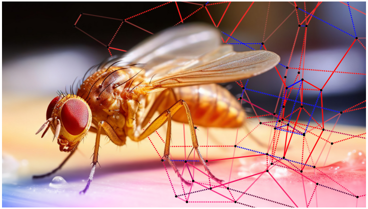
Flies and regular subdivisions

Michael Joswig

TU Berlin & MPI-MiS Leipzig

14 Dec 2023



©Thomas Endler

joint w/ Holger Eble
Lisa Lamberti
Will Ludington

1. Mathematics
   - epistasis
   - fitness landscapes
   - cluster partitions and dendrograms
2. Statistics
   - significance test
3. Biology
   - *E.coli* evolution
   - *Drosophila* microbiome

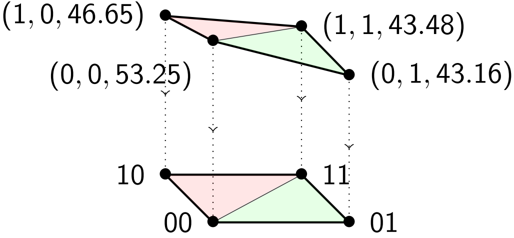
Regular subdivision of a point configuration

Epistasis [Bateson 1909]

- point set V in $\mathbb{R}^n$
  - n-biallelic genetic system $V = \{0,1\}^n$

- lift to $\mathbb{R}^{n+1}$ via height function $h : V \to \mathbb{R}$
  - phenoytpe

- take upper convex hull and project back
  - yields subdivision $\mathcal{S}(V,h)$ of $\mathrm{conv}(V)$



- generic height function ⇝ triangulation
  - lifted points coplanar $\iff$ no biological interaction

Epistasis and shapes of fitness landscapes

Beerenwinkel, Pachter & Sturmfels 2007

Consider n-biallelic system $V = \{0,1\}^n$ with phenotype $h : V \to \mathbb{R}$.

- (relative) population = map $p : V \to \mathbb{R}_{\geq 0}$ with $\sum_{v \in V} p(v) = 1$
- allele frequency vector $\rho(p) := \sum_{v \in V} p(v) v$ contained in $[0,1]^n$
- $\Delta_V :=$ set of all relative populations = simplex of dimension $2^n - 1$
- for fixed $w \in [0,1]^n$:

$$\begin{array}{ll} \text{maximize} & h \cdot p \\ \text{subject to} & p \in \Delta_V \text{ and } \rho(p) = w \end{array} \qquad (\mathrm{LP}(h,w))$$

- if h and w generic then $\mathrm{LP}(h,w)$ has unique optimal solution, the fittest population $p^* = p^*(h,w) =$ vertex of $\{p \in \Delta_V \mid \rho(p) = w\}$
- optimal value of $\mathrm{LP}(h,w)$ is $h \cdot p^* = \sum \lambda_i(h(v_i))$
- piecewise linear function $h^* \colon [0,1]^n \to \mathbb{R}$, $w \mapsto h \cdot p^*(h,w)$
- regions of linearity of h^* = maximal cells of $\mathcal{S}(V,h)$

The epistatic weight of a dual edge

Let V be vertex set of some n-polytope, equipped with generic height function h. Thus $\mathcal{S} = \mathcal{S}(V, h)$ is a triangulation. For

$$s \;=\; \operatorname{conv}\{v_1, v_2, \dots, v_{n+1}\} \quad \text{and} \quad t \;=\; \operatorname{conv}\{v_2, v_3, \dots, v_{n+2}\}$$

two adjacent n-simplices of $\mathcal{S}$ define

$$E_h(s,t) \;:=\; \begin{pmatrix} 1 & v_{1,1} & v_{1,2} & \dots & v_{1,n} & h(v_1) \\ 1 & v_{2,1} & v_{2,2} & \dots & v_{2,n} & h(v_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & v_{n+2,1} & v_{n+2,2} & \dots & v_{n+2,n} & h(v_{n+2}) \end{pmatrix} \;.$$

The epistatic weight of the dual edge (s,t) is

$$e_h(s,t) \;:=\; \left|\det E_h(s,t)\right| \cdot \frac{\operatorname{nvol}(s \cap t)}{\operatorname{nvol} s \cdot \operatorname{nvol} t} \;.$$

▸ statistics

Cluster partitions and epistatic filtrations

Consider $\mathcal{S} = \mathcal{S}(V, h)$, with dual graph Γ.

