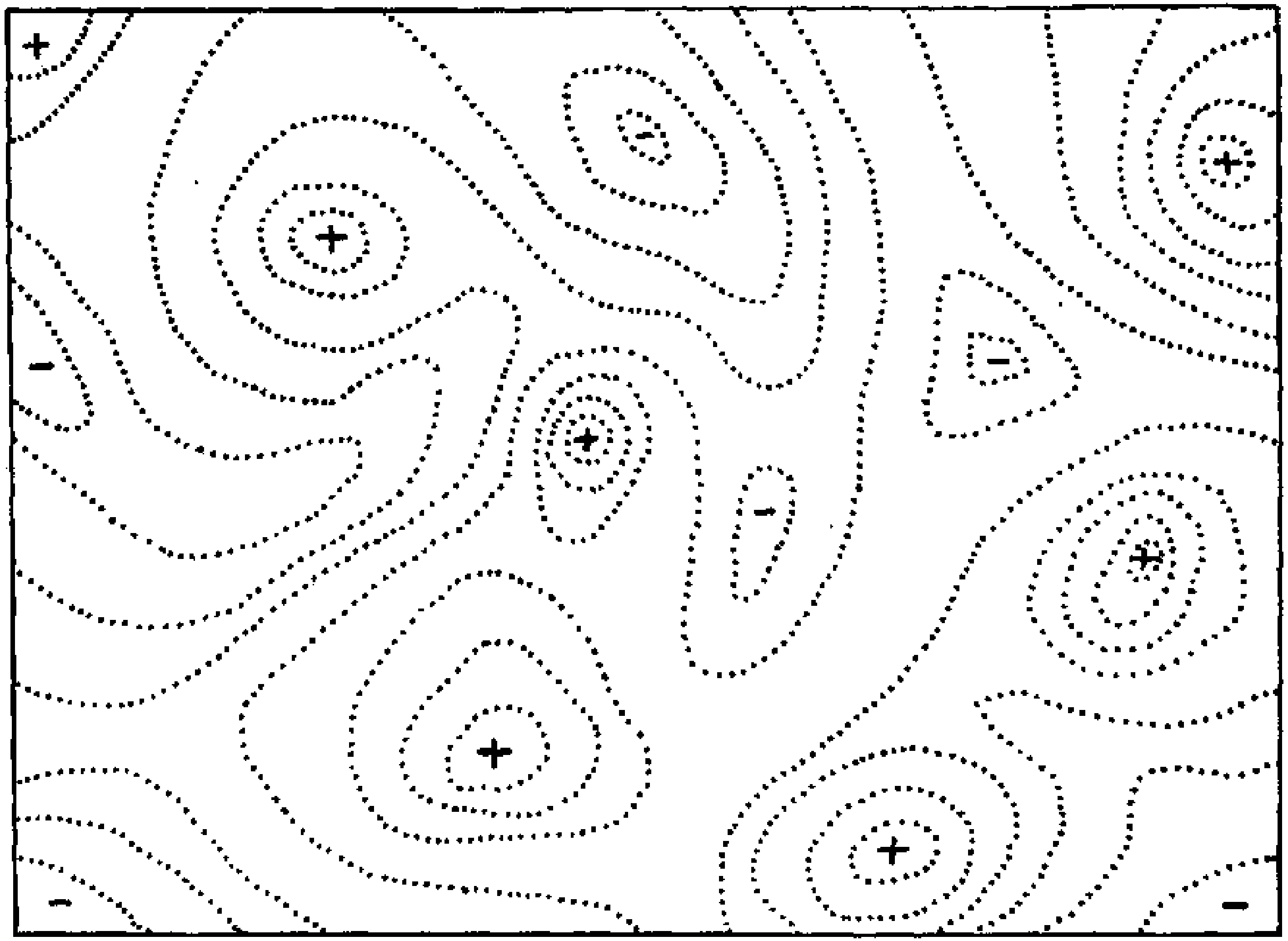
val his leu thr pro val glu lys

1 2 3 4 5 6 7 8

Mutant protein

Glutamate (glu), a negatively charged amino acid, is replaced by valine (val), which has no charge.

Fitness landscape concept

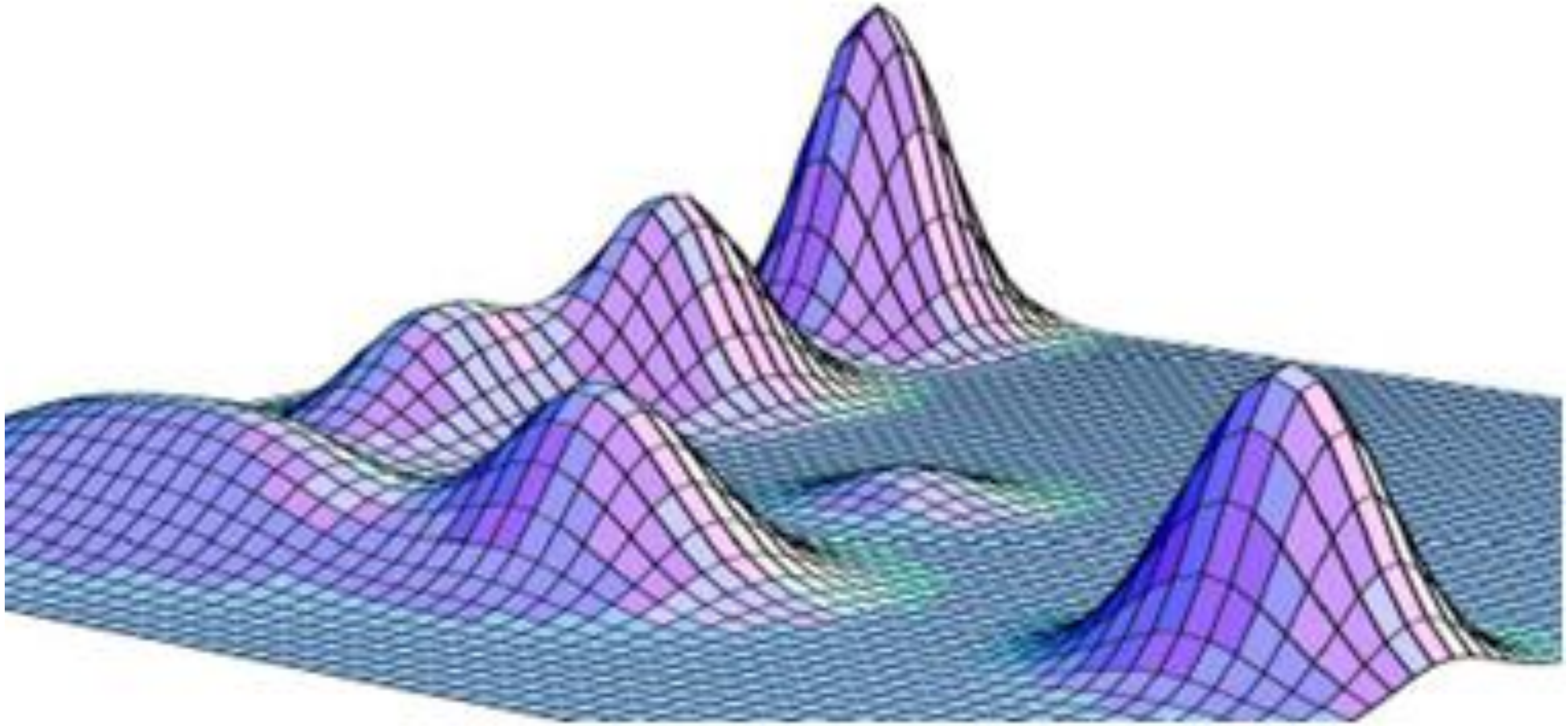


Wright 1932

Fitness landscape: side view

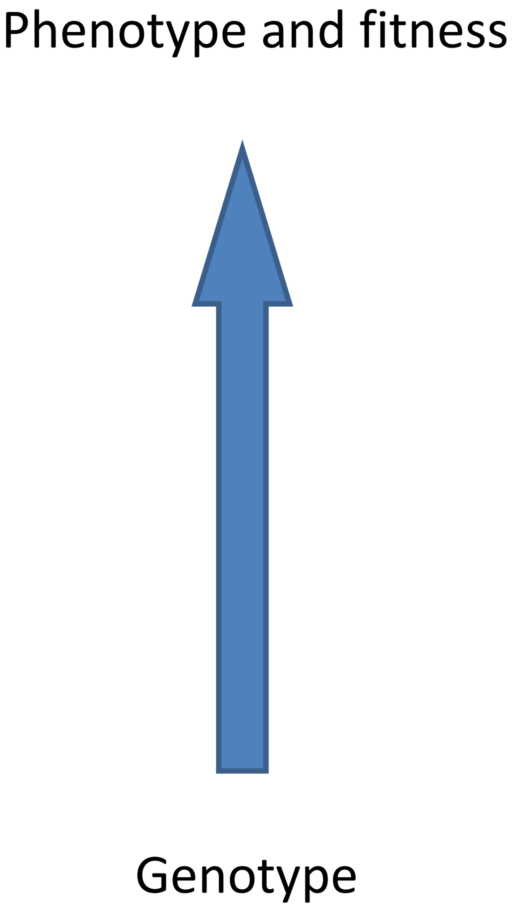
Size of genome: 3,200,000,000 nt

Size of genotype space: $4^{3,200,000,000}$ – impossibly large



Holy Grail of evolutionary biology

- Genotype-to-phenotype connection

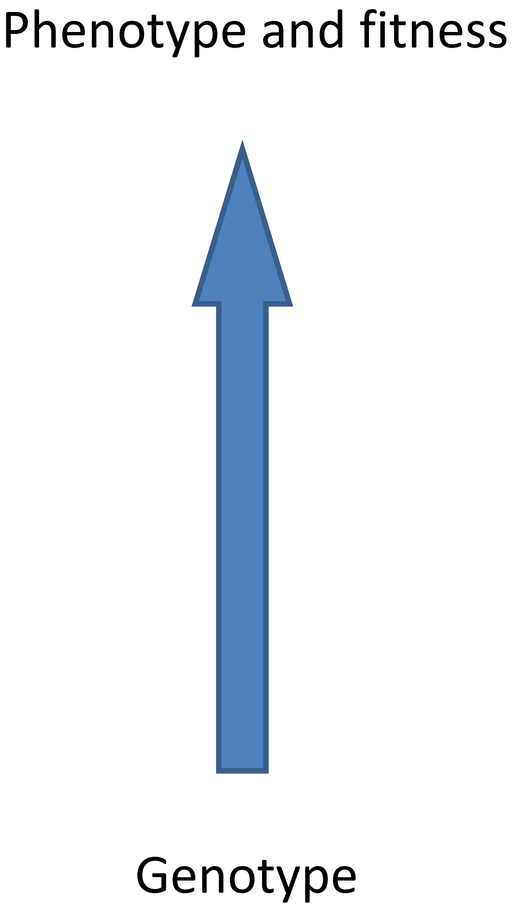


Phenotype and fitness

Genotype

Holy Grail of evolutionary biology

- Genotype-to-phenotype connection



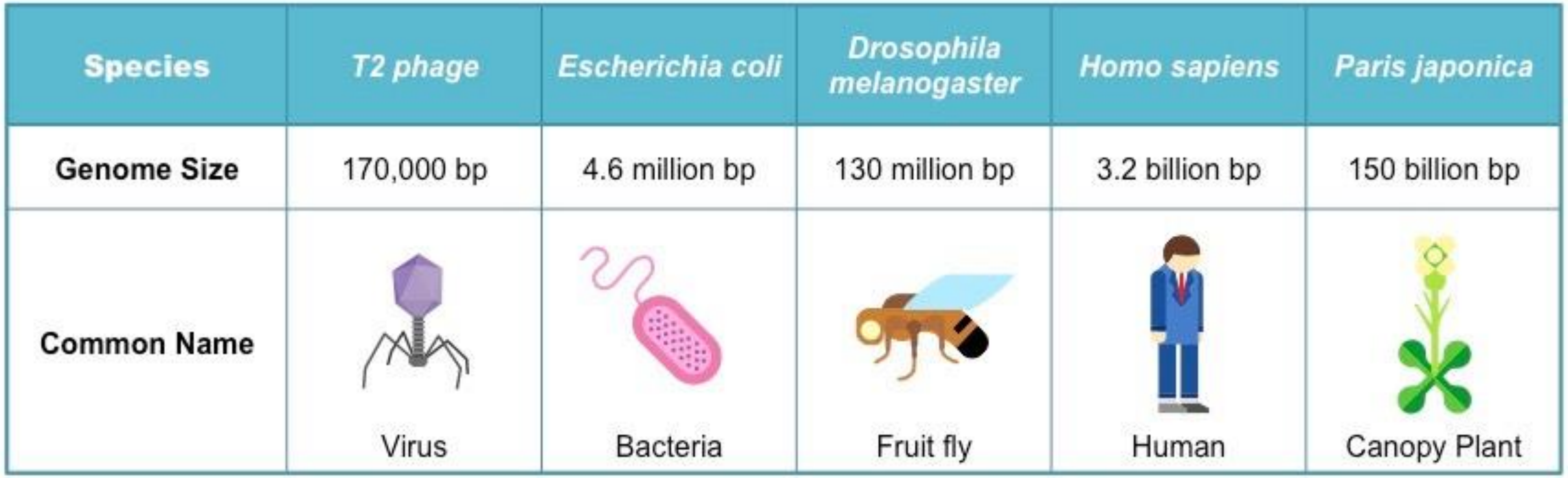
Phenotype and fitness

Genotype

…ACCGTAGTTGTGAAACTATAC…

Genome size

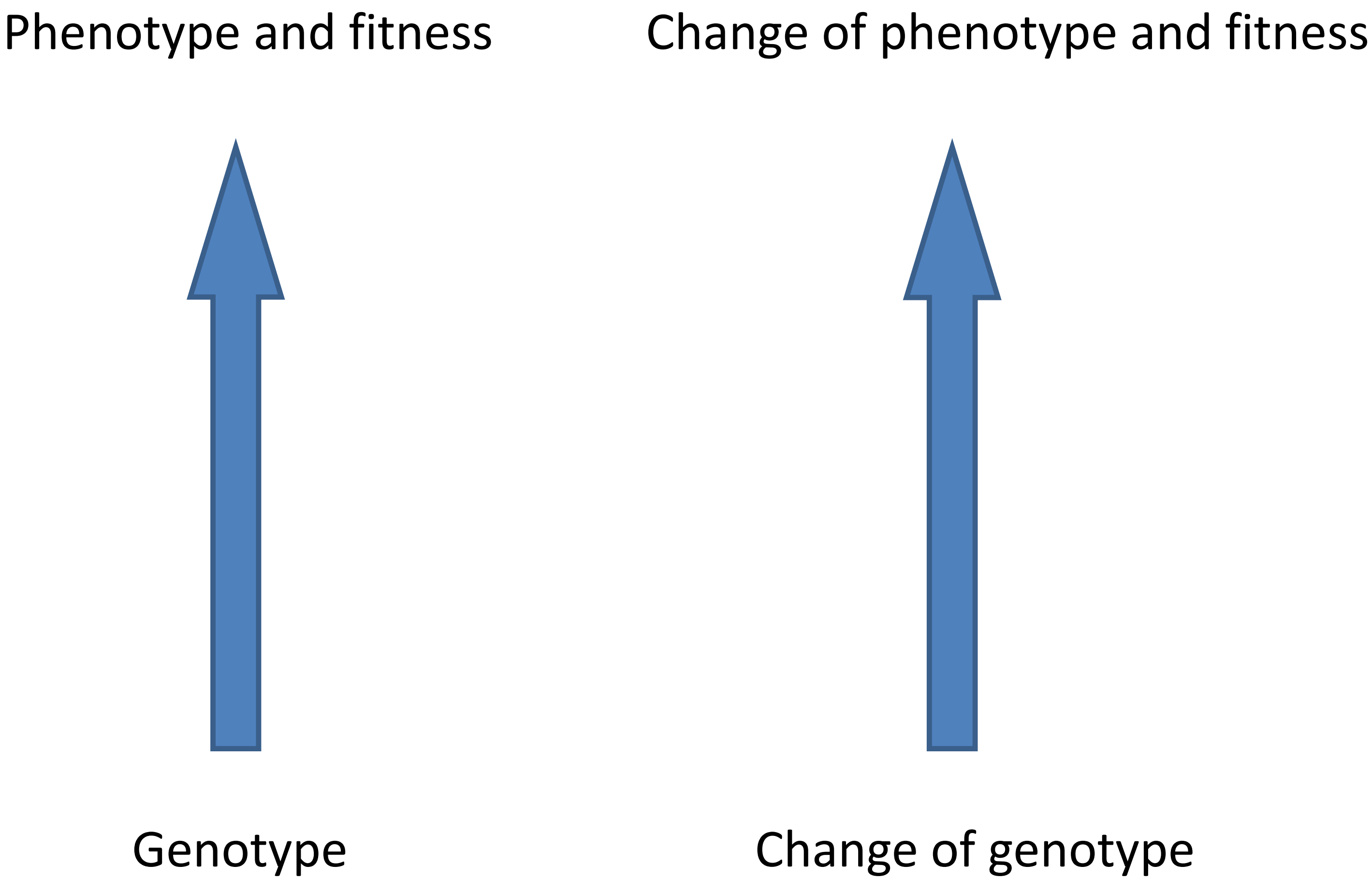
- However, the number of variants is huge:



Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	Virus	Bacteria	Fruit fly	Human	Canopy Plant

- For human-size genome it is $4^{3,200,000,000}$ variants
- Experimentally impossible, prediction – we don’t understand much

Approach to prediction



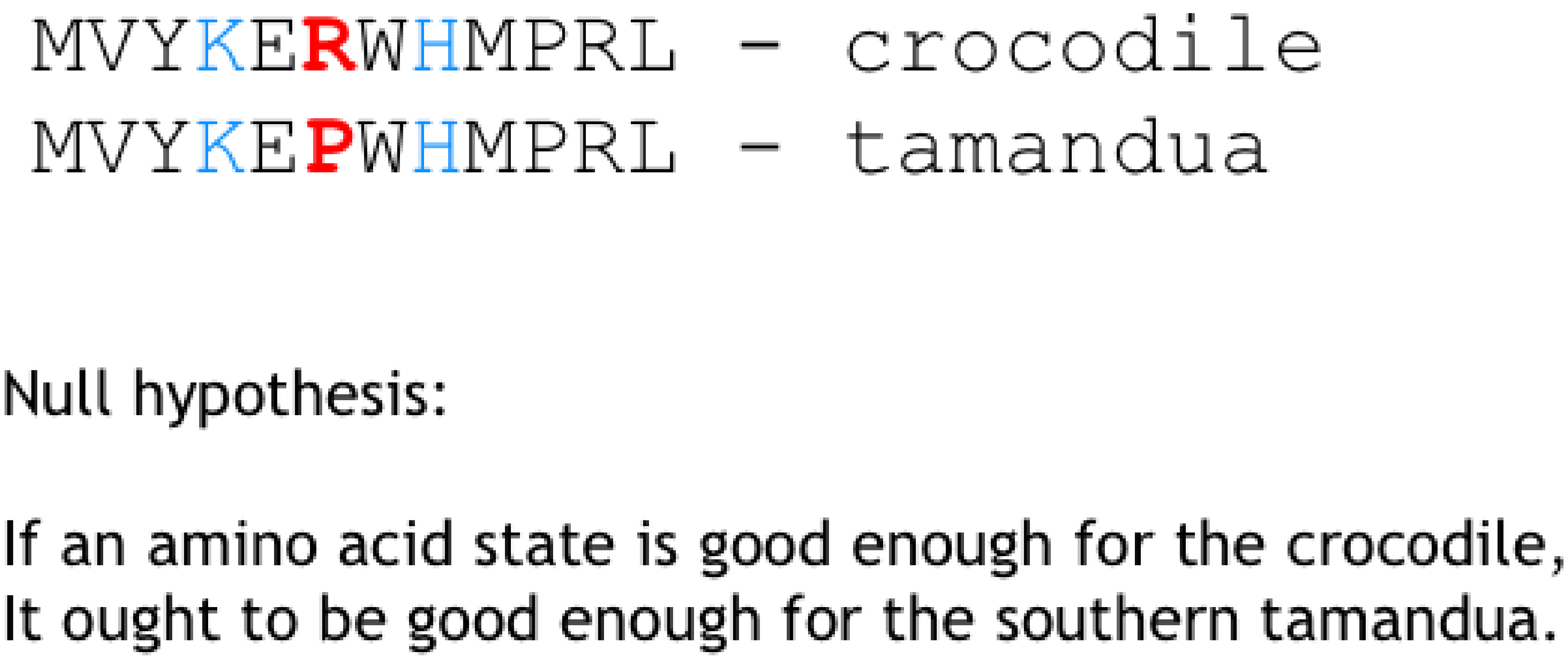
Phenotype and fitness

Genotype

Change of phenotype and fitness

Change of genotype

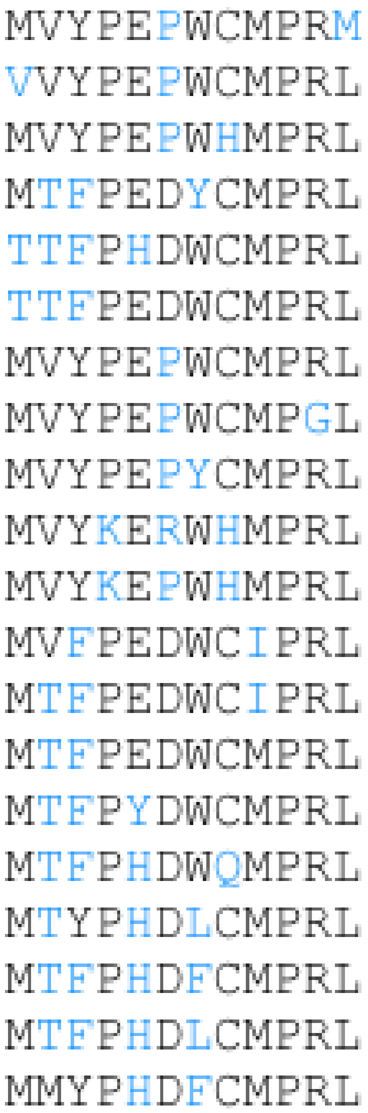
Predictability



```
MVYKERWHMPRL - crocodile
MVYKEPWHMPRL - tamandua
```

