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.

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

1. INTRODUCTION AND RESULTS

Epistasis is typically 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. Sign epistasis may be important for treatment strategies, both for antibiotic resistance (Goulart et al., 2013) and HIV drug resistance (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. We will investigate the relation between entropy (Shannon, 1948) and epistasis for such a system.

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). As conventional, 000 denotes the wild-type. For instance, consider the log-fitness values

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

The two-way interactions can be described by the following sign pattern.

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

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. The total 3-way epistasis is zero as well,

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

For more background on gene interactions, see Beerenwinkel et al. (2007 b). We will compare epistasis and entropy (e.g Streiloff et al., 2010; Gupta and Adami, 2015). The starting point for adaptation is 000, and we use standard assumptions for the evolutionary process in an infinite population. The genotypes $010, 001$ and 011 being rare, we approximate their proportions to zero. By assumption, we expect equilibrium proportions for the subsystem $100, 110, 101, 111$. Indeed, there is no epistasis for this subsystem.

Let p_{1**} be the proportion of genotypes with a substitution at the first locus, and p_{11*} the proportion with mutations at both the first and second loci at some point in time. By assumption, the expected proportion of each genotype is

$000 : 1 - p_{1**}, \quad 100 : p_{1**}\,(1 - p_{11*})^2 \quad 110 : p_{1**}\,p_{11*}(1 - p_{11*}) \quad 101 : p_{1**}p_{11*}(1 - p_{11*}),$
$111 : p_{1**}\,{p_{11*}}^2$

For instance, if $p_{1**} = 0.1$ and $p_{11*} = 0.1$ then the proportions are

$000 : 0.9, \quad 100 : 0.081, \quad 110 : 0.009, \quad 101 : 0.009, \quad 111 : 0.001.$

Consider the entropies for the second and third loci, $H(2)$ and $H(3)$, the joint entropy $H(2, 3)$, and the shared information H(2)+H(3)-H(2,3) (see the method section).

$$\begin{aligned} H(2) = H(3) =& -0.99 \log 0.09 - 0.01 \log 0.01 = 0.08079314 \\ H(2,3) =& -0.981 \log 0.981 - 0.009 \log 0.009 \\ & - 0.009 \log 0.009 - 0.001 \log 0.001 \\ =& 0.1563824 \\ H(2) + H(3) - H(2,3) =& 0.002145886 \end{aligned}$$

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. We would find shared entropy for $\binom{L}{2}$ pairs of loci, although there is 2-way epistasis $L - 1$ pairs of loci only.

2. 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 (as defined here) in a single environment. There are obviously other reasons for caution in interpretations of entropy. 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.

Entropy based approaches to epistasis are coarse. 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 is important for handling drug resistance data.

3. 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 $\quad H(x, y) - H(x) - H(y).$

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.

Goulart, C. P., Mentar, M., Crona, K., Jacobs, S. J., Kallmann, M., Hall, B. G., Greene D., 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.

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.

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

Weinreich, D.M, Watson, Chao. (2005) Sign epistasis