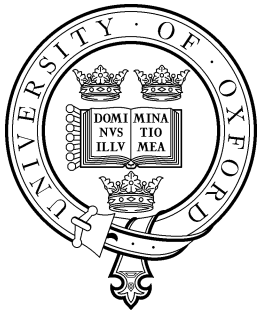
The evolutionary dynamics of neutral networks: Lessons from RNA



Mark Rendel

Magdalen College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2008

Acknowledgements

There are many people who I would like to thank in relation to this thesis. First, my supervisor, Alan Grafen, who has provided inspiration, and a fantastic sounding board. He has enabled me to turn this thesis from a jumble of ideas into something coherent, and I hope intelligible. His sharpness and insightfulness, once I had finally made myself clear, has made the experience both exciting intellectually, and a pleasure. I would also like to thank Jotun Hein, who filled in as my temporary supervisor while Alan was on sabbatical, and who, with his group, welcomed Alexis Gallagher and me warmly.

I would also like to thank Alexis Gallagher for many early discussions, and providing the initial code port to the RNAfold algorithm. Andrew Rendel, Marco Archetti, Jack Mellor, and Alexis Howe for helpful advice and discussions, along with the members of the Animal Behaviour Research Group and the Evolutionary Biology Group among others. Steven Young provided valuable advice in my attempts to fight through the (unintuitive for a biologist) world of grid computing, the NGS and the Globus Tookit. I must also thank David Rendel, Sara Engleka, Zan Goldblatt, and Alexis Howe for proof reading various sections and drafts.

I would particularly like to thank Walter Jetz and his Lab at UCSD for making me feel so at home, and providing an excellent work environment, in which the bones of chapter five were put in place.

Financially, I am indebted to the Wingate foundation, the Sir Richard Stapley educational trust, the Fish Mongers' Company of London, the Flemming family through Magdalen College, and my parents Sue and David Rendel, who provided financial support, a roof and patience for the last 9 months.

Does the road wind up-hill all the way?
Yes, to the very end.

Christina Rossetti,1861.

Abstract

The evolutionary options of a population are strongly influenced by the availability of adaptive mutants. In this thesis, I use the concept of *neutral networks* to show that neutral drift can actually increase the accessibility of adaptive mutants, and therefore facilitate adaptive evolutionary change. Neutral networks are groups of unique genotypes which all code for the same phenotype, and are connected by simple point mutations. I calculate the size and shape of the networks in a small but exhaustively enumerated space of RNA genotypes by mapping the sequences to RNA secondary structure phenotypes. The qualitative results are similar to those seen in many other genotype–phenotype map models, despite some significant methodological differences. I show that the boundary of each network has single point–mutation connections to many more phenotypes than the average individual genotype within that network. This means that paths involving a series of neutral point–mutation steps across a network can allow evolution to adaptive phenotypes which would otherwise be extremely unlikely to arise spontaneously. This can be likened to walking along a flat ridge in an adaptive landscape, rather than traversing or jumping across a lower fitness valley. Within this model, when a genotype is made up of just 10 bases, the mean neutral path length is 1.88 point mutations. Furthermore, the map includes some networks that are so convoluted that the path through the network is longer than the direct route between two sequences. A minimum length adaptive walk across the genotype space usually takes as many neutral steps as adaptive ones on its way to the optimum phenotype. Finally I show that the shape of a network can have a very important affect on the number of generations it takes a population to drift across it, and that the more routes between two sequences, the fewer generations required for a population to find an advantageous sequence. My conclusion is that, within the RNA map at least, the size, shape and connectivity of neutral networks all have a profound effect on the way that sequences change and populations evolve, and by not considering them, we risk missing an important evolutionary mechanism.

Contents

List of Figures

Chapter 1

Introduction

The opening lines from Christina Rossetti's poem 'Uphill' ask a poignant question about whether the road through life runs up-hill all the way. In this thesis I shall explore whether these lines are a good metaphor for evolution, which is often characterised as an up-hill struggle within an adaptive landscape, where each evolutionary step involves an increase in fitness. In fact I show that though the road taken by an evolving population can wind significantly, each step does not necessarily have to be up-hill.

The work presented here has been stimulated by a refocus on studying the dynamics of adaptive evolutionary change, after a period where the neutral theory dominated molecular evolutionary research. Much recent research has often focused on the generation of, as well as selection for, adaptive changes. These studies have usually taken one of two approaches: The first is fuelled by massive increases in computing power, which has allowed advances in analytical and simulation based methods, especially using whole genetic sequences rather than studying individual loci (see Grüner et al., 1996a,b; van Nimwegen et al., 1999; van Nimwegen and Crutchfield, 2000; Göbel, 2000; Stadler et al., 2001; Ebner et al., 2001; Smith et al., 2003; Kospach, 2003; Kutschera and Niklas, 2004, for examples). In the second, long term evolutionary experiments using organisms with short generation times have led to insights into the generation of and selection for advantageous mutations (e.g. bacteria: Lenski and Travisano (1994); Papadopoulos et al. (1999); Riley et al. (2001); Buckling et al. (2003), viruses: Burch and Chao (1999, 2000); Makeyev and Bamford (2004); Koelle et al. (2006) and fruit flies (*Drosophila* sp.): Houle and Rowe (2003); Alipaz et al. (2005); Rundle et al. (2006). See Elena and Lenski, 2003, for an overview). These new approaches supplement the already established fields of population genetics and phylogenetics, in turn built upon abstract models such as Fisher's geometric model (Fisher, 1930), and Wright's adaptive landscape (Wright, 1932).

Studying whole organisms or modelling many loci simultaneously has highlighted the affect that neutral mutations can have on the accessibility of adaptive mutants across the adaptive landscape. When neutral mutants are incorporated into a multi locus model, any neutral steps from an area of the landscape where there are no further adaptive mutations can act like walking across a ridge or plateau on the side of a mountain, potentially finding a further adaptive change on the far side. In this way, neutral drift may provide an important evolutionary mechanism by which a population can access a greater number of higher fitness mutants than might otherwise be possible. The work in this thesis examines the potential role that neutral mutations can play in adaptive change, by using a computational rather than empirical biological

approach, simulating the evolution of RNA sequences within a discrete, map-based model.

In the rest of this introduction, I shall start by discussing the broader theoretical evolution framework from a historical perspective. This builds towards a general explanation of the concept of discrete *genotype-phenotype maps* and *drift-based* evolution. *Neutral networks* are introduced as groups of closely related genetic sequences which all code for the same phenotype, and therefore fitness. I highlight their possible effects within evolution, and review what we already know about their structure and influence. I then provide a brief overview of the different methodological approaches that have been used to consider these ideas, outlining the benefits of and problems facing each. Finally I shall summarise the ideas discussed in this introduction, and address certain questions that arise.

1.1 Theories of adaptive evolution

The field is not a new one within biology; from the turn of the 19th century, and even before, biologists have been struck by the observable inter- and intra-species variation present in the natural world and have sought to explain this variation by a process of change over time (e.g. Lamarck, 1809). In 1859, Darwin presented his theory of evolution by natural selection, now established as the mainstay of adaptive evolutionary theory. This theory has been continually refined since its inception – often on the basis of new empirical data or discoveries. Most importantly, Fisher (1918) integrated natural selection with genetics by combining it with Mendel's (reprinted, 1951) work on the particulate nature of inheritance. This 'New Synthesis', became the basis of adaptive evolutionary theory as we know it today.

Since then different interpretations of limited models and/or data have fuelled controversies within evolutionary theory, some of which show no sign of abating, even if the areas of greatest contention have shifted over the years (e.g. Fisher, 1930; Wright, 1932; Kimura, 1968a; Eldredge and Gould, 1972; Dawkins, 1976; Gould and Eldredge, 1977; Coyne et al., 1997; Wade and Goodnight, 1998; Wilke et al., 2001; Comas et al., 2005). One of the longest lasting controversies has been over the importance of epistatic interactions (where two genes or loci combine to produce a result different from the sum of their parts). Wright's conviction in their importance led him to design the first maps exploring the relationships between genetic variation and change in fitness— his so-called *fitness* or *adaptive landscapes* (Wright, 1932).

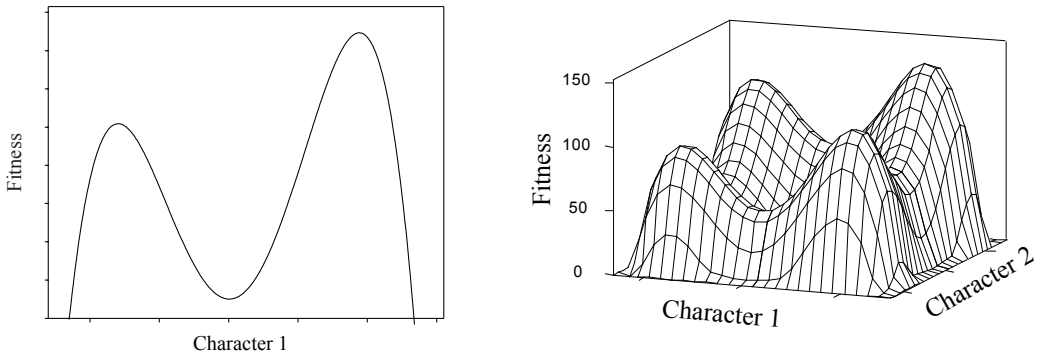
1.1.1 Fitness Landscape

The concept of the fitness landscape has had a pervasive and long lasting influence on evolutionary thinking, and neatly elucidated Wright's idea that evolution does not necessarily always achieve an optimum solution, but can get stuck in 'local peaks'.

In this he assumed the environment was static, or at least did not change in a way which affected a population's evolutionary potential. As usual, this contrasted sharply with Fisher (1930) who supposed a static environment to be unrealistic and had designed a geometric model in which a population evolves towards an optimum it never reaches. Fisher's argument was partly based on the assumption that as the number of dimensions in which a change could happen increased, it became more and more likely that a fitter mutant could be found in some other dimension – meaning that 'local peaks' become less and less likely and are better described as a shoulder on the way to the summit (Fisher correspondence to Wright, May 31, 1931 in Provine, 1986, p274). The controversy over whose approach is better has continued to rumble on to the present day and shows no sign of being resolved (see Skipper, 2002, 2004, for a very clear overview).

One of the greatest strengths of the landscape metaphor is that it is an intuitive way of visualising an abstract concept. However, it can also be subtly misleading, and has been used loosely over the years in many different ways. Thus, it is arguably most valuable when used simply as an illustration, and it is in this manner that I intend to use similar diagrams throughout this thesis. In fact, it has long been unclear what each axis was supposed to represent; indeed Wright himself changed them (Wright, 1932, 1970). Examples include gene frequencies (Ridley, 2004), continuous phenotypic characters (Lande, 1976; Ridley, 2004), or discrete gene combinations (Strickberger, 2000) on the horizontal axes, with mean population or individual fitnesses on the vertical. There is no consensus even between modern evolution textbooks (e.g. Strickberger, 2000; Ridley, 2004) on which is most appropriate. In fact, each can be valid, but has different implications. In a two–dimensional landscape, a local peak can be quite clearly seen during continuous change over a given characteristic or phenotype, but this becomes more complex as further dimensions are added (Fig. 1.1).

As the number of dimensions considered increases, epistatic interactions have the potential to play a defining role in the shape of the landscape. Although they are not *required* to form local peaks in a continuous fitness landscape (e.g. Fig. 1.1b), if they are present it becomes impossible to predict the landscape without data points from across the range of all combinations of character values (Fig. 1.2).