Null hypothesis:

If an amino acid state is good enough for the crocodile,
It ought to be good enough for the southern tamandua.



```
MVYPEPWCMPRM
VVYPEPWCMPRL
MVYPEPWHMPRL
MTFPEDYCMPRL
TTFPHDWCMPRL
TTFPEDWCMPRL
MVYPEPWCMPRL
MVYPEPWCMPGL
MVYPEPYCMPRL
MVYKERWHMPRL
MVYKEPWHMPRL
MVFPEDWCIPRL
MTFPEDWCIPRL
MTFPEDWCMPRL
MTFPYDWCMPRL
MTFPHDWQMPRL
MTYPHDLCMPRL
MTFPHDFCMPRL
MTFPHDLCMPRL
MMYPHDFCMPRL
```

Поиск эпистаза в экспериментальных данных

Skoltech

Predictability

It would be easier to predict if the effect of a substitution is universal

Then, for human we would have to measure the effect of just all

3 * 3,200,000,000 single mutations (universal effect)

which looks reasonable instead of

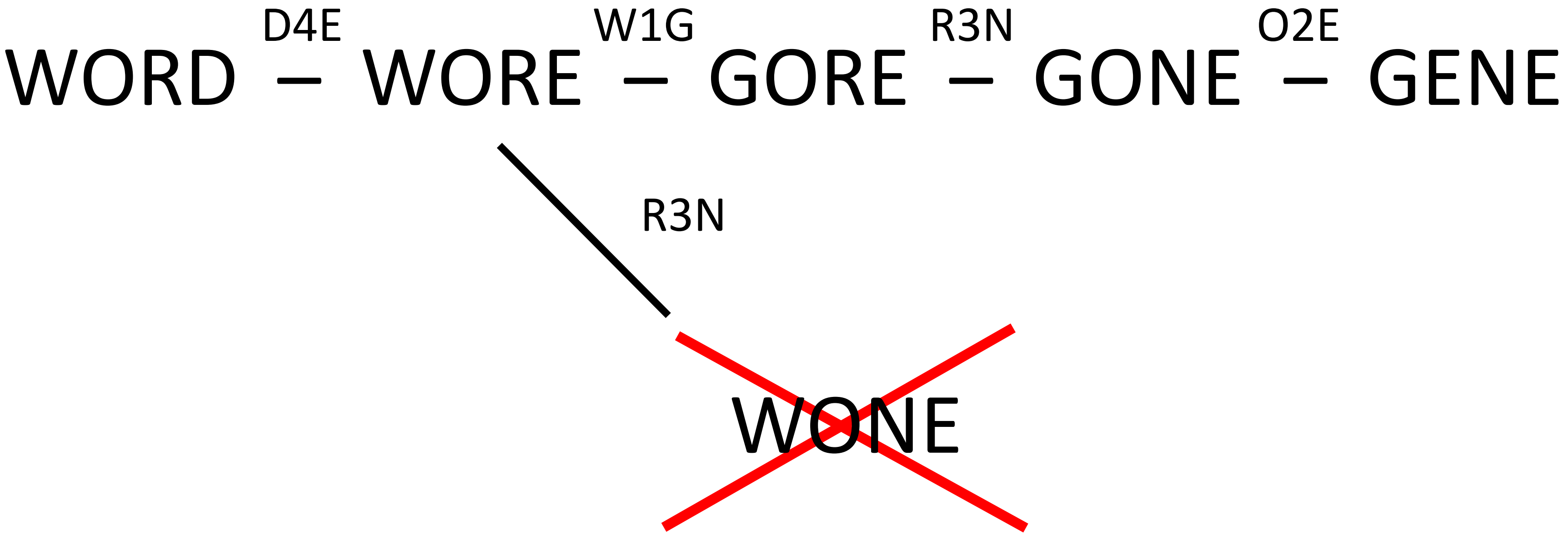
$4^{3,200,000,000}$ variants (full dependence)

Epistasis

However, it is not always true:

Epistasis – dependence of mutation effect on genetic context

Epistasis as a word game



Smith M. Nature (1970)

Life examples

- Non-epistatic logic (predictability) :
  - “If English was good enough for Jesus, then it will be good enough for Texas children” (Texas governor, ~1930)

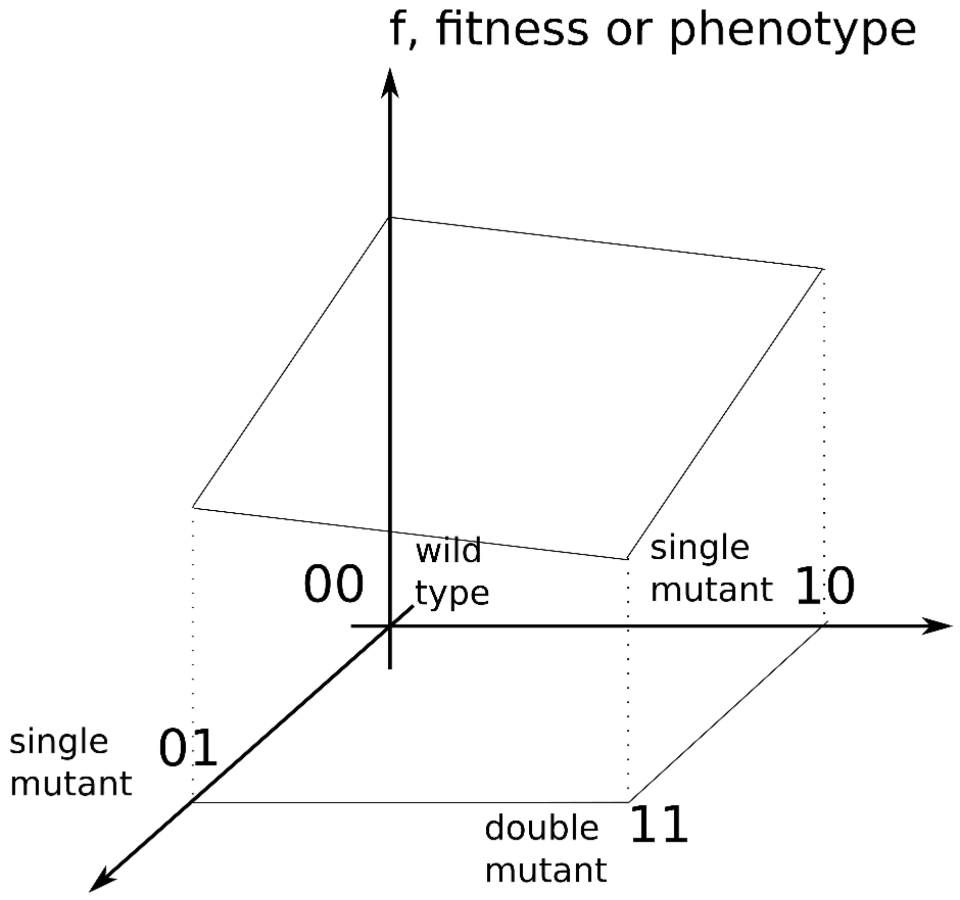
Life examples

- Non-epistatic logic (predictability) :
  - “If English was good enough for Jesus, then it will be good enough for Texas children” (Texas governor, ~1930)
- Epistatic logic (no predictability):
  - What is good for Russian, is mortal for German

  / Что русскому хорошо, то немцу смерть /
  - Spoon is good at lunchtime

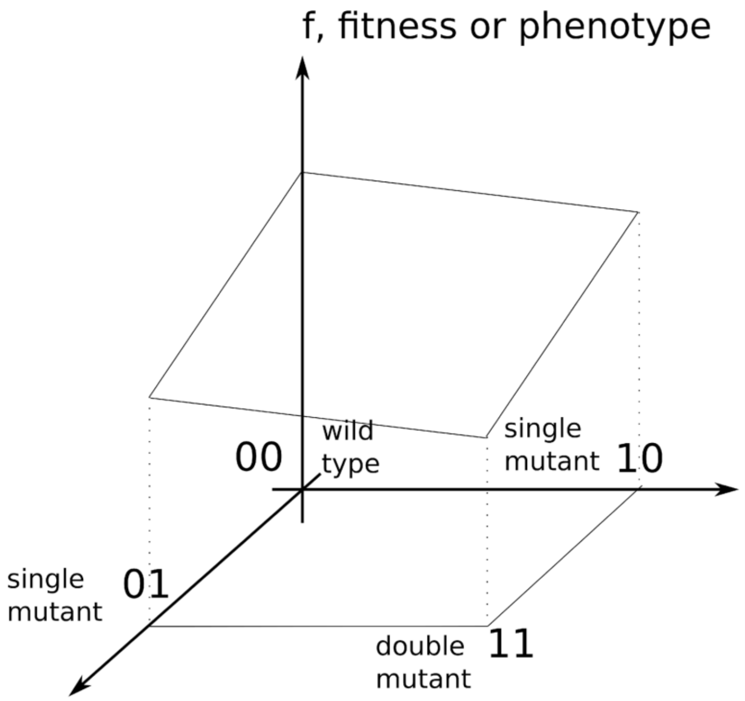
  / Хороша ложка к обеду /

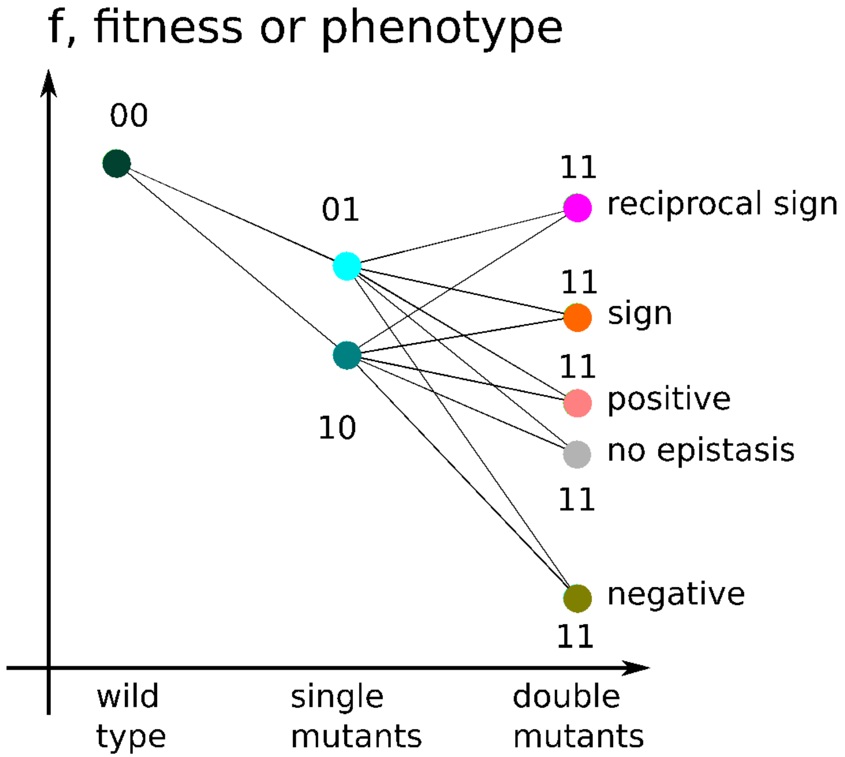
Visualization of epistasis



Types of epistasis

- Epistasis – non-additive effect of substitutions
- No epistasis – full predictability





Поиск эпистаза в экспериментальных данных

Skoltech

Formal definition of epistasis

effect of single substitutions

correction for triple substitutions

$$f(g) = const + \sum_{i=1}^{N} \alpha_i \delta_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} \delta_i \delta_j + \boxed{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_{ijk} \delta_i \delta_j \delta_k + \cdots}$$

reference level

correction for double substitutions

Higher-order epistasis

Here: $const = f(\mathrm{wt})$

$$\delta_i = \begin{cases} 1, & \text{mutation } i \text{ is present in genotype } g \\ 0, & \text{otherwise} \end{cases}$$

Поиск эпистаза в экспериментальных данных

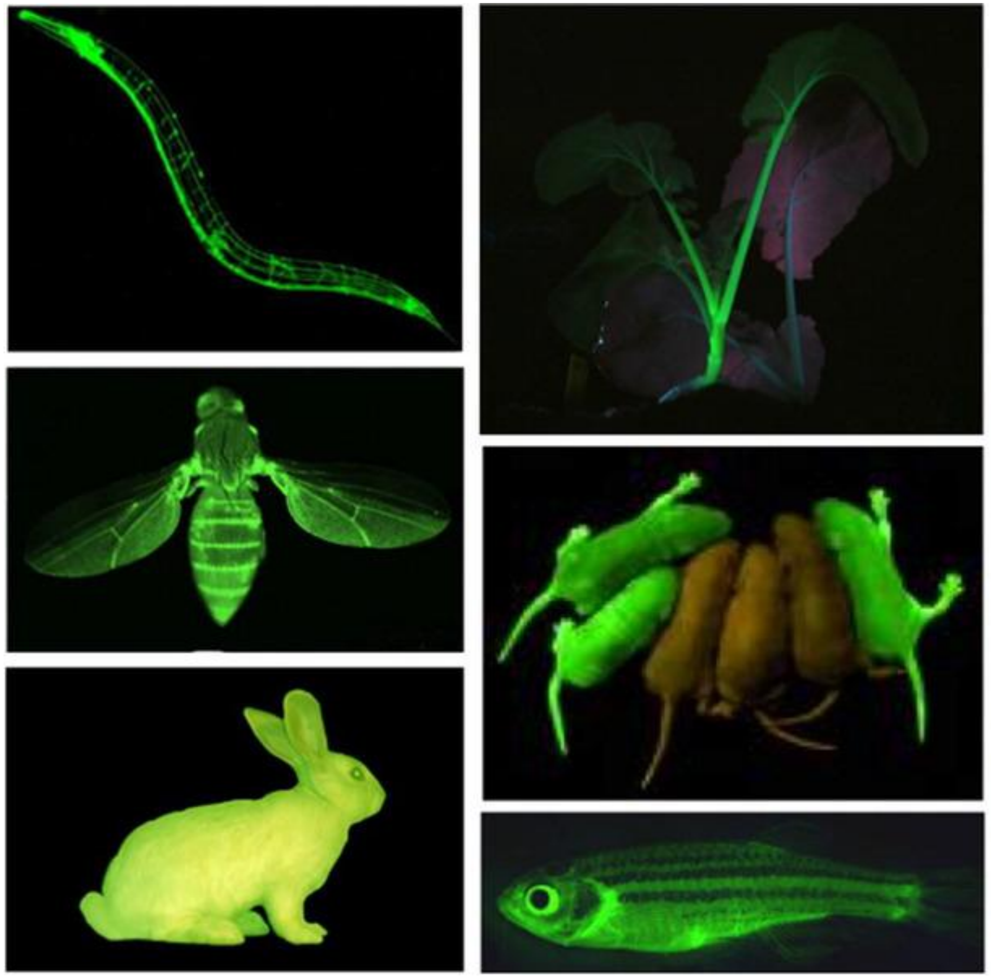
Epistatic terms

- N-order epistatic term => N-dimensional hypercube must be measured
- Experiments can be carefully designed to have all 2^N phenotypes
- What about random mutagenesis experiments?

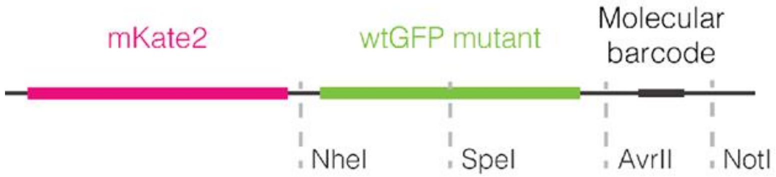
Random mutagenesis experiment in green fluorescent protein (GFP)

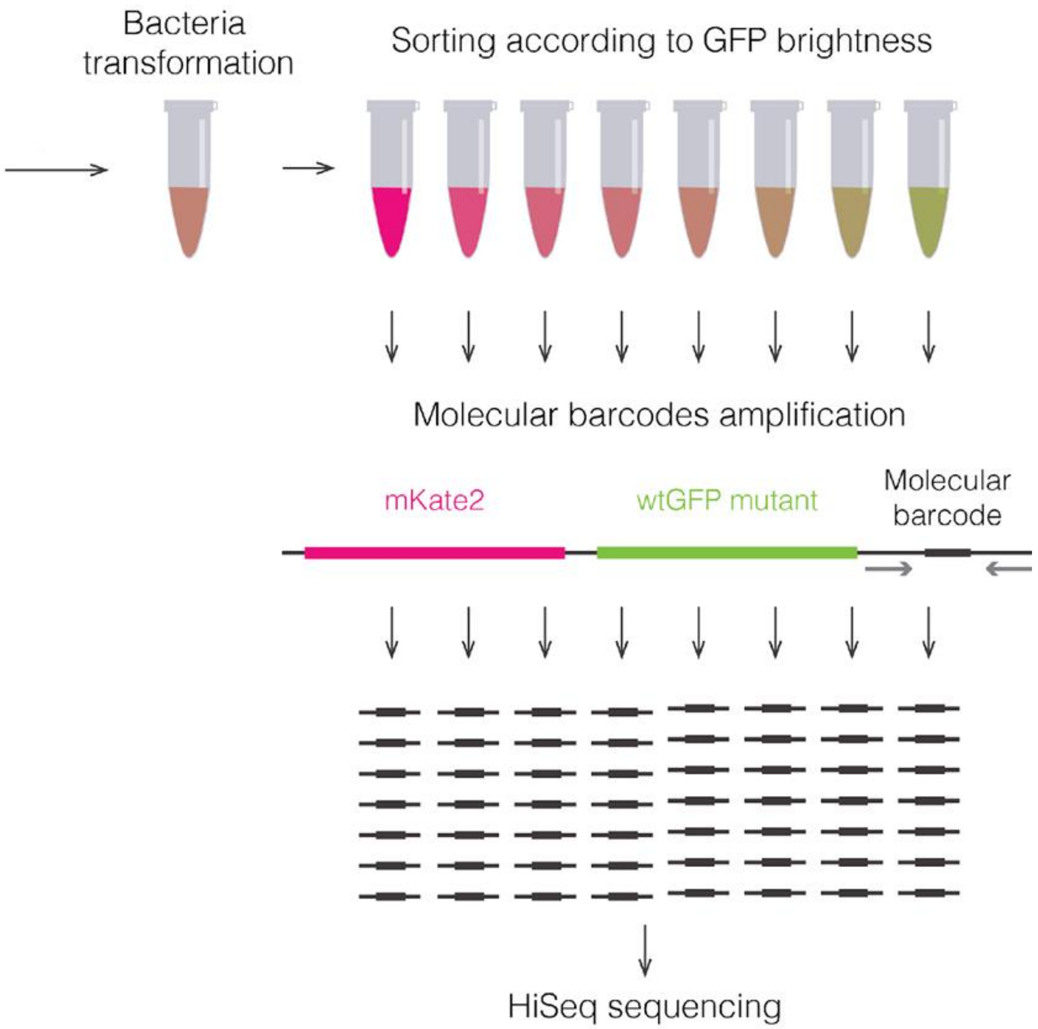
Sarkisyan K.S. et al. Nature (2016)

