

# Fitness Landscapes and Gene Location

*Peter Whigham*

Spatial Information Research Centre  
University of Otago, Dunedin, New Zealand  
Phone: +64 3 479-7391 Fax: +64 3 479-8311  
Email: pwhigham@infoscience.otago.ac.nz

**Presented at SIRC 2004 – The 16<sup>th</sup> Annual Colloquium of the Spatial Information Research Centre  
University of Otago, Dunedin, New Zealand  
November 29<sup>th</sup>-30<sup>th</sup> 2004**

## ABSTRACT

The paper outlines initial results into a study of the preference of gene location for diploid organisms under the presence of an environmental gradient. The work explores the properties of a spatially-explicit model of monoecious diploid individuals that evolve preferential coding of gene location. The individuals evolve over a space with a linear gradient, where the response to an individual's phenotype determines the age before breeding. Each individual has three chromosomes that determine the position and value of two genes. These genes combine to determine the resulting breeding response. The concept of an NK fitness landscape is used to represent two different scenarios of gene linkage. The results indicate that when gene linkage is related by a rugged fitness landscape, the genes cluster on the same chromosome, whereas when the fitness landscape is smooth the genes are more likely to be on separate chromosomes. The work has implications for understanding some of the possible mechanisms that lead to gene clusters.

**Keywords and phrases:** population genetics, functional clustering, gene location, diploid, genetic tradeoff

## 1.0 INTRODUCTION

Phenotypic traits of an organism are usually related to the effect of more than one gene. For example, coat colour in mammals is determined by a variety of interacting genes, resulting in a wide variety of possible combinations. It has also been known for some time that functionally interacting genes are often closely located on a chromosome. For example, the vertebrate major histocompatibility complex, which comprises between 20-100 functional genes, occurs on the same chromosome, under tight linkage conditions (Nei 2003). There are many other examples, such as conserved clusters of functionally related genes in bacteria (Tamames, Casari, Ouzounis et al. 1997), mycotoxin gene clusters from *Aspergillus flavus* (Zhang, Monahan, Takacz et al. 2004), histone genes from such divergent organisms as yeast, fly and human (Braastad, Hovhannisyanyan, Wijnen et al. 2004), individual yeast gene clusters (Zhu & Zhang 2000) and so on. Clearly there is a high degree of clustering within chromosomes of genes that are related in function or related through co-expression and the construction of genetic networks (D'haeseleer, Liang & Somogyi 2000).

The organization of genes within a genome can be considered in two ways (Hurst, Pal & Lercher 2004): between chromosomes and within chromosomes. There is now evidence that non-random gene order for both types of organization in bacteria and eukaryotes do occur. Until recently gene order in eukaryotes had been assumed to be random (Hurst, Pal & Lercher 2004), however with the advent of a number of complete genomes and a multitude of expression data there is now substantial evidence that this is not true. There are two main possibilities for how the clustering of these functionally-related genes comes about (Nei 2003): the first is based on a founder effect – they merely reflect the origin of new gene functions via tandem gene duplication and mutation. Since the original genes are physically close on the chromosome after duplication, and are likely to produce the same proteins, small changes via mutation produce new, but similar, functions. Hence the functions, through genetic drift, may change over time between these genes, but their origin gives them a functional relationship. The second main proposal is that natural selection maintains and promotes these linked

gene groups, and that therefore it must be an inevitable consequence of evolution and maintenance of species that functionally related genes evolve to become correlated on chromosomes.

This paper describes some initial experiments that test the role of evolution on clustering of gene combinations, using individual-based models of a simple genetic tradeoff based on a linear gradient. Two genes (A and B) are used to determine the final phenotype response for individuals, and the linkage between these genes is controlled by a gene relationship map, based on the concept of correlated fitness landscapes and the NK model (Kauffman 1995). The main question addressed is whether gene order varies under different types of correlated landscape, and if so whether selection alone produces separated genes between chromosomes or gene clusters. This paper is structured as follows: §2 introduces NK landscapes and their role in this work, §3 describes previous work with individual-based models of genetic tradeoff for a single gene, §4 describes the materials and setup for the experimental work, §5 presents the results of the models, and §6 draws some conclusions and directions for future work.