(a) **(b)**

Figure 1.1: Wright's concept of a fitness landscape is a useful one in many respects, but has been used loosely over the years in many different ways. **a) 2-D landscape:** as the character increases on the x axis, it reaches a local optimum before suffering a decline in fitness then recovering to the global optimum. **b) 3-D landscape:** the additive interaction between the two characters, both with two peaks, leads to 4 distinct optima.

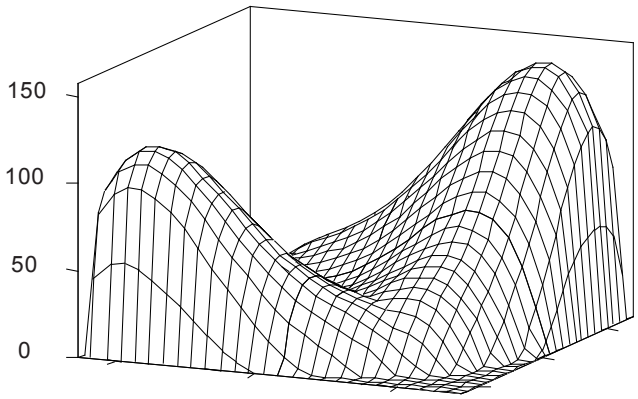


Figure 1.2: A three–dimensional graph showing how two continuous variables can interact epistatically to produce distinct optima which depend on the *combination* of the variables. (e.g. gene frequencies or phenotypic values)

Perhaps the greatest short–coming of the landscape metaphor is that it is limited to just three dimensions. In fact in a recent review Walter Fontana (2002) requested of his two–dimensional contour diagram:

> For a more accurate representation the reader ought to imagine at least a 100-dimensional space.

As I, like most people, have significant trouble visualising beyond three physical dimensions I shall not request the same! However, when using these kind of landscape diagrams we must be aware of the potential pitfalls of using an approach which

does not very accurately model the complex shapes and paths that it is supposed to highlight.

Despite these shortcomings, figures and diagrams remain the clearest way of expressing many of the central ideas within the topic and are used throughout the literature, albeit without being a true reflection of the actual space (e.g. Conrad, 1990). This thesis relies on diagrams as a heuristic tool to examine the ideas presented, and they are often followed by results presented in a far less digestible format than the preceding explanations because of the complex multi-dimensional nature of the spaces considered. Adaptive landscape diagrams can be particularly useful when considering the effect of neutral mutations on potential adaptive change.

1.1.2 Neutral Theory and Punctuated Equilibrium

With increasing electrophoretic data from proteins and later DNA sequence data, it became apparent that when no adaptive force drives genetic change, neutral variation builds up in the population (Kimura, 1968a,b; King and Jukes, 1969; Jukes and Kimura, 1984). The neutral theory contends that most molecular substitutions are due to the drift of neutral alleles in this manner. The chance changes in frequency of neutral mutants have led many biologists to dismiss neutral changes as unimportant to the study of natural selection (e.g. Grafen, 1988). However, the existence of neutral changes have the potential to play a defining role in adaptive change. Individually and in isolation, a whole range of mutations may be selectively neutral. If these mutations are not purged by an adaptive mutant occurring, they can build up over generations until several neutral mutations combine in one individual to confer some kind of selective advantage, as originally suggested by Wright (Wright correspondence to Kimura, in Provine, 1986, p.474). This idea has subsequently been shown to occur in various models by Huynen et al. (1996), Fontana and Schuster (1998a), van Nimwegen and Crutchfield (2000), Crutchfield (2002) and Smith et al. (2002, amongst others,).

In a landscape context, the neutral mutations can form 'ridges' or 'plateaus' between areas of equal fitness (Fig. 1.3). These are equivalent to flat versions of Fisher's 'shoulders' in the landscape with one key difference, that a mutation along the ridge is not selected for in the generation that it occurs. However, like Fisher's shoulders, ridges can reduce the number of local optima in a landscape. While fitness is *maintained* along a ridge, the more horizontal *neutral drift* that occurs the greater the chance of discovering a fitter phenotype, because more genetic combinations are encountered along a ridge than occur at a local peak. As a population drifts or diffuses

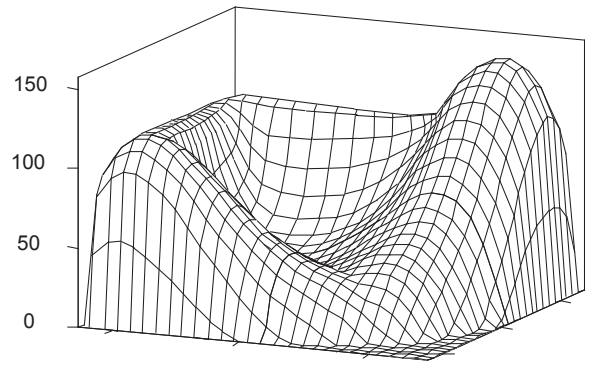


Figure 1.3: A three–dimensional graph showing how what was originally a local optimum in figure 1.2 can be linked via a neutral ridge to become the first point of a shoulder leading to a higher optimum.

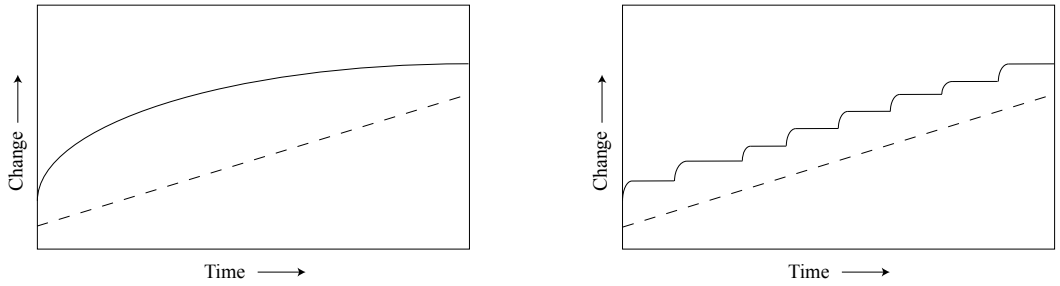
over a ridge or plateau, the *evolvability* [1] (i.e. the potential for *future* adaptive change) can change. If a particular neutral mutation means that an individual is *closer* to the next incline, their future offspring are more likely to reach that slope than the offspring of those unmutated individuals in the population.

Fontana and Schuster (1998a), van Nimwegen and Crutchfield (2000), Ebner et al. (2001), Smith et al. (2003) and Wolf et al. (2006) have all suggested that this kind of neutral drift provides an explanation of the punctuations seen in phenotypic evolution, which, in the fossil record, Eldredge and Gould originally attributed to environmental perturbations and which has also been explained by Wright's shifting balance theory by Lande (1986). Figure 1.4b shows that punctuated phenomena can be encountered if genotypic change is phenotypically silent but required to traverse a ridge in the landscape, before a mutant arises which changes the phenotype in an advantageous manner.

1.1.3 Discrete space

Much of the recent evolutionary research at a molecular level has focused on a discrete space of genetic combinations rather than continuous gene frequencies or phenotypic characters. It allows biologists to study evolution at a more fine–grained level than has been possible in the past, because at its lowest level the fundamental units of change (genetic code mutations) are discrete.

[1] The term 'evolvability' was coined by Dawkins in 'The Blind Watchmaker' (1986). For discussions on what evolvability is and its differing definitions see Nehaniv (2003) or Pigliucci (2008).



(a) **(b)**

Figure 1.4: Solid line represents phenotypic change, dotted line is genotypic change. **a)** Genetic changes occur at roughly the same rate assuming a constant mutation rate, but provide smaller and smaller fitness advantages as they approach the optimum. **b)** Local phenotypic stasis (the start of a ridge) is reached very quickly in the landscape. The genetic changes switch from having adaptive consequences to neutral drift along the ridge, but build up in the population at approximately the same rate, before a further bout of phenotypic change can occur.

Discrete space thinking was introduced first by Maynard Smith (1970) within the context of a discrete protein space. He used the metaphor of real and nonsense words for viable and inviable proteins. Sensical, 'viable' words make up a small percentage of all the possible letter combinations of a given length of word. In his example he changed letters one at a time (akin to single amino acids substitutions) to make a series of 'viable' words:

$$\text{WORD} \rightarrow \text{WORE} \rightarrow \text{GORE} \rightarrow \text{GONE} \rightarrow \text{GENE}$$

The total number of letter combinations is calculated from the length of the sequence and the number of possibilities at each position:

$$S = n^A \tag{1.1}$$

where S is the size of the set of sequences, n is the length of each sequence and A is the number of options at each position (the size of the alphabet). So for Maynard Smith's example using the English alphabet with 26 characters, the total number of possibilities is:

$$S_{four\ letter\ words} = 4^{26}$$

Maynard Smith's notion of single-step mutations (changing only one letter at a time) allows us to draw up a *local neighbourhood* of 'reachable' mutants, each differing at only one position from the original genotype (see Table 1.1a for an example using the word WORD). The local neighbourhood of any one sequence is made up of a

subset of all possible sequences. Of these local neighbours only some are likely to be viable (Table 1.1b). If each neighbour can only differ at one position, the number of

Character position			
1	2	3	4
AORD	WARD	WOAD	WORA
BORD	WBRD	WOBD	WORB
⋮	⋮	⋮	⋮
YORD	WYRD	WOYD	WORY
ZORD	WZRD	WOZD	WORZ

(a) All words

Character position			
1	2	3	4
CORD	WARD	WOAD	WORE
FORD		WOLD	WORK
LORD		WOOD	WORM
			WORT
			WORN

(b) Viable words

Table 1.1: a) The local neighbourhood of the word 'WORD'. **b)** The viable neighbours of the word 'WORD'

sequences in the local neighbourhood is:

$$N = (A - 1)n$$

For the WORD example the local neighbourhood is 25x4 sequences. A mutation is far more likely to result in a neighbour than in any other possible genotypic sequence. Gillespie (1984) has shown that in most realistic scenarios (a relatively low mutation rate) a single point mutation is possible, but the chance of getting a particular combination of two or more point mutations in a single generation is negligible. However, Gillespie also pointed out that if a single base change results in a viable mutant, the neighbourhood of the mutant includes $(A-1)n-1$ new sequences which *are* two base changes from the original sequence, increasing the search space for advantageous mutants.

As an introduction to higher dimensional spaces, it is worth noting here that if we ignore whether a word makes sense or not, no sequence requires more than four mutations to reach it from everywhere else in the space. Put another way, it is possible to directly traverse the whole (rather large) space of 456,976 different letter combinations (e.g. from AAAA to ZZZZ) in just 4 steps.

1.1.3.1 Hamming distance paths

The four character changes from WORD → GENE are the maximum number detectable by simply comparing the two sequences. However, it is quite possible that more than one change may occur at each position. In this case, only the last change at a given position is seen. The number of characters that differ between two strings

(sequences) of equal length is known as the Hamming distance (Hamming, 1950). This means that for each change that occurs at a *new* position the Hamming distance will increase by one. If there is a restriction on whether a letter can be changed based on creating sensical words then sometimes a path *longer* than the Hamming distance is the only way of reaching another ('fitter') solution. Consider my similar example:

APE APT OPT OAT MAT MAN

The maximum Hamming distance is 3 (the sequence length), but there is no Hamming distance route between APE and MAN whilst retaining sensical words. The shortest route in this case is two steps longer than the Hamming distance and requires two substitutions at the 1st and 3rd positions.