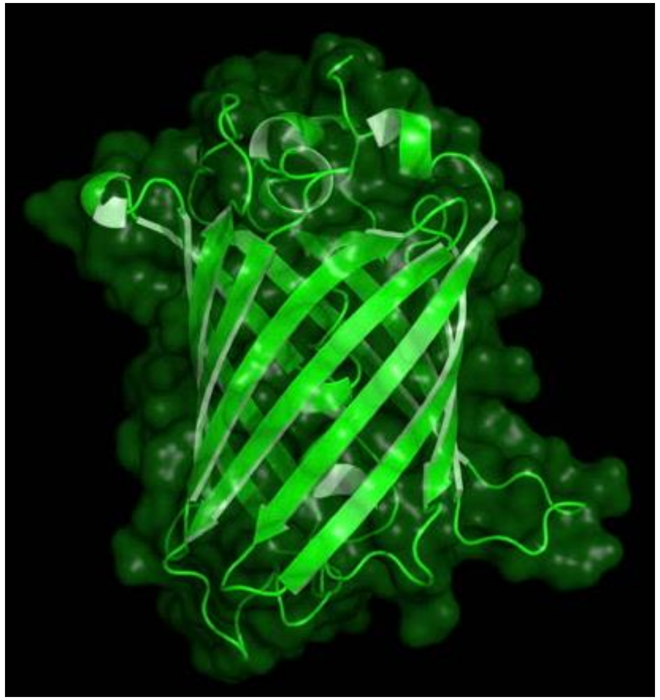
GFP, Nobel prize 2008



Experiment



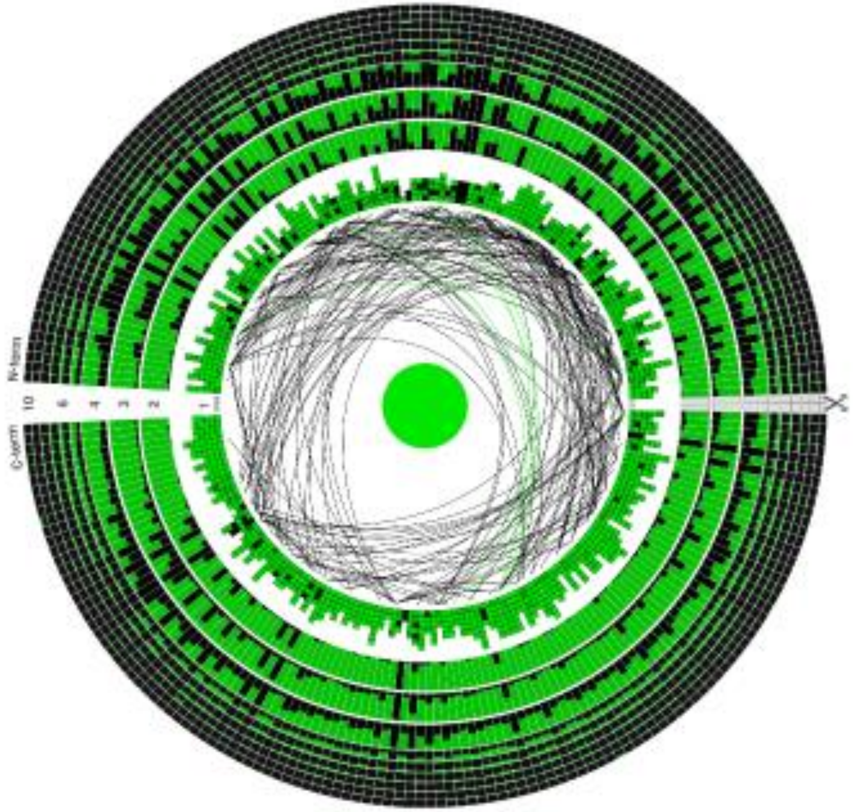




Sarkisyan K.S. et al. Nature (2016)

Random mutagenesis in GFP

- ✓ 56,086 unique nucleotide sequences
- ✓ 51,715 unique amino acid sequences
- ✓ 238 amino acid residues
- ✓ 1817 types of single mutations
- ✓ 50% of the population lose fluorescence after 5 mutations



Sarkisyan K.S. et al. Nature (2016)

Random mutagenesis data

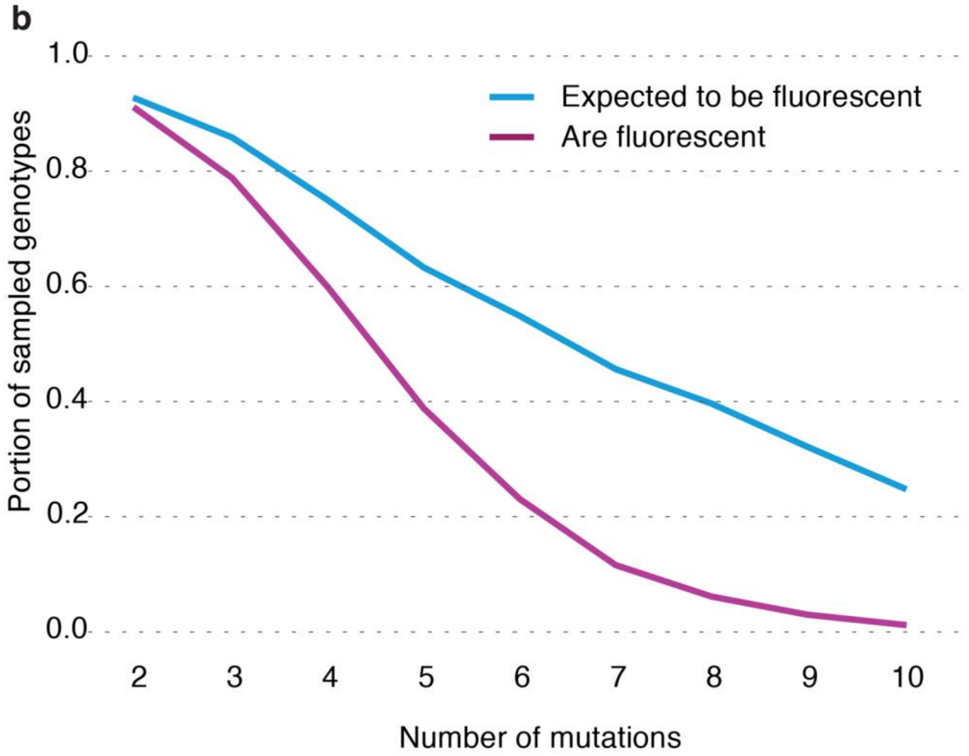
#	genotype	phenotype
1	A24G	0.95
2	S56T	0.1
3	A24G; C170M	0.97
4	A24G; S56T; C170M	0.07
…	…	…

# of measured genotypes:

- GFP: 51 715

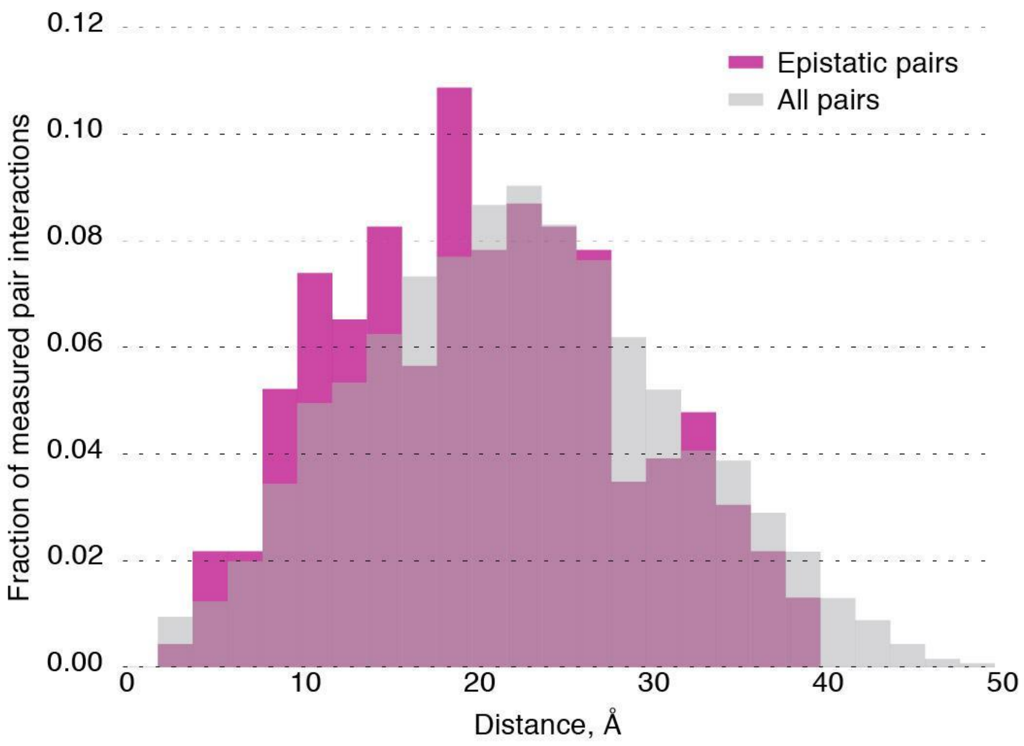
Sarkisyan K.S. et al. Nature (2016)

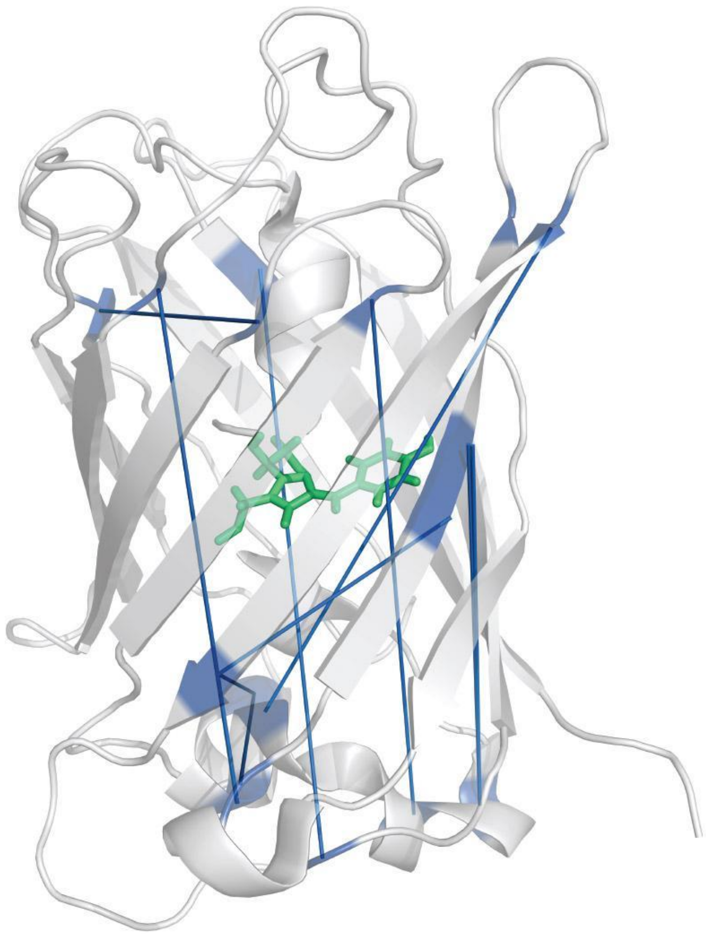
Expected vs. observed effects



Sarkisyan K.S. et al. Nature (2016)

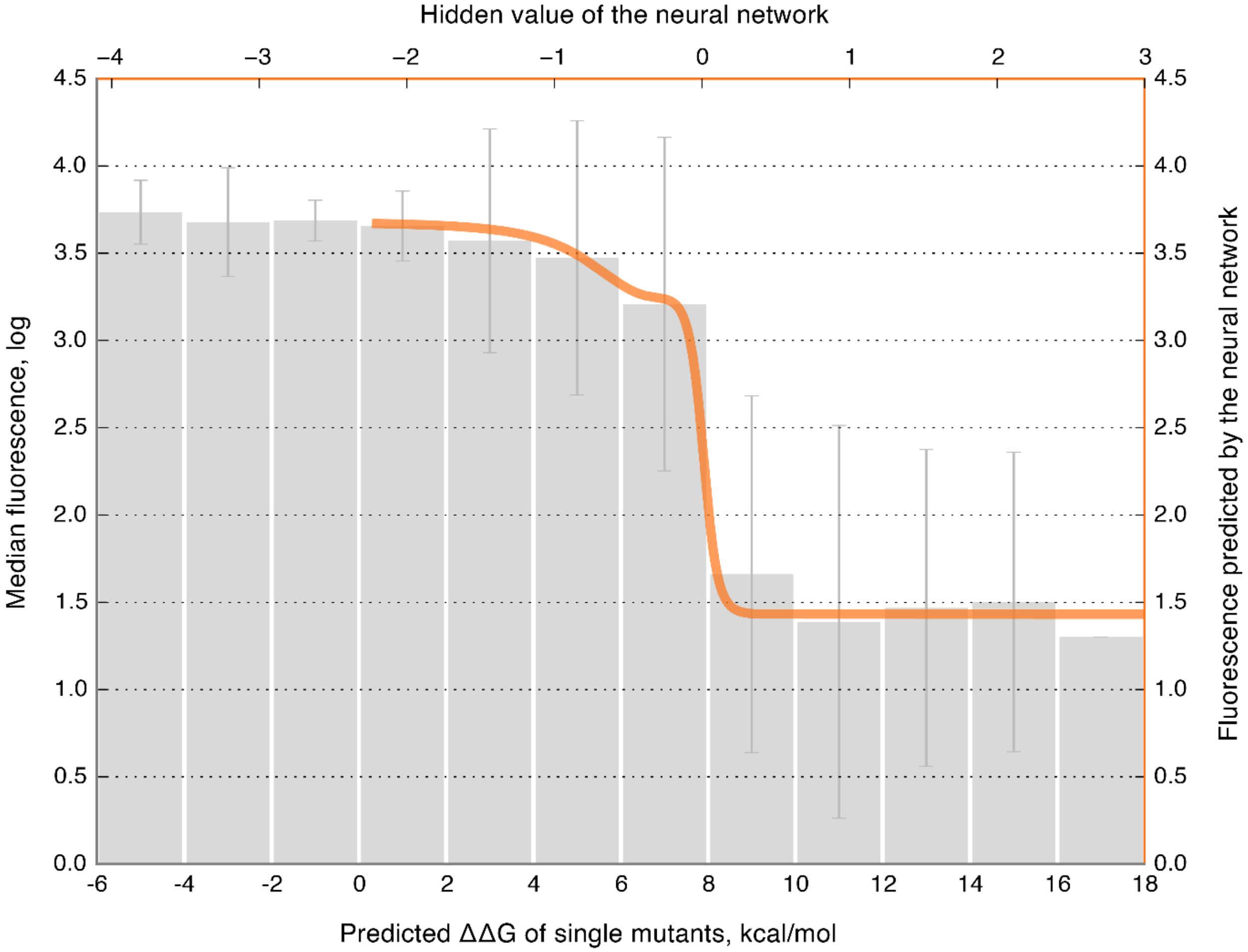
Epistatic pairs on GFP structure





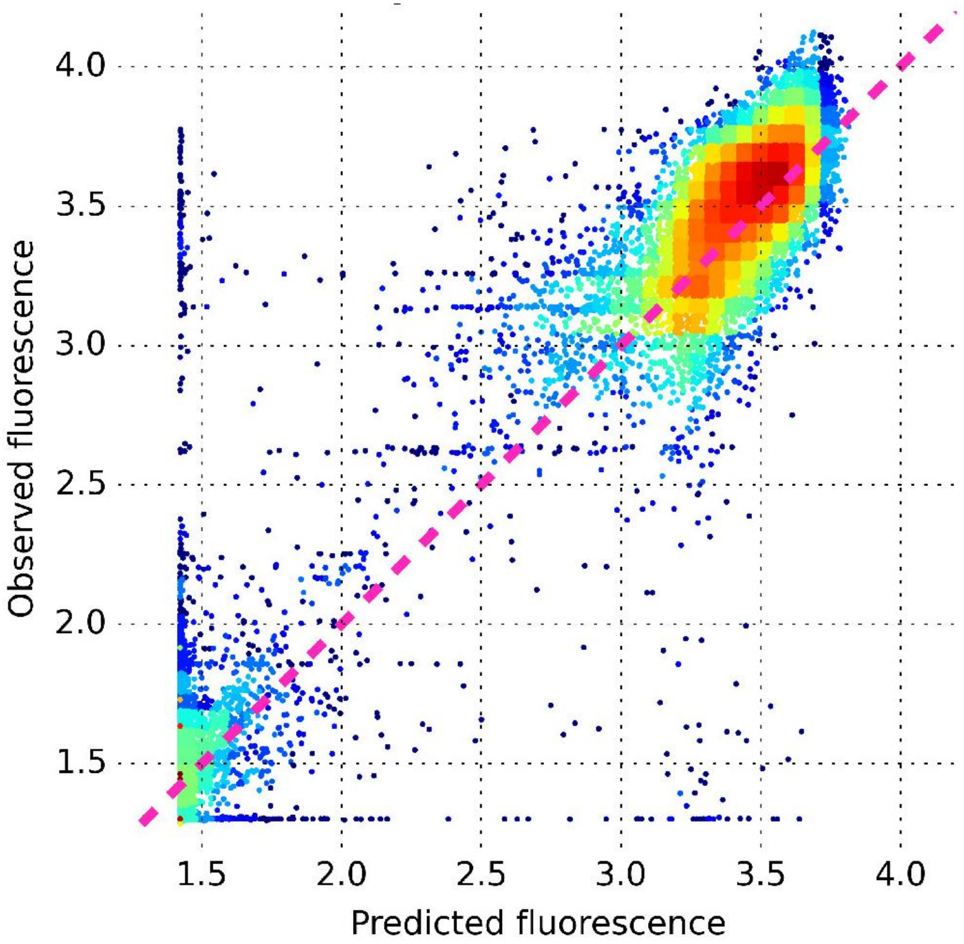
Sarkisyan K.S. et al. Nature (2016)

Protein stability as explanation



Sarkisyan K.S. et al. Nature (2016)

Observed vs. predicted



Sarkisyan K.S. et al. Nature (2016)

Conclusion #1

✓93% of epistasis was explained by change of protein stability

✓6% of epistasis remained unexplained

Sarkisyan K.S. et al. Nature (2016)

Quasi-random mutagenesis experiment in yeast HIS3 protein

Pokusaeva V.O. et al. PLoS Genetics (2019)