## 2.0 NK LANDSCAPES

The NK model was designed to show how different features of a genotype would behave under a variety of tuneable relationships between genes (Kaufmann & Johnson 1992). The model focused on the coupling between N genes (epistasis) by allowing the fitness of one gene to be related to K other genes. The results from the NK model showed that evolution on rugged (i.e. highly coupled) landscapes slowed exponentially, however there was no consideration of an explicit model of chromosomes or loci. In this paper a similar concept to an NK landscape is used to model the correlation or linkage between two genes.

## 3.0 SINGLE LOCUS GENETIC TRADEOFF MODEL

This work is based on a model of genetic tradeoff (Whigham & Green 2004) for a monoecious, diploid two allele, one locus system. In this model, a spatial grid with a stepped linear gradient was used to explore the behaviour of a population of individuals, as shown in Figure 1. The behaviour (phenotype) of each individual was modified based on their locus value and the underlying gradient where they were located. The work considered two forms of genetic tradeoff – one that altered the relative fitness of each individual, and one that altered the age before breeding. In both forms of tradeoff significant banding patterns evolved under a variety of initial conditions.

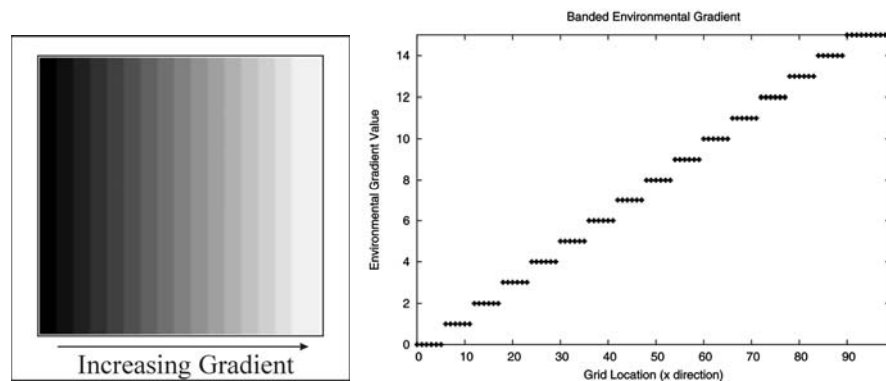


Figure 1. The stepped-gradient used in the genetic tradeoff model(Whigham & Green 2004)

The time-to-breeding tradeoff model from this study showed that even small differences in the average time to maturity produced clear differentiation, and that therefore this approach has a strong selection process in determining the final composition of the population. The results of this work showed that for tradeoff slopes greater than 1.0 differentiation occurred, and that the model would always produce strong banding for tradeoff slopes of 1.5 and greater.

## 4.0 MATERIALS AND METHODS

A square spatial grid of 100 x 100 cells is used to represent the physical space where each organism exists. The modeling over this space is based on the concept of a cellular automaton (Hogeweg 1988; Ermentrout & Edelstein-Keshet 1993). The upper and lower edges of the space are wrapped around, since the gradient of the environment increases in the x-direction but is assumed to be continuous in the y-direction. The left and right edges of the grid are discontinuous. At most one individual can occupy a cell at any time. Breeding of individuals is performed by creating a sub-population, based on the von Neumann neighbourhood (von Neumann 1966) and randomly selecting two individuals as parents from this sub-population. Individuals are assumed to be monoecious and hence, if the sub-population size is one, they are able to mate with themselves. Individuals are represented as diploid individuals, with six chromosomes (3 from each parent). There are 8 alleles, and 4 loci. The first chromosome determines the subsequent locus for each of the two alleles, A and B. Since it is difficult to combine the mapping between the two alleles representing the coding for location, an allele is used to determine which of the mother or father chromosome coding values is selected. This coding is then interpreted to give the allele loci on the second and/or third chromosomes. The value of alleles A and B range between 0 and 15, represented by 4 binary bits. The expressed value for each allele is the average value of their representation between the mother and father chromosomes. A graphical representation of this structure is shown in Figure 2.

