SIMULATING GENOTYPIC MERIT WITH HIGH-ORDER EPISTATIC INTERACTIONS

A.B. Kinghorn[1] and B.P. Kinghorn[2]

[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong
[2]School or Environmental and Rural Science, University of New England, Armidale, 2351 Australia

SUMMARY

The real map from genotype to phenotype is very complex indeed, and yet we use simple models to analyse it and simple models to simulate it. This paper illustrates a method to simulate phenotypes as a function of genotypes that aims to better emulate the underlying complexity involved, with multi-level epistatic interaction among all loci within large groups of loci. It is proposed that such simulated data will give a more realistic basis to test QTL detection, GWAS and genetic evaluation methods.

INTRODUCTION

We want to understand and exploit the relationship between genotype and phenotype. To do this we use simple models and methods that we hope will lead us to making good decisions. However, life is more complex than we can perceive, as it has not been designed, but has evolved in a random manner. How can we test the usefulness of these simple models? They might lead to what seems like good genetic progress, but do they miss something in the real complexity that alternative models and methods might capture for our benefit? In addition, our simple models often lead us to think that there are many hundreds of QTL affecting a trait, with relatively few QTL of large effect – could reality be that there are far fewer QTL that, because of their complex interactions, masquerade as many hundreds of QTL? If this were true and detectable, then we might take a different direction in QTL detection, GWAS analyses and genetic evaluations. Simulation can be used to test this. However, datasets that are simulated using the same or similar statistical models as will be used to analyse them are self-fulfilling and not appropriate. And of course, the real model is too complex for us to know and use. Instead we need a tractable approach that emulates the high complexity of true genotype-phenotype relationships, including the high-order epistatic interactions that are evident when gazing at a biochemical pathway chart.

The NK model (Kauffman and Levin 1987) is a theoretical fitness model that provides an objective function relating a sequence (genotype) to fitness score (phenotype). Each locus interacts with a given number of other loci that are either neighbours or randomly determined. Each locus is given an individual fitness score based on the loci with which it interacts. The individual loci fitness scores are summed to give a sequence's total fitness. This model is useful in that an NK fitness landscape's complexity can be tuned by altering the number of interactions at each locus. Cooper and Podlich added an extra layer of interaction to the standard NK model by introducing the concept of environmental dependent gene expression to simulate gene-by-environment interactions (Cooper and Podlich 2002). Although the *NK* model is useful, it has limitations in representing some biological systems. At higher interaction values that are biologically relevant the landscape descends into a chaotic surface on which additive adaptation is essentially not possible.

To solve this problem, Kinghorn and Tanner proposed an approach where the effects of groups of interacting loci ("phenotypic contributors") are added sequentially and in accordance with natural selection (Kinghorn and Tanner 2017), similarly to how gene networks probably evolved over time (Amoutzias *et al*. 2004). This approach is based on method for simulating the response surface of ligand/target molecule affinity as a function of DNA aptamer sequence (Kinghorn and Tanner 2017). We have used a similar approach to model SNP data from many genomes.

MATERIALS AND METHODS

The Selective Phenome Growth Adapted NK Model (SPANK) method of Kinghorn and Tanner (2017) operates on single DNA sequences (DNA Aptamers, typically 30 to 100 bases long). Our method follows the SPANK method quite closely, presented here briefly, in our context:

- N is the number of QTL
- PC_i is the i^{th} Phenotypic Contributor, this being a vector of indicator variables {0,1} that point to loci involved in generating value for that *PC*. A key concept is that the genotypic merit for a haploid is the sum of many *PC*s – many components of genetic merit that contribute to expression of phenotype.
- $\varphi_{i,s}$ is the value of PC_i for sequence or haplotype s.
- n_{PC} is the number of *PC*s. This is unbounded.
- K is the maximum number of loci that can be involved in determining a *PC*. All levels from 1 to K can be involved, but only one level per *PC*.
- k_i This is the actual number of loci involved in determining PC_i.

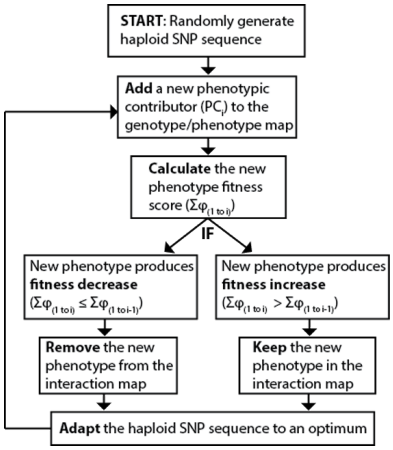

Figure 1. The SPANK method

There are three main parts to the method:

1. Generating the Genotype/Phenotype map (Figures 1, 2).
2. Analysing the SPANK Genotype/Phenotype map and comparing it to a randomly generated interaction map. To analyse the fitness landscapes we find 100,000 local optima and calculate their Hamming distance from the highest scoring optimum (Figure 3). The parameters used to drive the method can be changed to arrive at what is judged to be an appropriate fitness landscape, as indicated by such analysis.
3. For an implementation phase, the adopted Genotype/Phenotype map is used to generate phenotypes for the genotypes that are simulated into a real or simulated population.

