JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 51 (3): 365–369 (2012) doi: 10.1111/j.1759-6831.2012.00217.x

Research Article

A note on gene pleiotropy estimation from phylogenetic analysis of protein sequences

[1,2]Wen-Hai CHEN [1]Zhi-Xi SU* [1,3]Xun GU*

[1](*School of Life Sciences and Center for Evolutionary Biology, Fudan University,* Shanghai 200433, China)
[2](*College of Mathematics & Information Science, Wenzhou University,* Wenzhou, Zhejiang 325035, China)
[3](*Department of Genetics, Development and Cell Biology Iowa State University,* Ames, IA 50011, USA)

Abstract Recently, several statistical methods have been independently proposed for estimating the degree (n) of gene pleiotropy (i.e. the capacity of a gene to affect many phenotypes) without knowing measurable phenotypic traits. However, the theoretical limitation of these approaches has not been well demonstrated. In this short note, we show that our previous method based on the phylogeny of protein sequences is, in fact, an effective estimate of a parameter that can be written symbolically as $K = \min(n,r)$, where r is the rank of mutations at an amino acid site. Hence, understanding of r is crucial for appropriate interpretation of the estimated K, denoted by K_e (the effective gene pleiotropy). Indeed, when protein sequence alignment is used to estimate effective gene pleiotropy (K_e) by this method, K_e can be interpreted as an effective estimate of n when $n \leq 20$, as long as the phylogeny is sufficiently large. If $n > 20$, $K_e \to 20$, although the true n could be much higher.

Key words gene pleiotropy, molecular phenotypes, phylogenetic analysis, protein sequence.

Although the concept of gene pleiotropy (i.e. the ability of a gene to affect multiple phenotypes) has been generally accepted, a number of long-standing controversies remain unsolved (Fisher, 1930; Wright, 1968; Keightley, 1994; Hartl & Taubes, 1996, 1998; Waxman & Peck, 1998; Lynch et al., 1999; Bataillon, 2000; Poon & Otto, 2000; Wagner, 2000; Elena & Lenski, 2003; Welch & Waxman, 2003; Otto, 2004; Eyre-Walker et al., 2006). One of the most important issues is how to estimate the degree of pleiotropy. Although the description of pleiotropic effects is biologically straightforward, estimation of the degree of pleiotropy is phenotype specific and technology dependent (Dudley et al., 2005; Ohya et al., 2005; Pal et al., 2006; Cooper et al., 2007; Wagner et al., 2008). Consequently, when experimentally measurable phenotypes are used (Dudley et al., 2005; Ohya et al., 2005; Cooper et al., 2007; Wagner et al., 2008), the relationship between the number of phenotypic traits measured and the degree of pleiotropy is not always clear.

Nevertheless, from statistical genetics (Lande, 1980; Turelli, 1985; Wagner, 1989; Barton, 1990; Zhang & Hill, 2003; Martin & Lenormand, 2006), the degree of gene pleiotropy can be defined theoretically as the number (n) of distinct phenotypic traits affected by the gene, each of which corresponds to single non-trivial fitness component (hereafter referred to as "molecular phenotypes"). Based on phylogenetic analysis of protein sequences, Gu (2007a) developed a statistical method to effectively estimate the degree of gene pleiotropy (K), denoted by K_e. Follow-up studies showed that K_e is typically around 6–7 (Su et al., 2010, Zeng & Gu, 2010). However, how to interpret this estimate (K_e) appropriately in the context of development, cell biology, and evolution remains contentious (Gu, 2007b; Zeng & Gu, 2010; Wagner & Zhang, 2011). In the theory formulated by Gu (2007a), the claim $K = n$ holds by default. It has been realized that this claim actually assumes a full rank of mutational effects on n phenotypes. If this is not the case, the estimate of gene pleiotropy would be much less than n, as shown by Chevin et al. (2010) and Martin & Lenormand (2006). In this short note, we address this issue and show that the method of Gu (2007a) for the estimation of gene pleiotropy is actually for $K = \min(n,r)$, where r is the rank of mutational effects.