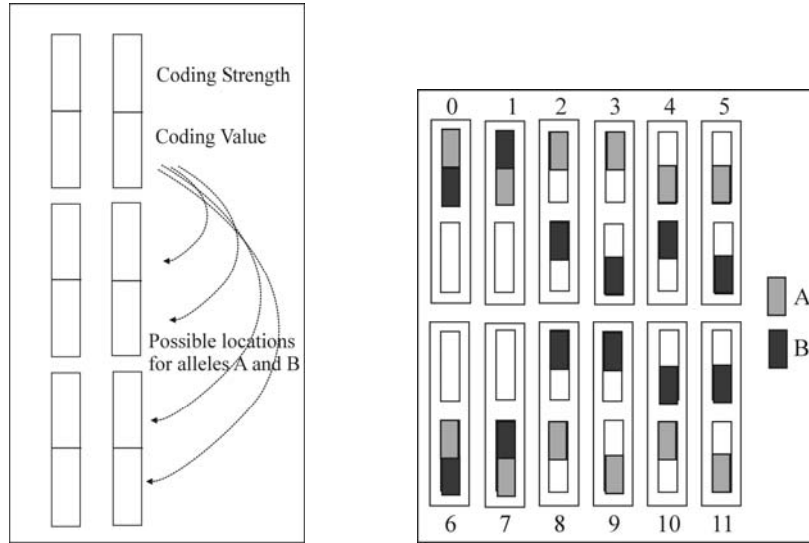


Figure 2. (a) An individual chromosome structure and (b) Coding Interpretation

Figure 2 shows an individual chromosome structure and the corresponding loci that are derived from the coding value. Note that Figure 2(b) shows the location of each allele for just one of the two chromosomes for an individual. Since 4 bits are used to represent the coding value, there are 4 (12-15) values that do not correspond to legal locations. In the simulation individuals with these coding values are assumed to be invalid, and die immediately. This is analogous to a fatal mutation.

Selection of individuals from a sub-population for breeding is random, since there is no explicit fitness measure. The tradeoff relates to the time to reach reproductive maturity, and therefore being able to live at higher environmental conditions implies that an individual breeds more slowly. The expressed value of the locus for an individual determines the age of maturity and the upper limit of the environmental gradient where the individual can live. Hence an individual that can live on the highest gradient value can potentially live at all locations. The actual value of the tradeoff, based on the expressed value of A and B, is determined by a matrix that gives a mapping between the values of A and B, and the subsequent expressed value. This is the analogy to an NK landscape, since this matrix can be used to produce a family of possible gene interaction landscapes. Given a final expressed value (phenotype), the age of breeding maturity ( $BM(v)$ ) is determined by the linear relationship:

$$BM(v) = v * 2.0 + 30 \quad (1)$$

The true breeding age was then derived by a Gaussian distribution  $N(BM(v),2)$ . The age to die for all individuals, independent of their breeding age, was  $N(80,5)$ .

All simulations were run as steady-state models, where for each time step every empty cell was considered as a possible location for a new individual. The sub-population for this empty cell as was then constructed from those individuals in the Von Neumann neighbourhood that had reached breeding maturity. Mating between individuals followed Mendelian rules for diploid individuals, with a mutation rate of 0.00001.

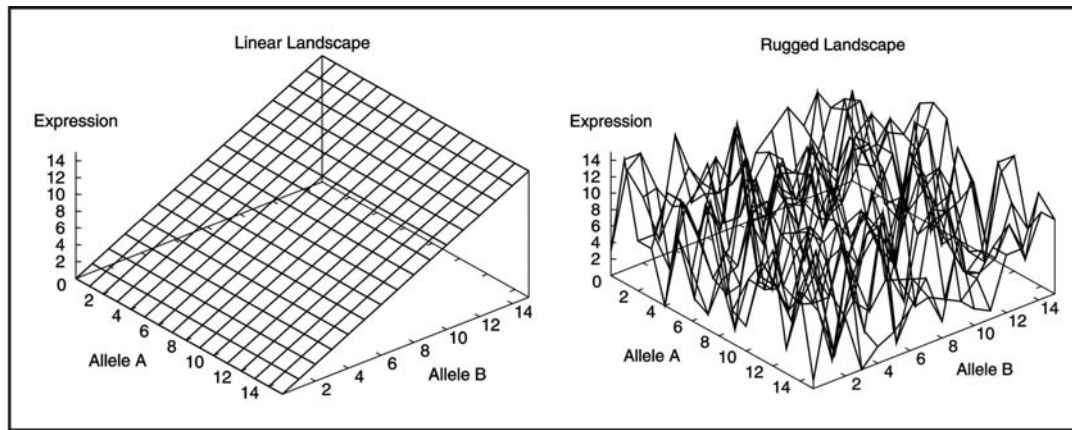


Figure 3. Two forms of landscape representing the coupling between Alleles A and B

Figure 3 shows the two types of matrix used to determine the mapping between the alleles A and B and the final phenotype for an individual. The linear landscape implies that there is loose coupling between A and B, since for a single value of (say) A, all possible expression values of the tradeoff can be derived by incrementing B. Hence this is a correlated landscape. The rugged landscape is a random assignment of the mapping from the allele values to the final expressed value for an individual. Since there is no correlation between adjacent values of A and B this landscape represents a tightly coupled genetic network. For all experiments, when the rugged landscape was applied it was randomly initialized with an equal number of expression values for all possible 16 values so that there was no bias between the mapping of allele values and final expression.