The method follows Figure 1. A single haploid sequence is generated. This is the current Lead Sequence, which will direct the genotype/phenotype map evolution. The phenotypic contributors that make up the genotype/phenotype map will be formed around this lead sequence such that the lead sequence will be an optimum. To add a new phenotype to the interaction map, k_i is uniformly sampled from {1 to K}, and k_i loci randomly sampled from {1 to N}. The Lead Sequence alleles at these loci are used to determine its φ_i value, which is taken from a matrix of previously randomly generated φ_i values. For the PC_i to be accepted there must be an increase in the average merit across all prior *PC*s. The fitness score of the new phenotype ($\Sigma_{(1..i)}$) is then calculated and if it is greater than the fitness score of the old phenotype ($\Sigma\varphi_{(1\text{-}(i\text{-}1))}$) then the new PC_i is accepted, else it is rejected. Finally, noting that the lead sequence does not necessarily have the best genotype for a newly added *PC*, it is adapted using allelic substitution until it is at a fitness peak before the cycle is repeated. This allelic substitution proceeds by selecting the fittest 1-step mutant neighbour in sequence space and continuing the allelic substitution until no fitter 1-step mutant neighbours can be found.

RESULTS

To make a small illustrative example, the SPANK method was invoked with N=20 loci, n_{PC}=20 phenotypic contributors, and K= a maximum of 10 loci interacting to generate a *PC*. The resultant epistatic map is shown in Figure 2A. For comparison, an epistatic map was generated with random interactions that were not selected using a classical NK model, with 6 interactions per phenotypic contributor. In Table 1, a selection of the 20 phenotypic contributors (1, 2, 3 and 20) from the epistatic

map (Figure 2) are shown. The interacting loci and their φ value for two test sequences are displayed. φ values were previously generated constants that are functions of the alleles mapped by the PC. The φ values for each *PC* are averaged to give the haploid fitness for each of the two sequences.

Our aim is to develop fitness landscapes that are of high order complexity, yet are reasonably smooth and not chaotic, to the extent that a practitioner might expect in real populations. The measure of landscape smoothness we have chosen is the Hamming distance from the fittest optimum (Figure 3). For each landscape, 100,000 sequences are chosen at random and from these sequences random mutational walks uphill are taken until a local optimum is reached for each starting sequence. The Hamming distance from each of these local optima to the fittest recorded optimum is calculated. It can be observed that for the SPANK generated fitness landscape (Figure 3A) there is a stronger relationship between fitness and distance from the fittest optimum. Additionally, the line of best fit shows that for the SPANK generated fitness landscape a greater number of random uphill walkers reach the fittest optimum, indicating a smoother landscape. For the random landscape (Figure 3B) there is a much weaker trend of having higher scoring optima closer to the fittest optimum. For the random landscape just 2709 random walkers reached the fittest optimum, whereas 16,328 random walkers reached the fittest optimum for the SPANK landscape.

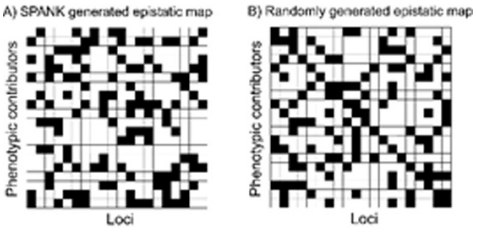

Figure 2. Epistatic maps generated A) by SPANK and B) at random without optimisation. The random map is a classical NK model with 6 interactions per phenotypic contributor. Each row is a phenotypic contributor and each column is a locus. Along each phenotypic contributor the interacting loci are denoted as dark shaded squares

DISCUSSION

As with the simple models we currently use to detect and exploit genotype-phenotype relationships, SPANK is not a model of the underlying biology. However, it does make a big step in the direction of emulating the biological complexity involved – permitting the involvement of multiple players (multiple loci and alleles) in each contribution of genotype to phenotype.

