EPISTASIS AND ENTROPY

KRISTINA CRONA

ABSTRACT. Epistasis is a key concept in the theory of adaptation. Indicators of epistasis are of interest for large system where systematic fitness measurements may not be possible. Some recent approaches depend on information theory. We show that considering shared entropy for pairs of loci can be misleading. The reason is that shared entropy does not imply epistasis for the pair. This observation holds true also in the absence of higher order epistasis. We discuss a refined approach for identifying pairwise interactions using entropy.

Department of Mathematics and Statistics, American University, Washington, DC

1. INTRODUCTION

Epistasis tends to be prevalent for antimicrobial drug resistance mutations. Sign epistasis means that the sign of the effect of a mutation, whether good or bad, depends on background Weinreich et al. (2005). Sign epistasis may be important for treatment strategies, both for antibiotic resistance and HIV drug resistance (Goulart et al., 2013; Desper et al., 1999; Beerenwinkel et al., 2007 a). For instance, there are sometimes constraints on the order in which resistance mutations occur. A particular resistance mutation may only be selected for in the presence of another resistance mutation. It is important to identify such constraints. A first question is how one can identify pairwise epistasis in a large system. We will discuss entropy (Shannon, 1948) and epistasis. Information theory has been used for HIV drug resistance mutations (Gupta and Adami, 2015) and more extensively for analyzing human genetic disease (e.g. Dong et al., 2008; Kang et al., 2008; Streiloff et al., 2010). For recent review articles on epistasis and fitness landscapes see e.g. Hartl (2014); Kondrashov and Kondrashov (2014), and for an empirical perspective (Szendro et al., 2012).

2. RESULTS

It is well established that genotypes are expected to be in equilibrium proportions if there is no epistasis in the system, i.e., if fitness is multiplicative. For instance, if two rare mutations have frequencies p and q, then the frequency of the genotype combining the two mutations is expected to be close to pq. This statement holds true regardless if recombination occurs or not (Otto and Lenormand, 2002).

We will explore the relation between entropy and epistasis for a system with constraints as described in the introduction.

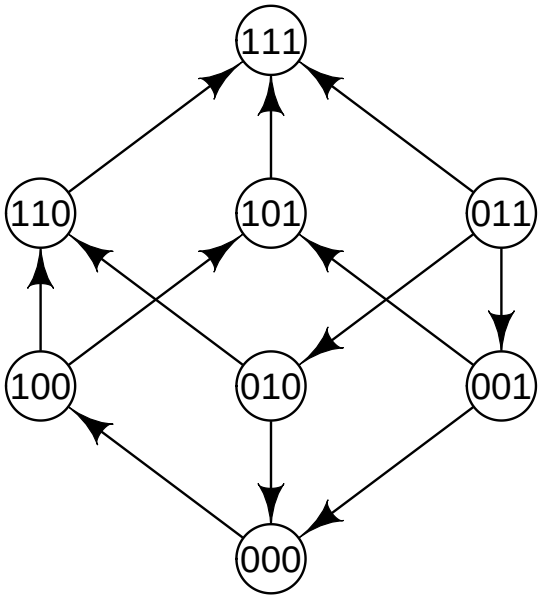

FIGURE 1. systems

Consider a 3-locus balletic system where a mutation at the first locus confers resistance, whereas mutations at the second and third loci are only selected for in the presence of the first mutation (otherwise they are deleterious). We represent the case with a fitness graph (Crona et al., 2013) (Figure 1). As conventional, 000 denotes the wild-type. For instance, one obtains a system with the fitness graph as in Figure 1 for the log-fitness values

$$w_{000} = 0, \quad w_{100} = 0.09531018, \quad w_{010} = -2.302585, \quad w_{001} = -2.302585,$$
$$w_{110} = 0.1906204, \quad w_{101} = 0.1906204, \quad w_{011} = -4.60517, \quad w_{111} = 0.2859305$$

The gene interactions for a 3-loci system can be described by the sign pattern of 20 circuits, or minimal dependence relations (Beerenwinkel et al., 2007 b). The relevant two-way interactions in this context be described by the six circuits corresponding to the faces of the 3-cube. Specifically,

$$
\begin{aligned}
w_{000} - w_{010} - w_{100} + w_{110} &> 0 \\
w_{001} - w_{011} - w_{101} + w_{111} &> 0 \\
w_{000} - w_{001} - w_{100} + w_{101} &> 0 \\
w_{010} - w_{011} - w_{110} + w_{111} &> 0 \\
w_{000} - w_{001} - w_{010} + w_{011} &= 0 \\
w_{100} - w_{101} - w_{110} + w_{111} &= 0
\end{aligned}
$$

The four inequalities express that there is positive epistasis for the first and second loci, as well as for the first and third loci. The two equalities show that there is no epistasis for the second and third loci, regardless of background. The total 3-way epistasis is zero as well,

$$w_{111} - w_{110} - w_{101} - w_{011} + w_{100} + w_{010} + w_{001} - w_{000} = 0.$$

Higher order gene interactions have also been described using Walsh coefficients (Weinreich et al., 2013; Poelwijk et al, 2015). For this landscape the Walsh coefficient $E_{011} = 0$, which indicates an absence of background averaged epistasis for the second and third loci.