The initial values for the population of individuals on the spatial grid were assigned in two ways: the first set all chromosome values to zero. Hence for the initial population individuals in the linear landscape could only exist on the environment with value zero (i.e. the left-most section of the grid). For the rugged landscape the expressed value depended on the random assignment of the mapping between the zero values for A and B, and hence there was some possibility that the individuals could initially fill the entire grid. The second initialization was random for all chromosome values. Figure 4 shows an example of these initial conditions on the grid. Note that the gaps in the initial random population are caused either by individuals with invalid codes, or where the expressed value is less than the environmental gradient value at the location of the individual.

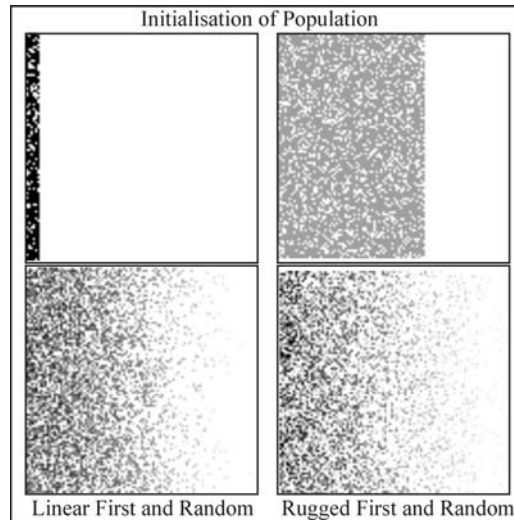


Figure 4. Examples of the initial distribution of individuals for linear and rugged couplings

Simulations were run for both the linear and rugged correlation landscapes, for initial population structures of first and random. For each run, a tradeoff gradient of 2.0 was used (see Equation 1) and the simulation halted after 5 million time steps. In addition, the same runs were performed with a tradeoff gradient of zero so that the affect of tradeoff pressure could be assessed. These runs were performed for 2 million time steps. All simulations were repeated 50 times.

## 5.0 RESULTS

The main question of interest with these simulations is whether there is a significant difference in the placement of the alleles A and B – in particular, under what (if any) conditions do the alleles reside on the same chromosome, or on different chromosomes? Referring to Figure 2, there are only 4 coding values that allow both alleles to reside on the same chromosome, whereas there are 8 codings that place A and B on different chromosomes. This will potentially introduce some bias to the results that will be discussed in §6. The resulting count of the number of individuals with a code selecting for the same or different chromosomes was summed for each final population, and this result averaged over the 50 runs. Figure 5 shows the average resulting gene expression for both the linear and rugged landscapes. Note that independent of the gene correlation distinctive banding has been formed. In Figure 5, the fill first initialization is the top row, with the random initialization on the second row.

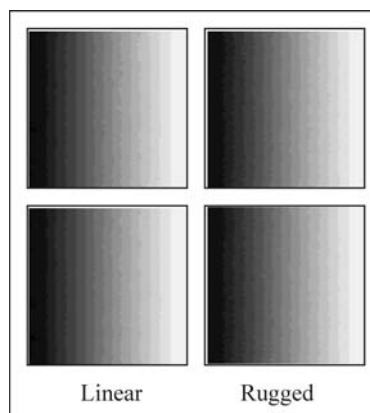


Figure 5. Resulting averaged gene expression values for linear and rugged landscapes.

Of more interest is how the gene coding behaviour in relation to the placement of A and B on the same or different chromosomes. Table 1 summarises the average number of same/different placings over the 50 runs.