1.1.4 Quasi-species model

The discrete space molecular quasi-species model was introduced by Eigen and Schuster (1977) based on Eigen (1971) as a way of modelling the evolution of the early molecules of life. It says that an infinitely large asexually reproducing haploid population undergoes selection based on the average fitness of a cluster of discrete points, rather than the fittest individual genotype. This is because any individual residing at a point in the discrete space has a chance of producing offspring either identical to itself or mutated to a closely related neighbour. As mutation rate increases, a population will spread out across a fitness peak in the space, producing more and more mutated and potentially less fit offspring. Purging selection removes the less fit mutant individuals, and a balance is struck with continual replacement by the newly mutated offspring of fitter individuals. One result of this spread, is an *error threshold* mutation rate, beyond which mutation away from the peak is more powerful than selection back to it, and the population can drift or diffuse away from it. The existence of a hard error threshold is still debated depending on the parameters of the model, and especially under more realistic assumptions of back mutations (Comas et al., 2005; Wilke, 2005; Takeuchi and Hogeweg, 2007).

When fitness does not change with distance because the population is on a ridge or plateau instead of a peak, the population is free to diffuse/wander across it depending on the population size and mutation rate van Nimwegen et al. (1999) and Sumedha et al. (2007a). The larger the plateau, the further a population can spread and mutate widely before falling off the edge and being purged. This concept has been extended to argue for selection for mutational robustness, especially in high

mutation environments (van Nimwegen et al., 1999; Bornberg-Bauer and Chan, 1999; Wilke et al., 2001; Wilke, 2001a). As mutation rate increases, large plateaux become increasingly advantageous, because fewer mutant offspring are likely to 'fall off' them, leading to Wilke et al.'s term 'survival of the flattest'.

The difficulty of biologically quantifying the parameters of the quasi-species model has meant that uncontentious quasi-species theory has remained mainly abstract and qualitative. Perhaps the most important impact of quasi-species theory in relation to this work has been to concentrate attention on mutation as a powerful and continuous force in evolution, not just providing the raw material for natural selection, but influencing evolutionary dynamics in its own right through selection of populations made up of 'clouds' of individuals in a discrete genotype space.

1.2 The genotype–phenotype map

Following on from Maynard Smith (1970), Gillespie (1984), Kauffman and Levin (1987) and Kauffman (1993), *genotype space* can be defined as the set of all combinations of possible alleles or bases for a given sequence length (i.e. the number given by equation 1.1). A function of some kind then maps each of these genotypes to a (set of) particular phenotype(s) or fitness(es). This can be anything from assigning random or correlated fitnesses, as in Kauffman and Levin's 'NK' landscapes, through the biophysical folding pattern of RNA sequences to the complex developmental pathways that transmit genetic information into higher vertebrates.

A mapping function as complex as the developmental pathways in vertebrates is likely to have modifiers that alter the phenotype (see West-Eberhard, 2003; Carroll et al., 2004, for recent reviews). This kind of developmental or phenotypic plasticity means that a genotype can map to more than one phenotype even within the relatively simple mapping of RNA sequences to structure (Schultes and Bartel, 2000). However, the opposite is very common when the function between genotype and phenotype is more limited – many different genotypes all mapping to the same phenotype and/or fitness.

In protein–coding genes, only the amino acid tryptophan has just one triplet base codon in the genetic code which maps to it (UGG), the other 19 amino acids are mapped to by up to six different codons each. For example, the codons UUA, UUG, CUA, CUC, CUG, and CUU all map to the amino acid leucine, meaning the underlying sequence can change significantly, while still coding for the same polypeptide

chain phenotype [2].

The many–to–one mapping applies equally to RNA genes which code for folded RNA structures. Swapping RNA base-pairs within a helical stack results in no change in structure, as long as the base-pairs remain matching (Fig. 1.5a). However, it only takes one substitution from a matching to a non-matching base to completely disrupt the structure (Fig. 1.5d). Protein coding and RNA structure genes have different mapping functions, but the complex many-to-one nature of the map remains (synonymous codons or conserved base-pairs).

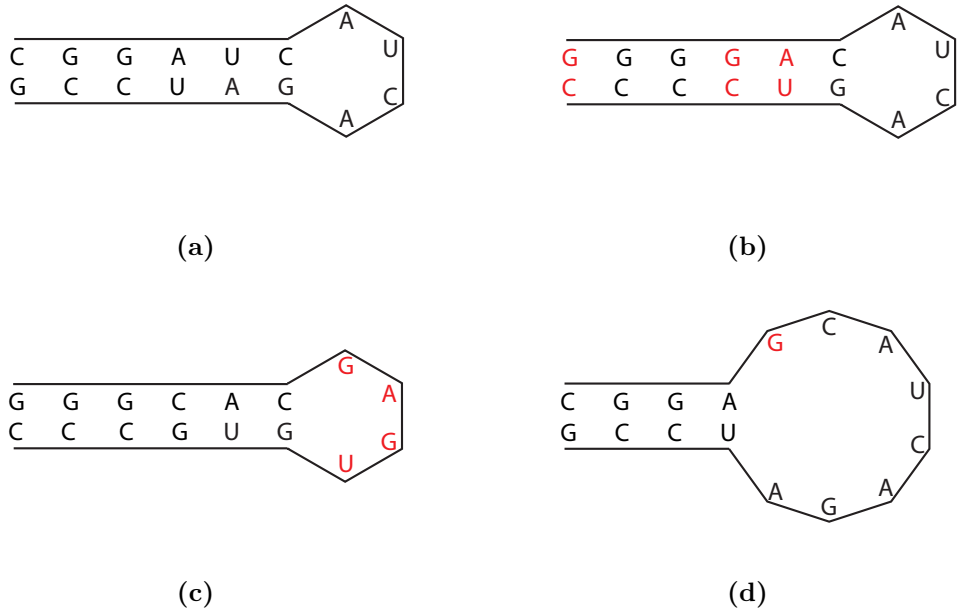


Figure 1.5: RNA genes produce single–stranded RNA sequences which can be folded back on themselves to form bonds between complementary bases. **a)&b)** Quite different sequences can produce the same secondary structure while the paired bases (shown in red) continue to complement each other. **c)** unpaired base pairs can often be substituted while retaining a structure, unless a mutated unbound base happens to provide a preferential binding site. **d)** It only requires a single substitution in the RNA strand to significantly change the phenotypic structure.

Models, such as those of Lipman and Wilbur (1991), Schuster et al. (1994), Huynen et al. (1996), Govindarajan and Goldstein (1997a), Bastolla et al. (1999),Held et al. (2003), Bardou and Jaeger (2004) and Anderson and Jensen (2005) which use

[2]Although there is evidence to show that codon usage bias (Miyata et al., 1979; Modiano et al., 1981; Kimura, 1981; Conrad et al., 1983; Archetti, 2006) occurs in many organisms, for the purposes of this thesis I shall assume these effects to be inconsequential at a genetic map level as long as the phenotype remains identical in other respects. With an arbitrary phenotype as discussed here, any biases at the level of the genotype could be included in a more complex phenotype.

a genotype–phenotype mapping, usually fix the mapping function between genotype and phenotype and concentrate on the links between different genotypes in the space (but see Ancel and Fontana, 2000; Wroe et al., 2007). They therefore complement and contrast with conventional population genetic models, which usually focus on the selective effects of polymorphisms already present in populations. As selection generally acts at the level of the phenotype (Weiss and Fullerton, 2000; West-Eberhard, 2005), in most population genetic models each genotype codes for a different phenotype in a one–to–one mapping.

At its most abstract, the genotype–phenotype map does not say anything about the fitnesses of individual genotypes or phenotypes. Instead, it relates the distances and paths between genotypes according to how they map to their respective phenotypes. By considering the map, we can get an insight into which genotypes and phenotypes are *available* as mutational fodder for the selection cow to chew.

1.2.1 Phenotypic frequencies and shape space covering

The mapping between genotypes and phenotypes depends on the specific function, but when that function maps many genotypes to one phenotype, Schuster et al. (1994), Bornberg-Bauer (1997), Göbel (2000), Aita et al. (2003) and Sumedha et al. (2007b) have all shown in simple biophysical models that the distribution of sequences into phenotypes is similar to that of a generalised Zipf's law – a power law distribution, where the number of sequences coding for a particular phenotype is inversely proportional to the rank of the number of sequences coding for that phenotype (Zipf, 1935). In other words, most genotypes map to just a few phenotypes, and the rest all map to different phenotypes.

These and other studies have shown the existence of many different phenotypes within the local neighbourhood of most sequences indicating that phenotypic changes are often directly accessible, and furthermore Schuster et al. (1994), Grüner et al. (1996a), Reidys et al. (1997) and Sumedha et al. (2007b) have also shown a phenomenon in the RNA genotype space that Schuster et al. (1994) called *shape space covering* where a high percentage (up to 80%) of all the phenotypes in the space can be found by changing as few as 20% of the positions from any random genotype in the space (values from Sumedha et al., 2007b). Fontana (2002) likened this finding to reducing the haystack of genotypic sequences in which to find a needle to just a small bale of straw. However, there is debate about whether this is a universal property of many–to–one genotype–phenotype maps, because its existence in protein lattice

models is questioned by Li et al. (1996), Bornberg-Bauer (1997) and Bornberg-Bauer and Chan (1999), but see Babajide et al. (2001).

Perhaps most importantly from a evolutionary point of view, it has also been pointed out by Sumedha et al. (2007b) that shape space covering assumes that an evolutionary search can wander arbitrarily away from a starting sequence, ignoring phenotypic changes. Furthermore, that to search through even a reduced genotype space containing a suitable fraction (their figure is 8.910^{-30}) of the possible sequences with a length of 100 bases would still take a realistically sized population of bacteria more than the age of the earth. Shape space covering gives us an interesting insight into the way that the map may be structured, but tells us little about the effect that this structure has on evolution.

Much of the work on the genotype–phenotype map uses haploid genotypes, simple point mutations and asexual reproduction. Factors such as dominance and recombination can be included, but muddy the clarity with which we can define local neighbourhoods. The same simplifying principles apply to the function mapping from genotype to phenotype, which in real life is far too complex to be able to predict accurately. Within a simple, haploid asexual (i.e. assuming no recombination) genotype–phenotype map, the distance between genotypes can be accurately measured in terms of the number of single point mutations.

This is particularly important when there are strong epistatic interactions between different sections of a genotype. If we return to the Maynard Smith's word example: whether a word continues to make sense when a particular letter at a particular position is changed is entirely dependent on the other letters around it in the word. For example, changing the 'o' to an 'a' in 'word' makes 'ward', but changing the same letter at the same position in 'woad' gives 'waad'. The effect that a genotypic change has on a phenotype often cannot be predicted by examining just one or a few loci, and therefore using a simple mapping function from each genotype as a whole can capture this kind of complex interaction between genotype and phenotype.