HIS3 as the next model

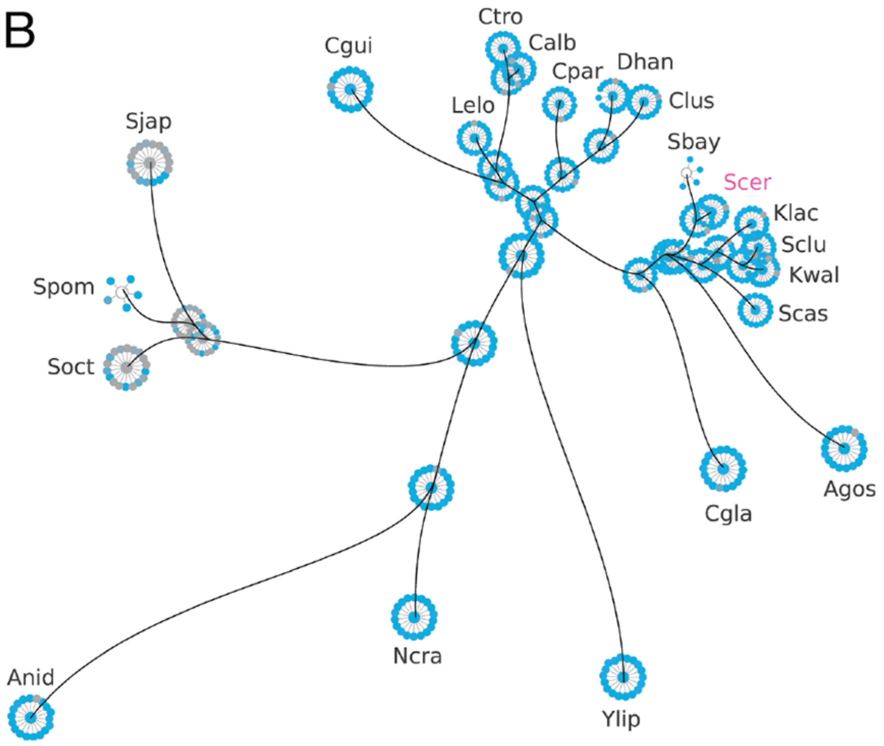
- ✓ 220 amino acids long
- ✓ Is essential for Histidine synthesis, conditionally essential for yeast growth
- ✓ Present in a single copy
- ✓ Relatively conservative protein sequence

```
-----MTEQKALVKRITNETKIQIAISLKGGPLAIEHSIF----PEKEAEAVAEQATQSQVINVHTGIGFLDHMIH
----MSETQQAFVKRYTPLRPSPNSLALNGGPFEIGQSIL-----GGAKTTVAHQASSSQVINVQTGVGFLDHMIH
-----MSEQKALVKRITNETKIQIAIALKGGPLALEHSIF----PAREADAVAEQATQSQVINVQTGIGFLDHMVH
---------MAFVKRVTQETNIQLALDLDGGSVSVRESIL----------GKEYASGDGQTIHVHTGVGFLDHMLT
MAQEQEQEQRALINRITNETKIQIAISLKGGPLTLQSSIF----PTKESSNVATQATSSQVIDIHTGVGFLDHMIH
---MTYPERKAFVSRITNETKIQIAISLNGGPISIENSIL-----QREESDAAKQVTGSQIIDIQTGVGFLDHMIH
------MAKTATIKRDTNETKIQIAISLEGGHIALEESIFKNSANETKDDSHATQATSTQVIQVQTGIGFLDHMLH
----MSSERKAFVKRDTNETKIQIALSLDGGAVSIPTSIL---PKNDKVEDHAIQKTGGQVINVQTGIGFLDHMLH
-------MRRAFVERNTNETKISVAIALDKAPLPEESNFI-----DELITSKHANQKGEQVIQVDTGIGFLDHMYH

ALAKHSGWSLIVECIGDLHIDDHHTTEDCGIALGQAFKEALGAVRGVKRFGSGFAPLDEALSRAVVDLSNRPYAVV
ALAKHSGWSLIVECVGDLHIDDHHTTEDCGLALGQALREAIGQVRGVKRFGTGFAPLDEALSRAVVDLSNRPYAVV
ALAKHAGWSLIVECIGDLHIDDHHTTEDCGIALGQAFKEALGAVRGVKRFGSGFAPLDEALSRAVVDLSNRPYAVV
ALAKHGGWSLILECIGDLHIDDHHTVEDCGIALGQAFKEALGSVRGIKRFGHGFAPLDEALSRAVVDFSNRPFAVV
ALAKHAGWSLIVECIGDLHIDDHHTTEDCGIALGEAFKEAMGVVRGVKRFGTGFAPLDEALSRAVVDLSNRPYAFI
ALAKHSGWSLIVECIGDLHIDDHHTTEDCGIALGQAFKEALGHVRGVKRFGSGYAPLDEALSRAVVDLSNRPYAVI
ALAKHSGWSLIIECIGDIHIDDHHTAEDVGITLGLAFHKALGQVKGVKRFGCGFAPLDEALSRAVVDLSNRPYAVI
ALAKHSGWSLVVECIGDLHIDDHHTSEDVGIALGMAFKDALGQIKGVKRFGHGFAPLDEALSRAVVDLSNRPFAVV
ALAKHAGWSLRLYSRGDLIIDDHHTAEDTAIALGIAFKQAMGNFAGVKRFGHAYCPLDEALSRSVVDLSGRPYAVI

ELGLQREKVGDLSCEMIPHFLESFAEASRITLHVDCLRGKNDHHRSESAFKALAVAIREATSPNGTNDVPSTKGVLM
DLGLRREKIGDLSTEMIPHFLQSFAESARVTLHVDCLRGTNDHHRSESAFKAVAVALGDALTRTGTDDVPSTKGVLM
ELGLQREKVGDLSCEMIPHFIESFAEASRITLHVDCLRGKNDHHRSESAFKALAVAIREATSPNGTNDVPSTKGVLM
ELGLKRERIGQLSTEMIPHFLESFATEARITMHVDCLRGTNDHHRSESAFKALAIAIREARTPTGRDDVPSTKGVLA
ELGLKREKIGDLSCEMIPHFLESFAEAARITIHVDCLRGKNDHHRSESAFKALAVAIREATSPNGTNDVPSTKGVLM
ELGLKREKIGDLSCEMIPHFLESFAEAARITLHVDCLRGFNDHHRSESAFKALAIAIKEAISSNGTNDVPSTKGVLM
ELGLKREKIGDLSCEMIPHVMESFAQGAAITIHVDCIRGFNDHHRAESAFKALAVAIKEATSSNGTDDVPSTKGVLF
ELGLKREKIGDLSTEMIPHVLESFAQLAAITMHVDCLRGFNDHHRAESAFKALAIAIKEAISKTGKDDVPSTKGVLS
DLGLKREKVGELSCEMIPHLLYSFSVAAGITLHVTCLYGSNDHHRAESAFKSLAVAMRAATSLTGSSEVPSTKGVL-
```

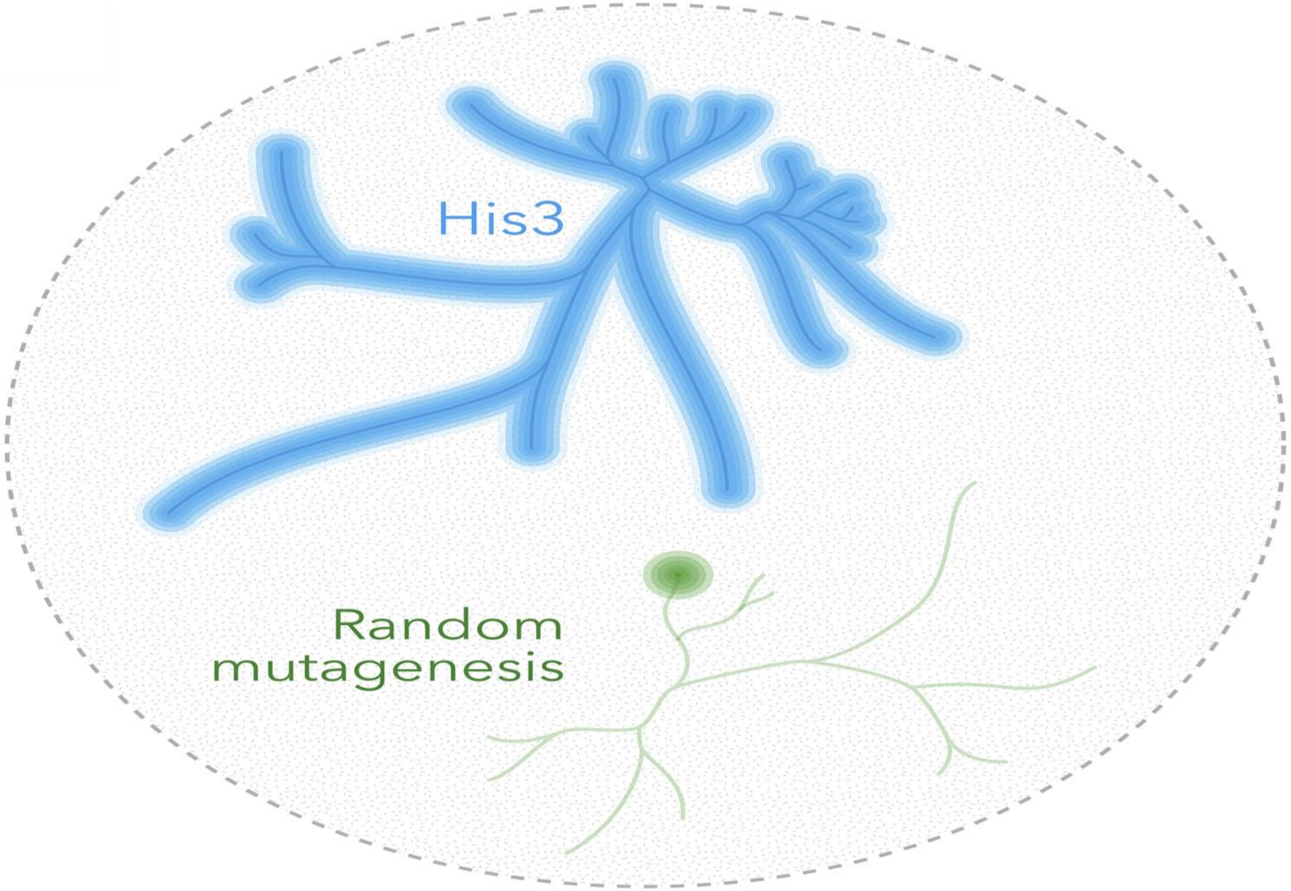
Pokusaeva V.O. et al. PLoS Genetics (2019)



Поиск эпистаза в экспериментальных данных

Skoltech

Quasi-random vs. random

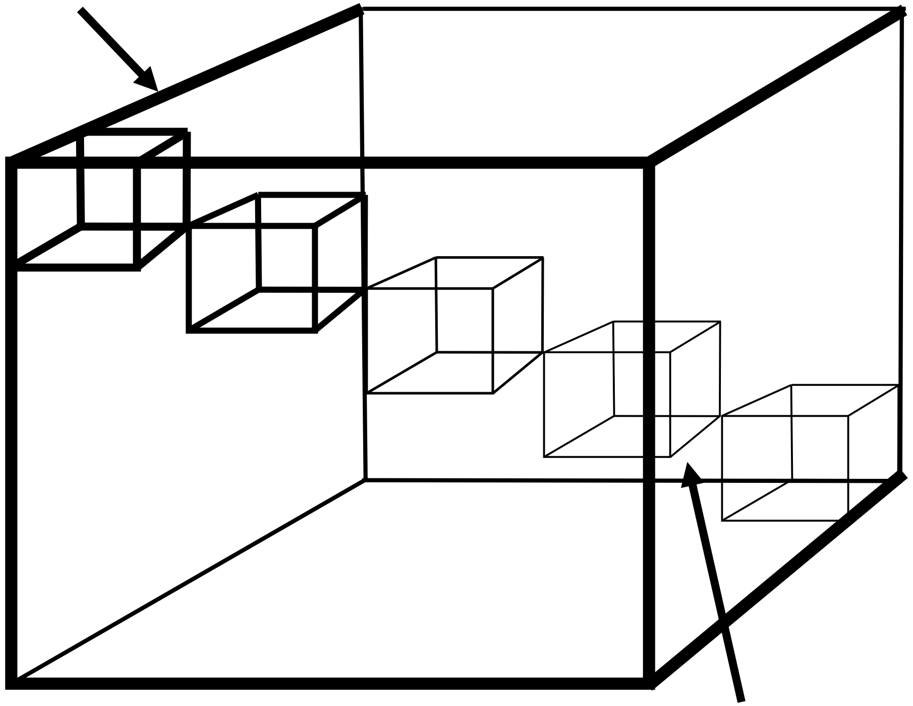


Pokusaeva V.O. et al. PLoS Genetics (2019)

Segments of HIS3

The entire sequence space of His3 has 220 dimensions and a volume of 20^{220}

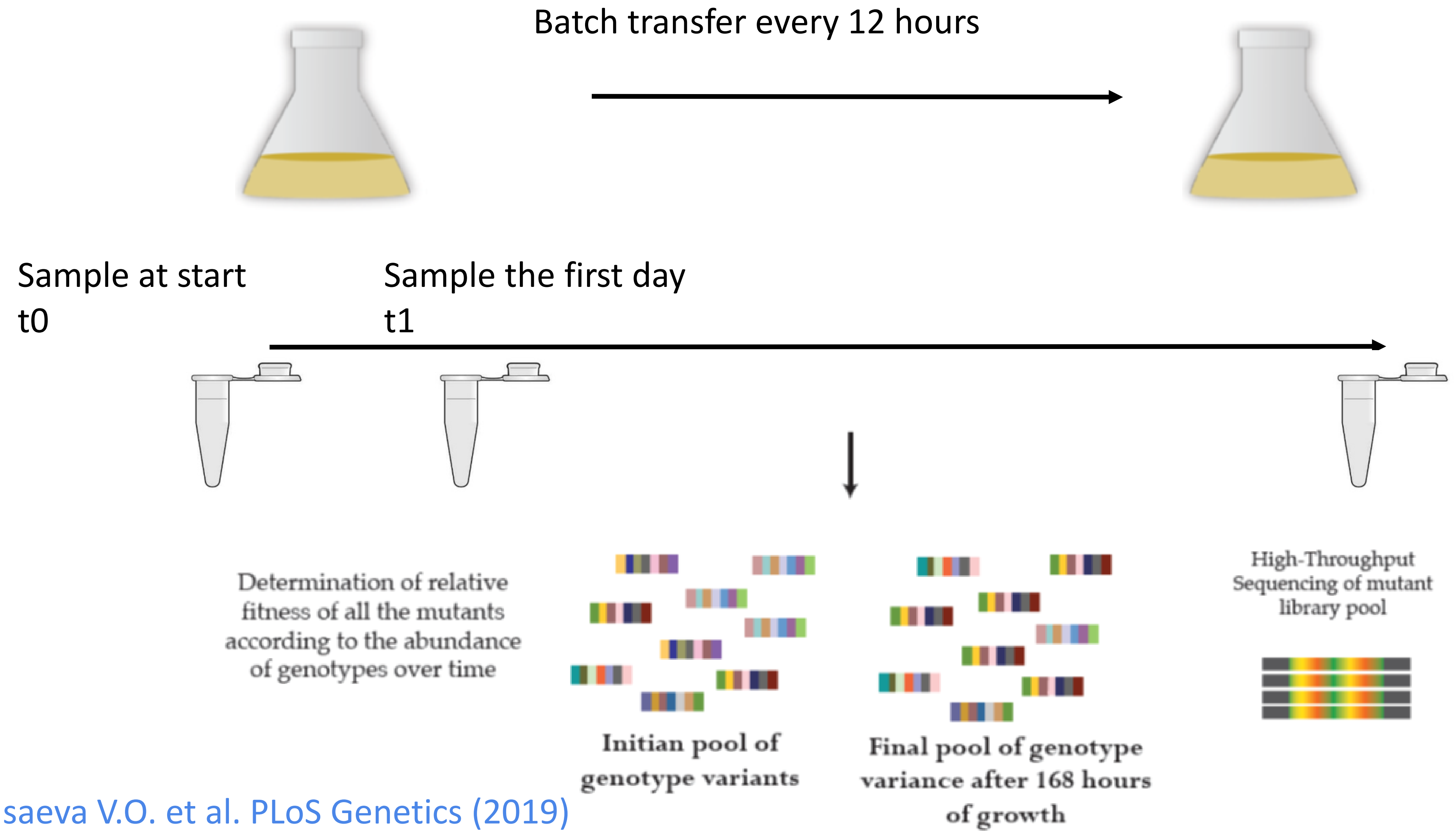
Fitness was measured for 721,791 genotypes



A subsection of His3 space is more manageable.
We study 12 segments, each 15-22 amino acids long.

Pokusaeva V.O. et al. PLoS Genetics (2019)

The HIS3 experiment



Batch transfer every 12 hours

Sample at start
t0

Sample the first day
t1

Determination of relative fitness of all the mutants according to the abundance of genotypes over time
Initian pool of genotype variants
Final pool of genotype variance after 168 hours of growth
High-Throughput Sequencing of mutant library pool

Pokusaeva V.O. et al. PLoS Genetics (2019)

Data structure is the same

#	genotype	phenotype
1	A4G	0.95
2	C6T	0.68
3	A4G:C10M	0.35
4	A24G:S56T:C170M	0.02
…	…	…

# of measured genotypes:

- His3: 721 791:
  - 12 segments, from 16 000 to 82 000 genotypes

Pokusaeva V.O. et al. PLoS Genetics (2019)

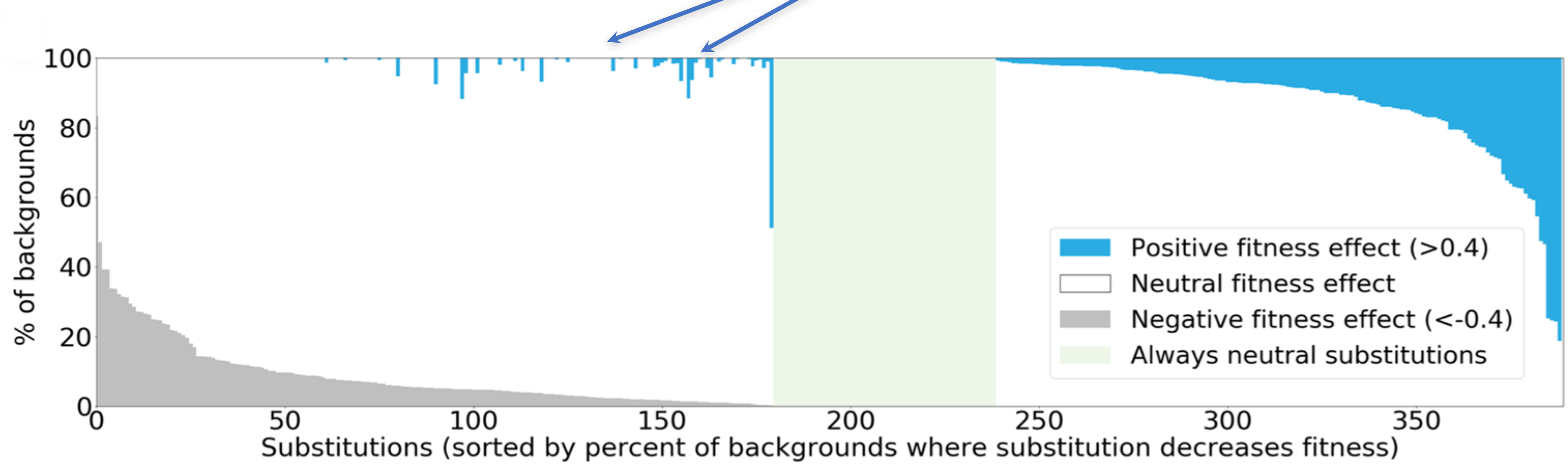
Поиск эпистаза в экспериментальных данных