Landscape	Same	Different	Non-Coding
Linear Fill-First	2875 ± 2805	6834 ± 2803	290 ± 15
Linear Fill-Random	3233 ± 3147	6476 ± 3144	290 ± 16
Rugged Fill-First	3845 ± 2018	5784 ± 2019	367 ± 284
Rugged Fill-Random	6864 ± 1464	2837 ± 1466	298 ± 19

Table 1. Average A and B allele positioning on chromosomes

Using a double-sided t-test for estimating whether there is a significant difference between means, there is a greater than 99% confidence that the linear and rugged averages are drawn from a different population and hence that there is a significant preference for genes to be located on the same chromosome with rugged landscapes, and different chromosomes when the gene correlation landscape is linear. Note that there is no significant difference between the initialization (first or random) for the linear results. However, when the populations are initialized with fill first, both rugged and linear landscapes have a bias towards placing the alleles on different chromosomes. However, this result must be taken with some caution, since there are twice as many possible ways to produce different chromosome codings than codings for the same chromosome. If we assume that the process is random, then we could halve the results for different chromosome position. This would imply that there is little or no significant difference in preference for placement on the same or different chromosomes for allele position when the environment commences with a single value (zero) for individuals.

The second set of runs set the tradeoff gradient to zero. As was shown in the previous single allele study (Whigham & Green 2004) individuals that have a gene expression that allow them to live at all grid locations would be expected to eventually take over the complete set of grid locations. Runs for both the linear and rugged landscapes were performed with a random initialisation. Results are shown in Table 2.

Landscape	Same	Different	Non-Coding
Linear	3069 ± 3976	6797 ± 3976	132 ± 9.4
Rugged	9133 ± 2327	733 ± 2325	133 ± 10

Table 2. Response to linear and rugged landscapes with zero tradeoff gradient

Table 2 shows a clear difference in behaviour based on the correlation landscape for alleles A and B. In particular, with a rugged landscape there is almost a complete selection towards coding the alleles on the same chromosome, whereas with the linear landscape there is little difference (assuming that the coding process is random, and therefore that the different value must be halved).

To address the behaviour of a random model, the system was finally run with a zero tradeoff gradient, and no environmental gradient (all values in the 100x100 grid were set to zero). For this null model there is no influence on the correlated landscape, and so the run was performed with only the rugged landscape. Since there is no selection pressure for the model it should drift to a random, fixed gene value. In this case it would be expected that coding for placement on different chromosomes would be twice as likely as coding on the same chromosome, due to the bias in the coding representation. The resulting coding values for a random and fill first initial population are shown in Table 3. There is no significant difference at the 99% level between the first or random initialization for same/different results, with the same coding occurring 53% as often as the different coding for fill-first, and 65% as often when the initialization is fill-random. The result for fill-first is as predicted, however the fill-random ratio is slightly biased towards the same coding. Further research is required to determine why the initialization causes this increase towards placement on the same chromosome, however this null model does not invalidate the previous results.

Initialisation	Same	Different	Non-Coding
----------------	------	-----------	------------

Fill-First	3443 ± 3441	6424 ± 3442	132 ± 12
Fill-Random	3912 ± 3407	5952 ± 3407	135 ± 10

*Table3. Simulation behaviour with genetic drift*

## 6.0 DISCUSSION AND CONCLUSION

The results for the linear and rugged landscape (Table 1) imply that through evolutionary processes there is strong selection pressure to place alleles on the same or different chromosomes, depending on the form of allele correlation and the initial population structure. A linear landscape allows the allele values to be altered in such a way that there is a smooth transition through the landscape of expression values. Since this results in the allele values being able to be treated independently even though they combine to produce the final expression, they are placed on separate chromosomes. A rugged landscape means that a change in either allele value produces a random change in expression, and hence the values for A and B are intimately associated – a small change in either allele produces a random change. This results in the alleles being tied together and therefore being associated on the same chromosome.

What are the implications from this preliminary study? The main result supports previous theoretical work and studies of genomic data in that gene order is not random, and that it is likely that functionally related genes are likely to cluster together on the same chromosome under some circumstances. In particular, through the mechanism of evolution alone these gene orders can be produced. Of more interest is the fact that two genes that are related in the phenotypic expression of an organism may not cluster if the landscape of interaction between the genes is smooth. This raises the issue of how gene interaction should be described, and whether it is possible to characterize gene relationships from real data in terms of landscape properties. Clearly an understanding of the relationship between linkage in genes and landscapes would allow a greater understanding of gene order. Since the location of a gene within a genome is often difficult to discover, understanding why genes are not randomly ordered, and where they are likely to cluster, is fundamental to the future study of bioinformatics. The other main issue that this work highlights is the founder effect of gene order – depending on the initialization of the population there is a significant difference in preferential placement of alleles. Clearly further work is required to understand how initialization changes the preferential placement of alleles due to selection. It may well be that the major observed clustering of genes is dominated by the growth of genomes and gene duplication, and that natural selection plays a minor role in the placement of correlated genes. A model that allowed the genome to increase in size during evolution would be one approach that would allow some of these issues to be explored in more detail.