1 Material and methods

Homologous protein sequences from eight vertebrates (human, mouse, dog, cow, chicken, *Xenopus*, fugu, and zebrafish) were obtained from Ensembl EnsMart (http://www.ensembl.org/Multi/martview, accessed June 2008). For those genes that have multiple protein sequences, the longest protein sequence was

Received: 5 April 2012 Accepted: 4 June 2012
* Author for correspondence. X Gu, E-mail: xgu@iastate.edu. ZX Su, zxsu@fudan.edu.cn. Tel.: 86-21-65642507. Fax: 86-21-65642507.

used for further analysis. Because there are several annotated homology relationships between human and other species in Ensembl, we only considered pairs of genes annotated as "Unique Best Reciprocal Hit" (UBRH), meaning that they were unique reciprocal best hits in all-against-all BlastZ searches, to be orthologous. For each set of orthologous genes, the number of synonymous substitutions per synonymous site (dS) and the number of non-synonymous substitutions per non-synonymous site (dN) between human and mouse orthologs were retrieved from Ensembl EnsMart. We then used the method of Gu (2007a) to estimate the effective gene pleiotropy, K_e.

2 Results and Discussion

Under the multidimension model of phenotypic evolution (e.g. Fisher, 1930; Wagner, 1989; Martin & Lenormand, 2006), the dimensionality (n) is the actual number of orthogonally selected molecular phenotypes, rather than an arbitrary number of measurable phenotypic traits. Here we restrict the phenotype dimensionality (n) with respect to a given gene. That is, n is the dimension of molecular phenotypes of a gene, all of which are under non-trivial stabilizing selections. There is no doubt that the number of molecular phenotypes (n) is a natural measure for the degree of gene pleiotropy. The central issue here is whether we can estimate n based on mutational variation (Martin & Lenormand, 2006; Chevin et al., 2010) or protein sequence divergence (Gu, 2007a).

Martin & Lenormand (2006) used the distribution of mutational effects on fitness from mutation accumulation experiments to estimate n and obtained a very low estimate (somewhere between 1 and 2) of n. Later, Chevin et al. (2010) showed that the result of Martin & Lenormand (2006) probably reflected the rank or dimensionality (r) of mutational effects rather than that of phenotypes. That is, the rank of a single mutation is virtually 1 ($r = 1$), although it may affect as many as all n phenotypes. Because the method of Gu (2007a) was based on multiple sequence alignment rather than single mutations, the problem is not so obvious. We have solved this problem as outlined below.

2.1 Gu's method estimates $K = \min(n,r)$

We refer to a class of phenotype–genotype models as Fisher-related models, including Fisher's (1930) original geometric models and various multivariate models (e.g. Lande, 1980; Turelli, 1985; Waxman & Peck, 1998). In the abstract sense, multiple functions of a gene, or pleiotropy, can be represented by n distinct components in the fitness, called molecular phenotypes. These n molecular phenotypes can be viewed as an n dimensional space in Fisher's model. Random mutations of the gene generate a mutational distribution for n molecular phenotypes. Theoretically, for molecular phenotypes denoted by $(y_1, \ldots, y_n)$, each y_i represents a non-trivial component of genetic variation in organismal fitness as a result of a specified (yet unknown) biological process. Molecular phenotypes may represent things ranging from subcomponents of protein function to distinct physiological processes in various tissues. Because these underlying biological processes are usually intractable, the concept of molecular phenotypes may avoid this difficulty.

To be concise in the following analysis of gene pleiotropy we used a simple stabilizing selection model. It is common to assume a Gaussian-like fitness function for n molecular phenotypes $y = (y_1, \ldots, y_n)'$ of a gene; that is:

$$w(y) = \exp\left[-\frac{y'\sum_w^{-1} y}{2}\right] \qquad (1)$$

where $\sum_w$ is a (positive definite) symmetric matrix characterizing the correlated stabilizing selections on n molecular phenotypes. The rank of selection matrix $\sum_w$ is n (full-rank), indicating that its n eigenvalues are all non-zero positive. The ith diagonal element $\sigma^2_{w,i}$ measures the strength of stabilizing selection on the ith molecular phenotype, whereas the ijth non-diagonal element $\sigma_{w,ij}$ measures the correlated stabilizing selection on y_i and y_j. When $n = 1$, Eq. (1) can be reduced to the simplest form of $w(y) = e^{-y^2/2\sigma_w^2}$.