Skoltech

Поиск эпистаза в экспериментальных данных

Sign epistasis in HIS3

Sign epistasis

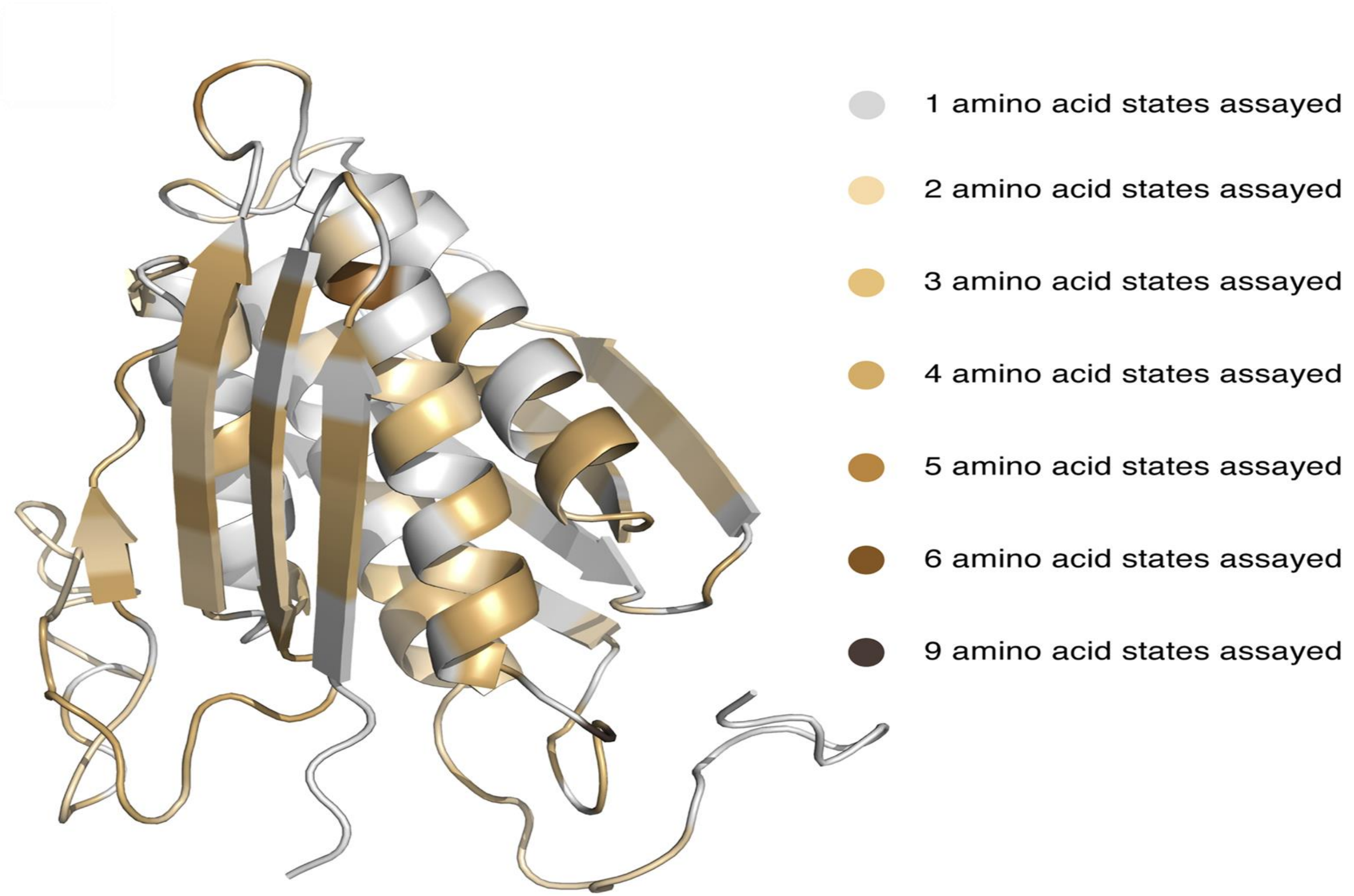


Only 15% of substitutions are universally neutral

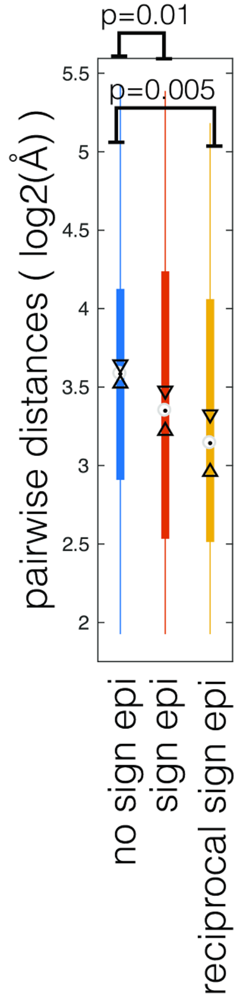
Pokusaeva V.O. et al. PLoS Genetics (2019)

Epistatic pairs on HIS3 structure

Поиск эпистаза в экспериментальных данных



1 amino acid states assayed
2 amino acid states assayed
3 amino acid states assayed
4 amino acid states assayed
5 amino acid states assayed
6 amino acid states assayed
9 amino acid states assayed



Pokusaeva V.O. et al. PLoS Genetics (2019)

Skoltech

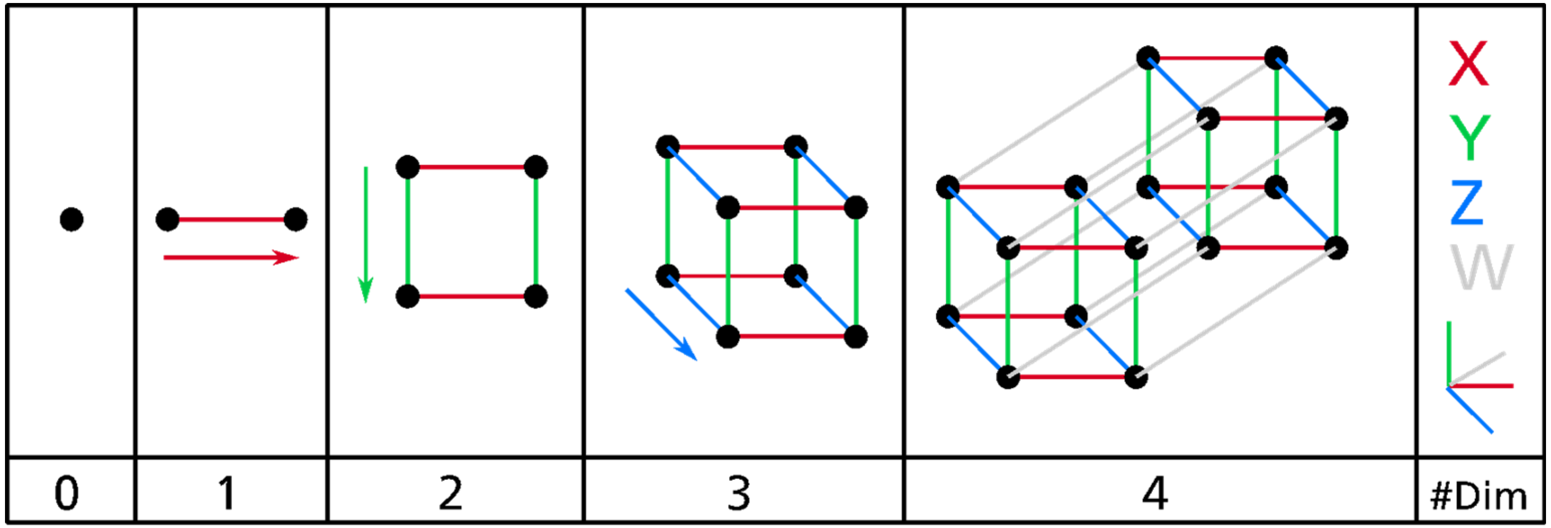
Conclusion #2

- ✓ Just 15% of amino acids found in yeast His3 orthologues were always neutral;
- ✓ The impact on fitness of the remaining 85% depended on the genetic background;
- ✓ Furthermore, at 67% of sites, amino acid replacements were under sign epistasis, having both strongly positive and negative effect in different genetic backgrounds;
- ✓ 46% of sites were under reciprocal sign epistasis.

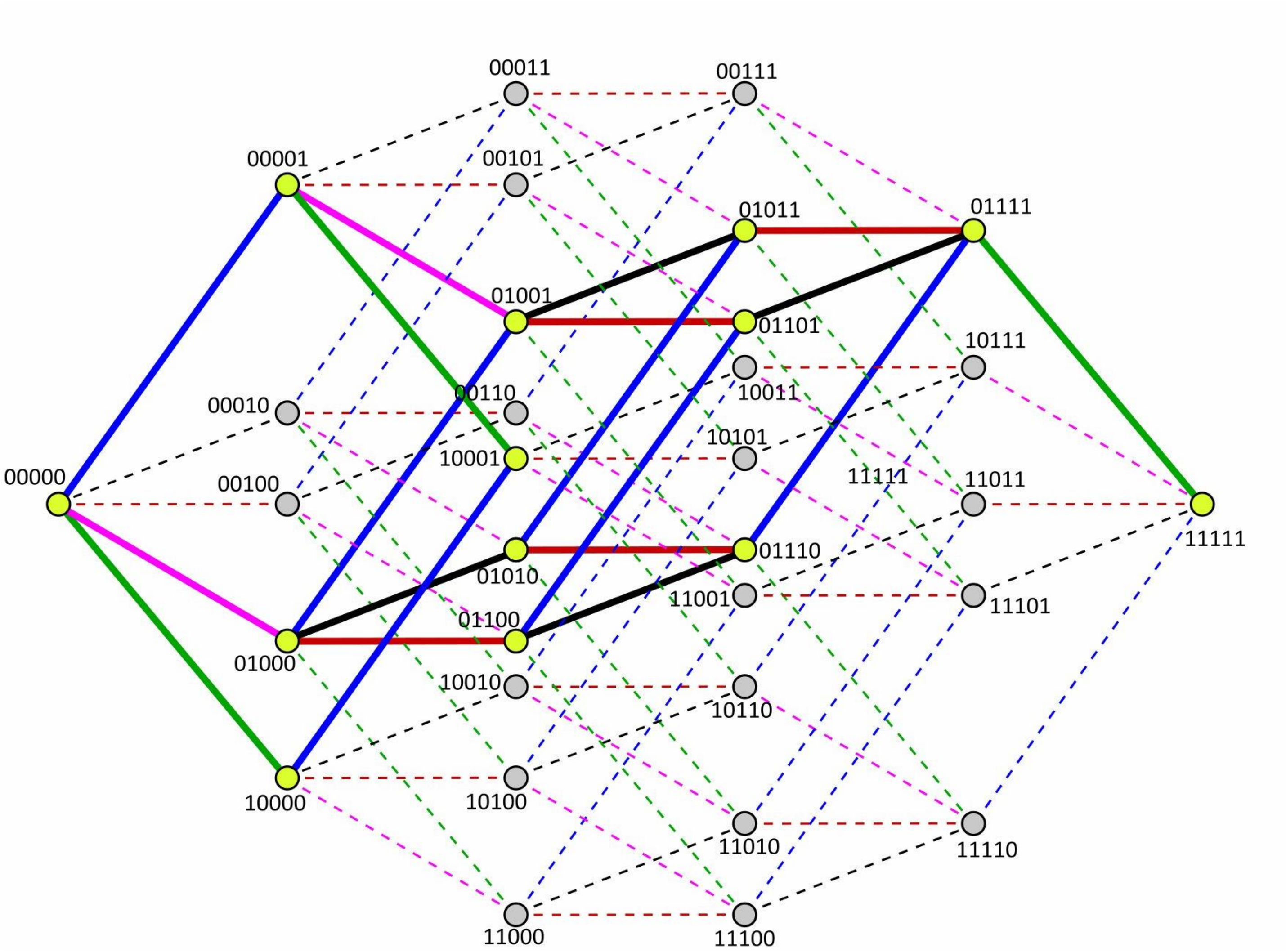
Pokusaeva V.O. et al. PLoS Genetics (2019)

How to find all hypecubes in random mutagenesis experimental data?

Creation of hypercubes

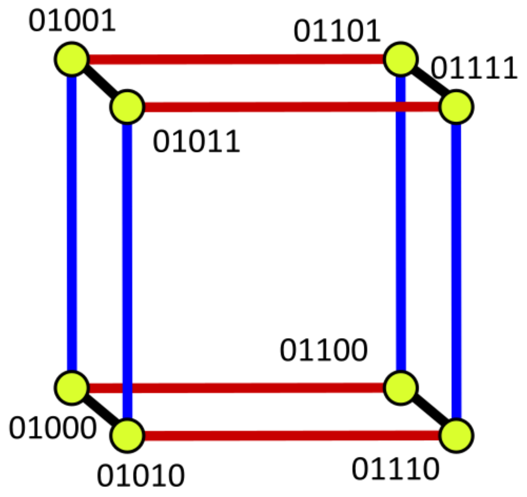


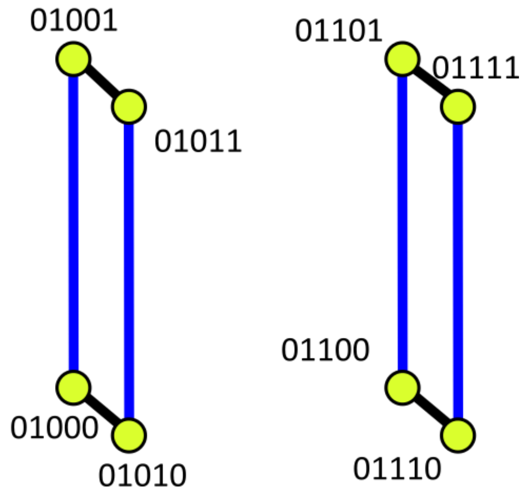
Finding all hypercubes



Idea for the algorithm

- We use the fact that N-dimensional hypercube consists of two parallel (N-1)-dimensional hypercubes





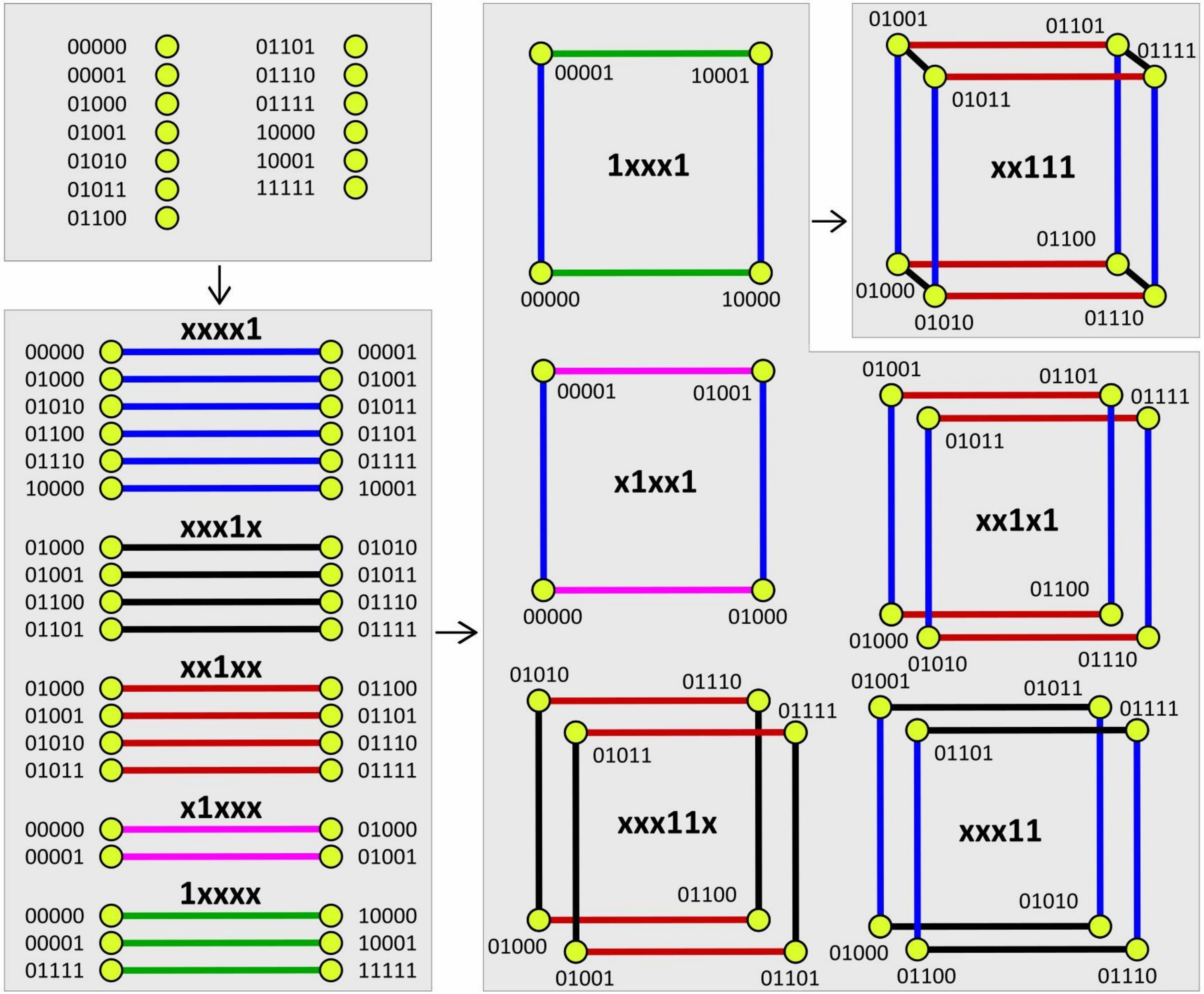
Поиск эпистаза в экспериментальных данных