We will consider entropy during the process of adaptation for this landscape. The starting point for adaptation is the wild-type 000. We use a standard Wright-Fisher model for an infinite population with mutation rate $\mu = 10^{-7}$. The gene frequencies and shared entropy after the given number of generations are listed in the table.

TABLE 1. Gene frequencies and shared entropy $I(2,3)$ for an infinite population with mutation rate 10^{-7}.

generations	000	100	010	001	110	101	011	111	I(2,3)
130	0.7692	0.1850	0	0	0.0214	0.0214	0	0.0031	0.003206041
140	0.4834	0.3015	0	0	0.0904	0.0904	0	0.0343	0.01736237
146	0.2723	0.3008,	0	0	0.1597	0.1597	0	0.1075	0.02335234
150	0.1569	0.2539	0	0	0.1974	0.1974	0	0.1944	0.0211462
160	0.0229	0.0959	0	0	0.1934	0.1934	0	0.4943	0.006950302
170	0.0020	0.0216	0	0	0.1132	0.1132	0	0.7501	0.001270666

The shared entropy for the second and third loci differs from zero. However, there is no 2-way epistasis for the pair of loci.

By extrapolation, consider an analogous system for L-loci. Then $L-1$ mutations are selected for only if the first mutation has occurred, but there are no other interactions. One would get non-zero shared entropy for $\binom{L}{2}$ pairs of loci, although there is 2-way epistasis for $L-1$ pairs of loci only.

2.1. **A pair with no epistasis and maximal shared entropy.** The landscape

$$w_{000} = 0, \quad w_{100} = 0 \quad w_{010} = -2.302585, \quad w_{001} = -2.302585,$$
$$w_{110} = 0.09531018, \quad w_{101} = 0.09531018 \quad w_{011} = -4.60517, \quad w_{111} = 0.1906204$$

is closely related to the previous example. Indeed, the two-way interactions can be described by the sign pattern

$$\begin{aligned} w_{000} - w_{010} - w_{100} + w_{110} &> 0 \\ w_{001} - w_{011} - w_{101} + w_{111} &> 0 \\ w_{000} - w_{001} - w_{100} + w_{101} &> 0 \\ w_{010} - w_{011} - w_{110} + w_{111} &> 0 \\ w_{000} - w_{001} - w_{010} + w_{011} &= 0 \\ w_{100} - w_{101} - w_{110} + w_{111} &= 0 \end{aligned}$$

and the total 3-way epistasis is zero:

$$w_{111} - w_{110} - w_{101} - w_{011} + w_{100} + w_{010} + w_{001} - w_{000} = 0.$$

Also in this case, there is no epistasis for the second and third loci. Mutations at the second and third loci are selected for only in the presence of a mutation at the first locus. However, this fitness landscape differs from the previous example in that a mutation at the first locus is neutral for the wild-type.

Suppose that 50 percent of hosts start a new treatment with 000 viruses, and 50 percent start with the 100 genotype. That could be realistic, for instance if the 100 genotype had some resistance to a previously used drug. By assumption, eventually one would have about 50 percent 000 genotypes and 50 percent 111 genotype in the total population. Then $I(2,3) = 2$ although there is no epistasis for the second and third loci. This example also points at a fundamental problem relating pairwise epistasis and entropy. At the time when we have 50 percent 000 genotypes and 50 percent 111 genotypes, obviously no method can reveal pairwise epistasis.

2.2. **A refined approach.** We will discuss a refined approach for identifying pairwise epistasis. Suppose that we have identified shared entropy for a particular pair of loci $\{k, l\}$. Let $S_1^{k,l}$ denote the set of loci such that the shared entropy

$$I(k : i) \neq 0 \text{ or } I(l : i) \neq 0.$$

Let $S_2^{k,l}$ denote the set of loci with non-zero shared entropy for some locus in S_1, and so forth. Let $S^{k,l} = \bigcup S_i \setminus \{k, l\}$.