Under the stabilizing model (Turelli, 1985; Waxman & Peck, 1998), the coefficient of selection for the molecular phenotype y can be generally defined as $\rho(y) = w(y) - 1$, reflecting the stabilizing (purifying) selection against deleterious mutations leading to a deviation from the optimum. A consequence of this model is that sequence evolution is dominated by the fixation of very slightly deleterious mutations (Kimura, 1983). In the following, the selection intensity $S(y)$, defined as:

$$S(y) = 4N_e\rho(y) \approx -2N_e\left(y'\sum_w^{-1} y\right), \qquad (2)$$

is more frequently used, where N_e is the effective population size.

A fundamental problem is how the degree of gene pleiotropy may affect the evolutionary rate of protein sequences. To derive the relationship between gene pleiotropy n and the evolutionary rate (λ), we first briefly

introduce the standard theory of molecular evolution (Kimura, 1983), which claims that the rate (λ) of protein evolution is given by the following:

$$\lambda = v\frac{S}{1-e^{-S}} \tag{3}$$

where v is the mutation rate, $S\ (=4N_e\rho)$ is the selection intensity, and ρ is the coefficient of selection. Hence, $\lambda = v$ when $S = 0$ (neutral selection), $\lambda < v$ when $S < 0$ (purifying selection), and $\lambda > v$ when $S > 0$ (adaptive selection). The pleiotropy model of molecular evolution (Gu, 2007a) suggests that the selection intensity of a gene is defined by the means of molecular phenotypes, as shown in Eq. (2). Therefore, the evolutionary rate of a mutant affecting the molecular phenotypes (y) can be written as follows:

$$\lambda(y) = v\frac{S(y)}{1-e^{-S(y)}} \tag{4}$$

Thus, given $p(y)$, the distribution of mutational effects that generates a variation of y, the (mean) evolutionary rate of a gene is given by:

$$\bar{\lambda} = v\int \frac{S(y)}{1-e^{-S(y)}}p(y)dy \tag{5}$$

The distribution $p(y)$ characterizes the effects of mutations on molecular phenotypes. Random mutations at an amino acid site of a gene may affect molecular phenotypes in a correlated fashion, which can be modeled by a multivariate normal distribution with the covariance matrix $\sum_m$. It should noted that the rank (r) of the covariance matrix $\sum_m$ is unknown, which could be less than, that is, $r \le n$. In the derivation of Gu (2007a), an assumption of $n = r$ was implied. Although this did not affect the method of estimation, as shown below, the biological interpretation needs to be revised.

We consider any kth moment of the evolutionary rate of a gene, defined as $\bar{\lambda}^k = \int \lambda(y)^k p(y)dy$. As a close approximation, Gu (2007a) showed that $\lambda/v \approx e^{-|S|}(1 + c\,|S|)$ for any y with $S(y) < 0$, where $c = 0.5772$ and $|S|$ is the absolute value of S. Note that $|S(y)| = 2N_e y' \sum_w^{-1} y$ is a quadratic form of y. Hence, for $k = 1, 2, \ldots$, we have:

$$\bar{\lambda}^k \approx v^k \int e^{-k|S(y)|}[1 + c|S(y)|]^k\ p(y)dy \tag{6}$$

After some calculations, one can show that the analytical solution of Eq. (6) depends on a series of matrix calculations (traces and determinants) generated from a series of matrices with the forms of $[\sum_w^{-1}]^i\left[\sum_m\right]^j$, for any $i,j = 1,2,\ldots, k$. According to matrix theory, and based on the assumptions that the rank of matrix Σ_w is n and the rank of matrix Σ_m is r, we claim that the rank of any matrix $[\Sigma_w^{-1}]^i\,[\Sigma_m]^j$ is $K = \min(n,r)$. It follows that any kth moment of the rate ($\bar{\lambda}^k/v^k$) depends on $K = \min(n,r)$ rather than n and r individually. For instance, the mean evolutionary rate of a gene is given by:

$$\bar{\lambda} = v\left[\prod_{i=1}^{\min(n,r)}(1+2B_i)\right]\left[1+\prod_{i=1}^{\min(n,r)}\frac{cB_i}{1+2B_i}\right] \tag{7}$$

where $B_i = 2N_e\alpha_i$ and values of α_i are non-zero positive eigenvalues of A, the matrix of $[\Sigma_w^{-1}]^i\,[\Sigma_m]^j$ when $i = j = 1$. We thus show that the parameter $K = \min(n,r)$ can be estimated by the method of Gu (2007a).