Algorithm

Поиск эпистаза в экспериментальных данных

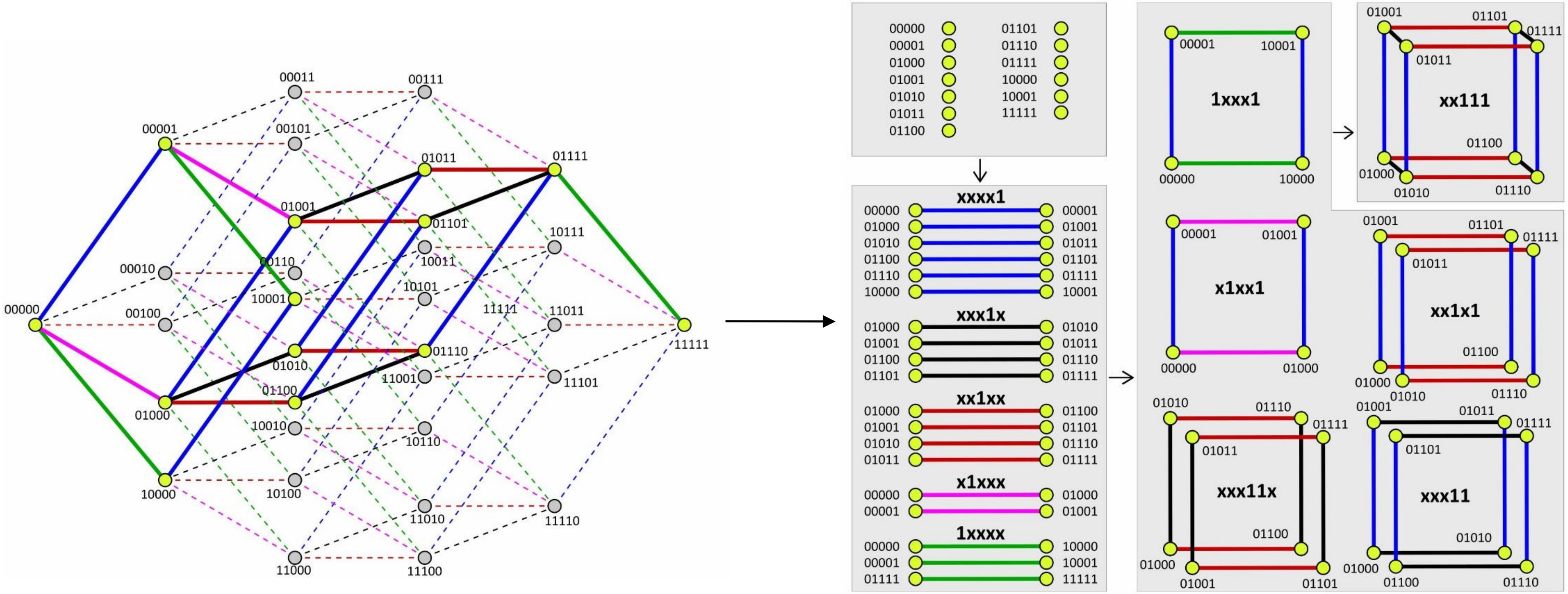
Bioinformatics, accepted

https://github.com/ivankovlab/HypercubeME



Application to HIS3 data

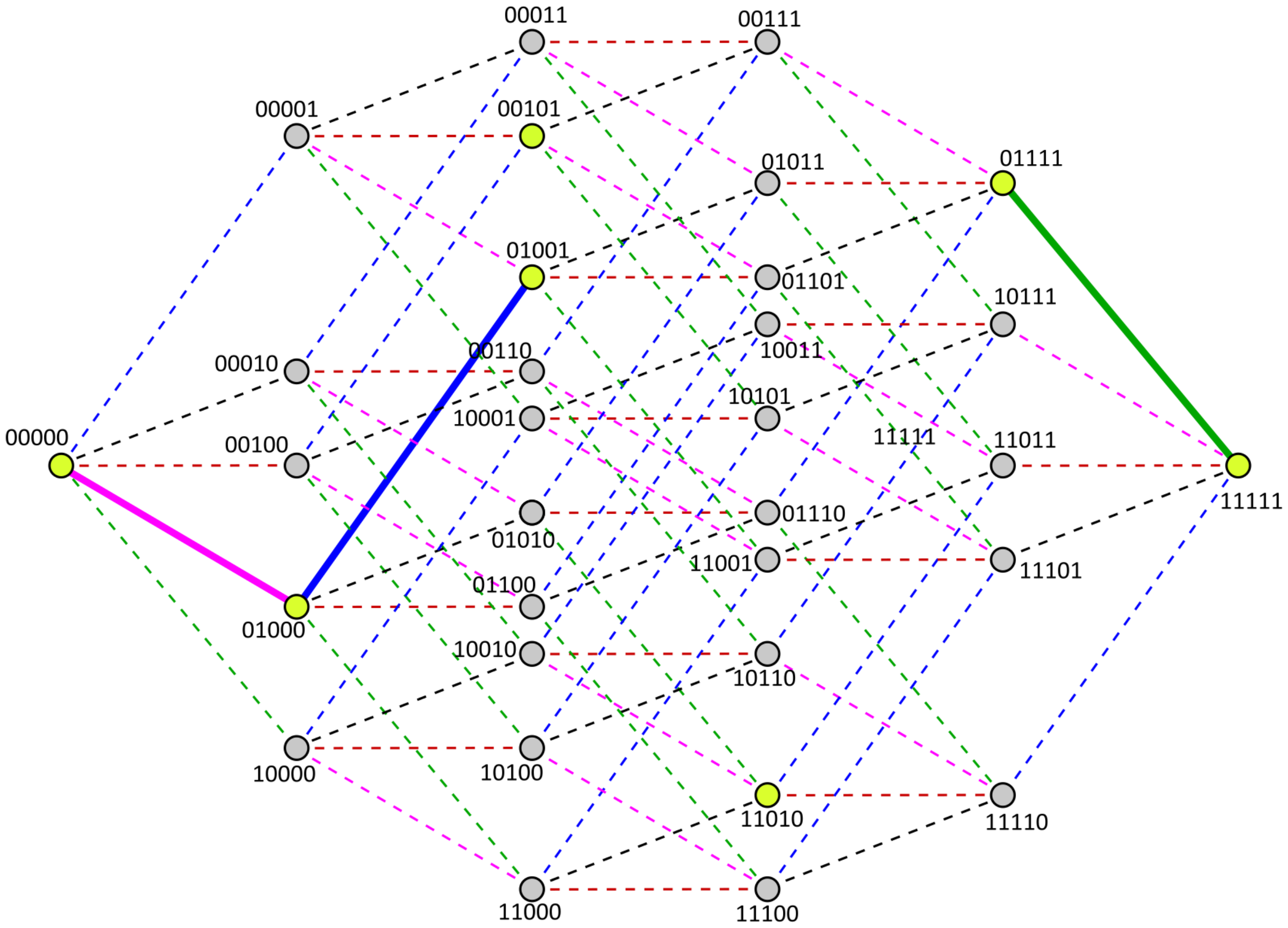
Поиск эпистаза в экспериментальных данных



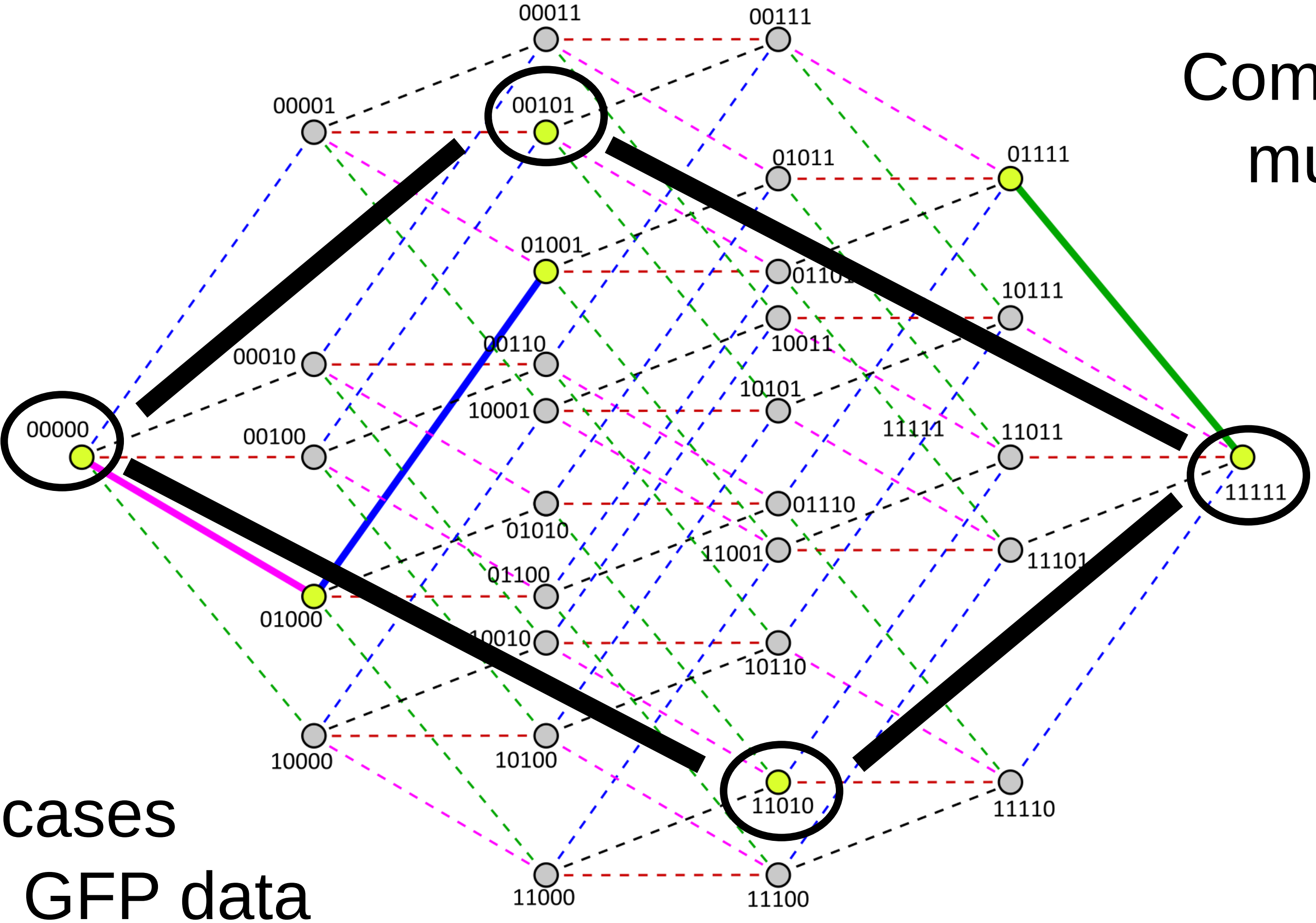
- All 199,847,053 hypercubes were found in the data from HIS3 experiment (Pokusaeva et al., 2019)
- 88% of them are of order 3 and higher

Can we study epistasis on sparse data?

Example of sparse data



Rectangles in genotype space



Composite mutations

>800,000 cases found in GFP data

Uni- vs. multi-dimensional epistasis

Issue

In GFP data 6% of variance was due to multi-dimensional epistasis

However, sign and reciprocal sign cases found were rare and did not explain that variance.

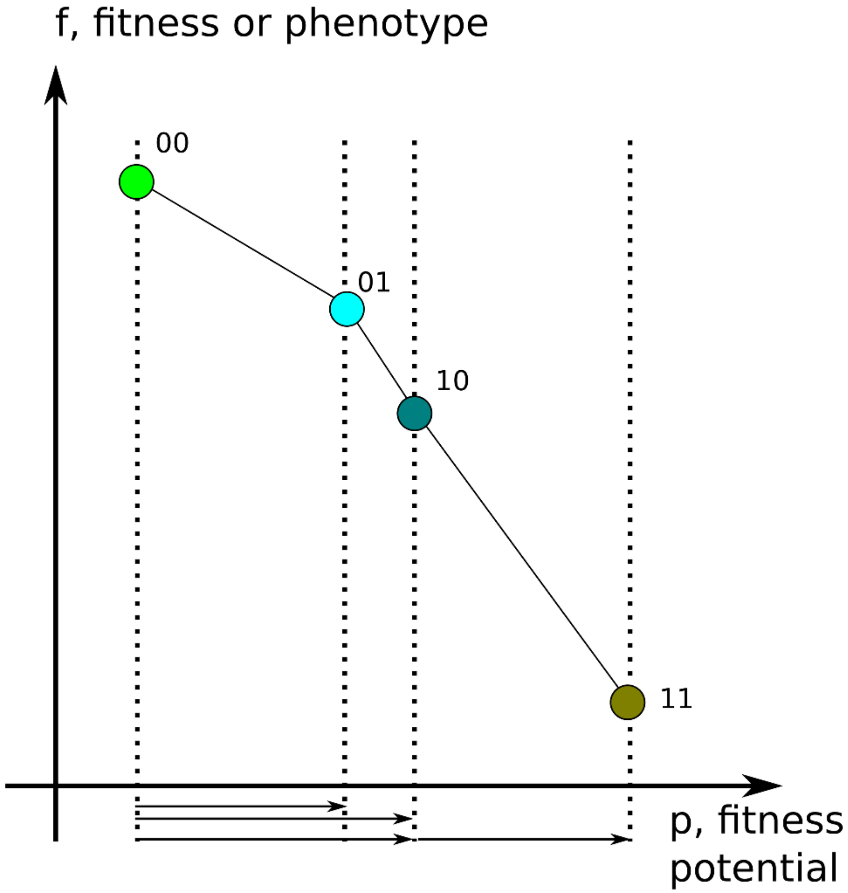
Issue

In GFP data 6% of variance was due to multi-dimensional epistasis

However, sign and reciprocal sign cases found were rare and did not explain that variance.

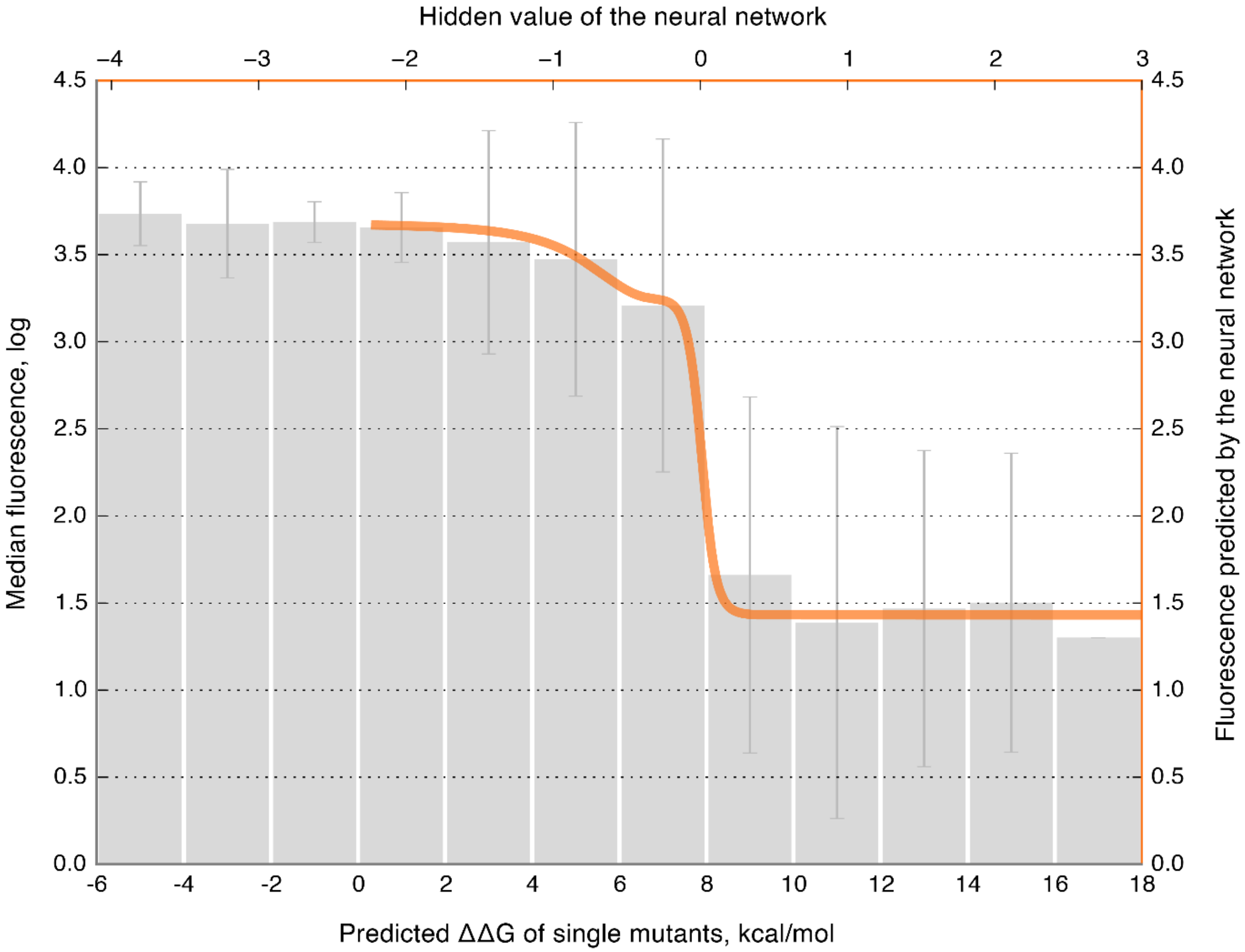
Are there other types of multi-dimensional epistasis?

Fitness potential concept



$$p(g) = const + \sum_{i=1} \alpha_i \delta_i$$

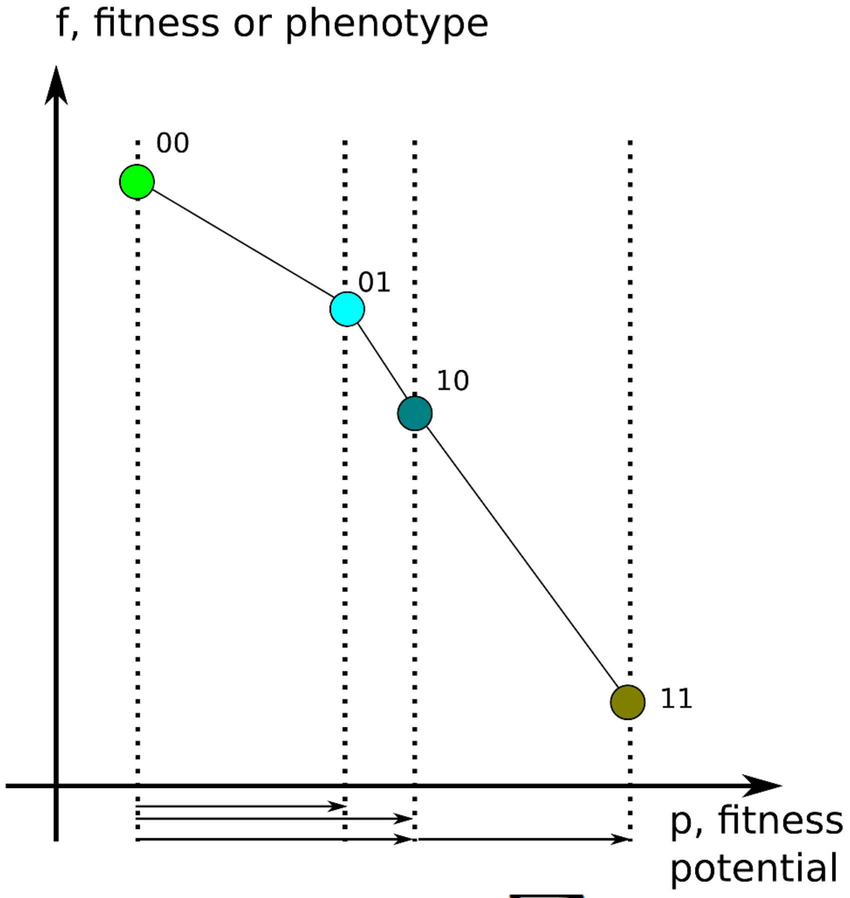
Protein stability – fitness potential?



MDE cases

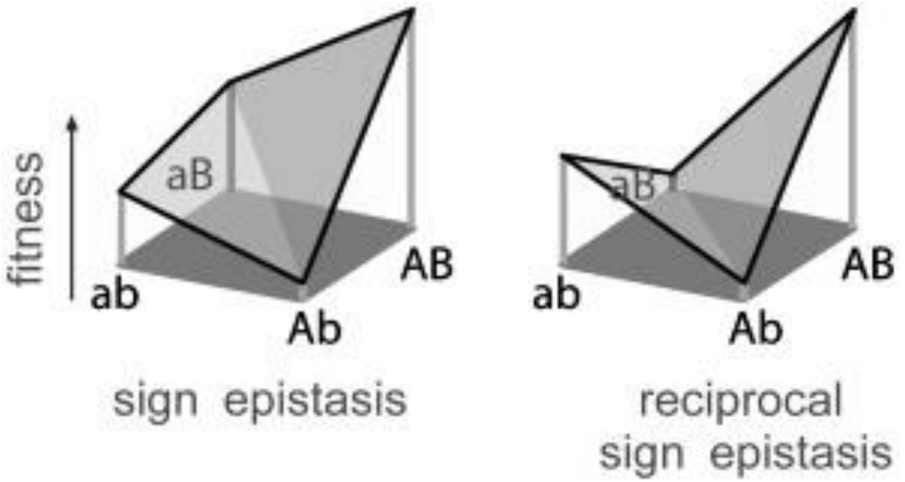
Поиск эпистаза в экспериментальных данных