Picking threshold value $\theta \geq 0$ yields

- $\Gamma(\theta) = \Gamma$ minus dual edges of epistatic weight $> \theta$

Definition

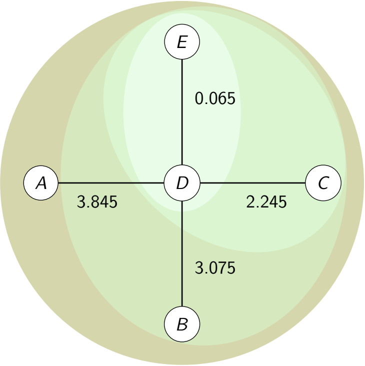
A θ-cluster of $\mathcal{S}$ is a connected component of $\Gamma(\theta)$.

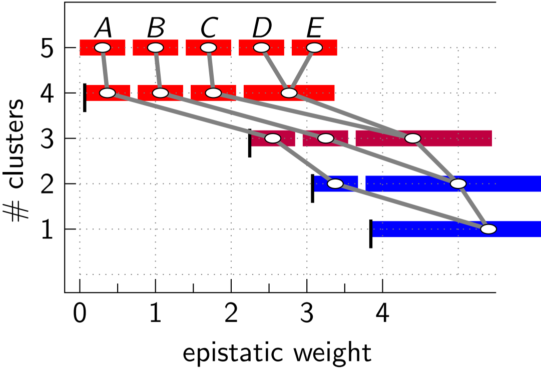
- induces partition of $\Gamma(\theta)$ into θ-clusters
- 0-cluster = single facet; ∞-cluster = all facets
- epistatic filtration $\Gamma(0) < \Gamma(\theta_1) < \cdots < \Gamma(\theta_\ell) = \Gamma$, linearly ordered by refinement

Example: nonunimodular triangulation of $[0,1]^3$

Consider triangulation $\mathcal{S}([0,1]^3, \texttt{ttd})$ with five maximal simplices:

$A = 000\ 100\ 110\ 101 \quad B = 000\ 001\ 101\ 011 \quad C = 000\ 010\ 110\ 011$
$D = 000\ 110\ 101\ 011 \quad E = 110\ 101\ 011\ 111$





Height functions as random variables

Fix simplices s and t with joint vertices $v_1, v_2, \ldots, v_{n+2}$ and random variables X_{v_i}. We set

◂ recall E

$$\lambda_i \ := \ (-1)^{n+i} \det(E_i) \cdot \frac{\operatorname{nvol}(s \cap t)}{\operatorname{nvol} s \cdot \operatorname{nvol} t} \ .$$

Then the expectation of the random variable $e_X(s,t)$ satisfies

$$\left| \sum_{i=1}^{n+2} \lambda_i \, \mathbb{E}\bigl(X_{v_i}\bigr) \right| \ \leq \ \mathbb{E}\bigl(e_X(s,t)\bigr) \ \leq \ \sum_{i=1}^{n+2} |\lambda_i| \, \mathbb{E}\bigl(X_{v_i}\bigr) \ .$$

If the random variables X_{v_i} are independent, then

- variance can be bounded, too.

If additionally, each random variable is normally distributed, then

- folded normal distribution

Significance test for one epistatic weight

Let (s, t) be a dual edge of $\mathcal{S}$.

- distribution mean $\mu = \mathbb{E}(e_X(s,t))$ of random variable $e_X(s,t)$ not known exactly
- wanted: one-sided test of significance with null hypothesis $\mu = 0$ vs. alternative $\mu > 0$

Assumption: random variables X_v normally distributed (and independent)

- for sample mean $Z = e_{\bar{X}}(s,t)$ then

$$P(X \geq Z) \ = \ \int_Z^\infty \frac{\sqrt{2}}{\sigma_{e_{\bar{X}}(s,t)}\sqrt{\pi}} \, e^{-\frac{1}{2}\left(\frac{x}{\sigma_{e_{\bar{X}}(s,t)}}\right)^2} dx$$