2.2 Determination of mutational rank (r)

Because K defined by $\min(n,r)$ measures the true degree of gene pleiotropy only when $n \le r$, knowing the mutational rank (r) is important for having a correct biological interpretation of the method of Gu (2007a). This means that the method of Gu (2007a) for gene pleiotropy is semiquantitative: the effective gene pleiotropy K_e is an effective estimate of n only when $n < r$. Otherwise (i.e. $n > r$) we can only claim that K_e is a low-boundary of the real gene pleiotropy. Similar to Chevin et al. (2010), we used computer simulation to illustrate this result. The degree of gene pleiotropy (n) is chosen from values 2 to 30, in increments of 2. For a fixed n, we performed the simulation according to the stabilizing selection model and the normal distribution of mutational effects with the rank $r = 10$. Figure 1 shows the estimated effective gene pleiotropy (K_e) increases with increasing values of n when $n < 10$. As expected, when $n > 10$, the estimated K_e becomes quickly saturated towards the line of $r = 10$, regardless of increasing n.

The crucial question for the method of Gu (2007a) is how we can determine r for a multiple sequence alignment (MSA) along a molecular phylogeny. This is a complicated issue that will be addressed in a subsequent paper. Here we discuss the matter briefly. First, for protein sequence alignment, each position can occupy potentially 20 amino acids. Hence, the maximum rank is $r_{max} = 20$ under the assumption that each type of amino acid has the rank of 1, regardless how many molecular phenotypes it may affect. Second, the related molecular phylogeny must be large so that each position has sufficient time to experience mutations towards all types of amino acids.

In fact, our empirical analysis is compatible to these claims. First we consider the dataset of Su et al. (2010). The histogram of effective gene pleiotropy (K_e) for 321

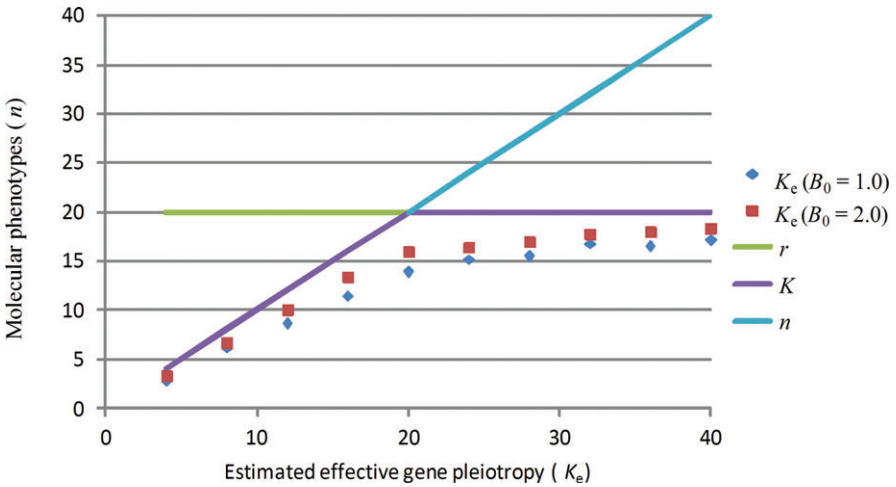

Fig. 1. Simulation study for the estimated effective gene pleiotropy (K_e) when the number of functional units of a gene is set to be $r = 10$. See text for details. We set the baseline selection intensity B_0 around 1.0–2.0, as suggested by the dataset of Su et al. (2010).

vertebrate genes clearly shows that most genes have an estimate of K_e that is <20 (Fig. 1). Indeed, the range (95% quantile) of K_e is approximately from 2 to 14, with a peak around 6–7. In particular, at the cut-off of the 99% quantile, the K_e in the data of Su et al. (2010) is 18.3, whereas the maximum value is 20.4. Recently, we have estimated the effective gene pleiotropy (K_e) for 600 single-copy genes based on the protein sequence alignment from eight completely sequenced yeast genomes, as well as 900 single-copy genes based on the protein sequence alignment from 12 completely sequenced fruitfly genomes. Consistently, in these two new datasets, the cut-off of the 99% quantile of K_e is around 20. The similarity of three datasets is impressive, suggesting that the mutational rank r largely relies on the type of genetic variation/divergence. Because genes with K_e close to 19 may have a much higher true degree of pleiotropy (n), the proportion of very highly pleiotropic genes ($n > 19$) can be evaluated by examining whether there is peak for $K_e \geq 19$ in the histogram of K_e. We did not find such a peak in any of the three datasets, indicating that the number of very highly pleiotropic genes ($n > 19$) in the genome is quite small (Table 1).