1.2.2 Neutral networks

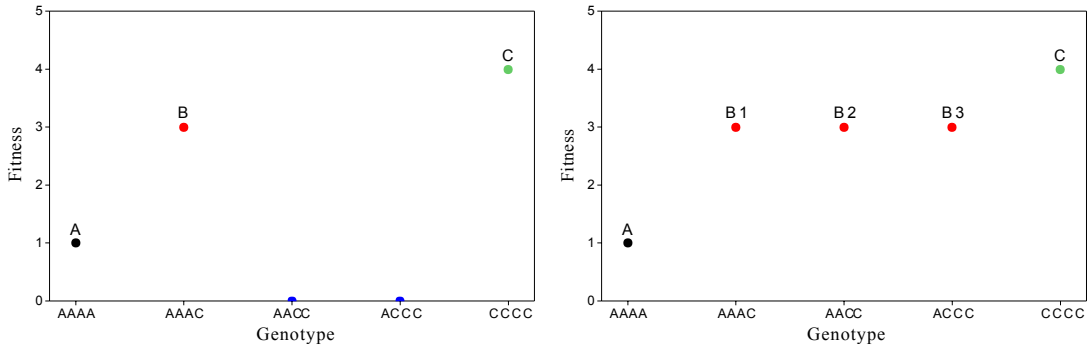
Maynard Smith (1970) talked about a network of connected genotypes which all code for viable phenotypes. The single large network has subsequently been broken down into a set of networks, where all the genotypes within each network code for the same phenotype. This was initially done by Lipman and Wilbur (1991) using a protein lattice model, but Schuster et al. (1994) first coined the term *neutral network* in relation to a model of RNA sequences mapping to the same secondary structure. Since

then, the existence of extensive neutral networks in genotype space have been shown in RNA (e.g. Schuster et al., 1994; Huynen et al., 1996; Held et al., 2003), protein lattice models (e.g. Lipman and Wilbur, 1991; Govindarajan and Goldstein, 1997a; Bastolla et al., 1999), virus models (e.g. Koelle et al., 2006) and Kauffmanesque 'NK' landscapes (e.g. Barnett, 1998; Newman and Engelhardt, 1998; Smith et al., 2002).

Many of these models have often been abstract or very mathematical, they have therefore not penetrated fully into wider evolutionary thinking. For this reason, here I present a simple explanation of what neutral networks are, and why they are important. I end the section by examining some different potential network structures, particularly pointing out those which have a limited or negligible effect on evolutionary dynamics.

First consider the simplest of networks: Where two discrete genotypes result in the same phenotype, they are neutral with respect of each other. If the second genotype also happens to be in the *local neighbourhood* of the first (that is the $3n$ mutant sequences that differ from the first by just a single base), the two can be grouped together into a neutral network because a simple point mutation can change one into the other. Huynen et al. (1996), Fontana and Schuster (1998a), van Nimwegen and Crutchfield (2000), Ebner et al. (2001), Smith et al. (2003) and Wolf et al. (2006) (among others) have suggested that neutral networks can have adaptive consequences. This is because if a population can drift across a flat neutral network in a discrete space, then it can potentially undergo many changes to its genotypic sequences without altering the phenotype. Each new neutral mutant brings the population into contact with more local neighbours, and hence increases the number of genotypes which can be searched. The more genotypes that become accessible as mutants, the higher chance that one of them will be adaptive.

This is the discrete equivalent of the idea of neutral drift along ridges in a continuous landscape (Fig. 1.6b c.f. Fig. 1.4b). The size and shape of the network will limit the neutral drift of a population until a fitter phenotype is encountered. It is this kind of *drift–based search* which Huynen et al. and others suggest provides an alternative form of evolutionary progression to either making an improbable multi-mutational jump, or mutating through a series of deleterious intermediates (Huynen et al., 1996). In terms of a fitness landscape, traversing a flat ridge is easier than jumping from peak to peak or descending from a local peak into a col or valley before climbing the higher peak on the far side.



(a) Simple mapping **(b)** Neutral-network

Figure 1.6: Each point on the x axis represents a discrete genotypic sequence. They differ from the sequences on either side by just one letter. Each genotype codes for a phenotype which is either viable (A, B and C) and has a fitness value associated with it, or is inviable. **a)** Starting from the left hand end, a point mutation at the 4^{th} position is required to mutate from $A \Rightarrow B$ (AAAA→AAAC). This is selected for because the fitness of B is higher than A. The intermediate sequences (AACC and ACCC) between B and C are inviable. B is very unlikely to mutate directly to C because it requires a number of simultaneous base changes in a single generation (AAAC→CCCC). **b)** A neutral network, where AAAC, AACC and ACCC all code for the same phenotype (B). Starting from the left again, a mutation from $A \Rightarrow B$ (AAAA→AAAC) is selected for. However, a mutation from $B1 \Rightarrow B2$ is also possible, as there is no selective advantage in either direction (AAAC↔AACC). Over a period of time, a population initially centred around $B1$ is likely to spread across the network to $B3$, and eventually mutate from $B3 \Rightarrow C$ (ACCC→CCCC). This is selected for, and will result in the population becoming centred around the fitter genotype CCCC (phenotype C). Any genotype with a direct link to a different phenotype i.e. A, $B1$, $B3$ and C is known as a *portal genotype*. Genotypes $B1$ and $B2$ (AAAC and AACC) can reach the fitter phenotype C via neutral mutations, but not directly.

1.2.2.1 Direct and indirect connections and network portals

Figure 1.6b shows an example of how neutral networks can connect distant phenotypes. In this case there are five genotypes coding for three phenotypes – A, B and C. An individual with genotype $B3$ (ACCC) has the same fitness but higher evolvability than $B1$ (AAAC). This means that $B3$ has an increased chance of undergoing a beneficial mutation in the future because it is *closer* on the network to the advantageous mutant C (CCCC). Although it holds no selective advantage in terms of its phenotype, an individual with a $B3$ genotype is more likely to be the ancestor of a descendant population than any that still have the $B1$ genotype. Any genotype which includes a different phenotype in its neighbourhood can be said to be *directly* connected to that phenotype and here I shall use van Nimwegen and Crutchfield's term *portal genotype* to describe that connection (e.g. A & $B1$ and $B3$ & C). If two sequences are connected by a series of single–mutant neighbours, then we can say that they are *indirectly* connected (e.g. $B1$ & C, because it is possible to move from

one to the other via single point mutations through B, but would require improbable multiple base substitutions to jump directly between them).

Neutral network structure in genotype space can perhaps be more clearly elucidated by using the analogy of a 'fitness sky scraper' rather than a fitness landscape. Consider a building, where each room (network) on a floor (phenotype) is connected to other floors by many set of staircases leading up and/or down (portals between networks). When a crowd of blind people (without guide dogs) arrive on the ground floor, the crowd starts to disperse across the floor. They can jump, but are more likely to fall out of a window, so most take one discrete step after another. Each step can take them closer to or further away from stairs, or even make no difference to the distance they have to go. However, once a set of stairs has been found to a higher floor, the individual(s) who found the steps can quickly leave all their compatriots behind. At the next level they multiply (where the analogy breaks down slightly!), and a similar process occurs until the top floor is reached. At each level, the shape and size of rooms on that floor, and the position of the staircases all have a profound effect on how quickly (if at all) a higher floor can be found.

In the last 10 years, neutral networks have been shown to percolate through genotype space in different model systems (e.g. Huynen et al., 1996; Grüner et al., 1996a; Bornberg-Bauer, 1997; Fontana and Schuster, 1998a; Aita et al., 2003), to such an extent that the sequence information can be completely lost, while the phenotype is still retained (Huynen et al., 1996). Furthermore, neutral networks can change the genotypic sequence in such a way that advantageous phenotypes, which were not in the initial local neighbourhood, become available after a number of neutral changes (Huynen et al., 1996; Schuster and Fontana, 1999; Smith et al., 2002; Crutchfield, 2002) (Fig 1.6b). This means that there is less chance of a population becoming stuck in a local fitness optimum and the landscape is less 'rugged' (i.e. full of local optima (Kauffman and Levin, 1987)) than might otherwise be predicted.

This kind of drift, where neutral mutations combine to create an adaptive change, can be seen as a type of epistatic interaction, where the combined (positive) effect of the mutations, is different from the sum of its parts (zero fitness effect). A change in the direction of an epistatic effect (in this case from neutral to advantageous) has been termed *sign epistasis* by Weinreich et al. (2005). In the example in figure 1.6, mutation from ACCC→CCCC is contingent on AACC→ACCC having previously occurred but together they code for a fitter phenotype. This sort of epistatic effect is made more transparent when considered within a genotype–phenotype map framework, and while figure 1.6 is useful as an illustration, showing the shape and structure of a neutral

network is limited in the two–dimensional illustration. In a three–dimensional map, it is possible to visualise epistasis as mutations in alternate dimensions (Fig 1.7).

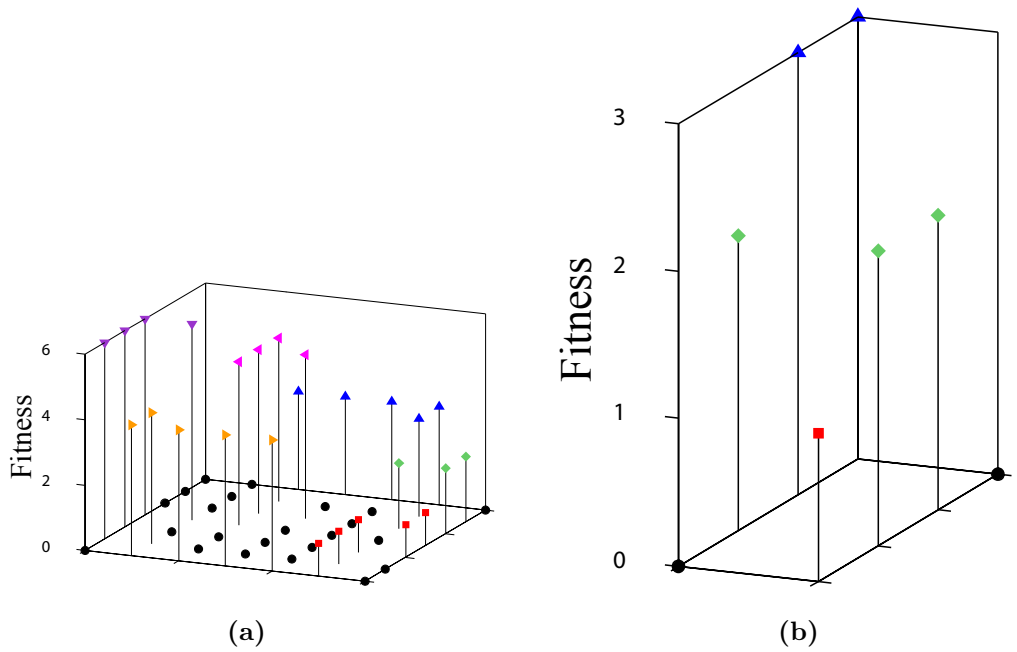


Figure 1.7: a) A more complex three–dimensional map with neutral networks of genotypes at different fitness levels. b) A subset of the (far right) section of the map, visualising how mutations in different dimensions can interact together to increase fitness, but individually are neutral.

Figure 1.7b (network ◆), shows that each mutation need not be contingent on another (as in Fig. 1.6b), but may occur independently and remain neutral. A step along either axis changes the genotype at a single position. When the second position is also mutated, and they therefore occur together, a fitness advantage is realised. It does not matter which mutation occurs first, either is neutral alone, and advantageous when in combination with the other. In a multi-dimensional map, mutation at each position in the sequence might be independent, but the resulting phenotype is contingent on certain mutations at other positions. This means that the speed at which evolutionary progression can occur is not based purely on the mutation rate (Kimura and Maruyama, 1966) but also on the connections within the underlying map. If the two mutations necessary for a phenotypic change can occur in any order, then on average they will occur together after fewer generations than if they have to occur in a particular order (see Chapter 5). An example would be mutations within an RNA secondary structure. If two changes are required in the unbound bases within the loop at the top of a hairpin (Fig. 1.5c) they could occur in any order. However, a

neutral change in the stem between bound bases is constrained by the need to maintain secondary structure(Fig. 1.5b). In this case, a change from an A-U base-pair to a G-C, must occur via a G-U intermediate, whose sub-optimal bonding can nevertheless maintain secondary structure. In contrast an A-C intermediate could not.