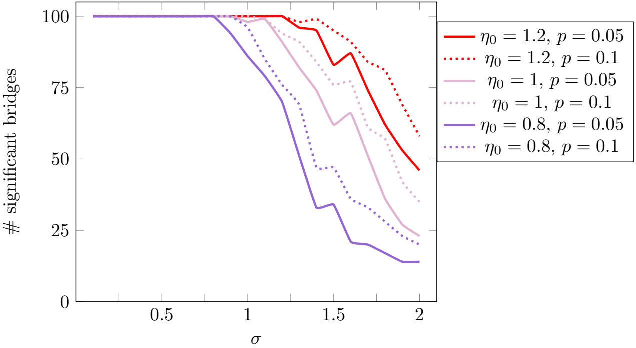
Definition

dual edge (s, t) significant if $P(X \geq Z) < 0.05$

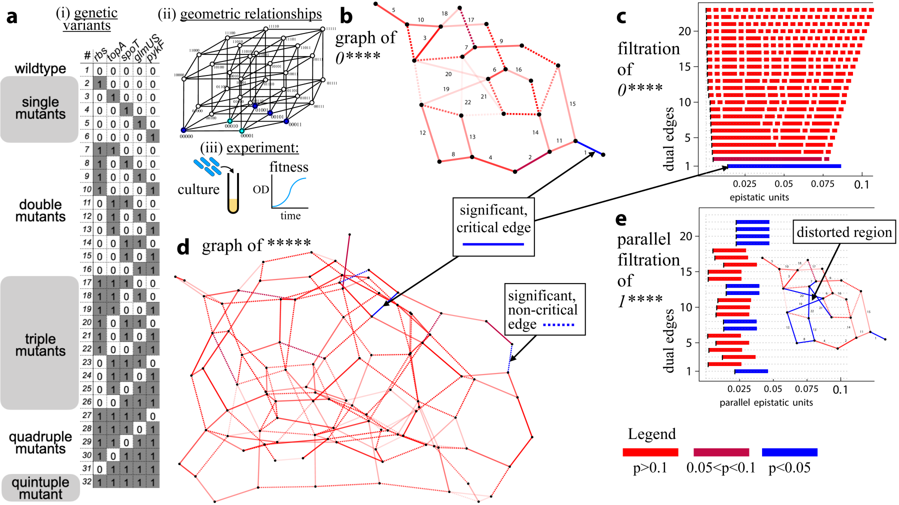
A synthetic experiment

For $V = \{0,1\}^5$, $\eta(v) = 5$ (for $v \neq 0$), $\eta(0) = 5 - \eta_0$, $0.8 \leq \eta_0 \leq 1.2$ the regular subdivision $\mathcal{S}(V,\eta)$ is a vertex split.

- to each vertex we assign normally distributed random variable with mean $\mu = 0$ and standard deviation $0.1 \leq \sigma \leq 2.0$
- 100 realizations per vertex
- for fixed (η_0, σ) repeat experiment 100 times; try $p \in \{0.05, 0.1\}$



E.coli evolution. Data set: Khan et al. 2011

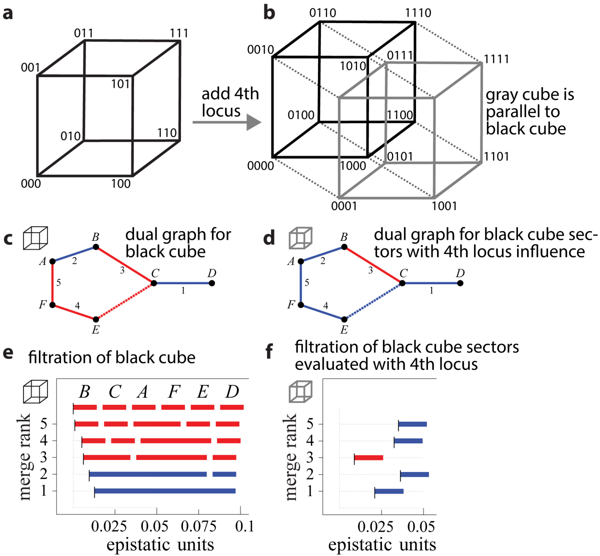


#	rbs	topA	spoT	glmUS	pykF
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1
7	1	1	0	0	0
8	1	0	1	0	0
9	1	0	0	1	0
10	1	0	0	0	1
11	0	1	1	0	0
12	0	1	0	1	0
13	0	1	0	0	1
14	0	0	1	1	0
15	0	0	1	0	1
16	0	0	0	1	1
17	1	1	1	0	0
18	1	1	0	1	0
19	1	1	0	0	1
20	1	0	1	1	0
21	1	0	1	0	1
22	1	0	0	1	1
23	0	1	1	1	0
24	0	1	1	0	1
25	0	1	0	1	1
26	0	0	1	1	1
27	1	1	1	1	0
28	1	1	1	0	1
29	1	1	0	1	1
30	1	0	1	1	1
31	0	1	1	1	1
32	1	1	1	1	1