Table 1 Effective gene pleiotropy restricted by the theoretical value $K = \min(n,r)$, where the maximum value of r is close to 20, as expected for the protein sequences

Datasets	No. species	No. genes	K_e cut-off	
			95% quantile	99% quantile
Vertebrates	8	321	12.4	18.8
Drosophila	5	437	13.1	20.2
Yeasts	12	580	12.6	19.4

3 Concluding remarks

Based on the statistical model of phenotype–genotype mapping, the degree of gene pleiotropy (n; i.e. the capacity of a gene to affect many phenotypes) can be estimated without knowing measurable phenotypic traits (Martin & Lenormand, 2006; Gu, 2007a), but the effectiveness of these methods may be limited by the rank of mutational effects. When protein sequence alignment is used to estimate the effective gene pleiotropy (K_e) according to the method of Gu (2007a), K_e can be interpreted as an effective estimate of n when $n \leq 20$, as long as the phylogeny is sufficiently large. If $n > 20$, $K_e \to 20$, although the true n could be much higher.

Acknowledgements This work was supported financially, in part, by Fudan University.

References

Barton NH. 1990. Pleiotropic models of quantitative variation. Genetics 124: 773–782.

Bataillon T. 2000. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? Heredity 84: 497–501.

Chevin LM, Martin G, Lenormand T. 2010. Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. Evolution 64: 3213–3231.

Cooper TF, Ostrowski EA, Travisano M. 2007. A negative relationship between mutation pleiotropy and fitness effect in yeast. Evolution 61: 1495–1499.

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. Molecular Systems Biology 1: 2005.0001.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature Reviews Genetics 4: 457–469.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Oxford University Press.

Gu X. 2007a. Evolutionary framework for protein sequence evolution and gene pleiotropy. Genetics 175: 1813–1822.

Gu X. 2007b. Stabilizing selection of protein function and distribution of selection coefficient among sites. Genetica 130: 93–97.

Hartl DL, Taubes CH. 1996. Compensatory nearly neutral mutations: selection without adaptation. Journal of Theoretical Biology 182: 303–309.

Hartl DL, Taubes CH. 1998. Towards a theory of evolutionary adaptation. Genetica 102–103: 525–533.

Keightley PD. 1994. The distribution of mutation effects on viability in *Drosophila melanogaster*. Genetics 138: 1315–1322.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Lande R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. Genetics 94: 203–215.

Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, Willis JH. 1999. Perspective: spontaneous deleterious mutation. Evolution 53: 645–663.

Martin G, Lenormand T. 2006. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60: 893–907.

Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latge JP, Francois JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S. 2005. High-dimensional and large-scale phenotyping of yeast mutants. Proceedings of the National Academy of Sciences USA 102: 19 015–19 020.

Otto SP. 2004. Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proceedings of the Royal Society B: Biological Sciences 271: 705–714.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nature Reviews Genetics 7: 337–348.

Poon A, Otto SP. 2000. Compensating for our load of mutations: freezing the meltdown of small populations. Evolution 54: 1467–1479.

Su Z, Zeng Y, Gu X. 2010. A preliminary analysis of gene pleiotropy estimated from protein sequences. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 314: 115–122.

Turelli M. 1985. Effects of pleiotropy on predictions concerning mutation-selection balance for polygenic traits. Genetics 111: 165–195.

Wagner A. 2000. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. Genetics 154: 1389–1401.

Wagner GP. 1989. Multivariate mutation–selection balance with constrained pleiotropic effects. Genetics 122: 223–234.

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the 'cost of complexity'. Nature 452: 470–472.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nature Reviews Genetics 12: 204–213.

Waxman D, Peck JR. 1998. Pleiotropy and the preservation of perfection. Science 279: 1210–1213.

Welch JJ, Waxman D. 2003. Modularity and the cost of complexity. Evolution 57: 1723–1734.

Wright S. 1968. Evolution and the genetics of populations. Chicago: University of Chicago Press.

Zeng YW, Gu X. 2010. Genome factor and gene pleiotropy hypotheses in protein evolution. Biology Direct 5: 37.

Zhang XS, Hill WG. 2003. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. Evolution 57: 1761–1775.