Uni-

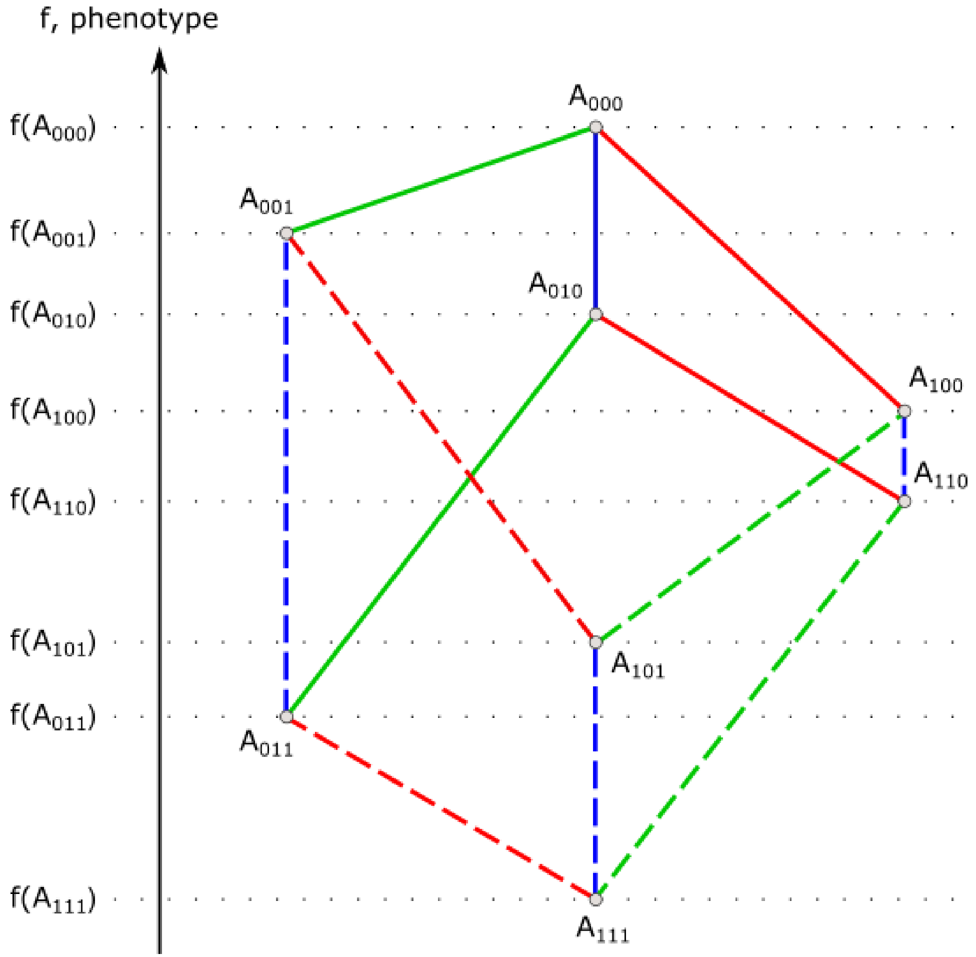


$$p(g) = const + \sum_{i=1} \alpha_i \delta_i$$

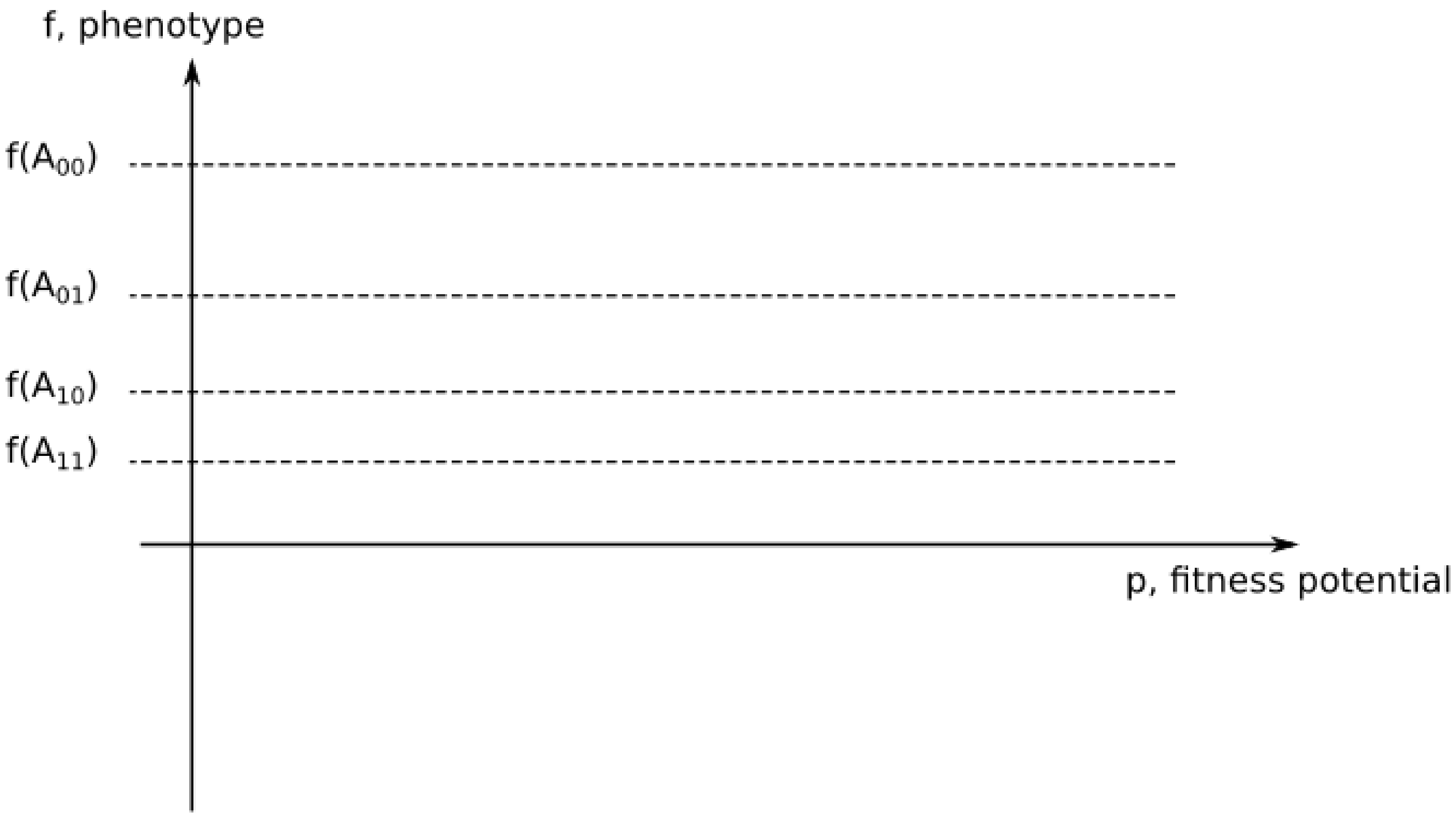
Multi-dimensional epistasis



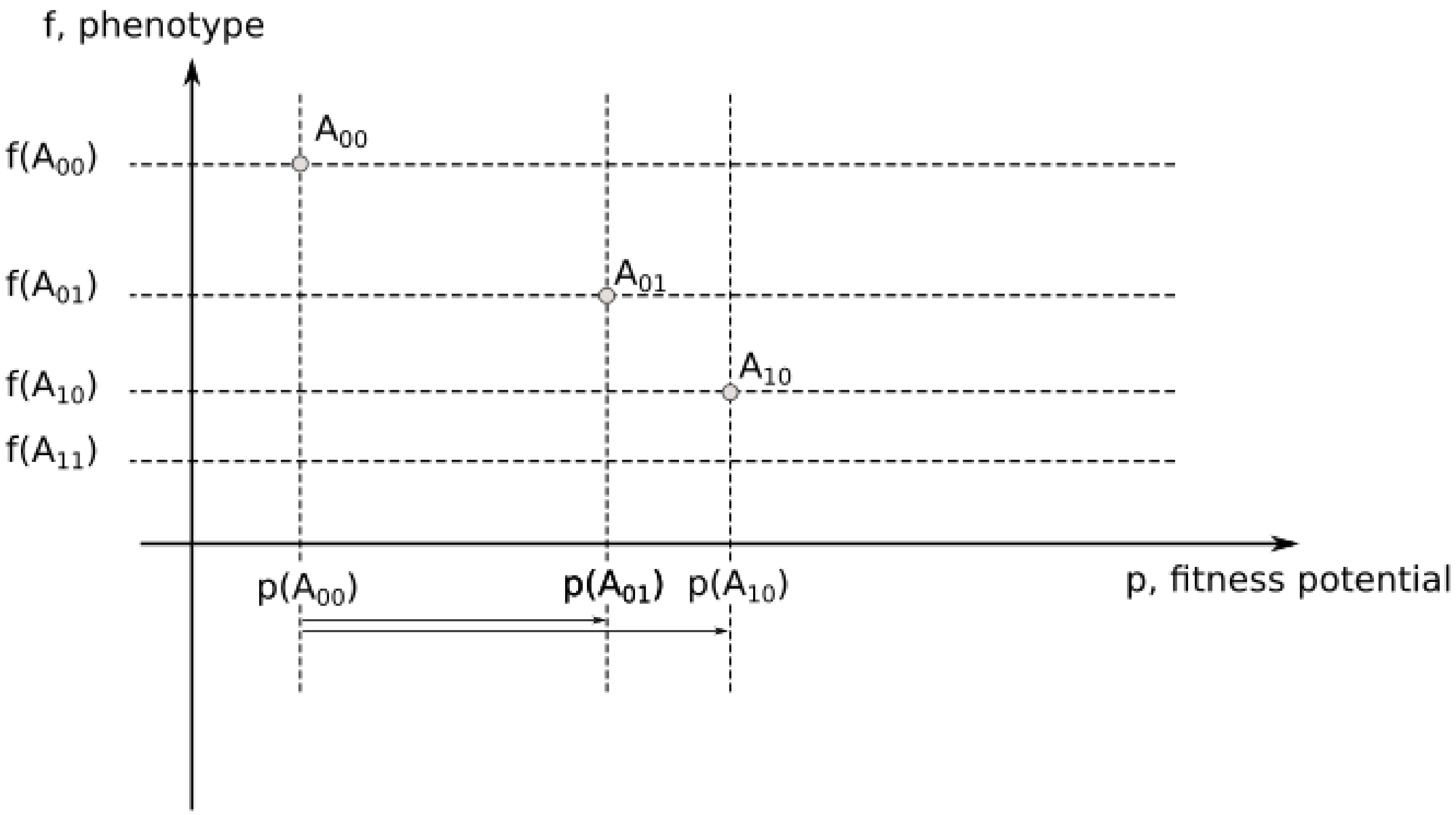
New type of MDE



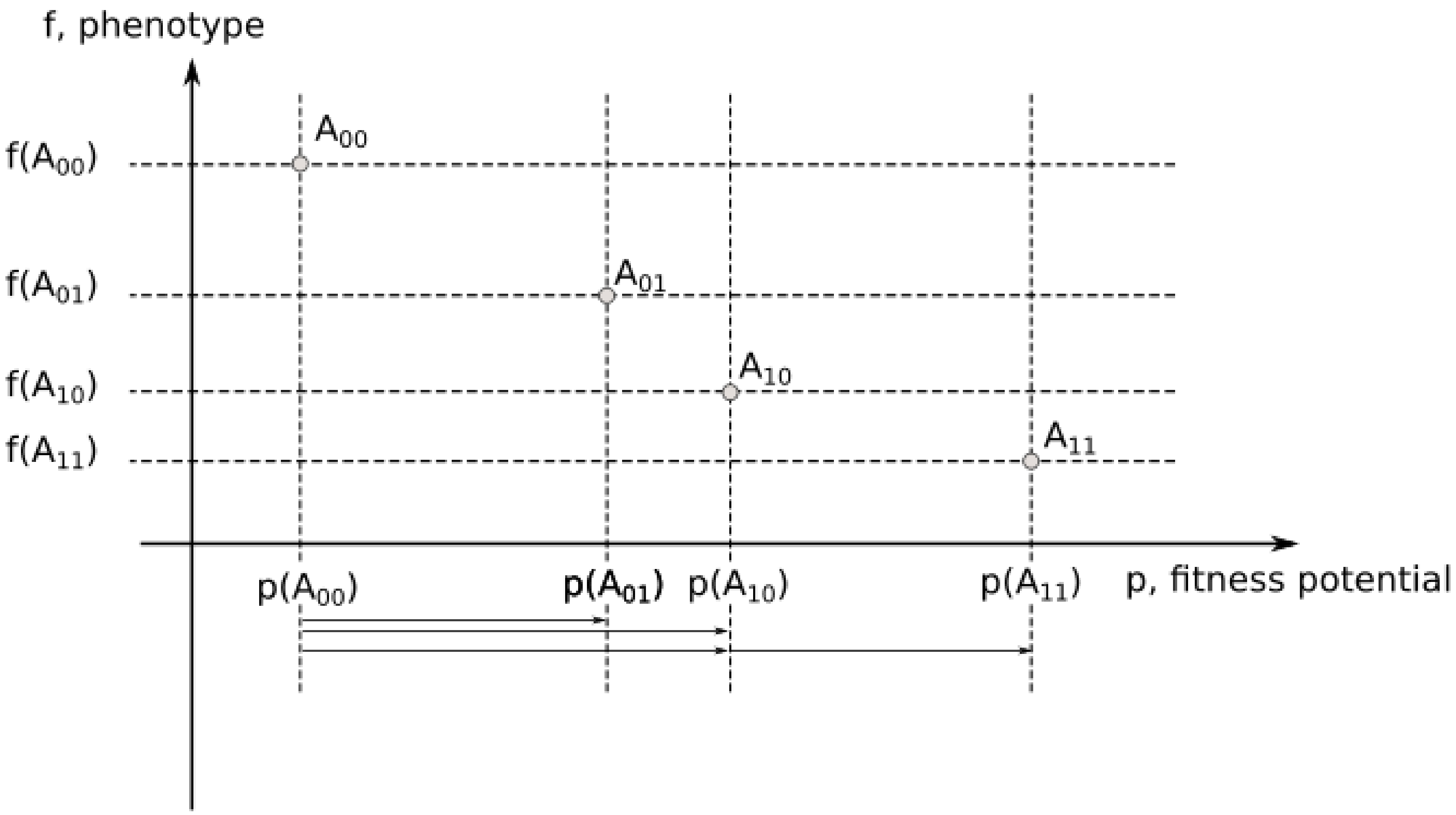
Another example of new MDE



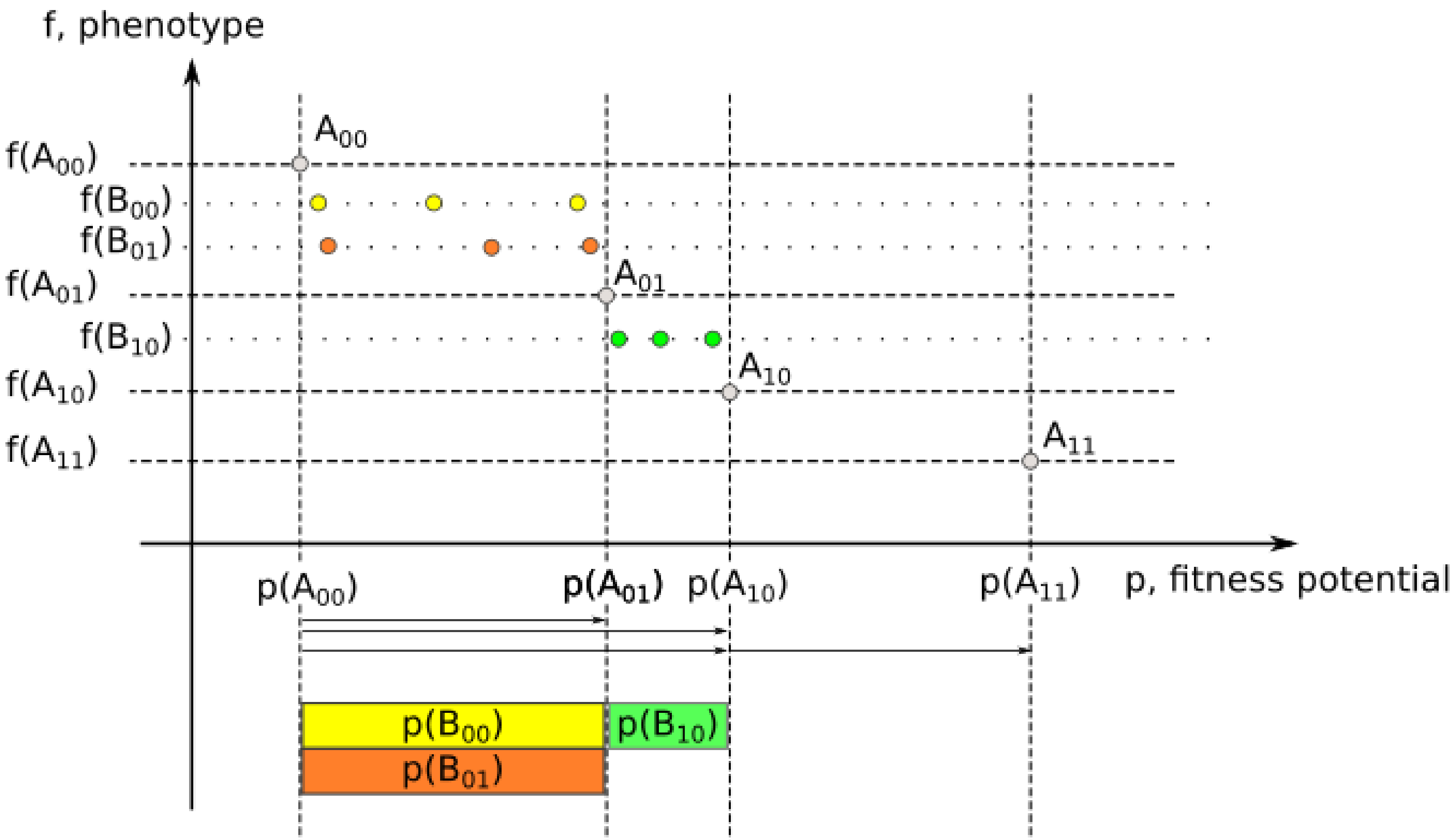
Another example of new MDE



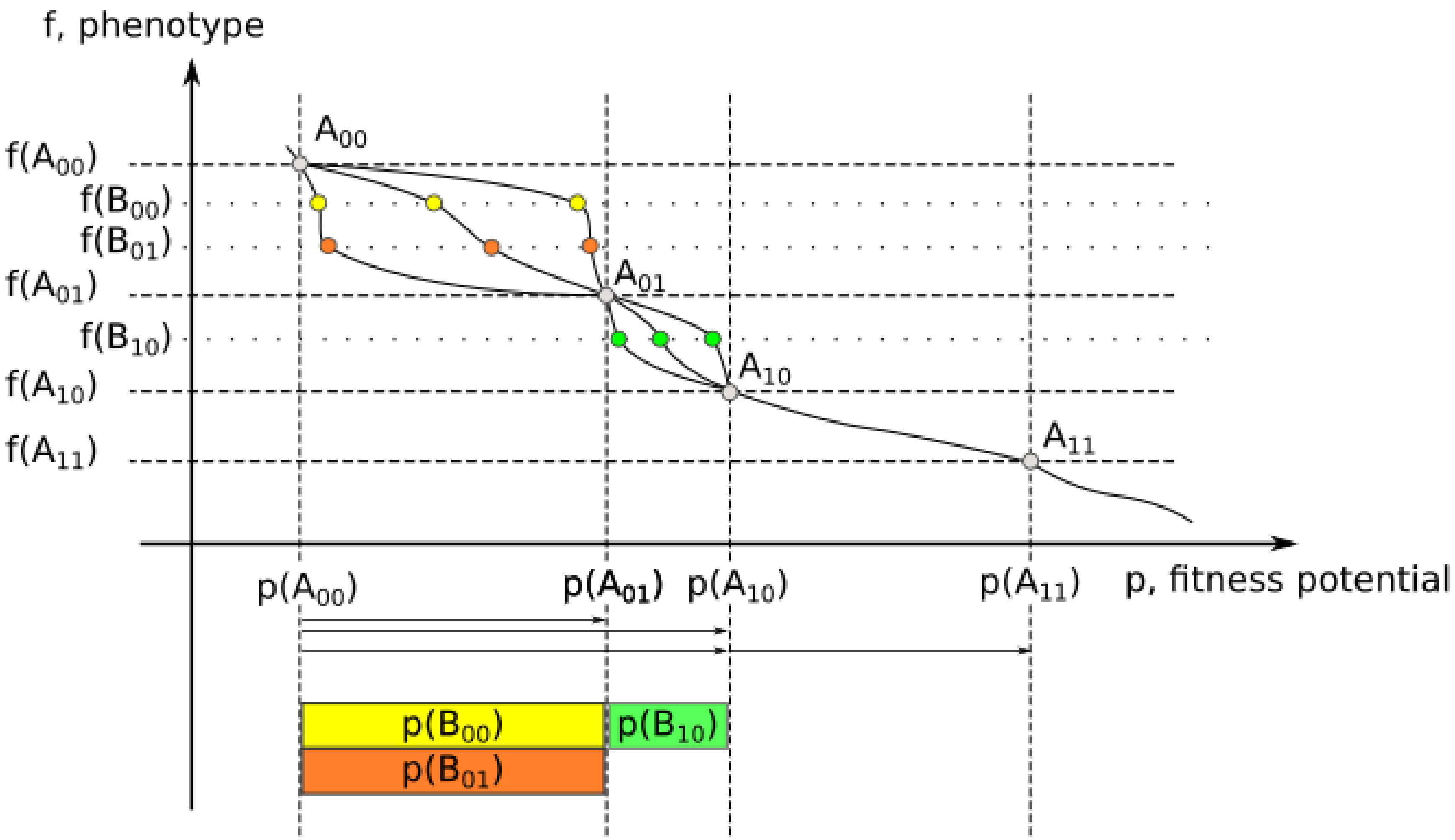
Another example of new MDE



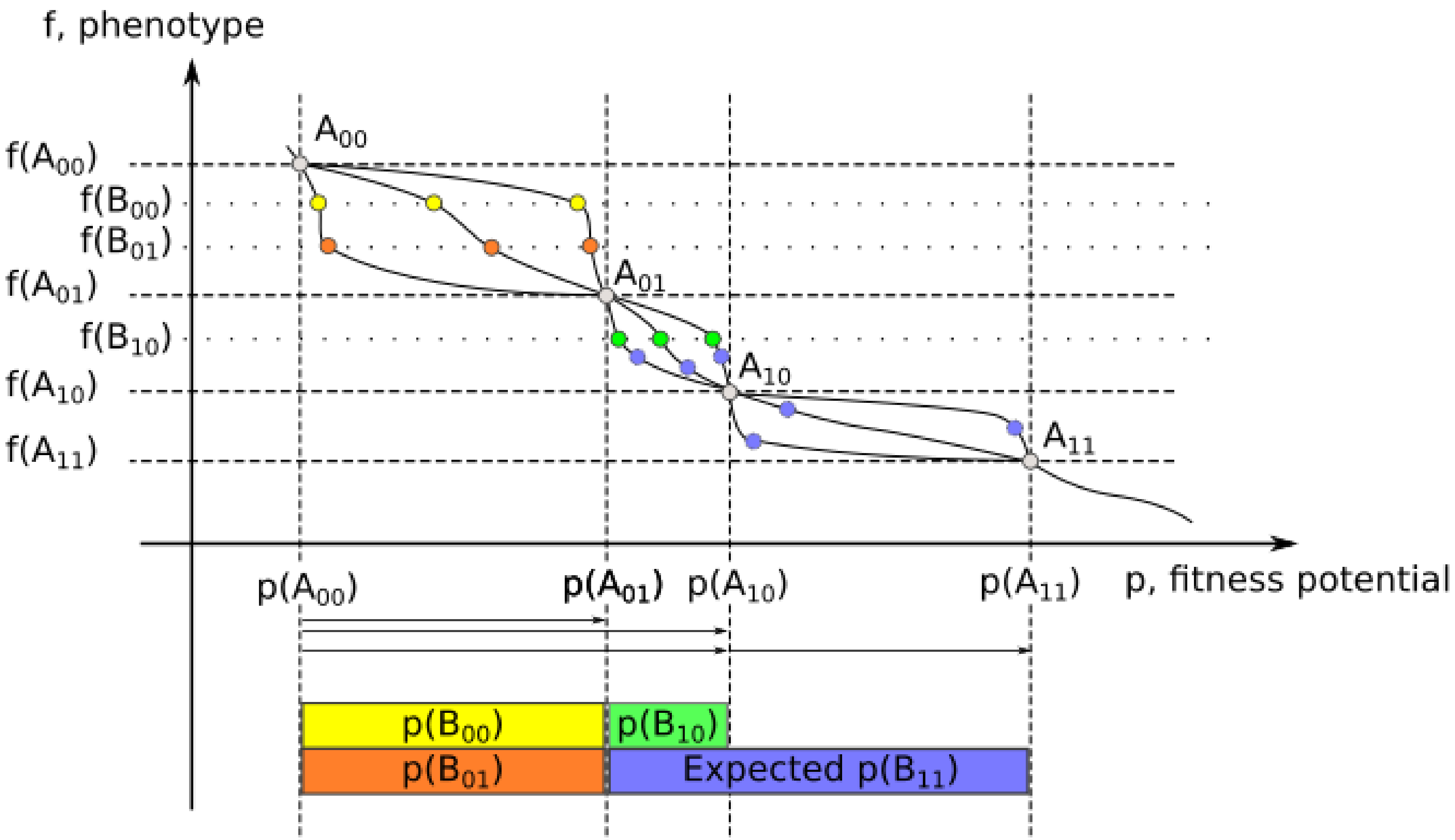
Another example of new MDE



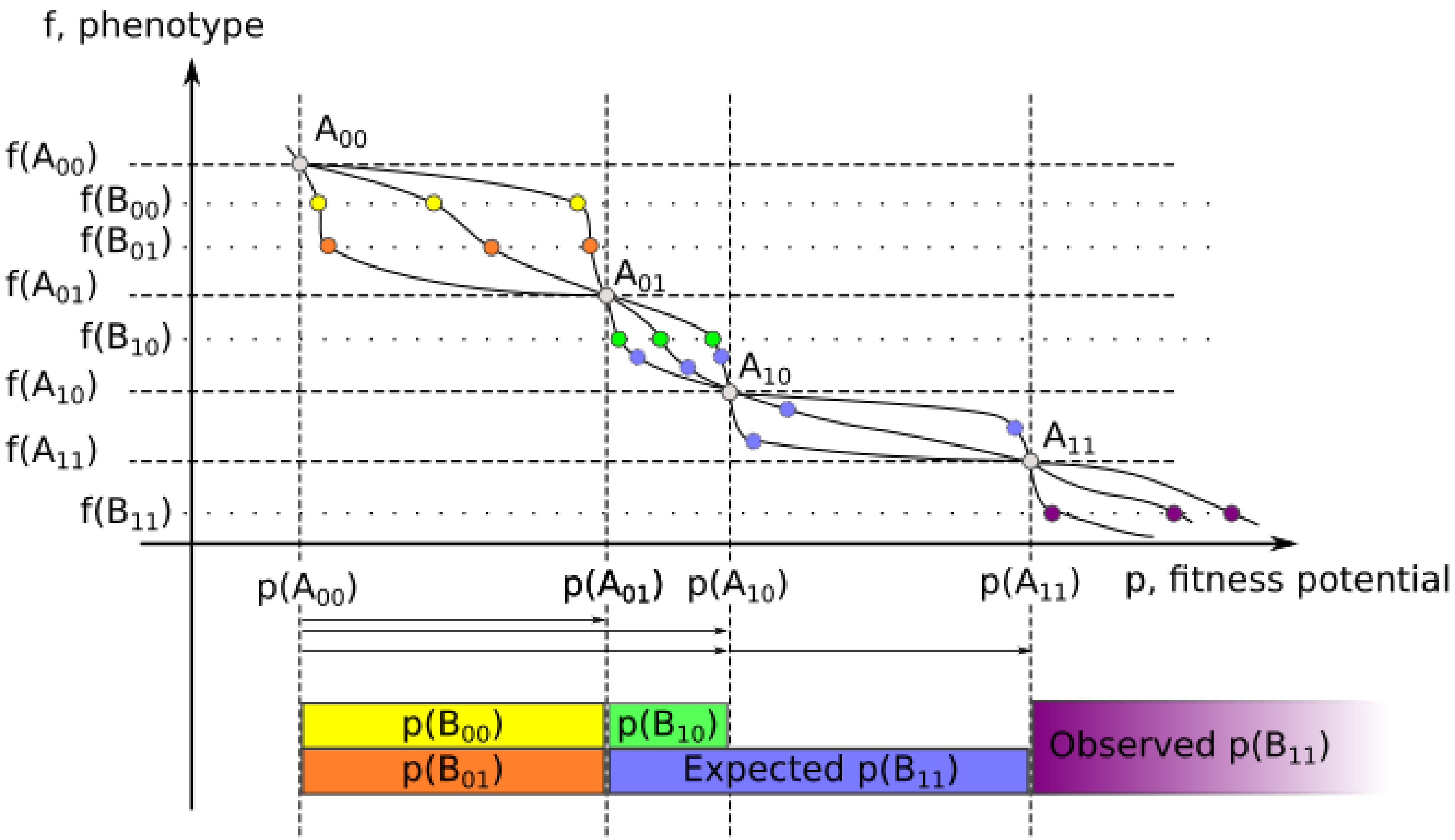
Another example of new MDE



Another example of new MDE

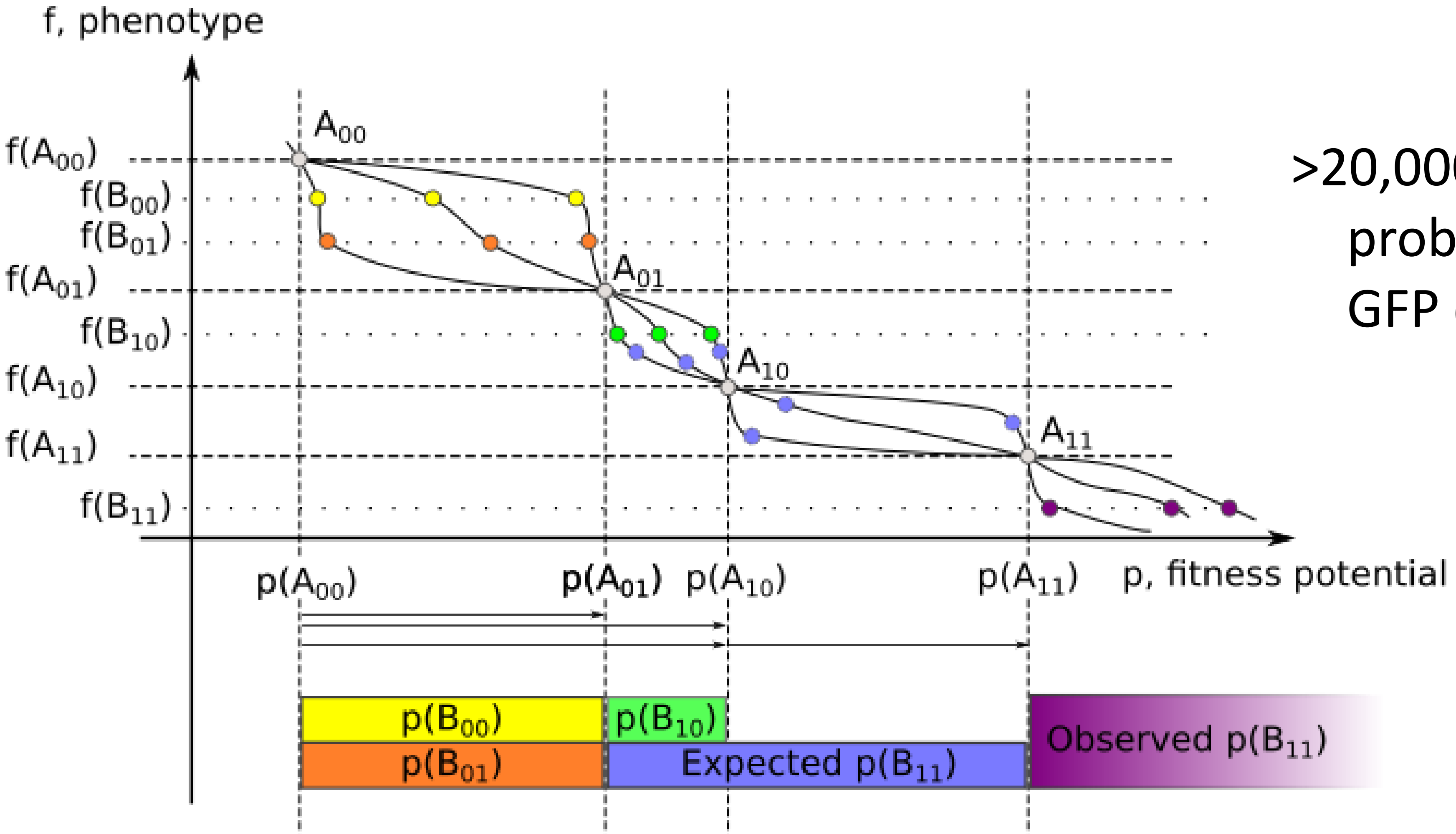


Another example of new MDE



Поиск эпистаза в экспериментальных данных

Another example of new MDE



>20,000 cases with probability >95% in GFP data

Поиск эпистаза в экспериментальных данных

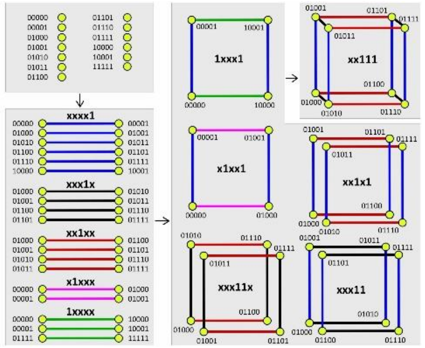
Skoltech

Hyperrectangles

Поиск эпистаза в экспериментальных данных

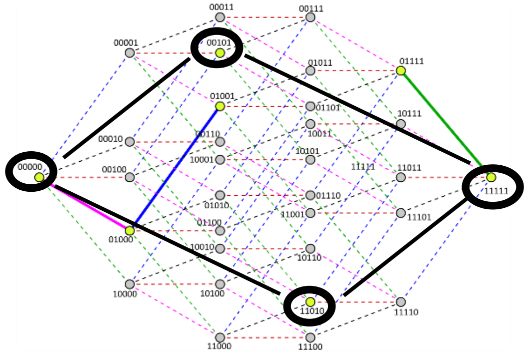
Further tool development for finding epistasis:

Hypercubes



+

Composite mutations

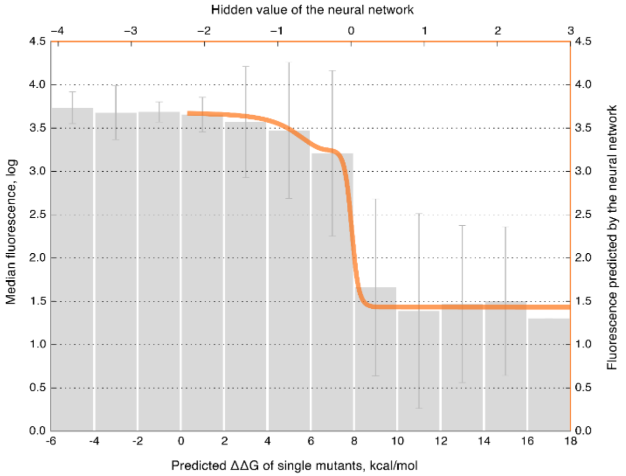


= Hyperrectangles

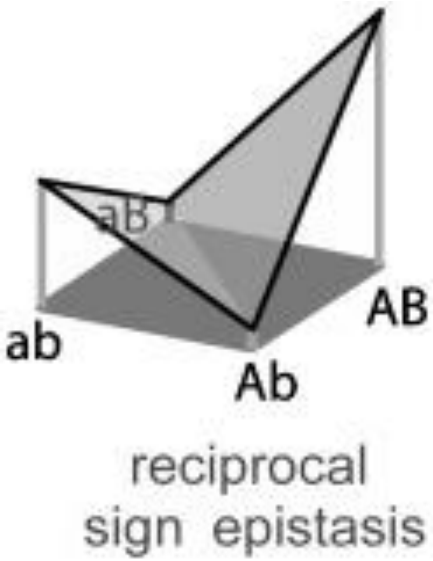
Поиск эпистаза в экспериментальных данных

Relationship between HOE & MDE

Higher-order but
not multi-dimensional



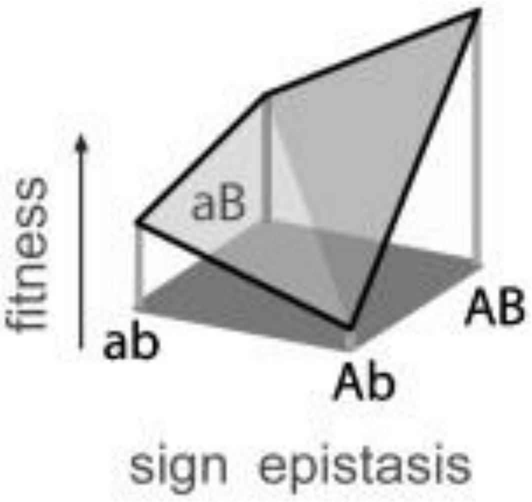
Multi-dimensional but
not higher-order



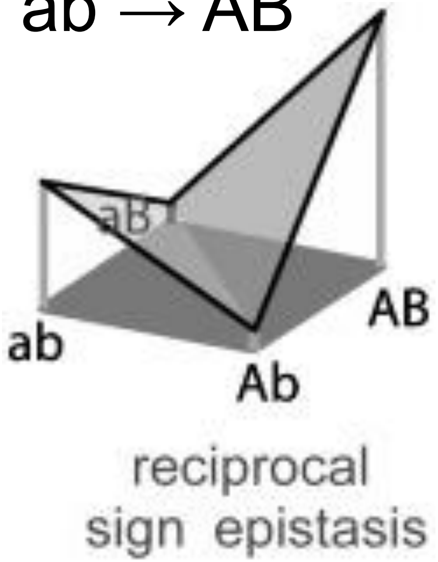
Is it true that higher-order epistasis results from non-linearity while multi-dimensional results from physical interactions?

Pathway accessibility and MDE

Two-dimensional:
50% of pathways
ab → AB



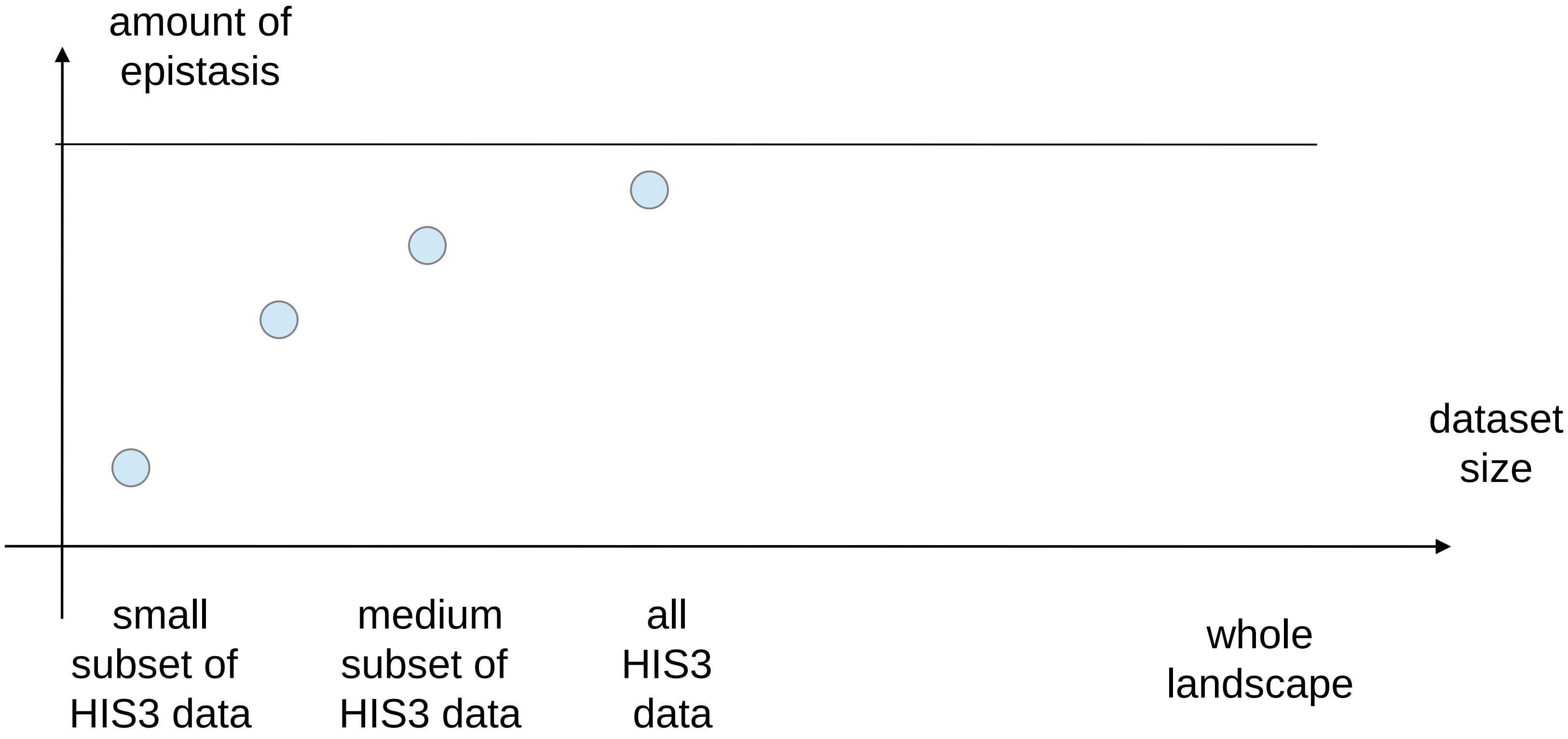
Two-dimensional:
0% of pathways
ab → AB



- ✓ New type of multi-dimensional epistasis: from 0% to 100%
- ✓ Correlation on expirimental data?

Extrapolation of epistasis

Find epistasis for fitness landscape subsets of different sizes and extrapolate. Can we find a limit?



Acknowledgements

Fyodor
Kondrashov

Grigoriy Leleytner

Daniil Bobrovskiy

Liubov Lonishin

Natalya Bogatyreva

Laura Aviñó Esteban

thx.

Skoltech