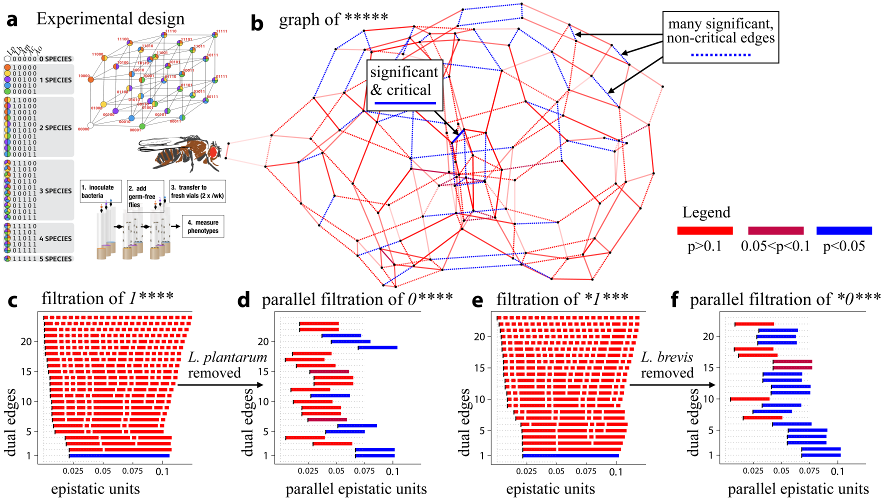
- significant 4D interaction: $00001 + 00000|01001|00101|00011 + 00010$
- *ribosome-binding site* mutation = master regulator

Marginal and conditional epistasis

parallel epistatic filtration



Drosophila microbiome. Data set: Ludington lab



a Experimental design

b graph of *****

significant & critical

many significant, non-critical edges

c filtration of *1*****

d parallel filtration of *0*****

L. plantarum removed

e filtration of **1****

f parallel filtration of **0****

L. brevis removed

- Lactobacilli = master regulators

Conclusion

- new method to process epistatic data in biology
  - ties in with previous approaches
  - provides a test for statistical signifance
  - agrees with established biological results
- new way to visualize high-dimensional data
  - works for arbitrary regular subdivisions
  - e.g., tropical hypersurfaces (which are dual to regular subdivisions)

Holger Eble, Michael Joswig, Lisa Lamberti, and William B. Ludington, Cluster partitions and fitness landscapes of the *Drosophila* fly microbiome, J. Math. Biol. **79** (2019), no. 3, 861–899.

———, Master regulators of biological systems in higher dimensions, Proc. Natl. Acad. Sci. USA **120** (2023), no. 51.

Epistatic Filtrations Calculator

This is an online client for computing higher-order epistatic interactions as detailed in the articles [1] and [2]. It was implemented as polymake extension and can be found online on GitHub. If you found this useful for your scientific work, please cite our paper [1].

The input is a sequence of genotype-phenotype maps, where several phenotypes for the same genotype are considered as independent measurements, thus giving rise to a distribution of phenotypes. Genotypes are 0/1-vectors (i.e., here we are treating the biallelic case only), and phenotypes are real numbers. The entire dataset is supposed to be contained in a single file of type csv (ASCII text, comma separated values). Such files can be exported from standard spreadsheet software.

Upload csv file

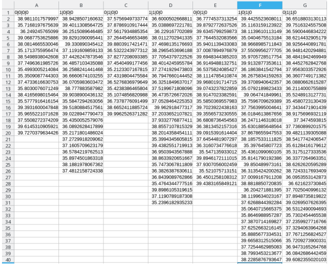
The input csv file must be in the precise format shown in the exemplary screenshot on the right hand side:

- The genotypes are placed in the first data row. Their coordinates are separated by vertical bars, e.g. 0|0|1|0.
- Right below the genotypes, the measured data is placed accordingly. The columns are allowed to be of varying size.