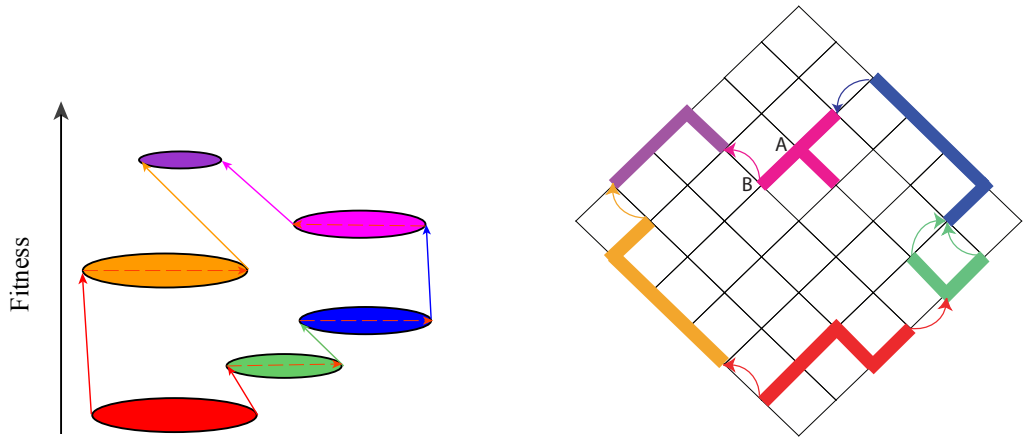
1.2.2.2 Visualisation

Although figure 1.7 uses a three–dimensional representation to highlight the epistatic interactions possible at two discrete loci, it is sometimes more instructive to highlight specific features using different clearer diagrams. In figure 1.8, I introduce two visualisations to the reader. They both represent the same space as Fig. 1.7a. The first is based on Crutchfield and van Nimwegen's 1999 visualisation of neutral networks as 'subbasins' connected by narrow 'portals'. This displays the genotype-phenotype map as a series of neutral-network discs ordered vertically by fitness. Arrows between discs indicate portal connections between networks. It is an attempt to produce a clearer illustration of the links *between* neutral-networks rather than the structure within them.

Figure 1.8b is a top view looking down on the network structures of figure 1.7a, and is based Wright's original fitness contour maps. In this representation, each node on the grid represents a different genotypic (DNA or RNA) sequence: each of these codes for a phenotype. If the node is coloured, the phenotype is viable; if empty, the phenotype coded is inviable. The more similar the sequences (but not necessarily the phenotypes), the closer they are on the grid. Sequences that have no nodes between them (e.g. A and B) are one simple point substitution apart. Sequences with the same colour have the same phenotype and have identical fitness. The coloured networks can be thought of as being contour lines on a topographic map, with fitness increasing out of the page.

1.2.3 Alternative network structures

There is evidence to suggest that neutral networks exist and percolate through large areas of genotype sequence space. If these networks share boundaries, mutations between them can be more available from some parts of the network than from others; however, there are arrangements of a many–to–one mapping where neutral genotypes have no overall effect on adaptive evolution. I review these ideas briefly here to give the reader an idea of the alternative shapes and structures possible within a genotype space. First, if sequences that code for the same phenotype are not adjacent,



(a) **(b)**

Figure 1.8: These two diagrams represent the same genotype–phenotype map from 1.7a. Each has its advantages, and both shall be used in an attempt to clarify the ideas presented. **a)** network structure is ignored but links between networks can be seen clearly. **b)** Network structure is clearer, but links between networks are less so. In this diagram B represents a *portal genotype* from pink to the purple phenotype.

they do not form networks, and there is no opportunity for drift-based searches of new areas. Second, if sequences which code for the same phenotype are adjacent to each other, but sequences which code for different networks are genetically distant from each other, there are no portals to find by drift–based search across a network. Third, if every genotype has access to the same set of viable phenotypes in its local network, drift–based search does not occur because each genotype on the network acts as a portal to the same set of phenotypes (Fig. 1.9). Figure 1.10a is a less extreme example of the structural example of figure 1.9a. The genotypes coding for a particular phenotype form small local disjunct networks within the space, rather than being part of one large network. This kind of break-up of networks coding for the same phenotype has been shown in phenotypes in the RNA genotype space (Reidys et al., 1997; Fontana and Schuster, 1998a), and is strongly influenced by the concepts of local neighbourhoods as described in chapter 2. The result is that each disjunct network can have a different set of connections to different phenotypes, and therefore different evolutionary potential (Fig. 1.10a).

In some circumstances the effect of a mutation can be conditional on mutations at other positions, as in Maynard Smith's word analogy. In the most extreme case of this kind of conditional epistasis , a mutation at one position changes the effect of a second mutation, from being deleterious to neutral, which in turn is necessary

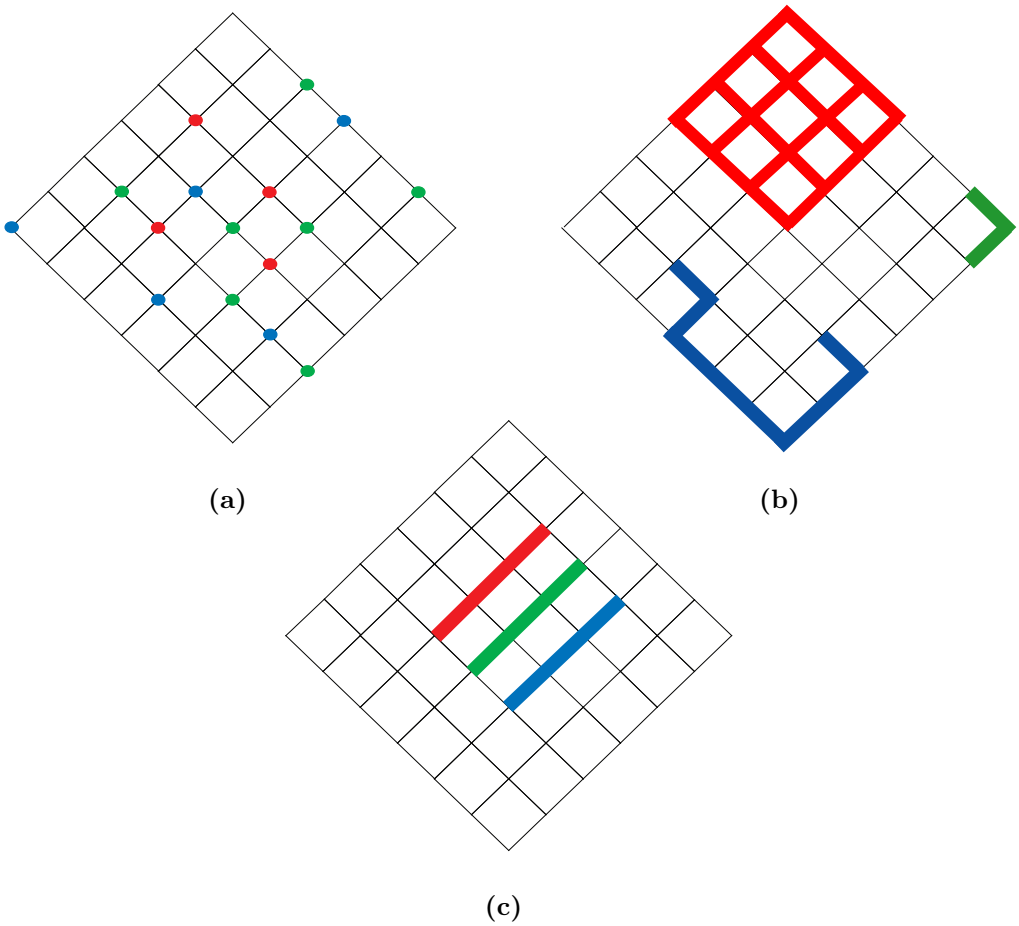
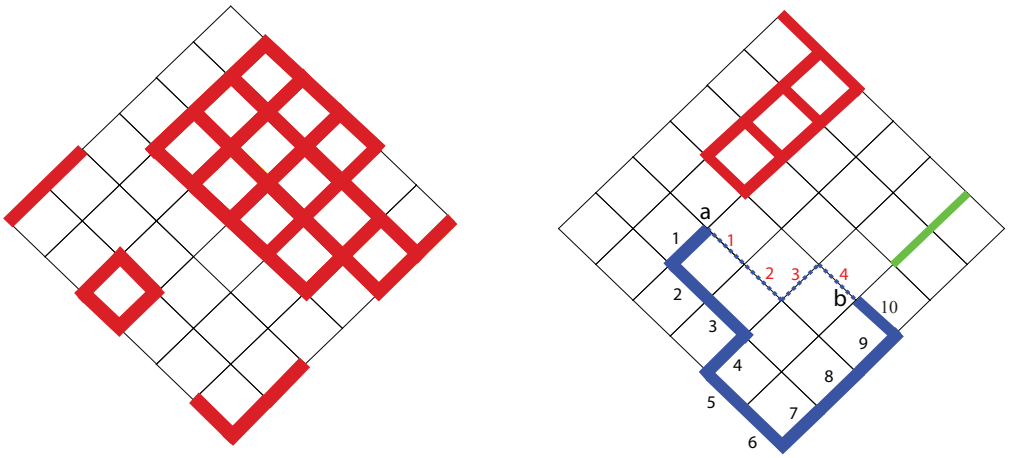


Figure 1.9: Three examples where neutral mutants have no effect on adaptive evolution: **a)** Neutral mutants with the same phenotype exist, but do not form networks within the map. **b)** Networks exist in the map, but the boundaries do not contact each other, making transitions very unlikely. **c)** Networks exist and contact, but each point is connected to the same set of neighbours, so if an adaptive step is possible it is immediately available.

to allow a further mutation at the original position (Fig. 1.10b). Consider again the words:

APE APT OPT OAT MAT MAN

The mutation of the E→T allows the mutation from A→O. Eventually this paves the way for the T→N. However, any mutation to an 'N' earlier would have resulted in an inviable 'phenotype'. Conditional neutral mutants are common in RNA genes (Fontana, 2002), and have been found in viruses (Quer et al., 2001). Therefore it is quite possible that in some sparsely populated networks the minimum path length required to reach a portal to a fitter phenotype is larger than the Hamming distance (Fig. 1.10b).



(a) **(b)**

Figure 1.10: More complex networks: **a)** Neutral mutants with the same phenotype form disjunct networks within the map. **b)** The path between two phenotypes across the network involves more changes than the total number of positions at which mutations can occur, i.e. the shortest path requires more changes than the Hamming distance. In contrast the shortest path between any two points on the red and green networks is always the Hamming distance.

While I used a two dimensional representation in figure 1.10b, it is important to note again that this is not a truly accurate representation of the genotype space: each position in the sequence is independent, and a mutation at that position can be to any base. In other words it is impossible to move more than one step in any one dimension in the genotype space, unlike in a traditional Wrightian landscape, and Fig. 1.10b above. Assuming that the probability of a transition between any two bases is equal, each must be equidistant from the other three. This cannot be drawn in fewer than three dimensions, forming a tetrahedron, and does not leave a dimension for a fitness scale as in the previous diagrams. The most simple example can be seen in Fig. 1.11a where in one dimension A and U seem to require intermediate steps through C and G.

Given that it is difficult to visualise long paths through a space where the maximum distance is one change in each dimension, it is interesting how closely the ideas and diagrams laid out in two and three dimensions in this section are mirrored by the results presented from interrogating the RNA genotype–phenotype map in the remainder of this thesis.