PC	k_{PC}	TS	Phenotypic contributors (PC) or Allele (TS)	φ
1	6		**01000101011000000001**	
		1	1**1**221**21**2**22**122111221**1**	f(122221) = 0.7843
		2	2**1**221**12**11**12**211222221**2**	f(111122) = 0.8342
2	9		**10000011001011011010**	
		1	**1**12212**1**222**1**2**2**11**2**21	f(112222112) = 0.9534
		2	**2**12211**2**111**2**2**1**1**2**2212	f(221211221) = 0.6934
⇩			⇩	⇩
20	10		**10010110000111011100**	
		1	**1**12**2**12122221**2**2**1**11**2**21	f(1221122112) = 0.8133
		2	**2**12**2**11211**2**2**1**1**2**222**1**2	f(2212211222) = 0.2643
		1	11221211222122111221	Fitness = 0.8830
		2	21221121112211222212	Fitness = 0.7985

Table 1. Results from a small illustrative example. Using the Epistatic map in Figure 2, *N* is 20 loci and n_{PC} *is* 20. For phenotypic contributors (*PC*), 1 denotes a locus involved in the *PC*, else 0. For the two test sequences (*TS*) shown, 1 and 2 are the two alleles at each locus. φ is the value component for the *TS* under the prevailing *PC*, these being randomly generated constants.

Haploid Fitness (Average of $PC_{(1\text{-}20)}$)

Notice that f(122221) = 0.7843 appears twice, by chance

The functions in the φ column of Table 1 that allocate value to genotypes are determined by allele pattern jointly across loci – for example f(122221) = 0.7843 appears twice in the table, for different sets of loci. This departs from what we might model biologically. An alternative would be to require specific genotypes over many fixed loci, but this would suffer from such complexes being rare to occur and rare to transmit. In conjunction with the selective adaptation steps described, the current approach leads to sensible patterns of fitness across genotypes (Figure 3), without eg "witch's hat" peaks of extreme fitness (Kinghorn and Tanner 2017).

For implementation, attention has to be paid to diploidy and its effect on the expression of single-locus dominance as well as epistasis. For the latter, it is possible to assume dominance of epistasis by stipulating that a *PC* function is expressed if each locus is represented by either one or two of the enabling alleles. In a similar manner we could assume recessive inheritance of epistasis, or a mixture.

In addition, single locus effects need to be addressed. The method proposed can handle that by allowing k=1, which is not represented in Table 1, or indeed by using a classic approach to generate these components. Sampling of k from a Poisson or adapted Gamma distribution might give a presumed sensible weighting to the different levels of epistatic interaction, including k=1 for PCs involving no interactions. Additive and dominance single-locus effects could be conventionally simulated separately for each locus, then, for an interaction set involving k loci, the overall effect taken as the average across the single locus effects multiplied by the φ function shown in Table 1. This would diversify the single-locus effects from the relatively narrow sampling the method provides, and increase diversity of effects for higher-order interacting groups of loci.

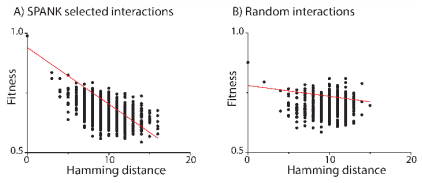

Figure 3. Hamming distance to fittest optimum. A) SPANK generated landscape B) Randomly generated landscape. The number of interacting loci for each landscape is matched. There is much superimposition of points, especially at the fittest optima (see text)

The SPANK model aims to mimic the complexity of genetic systems not from a top down approach, but from a bottom up approach that facilitates the emergence of complex interactions without devolving into chaos. Having a system that mimics gene interactions, in which the resultant interactions are explicitly recorded, we can evaluate the extent to which simple additive models exploit these interactions despite no specific fit to accommodate them. Many questions about the impacts of genetic interactions and our ability to detect and exploit them could be answered. Can a small number of complex interacting QTL masquerade as a large number of QTL? What number and strength of minimally interacting QTL are required to deviate observed causality from major QTL? Where is that missing heritability? Future work might be directed towards answering these and other questions regarding genetic interactions. This paper has only outlined and illustrated an approach to simulating phenotypes. If properly applied, the resulting genotype to phenotype map might approach the complexity of reality. This in turn could help provide insights to what we might be missing by using relatively simple models for QTL detection, GWAS and genetic evaluation.

REFERENCES

Kauffman S. and Levin S. (1987) *Journal of theoretical Biology* **128:** 11.

Cooper M. and Podlich D. W. (2002) *Complexity* **7**: 31.

Kinghorn A.B. and Tanner J.A. (2017) *Complexity* Article ID 6760852, 12 pages. https://doi.org/10.1155/2017/6760852.

Amoutzias G. D., Robertson D. L., Oliver S. G. and Bornberg-Bauer, E. (2004) *EMBO reports*, **5:** 274.