Browse... No file selected.

UPLOAD

Please upload a file.



	A	B	C	D	E	F	G
1	0|0|0|0	1|0|0|0	0|1|0|0	1|1|0|0	0|0|1|0	1|0|1|0	0|1|1|0
2	38.9811017579997	38.9428507160632	37.5756949733774	36.6000502668811	36.7774537313254	39.4425523608011	36.6518803130113
3	35.7168197675639	39.4011308564725	37.8786910917444	35.0388697221781	39.8792772637526	35.1163159123922	39.7516324557508
4	36.249245765099	36.2515089646485	37.5617934885354	36.229167702089	39.6345799259873	38.1139610113149	36.5900446834222
5	39.0687753625886	39.8291090095441	37.2644546653486	38.0112702941335	37.7644532083566	35.0404675513184	38.6214432905179
6	38.0814665530046	39.330890345412	39.8803917421471	37.4698135176693	35.9401139433083	38.9668985711843	38.9256440891781
7	35.1713755956474	37.1191609859133	36.5322243977312	38.2965453696188	37.0087894976879	37.5509956277705	36.9481420294881
8	38.5498938042608	37.4426247873546	37.8277208093385	37.7054379722526	39.6948344385203	35.9705728517754	38.4841942469949
9	37.7496361985726	36.4857103435088	37.4540499177456	38.4014245955784	36.9149488132751	39.9132877353611	38.4452762842768
10	35.4993005214084	36.2588241441468	35.2123307167815	37.2741929473803	36.5375824085427	36.1863841542794	37.9563033572926
11	35.3509087744303	36.6660674103255	37.4319804475584	36.7947660144452	38.1114785410874	36.2675834159263	36.3607749171382
12	37.4733616630753	36.0705983603472	36.5276836979649	36.3251849637017	39.9680191714715	39.3708940642357	36.0880662615287
13	35.8030076071249	38.777883587982	35.4238386465804	37.5199671808096	39.0743237822859	35.0792189823433	36.2114000755889
14	38.4165698015464	39.9038900436132	35.1074856820988	36.4735726672026	38.9147023382591	39.064741849961	35.5248913127731
15	35.5777916416154	35.5847294263056	36.7378776091499	37.0528464225353	35.5850369557983	35.7596709629389	35.4580723130439
16	39.3931600047848	39.5180884517561	38.6652411885724	39.962918477317	39.7023922438163	37.7563995004641	37.3434471901439
17	35.9655221071628	39.0228947790473	39.9962526371282	37.2033652107821	39.3556573230555	36.0184613867656	38.9175696932119
18	37.5508272374209	35.4350052579076		37.9332776877411	36.6808736454563	36.2471146318018	37.3474593815
19	39.6145310905921	36.0892628417899		38.8557107815329	36.3813452157316	35.6301885648564	37.7360899201575
20	39.7270379634426	35.2171801486024		38.2014358454111	39.0915391614404	37.8678655947553	39.4821139305088
21		37.2729918209062		35.3994345605815	37.6454481907297	38.1857533111825	38.5417742406547
22		37.1605709623179		39.4382551719913	36.3160734776618	35.397645807723	35.6128416179612
23		36.5784219762513		36.9503943567888	35.547135933012	35.4361099060105	35.3175127333536
24		35.8974501863318		38.8633920651667	39.8946172111015	35.8141790192386	36.3377264963351
25		38.1881978067362		35.7473067811809	37.9307056002459	39.850489973161	38.6262620595289
26		37.4812158724338		36.3826387630611	35.521075713151	36.3135424200262	38.7243317693409
27				36.8439089762866	36.4501258108312	37.0099167911208	36.0953553142873
28				35.4764344777516	39.4383165849121	38.8818850720835	36.621623730845
29				39.8986105319615		36.204271881395	37.7025040996132
30				37.1190789187308		38.1199634023167	37.8948735819822
31				35.2396182935233		37.6268844392284	39.0269507626395
32						35.0640715665375	36.5312490094693
33						35.8646988957287	35.7302454465538
34						37.3870714169827	37.2359927716766
35						37.6252663216145	37.3294063964268
36						35.8885677334531	37.7671256824527
37						39.6658312515066	35.7209273900331
38						37.7254462985083	36.9473165264768
39						38.7993453213677	36.0842686442459
40						38.2285876793647	39.6082355020103