A significant barrier to establishing the importance of neutral networks in real life lies in the fact that it is difficult to use real or simulated organisms. This is because there is a combinatorial explosion of genotypes and interactions even when viewed

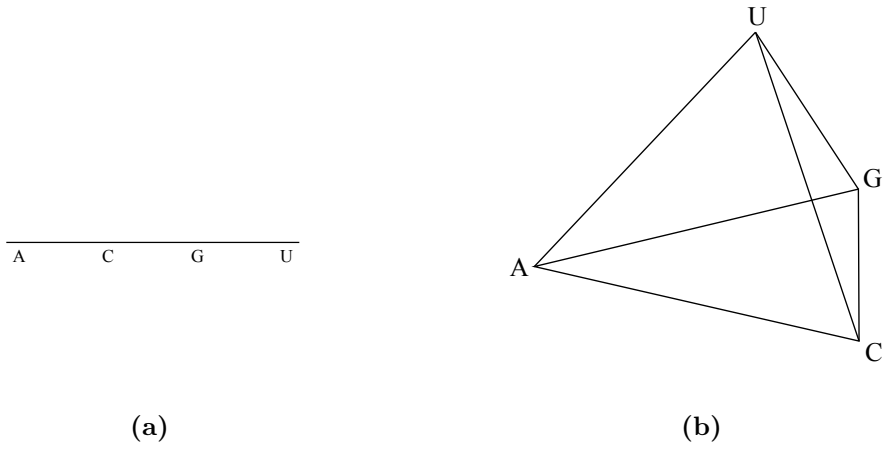


Figure 1.11: a) 2-D landscape: no matter which order the bases are positioned, some will always be further away from each other than others. In this case the distance between A and U is 3 units, whereas A and C is just 1. **b)** 3-D landscape: each base sits equidistant from every other. The chance of any base mutating to any other is equal.

at a coarse scale. In addition, sequencing large sections of genetic code from many individuals is infeasible, meaning that the method of evolutionary innovations can be missed if the intermediate steps are rare in the population (but see Poelwijk et al. (2007) for a review of recent work on tracking intermediate forms).

The function mapping realistic genotypes to phenotypes is also still far beyond the reach of current research, even at such a relatively simple molecular level as protein folding. In fact perhaps the largest trade–off between simplicity and reality is in the use of a limited genotype–phenotype function, potentially leading to significant differences between the impact of neutral mutations in model systems and their affect on real organisms. I shall postpone a discussion of the effect of using a simple mapping function, until I consider all of the assumptions of the genotype–phenotype map model, after the main body of results have been presented (see section 6.1). Despite the difficulties, numerous attempts using a variety of models have been made to assess the effect of neutral mutations, and these are now briefly reviewed.

1.2.4 The development of neutral network research methods

Computational simulation studies (Fontana and Schuster, 1998b; Ebner et al., 2001; Smith et al., 2002, 2003), mathematical theory (van Nimwegen et al., 1999; van Nimwegen and Crutchfield, 2000, 2001; Stadler et al., 2001), molecular biochemistry models (Grüner et al., 1996a,b; Göbel, 2000; Babajide et al., 2001; Kospach, 2003; Bardou and Jaeger, 2004) and micro-organism studies (Elena et al., 1996) have

been used to hypothesis or model the existence of neutral networks and/or assess their effect on evolution. The idea of neutral genotypic changes triggering occasional phenotypic shifts provides an appealing and parsimonious explanation of the punctuations seen in many natural and artificial evolutionary systems. However, there are huge challenges to be faced when monitoring or modelling such a complex system. For example, if the intermediate generations are not monitored between novel fitter phenotypes, it is impossible to distinguish between a rare beneficial mutation occurring after a long period of time, and a population traversing a neutral-network (reviewed Poelwijk et al., 2007).

This lack of transparency is particularly apparent with the top–down approach necessary when studying real organisms (e.g. Elena et al. (1996)). Even if sufficient changes occur over a monitored time-scale, whole organisms provide such a complex set of interactions between their genes, and within their environments, that it is usually impossible to tease out sufficiently whether a change is adaptive in the way we understand it to be (Gould and Lewontin, 1979). It is also virtually impossible to track each individual, and therefore to track exact genetic evolutionary trajectories.

Most neutral network research has thus been based on mathematical or simulated evolution models, with a focus on extending the effectiveness of simulated evolutionary systems by more closely mirroring 'real life'. It is telling that much of this work has only really been of interest to the Artificial Life (ALife) community. Barton and Zuidema (2003) have pointed out that this will remain the case until the ALife researchers "take more seriously the tools and insights from population genetics", and the converse also applies. Misunderstandings between 'mathematicians' on one side, and 'biologists' on the other, have led to each regularly talking past the other (see Wilke, 2005; Grafen, 2007). This problem is only slowly being rectified (Wilke, 2005).

Among ALife researchers the assumption that neutral networks increase evolvability or enable escape along a ridge in the adaptive landscape that would otherwise have been a local fitness optimum has often been taken for granted, based on a physical rather than biological sciences framework, and/or based on incomplete empirical data. As with all models there is a trade–off between biological realism, simplicity and approachability. We are walking a tightrope between these factors, with proponents on each side arguing that their method offers the greater insight.

Most artificial neutral network research tends to take one of two approaches; the first uses simulations of digital organisms, the second uses maps explicitly.

1.2.4.1 Digital evolution

With increasing computing power the kind of complex probabilistic calculations necessary to study selection over many generations have become easier to perform. Much effort has focused on using 'digital organisms' to simulate evolution (e.g. Dawkins, 1986; Ray, 1991; Adami, 1995; Yedid and Bell, 2002). These models can offer insights into various phenomena involving interactions too complex to study in biological models (e.g. epistatic effects: Lenski et al. (1999), evolution of parasites: Ray (1991) and evolution of mutational robustness: Edlund and Adami (2004); Comas et al. (2005); Elena et al. (2007)). The advantage of evolving digital organisms without explicitly modelling the genotype space is that evolution can be open ended or use a complex genotype–fitness function where fitness is based simply on the successful production of offspring. This can mean that many of the problems which underlying biologically based model systems also apply here. Digital organisms can present a very useful model as more controlled and repeatable system than using real ones, but the complexity which makes them ideal for studying higher level phenomena like the evolution of parasites means that they are not the ideal tool for researching the underlying relationships between genotypes and their respective phenotypes. The alternative approach focuses on simpler models where the genotype–phenotype map is specified.

1.2.4.2 Mapping approaches

A simple genotype–phenotype mapping approach can be used to investigate the underlying genotype space by considering the *possible* variation within the space. However, because of the vast number of genetic combinations, it is difficult to form a model except when using a very simple mapping function. An explicit mapping is therefore unable to provide such rich and complex dynamics as are observed in real or digital organism studies.

One direct mapping approach designs a mathematically–based space or graph, and attempts to draw out important overviews from analysing that graph. One example is based around Kauffman and Levin's N-dimensional hypercube (the 'NK' model – which models the number of loci (N) and the number of epistatic influences on each locus's fitness (K)). An other example are those of Gillespie (1984, 1991) and Orr (2002, 2006a,b). They have the advantage of being abstract, but tunable and can give an overview of broad dynamics as well as exact solutions to the questions posed. However, these kinds of mathematical models inevitably come at the price of

biological realism, and it has been argued that they may be too far removed from biological evolution to allow conclusions to be applicable. For example Kauffman and Levin (1987) and Kauffman (1993) first introduced the NK model for *rugged* adaptive landscapes i.e. those with many local optima without neutrality. The model has been extended to include neutrality by Barnett (1998) and Newman and Engelhardt (1998), but according to Geard et al. (2002) the results of these have "significantly different structural properties from each other".

A different approach uses molecular sequences mapping genotypes to RNA or protein based phenotypes. There is a pay–off between using a large, well–defined map, across which evolutionary steps and populations can be simulated (Huynen et al., 1996; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Ebner et al., 2001; Forster et al., 2006), and using a smaller space in which every single genotype can be exhaustively calculated to assess the structure and shape of the neutral networks within the map (Grüner et al., 1996a,b; Göbel, 2000; Kospach, 2003).

Within the exhaustively calculated map models, the number of sequences involved in even a small dataset has limited most previous work to considering the statistical properties of the phenotypic distributions of sequences, and the abstract properties of networks such as their size and the number of component parts. On the other hand, simulations across neutral networks have the benefit of using longer, more biologically plausible sequences, at the price of a lack of precise knowledge about the space, because even multiple runs of a simulation can cover only a tiny fraction of it.

Some approaches have combined simulation maps with analytical studies, notable successes have been the models of van Nimwegen et al. (1999), van Nimwegen and Crutchfield (2000) and Crutchfield (2002). Though they can be initially quite inaccessible to a non-mathematically trained biologist, they use a very simplified genotype–to–phenotype map to simulate and analytically calculate the effects of population size, mutation rate and the size of a network on the number of generations that it takes a population to drift to an advantageous phenotype, and conclude that it takes significantly fewer generations to drift across a neutral network than to traverse even a short or shallow valley of less fit intermediates.

1.2.4.3 The RNA map

RNA models have been favoured by modellers, often in relation to simulated evolution of tRNA shapes (e.g. Fontana and Schuster, 1998a). They have been used to show many evolutionary phenomena e.g. punctuated evolution (Fontana and Schuster,

1998a; Forster et al., 2006), evolution of mutational robustness (Ancel and Fontana, 2000; Wilke, 2001a), epistatic interactions (Wilke and Adami, 2001) and the distribution of advantageous mutations (Cowperthwaite et al., 2005; Sumedha et al., 2007b). Within these models, secondary structure is used as a proxy for fitness. Each secondary structure generally maps to fitness in a one–to–one manner, and is calculated using a folding algorithm *in silico* (e.g. Zuker and Stiegler, 1981; Hofacker et al., 1994; Knudsen and Hein, 1999).

RNA secondary structure based models have several advantages over their protein lattice model cousins. Accurate protein models have a much larger search space (20^n), which makes exhaustive enumeration infeasible. Protein lattice structure is also further removed from biological reality than the secondary structure of RNA folding models, because the bonding patterns of bases are more regular and easy to predict than those of proteins, and thus require fewer computational resources.

The major advantage of using RNA to model evolution is that it combines the heritable sequence coding nature of DNA with the structural complexity of proteins in a single molecule. The simple biophysical minimum free energy (Mfe) genotype–phenotype mapping seen in an RNA model reduces the number of steps at which evolutionary forces and environmental effects can have an impact compared to the more complicated pathway from genotype to phenotype for proteins. The Mfe algorithm allows the precise prediction of the effect of any mutation on a major aspect of the sequence's phenotype; but more than that, it allows the precise prediction of the effect of any epistatic or compensatory interaction between genetic mutations.

As well as the tRNA models mentioned at the start of this section, simulations of small sections of RNA folded into hairpin loops (with or without additional bulges) can effectively model the small loops of 18-30 bases found in the untranslated region (UTR) of protein coding mRNAs. These secondary structures play an important role in binding with certain regulatory proteins controlling the translation of mRNA (Harrell et al., 1991; Kikinis et al., 1995; Addess et al., 1997; Allerson et al., 2003). They exhibit neutral base changes within the stalk of the loop, and different binding affinities based on structure *in vivo* (Hall and Williams, 2004). The length of these regulatory loops are just beyond those used in exhaustive searches such as Göbel (2000) and Kospach (2003), and the model presented in this thesis (up to 16 bases).

At these lengths, exhaustive RNA genotype-phenotype maps are reaching the stage where in the not–too–distant future, we shall be able to predict the evolutionary options for RNA shapes *in vivo*.

For these reasons, the work in this thesis is based on an exhaustive search of the RNA map. The methods of calculating the map, and the preliminary results generated from finding the networks within the map will be outlined in Chapter 2. However, first I summarise the main points of this chapter, with reference to the approach taken in the rest of this thesis.