Let v denote one of the $2^{|S|}$ possible states for S, and consider the subsystem of genotypes determined by v. If the shared entropy $I^v(k : l) = 0$ for all v, then there is no indication of of epistasis for $\{l, k\}$.

We can apply the refined approach for the second and third loci in our example where $I(2,3) = 2$. Then

$$S = \{1\}, \quad I^{(0)}(2:3) = I^{(1)}(2:3) = 0.$$

Consequently, there is no indication of epistasis for the second and third loci.

The described method could be useful for identifying cases with shared entropy and no epistasis. However, it remains to explore to what extent the method is useful in a more general setting.

3. DISCUSSION

We have demonstrated that shared entropy for two loci does not imply epistasis for the pair. This observation holds true also in the absence of 3-way epistasis in a single environment. Entropy based approaches to epistasis are coarse. We have discussed a refined approach which filters out some cases where shared entropy depends on states at other loci.

There are obviously other reasons for caution in interpretations of entropy for drug resistance mutations. Different drugs constitute different environments. Some resistance mutations may be correlated if they are beneficial in the presence of a particular drug, but not for other drugs. In such cases entropy would not not imply epistasis.

Our results show that observations on entropy and epistasis based on 2-locus systems can be misleading for general systems. From a theoretical point of view, a better understanding of large systems would be useful for handling drug resistance data.

4. METHODS

Let x and y be discrete random variables with states x_1, x_2 and y_1, y_2. Let p_i denote the frequency of x_i, and p_{ij} the frequency for the combination of x_i and y_j. The entropy (Shannon, 1948) $H(x)$ and the joint entropy $H(x,y)$ are defined as

$$\begin{aligned} H(x) &= -\,p_1 \log(p_1) - (1-p_1)\log(1-p_1) \\ H(x,y) &= -\,p_{11}\log p_{11} - p_{12}\log\left(p_{12}\right) \\ &\quad -\,p_{21}\log p_{21} - p_{22}\log\left(p_{22}\right) \end{aligned}$$

The shared entropy is the quantity $I(x:y) = H(x) + H(y) - H(x,y)$.

In general $I(x:y) \geq 0$, and the shared entropy is a measure of dependence.

REFERENCES

Beerenwinkel, N., Eriksson, N. and Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli*; 13:893–909.

Beerenwinkel, N., Pachter, L. and Sturmfels, B. (2007). Epistasis and shapes of fitness landscapes. *Statistica Sinica* 17:1317–1342.

Crona, K., Greene, D. and Barlow, M. (2013). The peaks and geometry of fitness landscapes. *J. Theor. Biol.* 317: 1–13.

Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H. and Schäffer, A.A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Comput. Biol* 6 37–51.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L. (2008). Exploration of gene-gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* 16: 229-235

Goulart, C. P., Mentar, M., Crona, K., Jacobs, S. J., Kallmann, M., Hall, B. G., Greene D. and Barlow M. (2013). Designing antibiotic cycling strategies by determining and understanding local adaptive landscapes. *PLoS ONE* 8(2): e56040. doi:10.1371/journal.pone.0056040.

Gupta, A. and Adami, C. (2015). Changes in epistatic interactions in the long-term evolution of HIV-1 protease. `arXiv:1408.2761`.

Hartl, D. (2014) What can we learn from fitness landscapes? *Current Opinion in Microbiology* 21 (2014): 51-57.

Kang, G., Yue, W., Zhang, J, Cui,Y., Zuo, Y., Zhang, D. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases, *J. Theor. Biol.* 250

Kondrashov, D. A. and Kondrashov , F. A. (2014). Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* vol. 27, pp. 379-423 and 623-656, July and October, 1948.

Otto, S. P. and Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nature Reviews Genetics*; 3:252-261.

Poelwijk, F., Krishna, V. and Ranganathan, R. (2015). The context-dependence of mutations: a linkage of formalisms. arXiv:1502.00726 (2015).

Strelioff, C. C., Lenski R. E. and Ofria, C. (2010) *J. Theor. Biol.* 266 (4), pp. 584-594

Szendro, I. G., Schenk, M. F., Franke, J. Krug, J. and de Visser J. A. G. M. (2013). Quantitative analyses of empirical fitness landscapes *J. Stat. Mech. P01005.*

De Visser, J. A. G. M, and Krug, J. (2014) Empirical fitness landscapes and the predictability of evolution." *Nature Reviews Genetics* 15.7 (2014): 480-490.

Weinreich D. M., Lan, Y., Wily, C. S., Heckendorn, R. B. (2013) . Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Gen Dev* 23: 700-7.

Weinreich, D. M, Watson, R. A. and Chao, L. (2005) Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution.*, 9(6) pp 1165-74.