This paper has described a first attempt at characterizing a model for exploring the interaction between evolution, gene placement and gene order. The model has some limitations and biases that should be addressed. The first issue is regarding the use of a coding strength to select one chromosome to determine where the alleles are placed. Although this has some analogies with aspects of real biology (Jackson 2003) a simpler model that excludes the coding strength would be desirable. One approach is to just take the average of the coding values from the mother and father chromosomes, although this doesn't necessarily make sense in terms of a biologically reasonable response. Further work is required to explore more realistic code representations that are biologically plausible. A second area that needs addressing is the bias introduced by having only 4 codes for placement on the same chromosome, and 8 for different chromosomes. This would imply that there is a bias towards placement on different chromosomes in the model, and that therefore the results with same chromosome placement are even stronger than has been demonstrated here. A simple solution may be to have just two possible coding placements, and ignore the permutations of gene order.

As stated by (Hurst, Pal & Lercher 2004):

“..there is a need to develop a theory of genome-organization evolution, taking into account the mechanisms of genome rearrangement, mechanisms of control of gene expression and the evolutionary forces that result from different interactions of loci. ...Modelling the evolution of gene order, from both selective and neutralist perspectives, represents a considerable challenge.”

## ACKNOWLEDGEMENTS

The author would like to express his thanks to Colin Aldridge for computing support during the simulation work. Other members of Otago University, including Hamish Spencer, Jo Stanton and David Green must also be thanked for fruitful discussions and background information.

## REFERENCES

Braastad, C. B., H. Hovhannisyan, A. J. v. Wijnen, J. L. Stein & G. S. Stein (2004) Functional characterization of a human histone gene cluster duplication. *Genetics*, 342:1, pp. 35-40.

D'haeseleer, P., S. Liang & R. Somogyi (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16: pp. 707-726.

Ermentrout, G. B. & L. Edelstein-Keshet (1993) Cellular automata approaches to biological modeling. *J. Theor. Biol.*, 160: pp. 19-133.

Hogeweg, P. (1988) Cellular automata as a paradigm for ecological modeling. *Applied Math. Comput.*, 27: pp. 81-100.

Hurst, L. D., C. Pal & M. Lercher (2004) The Evolutionary Dynamics of Eukaryotic Gene Order. *Nature Reviews*, 5: pp. 299-310.

Jackson, M. (2003) Duplicate, decouple, disperse: the evolutionary transience of human centromeric regions. *Current Opinion in Genetics & Development*, 13: pp. 629-635.

Kauffman, S. (1995) *At Home in the Universe: The Search for Laws of Self-Organization and Complexity*, Penguin Books, London, England, pp. 321.

Kaufmann, S. A. & S. Johnson 1992, 'Coevolution to the edge of chaos: coupled fitness landscapes, poised states and coevolutionary avalanches', in *Artificial Life II*, Eds C. G. Langton, C. Taylor, J. D. Farmer & S. Rasmussen, Reading, MA: Addison-Wesley, pp. 325-370.

Nei, M. (2003) Let's stick together. *Heredity*, 90: pp. 411-412.

Tamames, J., G. Casari, C. Ouzounis & A. Valencia (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol*, 44:1, pp. 66-73.

von Neumann, J. (1966) *Theory of self-reproducing automata*, University of Illinois Press, .

Whigham, P. A. & D. Green (2004) A Spatially-Explicit model of Genetic Tradeoff. (to appear) *The 7th Asia-Pacific Conference on Complex Systems*, Australia, 6-10th December.

Zhang, S., B. J. Monahan, J. S. Takacz & B. Scott (2004) Indole-Diterpene Gene Clusters from *Aspergillus flavus*. *Appl Environ Microbiol.*, 70:11, pp. 6875-6883.

Zhu, J. & M. Q. Zhang (2000) Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomput.*, 5: pp. 476-487.