1.3 Summary

There has been an increasing amount of research into adaptive evolutionary dynamics from a discrete molecular genetic perspective. Molecular data have shown that in most circumstances the genetic code provides at least some degeneracy to produce neutral changes. This has led to the hypothesis that neutral mutations can play a role in adaptive evolutionary pathways.

Models based on biochemical or biophysical algorithms provide a simple, repeatable function to map the effects of genetic mutation on phenotype and fitness. Many of these have a large amount of degeneracy between the number of genotypes and the number of phenotypes they code for. The result is the existence of *neutral networks* within a genotype–phenotype map, which can act like ridges in a mountainous landscape by connecting areas of high fitness together and providing a way of negotiating a genetic neighbourhood in which there were initially no further accessible adaptive mutants. A population drifting across the map can build up a series of small neutral genotypic changes, until a particular combination of mutations interact epistatically to produce a different phenotype, and therefore fitness, which was not accessible before.

In this way, neutral steps can be a necessary intermediate for future adaptive change, even though they are not selected for directly in the generation that they occur. The existence of neutral networks seem to be a very robust phenomenon within most biochemical genotype–phenotype maps at least. RNA $\Rightarrow$ secondary structure contains the same kinds of neighbourhood degeneracy seen in the DNA $\Rightarrow$ RNA $\Rightarrow$ protein map, and therefore exhibits similar dynamics to those seen in the protein lattice models.

The concept of neutral networks and the effect they can have on adaptive evolution has been relatively simple to prove abstractly, and evidence is overwhelming as to their existence in models based around mathematical landscapes, and RNA and protein lattice folding models. However, whether they are important in a standard genetic system, and the role they play in biological evolution is still very much open to debate.

For example, whether the environment remains constant for long enough time periods for evolution across a map or landscape to hold has been questioned from the first (Fisher, 1930), but in many models is often not even mentioned as an assumption.

The RNA genotype space has been shown to exhibit *shape space covering*, where all common phenotypes are found within relatively few mutational steps of any random sequence in the space. The neutral networks have been shown to be large, and relatively pervasive – a set of genotypes vary massively at many different positions, while still retaining the same phenotype; however, their potential effect on biological evolution has been limited to a series of observations within simulations and have yet to be mapped out within an exhaustive model.

In this thesis I aim to quantify the biologically inferred effects of neutral networks within an exhaustive map model. The need to maintain a functional phenotype is as important biologically as taking into account the geometric distance is between any two random sequences in the space. When this kind of phenotypic consideration is included in a model, far greater path lengths through the space are recorded than might be predicted by purely geometric statistical analysis of the landscape.

With a small but exhaustively searched map, it is possible to interrogate the structure of the map in a different way to that of most other models, which are based on random walks or sampling subsections of space and where all combinations of bases at all positions are not calculated. Because most previous models are not exhaustively calculated, much of the work on neutral networks so far has concentrated on the most common phenotypes found in the space, even though there is often an assumption that fitter phenotypes are also more rare (van Nimwegen et al., 1999; Crutchfield, 2002).

As the very existence of life is highly improbable, taking samples or subsections of the space may miss the rare or improbable results which are in fact amongst the most important in evolution precisely because they do only happen very rarely. Using an initial exhaustive search of the space means that the chance of missing rare events is reduced, and can help to inform us as to where to target, and how to assess, the results of more complicated simulations. For example, in chapter 4, I perform simulations through the exhaustively mapped space which allow me to calculate the fraction of evolutionary trajectories that are able to reach the global optimum from a random starting point within the map. This would be far more difficult in a simulation where the entire space is not mapped, because the range of possible phenotypes and therefore fitnesses are not known or extremely difficult to calculate.

The results presented in the following chapters provide a glimpse of the level of biological complexity to be considered when trying to combine fairly abstract discrete genetic spaces with classical evolutionary theory. In attempting this, I also aim to extend the genotype–phenotype map framework and make it more accessible to biological scientists, in the hope that it will provide a resource for answering questions of evolutionary theory, population dynamics and adaptive landscapes.

Chapter 2

RNA genotype–phenotype map

2.1 Introduction

A neutral network can be defined within a discrete genotype space as a set of unique genotypic sequences each mapping to the same phenotype *and* each connected to at least one other genotype in the network by a simple mutation. They arise because of degeneracy in the genetic code, and have been shown to exist in many model genotype–phenotype maps (e.g. Govindarajan and Goldstein, 1997b; Bastolla et al., 1999; Ebner et al., 2001; Aita et al., 2003). It has been postulated by Huynen et al. (1996); Fontana and Schuster (1998a); van Nimwegen and Crutchfield (2000) and Smith et al. (2002), among others, that these networks might provide a method of escaping from apparent phenotypic stasis by random drift across the network.

In this chapter I shall lay out the methodology behind calculating the RNA genotype–phenotype map presented here. After giving a brief account of the mapping function, designed by Hofacker et al. (1994), I discuss the relationship between mutants within the genotype space, and define the type of mutation which constitutes a simple single–step jump, highlighting the effect this relationship has on local neighbourhoods.

The next section introduces the algorithms and computational methods that I have used to calculate the resulting neutral networks . Within this section, two main strategies are developed in an attempt to cope with the constraints imposed by the huge spatial complexity of such a large dataset and the resultant high computation time requirements. The first is a new and somewhat indirect method of calculating neutral networks based on sorting the sequences alphabetically. This algorithm is highly efficient at calculating neutral networks , especially at longer sequence lengths, because it only depends on the number of sequences that code for each phenotype. However, it is less good at establishing how networks are connected to each other across the space.

In contrast, the second method is a fast and efficient data array based on using decimal integers as a space saving technique to store sequences, combined with bitwise operations to quickly calculate local neighbourhoods. As sequence length increases, this methods becomes less efficient at calculating neutral networks than the sorting method. However, because the whole map is instantly accessible, interrogating the map to generate interesting results is much simpler. Readers not interested in the details of the computational methods employed to calculate the neutral networks, are invited to skip this section and resume at section 2.3.

In the second half of the chapter I introduce the network structures within the genotype–phenotype map in terms of the number of sequences, neutral networks and phenotypes. The aim here is not to attempt to fully characterise each network's shape and connectivity, but instead to give the reader a feel of how they are structured.

One of the principal issues with analysing such a large set of data is providing measures which simplify the vast quantity of information into something clear and meaningful. In the previous chapter, I used pictorial visualisations to explain some of the concepts, and these are combined here with histograms and graphs of the data which summarise the space. The graphs highlight relevant information such as network size distributions, shape and connectivity. To this end, some of them fill the role of elucidating the space itself rather than explicitly tackling a biological question.

However, it always pays to bear biological significance in mind – Schuster et al. (1994), Bornberg-Bauer (1997), Göbel (2000), Aita et al. (2003) and Sumedha et al. (2007b) have all shown that most phenotypes can be found within a few mutations of any random sequence in the space. While this is an interesting aspect of the geometry of the space, Sumedha et al. (2007b) point out, correctly in my opinion, that shape space covering is actually probably of little relevance to evolutionary innovation. Evolutionary pathways are not simply based on the geometry of mutational connections; it is by following the restrictions in the way that genotypes map to phenotypes that we can see that evolving towards a new phenotype follows the structure imposed by genotypes forming extensive neutral networks.

With this in mind, I then shift focus onto the more interesting questions that I ask in later sections and chapters, such as how connected the neutral networks are to each other? How well connected neutral sequences are to each other within a network? And what increase in evolvability, if any, is gained by being part of a large network which provides indirect access to a greater range of phenotypic options. Throughout the results section I draw out a few general trends across all sequence lengths, to suggest how the space might be extrapolated to longer sequence lengths.

The final section briefly compares the results of this model with similar maps of other authors (Grüner et al., 1996a,b), Reidys et al. (1997), Fontana and Schuster (1998a,b), Göbel (2000), Kospach (2003). It becomes clear that the qualitative similarity of the neutral networks described are robust to many changes in the initial parameters of the mapping function. The similarity in networks extends to those seen in protein lattice models (Govindarajan and Goldstein, 1997a; Hirst, 1999; Aita et al., 2003) and provides evidence that genotype–phenotype map models may be widely applicable within evolutionary systems, especially at the molecular level.

2.1.1 RNA genotype–phenotype function

Throughout this thesis the mapping function uses the `RNAFold` application designed by Hofacker et al. (1994), which in turn is based on an RNA minimum free energy (Mfe) folding algorithm (Zuker and Stiegler, 1981). It uses biophysical bond energy parameters to predict the likely base–pair bonds, and hence the lowest energy secondary structure of a given sequence of nucleic acid. By using secondary structure as a proxy for phenotype, these folding algorithms provide a simple, efficient and biologically grounded genotype–phenotype mapping with which to study evolution in a discrete space. What they do not provide is the level of complexity seen in the genotype–phenotype pathways commonly found in living organisms. Regulation of expression, phenotypic plasticity, and gene multiplication effects on phenotype are all excluded from this model for example.

The RNA sequences which perform regulatory functions within cells are generally longer than those mapped here. However, computation time and memory restrictions have limited this study to sequence lengths of up to 16 bases long. A factor in favour of short chain lengths is that prediction of secondary structure is more accurate: there is usually only one possible structural configuration or at most one or two other viable alternatives with similar minimum free energy (Göbel, 2000). Secondary structure accounts for the majority of an RNA molecule's free energy (Huynen et al., 1996) and furthermore, at short sequence lengths tertiary structures are very rare because the chain is not long enough to fold back on itself to form anything more complex such as pseudoknots. This simple, limited mapping can also be seen as a starting point from which more complex models can be derived.

2.1.1.1 Genotype

On the genotype side of the map, the space is made up of a complete set of RNA sequences. Each sequence has a unique combination of bases (Adenine, A; Cytosine, C; Guanine, G; Uracil, U), giving 4^n genotypic sequences per space (where n is sequence length). This means the key limitation in any exhaustive genetic model is the exponential increase in the number of genotypes as sequence length increases. It is clear that anything but short sequences will produce such a large number of base combinations that exhaustive enumeration becomes impracticable.

2.1.1.2 Phenotype

This model follows most other studies (e.g. Schuster et al., 1994; Grüner et al., 1996a,b; Bornberg-Bauer, 1996; Cupal et al., 2000; Wilke et al., 2003) in using secondary structure as the only factor in defining the phenotype and therefore fitness. The resulting set of phenotypes is orders of magnitude smaller than the 4^n sequences which map to them. *In vivo*, Hall and Williams (2004) showed that specific unbonded bases in the loops and bulges of an RNA structure have an important impact on the molecular recognition of a phenotype, but they also showed that changes in *structure alone* can significantly influence the binding affinity of RNA molecules. Therefore I suggest that using secondary structure alone is a valid simplification.

The alternative is to combine Mfe structure with other phenotypic factors. For example Cowperthwaite et al. (2005) used a combination of structure and thermodynamic stability. However, I suggest that using a highly simplified phenotype criterion is not just valid but valuable, because it reduces the complexity of epistatic interactions to a manageable level. As mentioned in the last chapter (see section 1.2.4.3) modelling interactions solely in the form of base–pairs allows us to calculate the epistatic interactions between bases in the genome in a highly regular way. Using this simple starting point means it is possible to highlight the most interesting and important points to focus on when computing power allows comparison with a particular known *in vitro* or *in vivo* model of RNA. Eventually, as protein structure prediction becomes more reliable and less computationally expensive, it may also become possible to use this framework to assess the evolutionary map between DNA and protein.

2.1.2 Mutations within genotype space

The mutations considered within this model are also the simplest possible: single nucleotide point substitutions. It gives a local neighbourhood of $3n$ sequences, each neighbour differing from the original sequence by one base at one position, similar to Maynard Smith's word analogy (Section 1.1.3). This contrasts with many other RNA models (e.g. Grüner et al., 1996a; Reidys et al., 1997), including those closest to this work (Göbel, 2000; Kospach, 2003). They include double (complementary) mutations at paired positions as neighbours. The biological and methodological reasons why I have chosen not to consider complementary mutations are outlined now.

First, across *Drosophila* species Kirby et al. (1995) found that mRNA regions with conserved secondary structure showed little linkage disequilibrium. They postulated

that this was due to the selective disadvantage of an initial mutant, which was then purged from the population before a compensatory mutation could occur. This points to so called 'base–pair' mutations being nowhere near as common as single nucleotide substitutions in natural populations. I argue therefore, that including base–pair mutants as part of the same *neutral* networks as single nucleotide mutants is unjustified. In contrast Göbel (2000, pg. 27) argues that base–pair mutants should be included because:

> [...] any primary mutation is likely to create a mismatch, in which case any compensatory mutation which restores the correct pairing will have a very strong effect, and will be selected for.

This is true, however there is also the possibility of the primary mutation leading to a mismatched phenotype which is actually fitter. This would result in any compensatory mutation restoring the original phenotype at that point being selected against! Furthermore, one can argue that a 'compensatory' mutation which leads to a fitter *different* phenotype than the original will also have a very strong effect and will be selected for. By this justification, every double mutant which codes for a different (fitter) phenotype must also be considered an accessible neighbour, potentially increasing the size of the neighbourhood from $3n$ to $\frac{9n!}{2[(n-2)!]}$

This leads on to my third reason for not considering base–pair mutations: that if they are included, the size of the local neighbourhood varies according to the number of base pairs in the phenotype.

If base–pair mutations were to replace single point substitutions at the paired positions in a sequence, the neighbourhood size is reduced in line with the number of base pairs (x), giving $3(n - x)$ neighbours instead of $3n$. The paired positions share the same reduced set of 3 alternate mutants (Table 2.1, column 3 'Replaced'). It also leads to single point mutants not being considered at paired positions. The result is that many alternate phenotypes which lie just one simple single nucleotide point mutation away are ignored as a neighbour.

In contrast, if base–pair mutations are counted in *addition* to the possible single point mutations, the size of the neighbourhood increases to $3n + 3x$ (Table 2.1, column 2 'Extra'). Under this scenario it is then necessary to assign a probability to paired positions for when a mutation involves a single base change, and when its complement is also mutated. Furthermore, having more than one mutation type at some positions and not others (e.g. those in a loop region) means that mutation rates must be different at different positions in the sequence. If there is an equal chance of

	Neighbourhood of AAU — phenotype — (.)[a]		
Position	Single substitution	Extra	Replaced
1	AAA	AAA	-
	AAC	AAC	-
	AAG	AAG	-
2	ACU	ACU	ACU
	AGU	AGU	AGU
	AUU	AUU	AUU
3	CAU	CAU	-
	GAU	GAU	-
	UAU	UAU	-
1&3	-	GAC	GAC
	-	CAG	CAG
	-	UAA	UAA

Table 2.1: The local neighbourhood of the genetic sequence AAU under different 'single–mutation' definitions, where the phenotype includes one base pair between the 1st and 3rd bases. 'Extra' means that base–pair mutants are considered in the neighbourhood in addition to single point substitutions, while 'replaced' means that only mutations that do not disrupt the structure are considered, but are not limited to a single point substitution.

[a]This phenotype could never form, because a loop must consist of a minimum of three unbound bases, but is shown here for simple illustrative purposes

mutating at any one position, the chance of getting a particular mutant in a paired region is lower than at other positions, because there are more alternatives. However, if the chance of mutating to each neighbour is constant, the probability of mutation at a particular position is higher at a paired one. Whether base–pair mutations are considered instead of, or in combination with, the mis-match mutants at a position, the number of neighbours changes with the number of base pairs in the phenotype.

If base–pair mutations were biologically plausible as part of an extended neutral network, then it would be computationally feasible to consider networks with different local neighbourhoods and internal mutational dynamics. However, when there is a biological reason not to include them, doing so risks muddying further the already complex inter-relationships of the networks within the space. In the future, neighbourhood relations could be extended to include other more plausible types of mutation such as deletions/insertions, inversions and duplications and would be a natural progression of the model. Recombination of genotypes has also not been considered in this mapping, but could easily be included in a future model.

2.1.3 Network neighbourhood

The local neighbourhood of a genotype is defined as the $3n$ sequences that differ from it by a single point mutation (Section 1.1.3). Any genotype or phenotype within the local neighbourhood of a genotype was defined in the last chapter as being *directly* connected. Here, an extended *network neighbourhood*, is defined as all the genotypes or phenotypes which are *indirectly* connected. That is, there must be at least one direct connection from a sequence somewhere in the network.

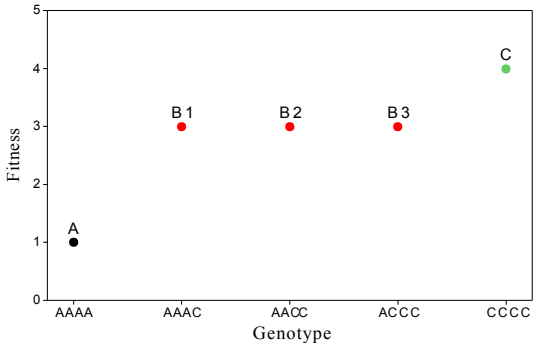


Figure 2.1: Genotype $B2$ has a *local* neighbourhood consisting solely of neutral neighbours $B1$ and $B3$. However, the network B contains portals to A and C ($B1$ & $B3$ respectively) meaning that the *network* neighbourhood of all the genotypes in B contains A & C. We can say that $B2$ is connected to A and C indirectly.

At this level, it only takes one direct connection between two portal sequences in different networks to indirectly connect all the sequences between those networks. This means that even if an individual genotype has no alternative phenotypes in its local neighbourhood, it may have access to many alternative phenotypes via neutral drift in its *network neighbourhood* (Fig. 2.1).

The next section discusses the methods used to calculate the neutral networks in the space, a prerequisite of finding the neighbourhood of a network.

2.2 Neutral network finding methods

The genotypes used in this model map are short sequences of RNA (10–16 bases long). They are folded *in silico* into their Mfe structure using the `RNAfold` application (vers. 1.5beta-15) (Hofacker et al. 1994, `http://www.tbi.univie.ac.at/RNA/`). It was parametrised for all sequence lengths with default values and with a temperature of $30°C$. Later the maps were recalculated with a temperature of 37°C for lengths 10-14. For the data set at a temperature of 30°C text files containing the sequence

and the folded phenotype were obtained from Alexis Gallagher. For each sequence length, the plain text file contained a complete alphabetical list of sequences. Each line consists of sequence, followed by the phenotype expressed using standard dot–bracket notation, where a period codes for an unbound base, and parentheses for paired bases:

$$\mathrm{CC}\mathbf{C}\mathrm{CCCCGGGGGGG} \Rightarrow (((((....)))))$$
$$\mathrm{CC}\mathbf{A}\mathrm{CCCCGGGGGGG} \Rightarrow ((.((....)).))$$

The data for the later 30°C sets was also obtained directly via a Java native method to the `RNAFold` C classes. All the software was coded in Java (version 1.6.0) and tested and run on an Apple eMac 1Ghz, with 1GB RAM (Random Access Memory). The sorting algorithm was calculated on the National Grid Service (NGS) primary clusters on dual processor Intel Xeon 3.06 GHz nodes with 2 or 4GB memory, running RedHat ES 3.0 (`http://e-science.ox.ac.uk/ngs/`). The integer array was calculated on a Compaq desktop with an Intel Pentium 4 2.66Ghz processor, and 768MB of RAM. Only the results for the even sequences are presented below, as there was no qualitative difference between even and odd sequence lengths.

Finding the neutral networks in such a large genotype space involves keeping track of large amounts of data. It is not, therefore, a trivial task. During this project I considered two main methods for tackling the problem: a neighbourhood search and a sorting algorithm. The neighbourhood search was used in the integer array algorithm as well as the initial phenotype set methods outlined below. I shall discuss the major benefits and drawbacks of each. Readers interested only in the results are advised at this point to skip to section 2.3.

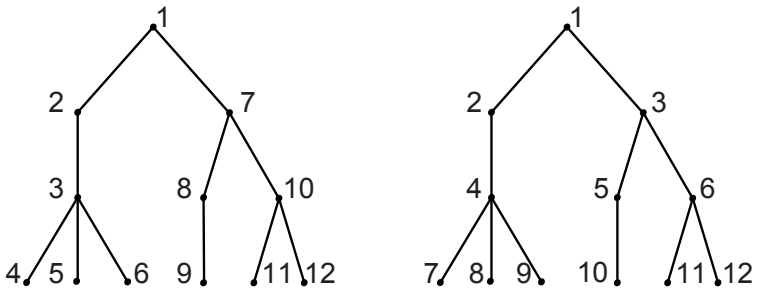
2.2.1 Neighbourhood searches

Neighbourhood searches involve calculating the one-mutant neighbourhood of a sequence and recording those neighbours that are neutral as part of the neutral network. This is repeated until all the neighbourhoods for each new addition to the network have been examined. There are two different methods of conducting this type of expanding search, depth–first and breadth–first.

1. **Depth-first search** The depth–first search pattern is outlined in Figure 2.2a. It involves searching the neighbourhood of the first sequence until the first neutral neighbour is encountered. At this point, the search switches to the neighbourhood of that neighbour. The process is repeated until a sequence is

reached that has no neutral neighbours. The search then backtracks to the last sequence whose neighbourhood has not been exhaustively searched, and continues from there.

2. **Breadth-first search** Figure 2.2b shows by contrast a breadth–first search, in which all of the neutral sequence neighbourhoods are exhaustively searched at each Hamming distance from the original sequence before moving on to the next level.

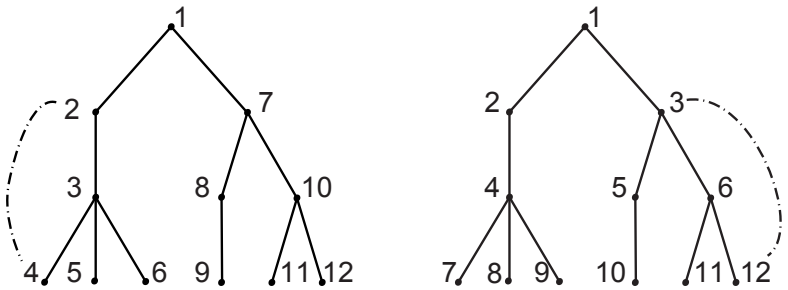


(a) Depth–first search **(b)** Breadth–first search

Figure 2.2: a) Depth first search, where the first neutral neighbour found is pursued until all new options are exhausted, before returning to the last incompletely searched neighbourhood. **b)** Breath first search, where each level of removedness is completely searched before progressing.

These methods are particularly effective on tree shaped structures, in which there are no connections to nodes that have been encountered before. However, if the structure is a network with back connections between nodes, then each searched node must be recorded and monitored. This is necessary to avoid returning to and researching nodes that have already been searched before. If searched nodes are not recorded, an infinite loop is constructed, and the network calculation never completed (Fig. 2.3).

Herein lies the computational problem with both neighbourhood searches: when searching large sets, recording and monitoring all the sequences encountered becomes too large to fit into RAM, and prohibitively slow to write to and read from a hard disk. It is possible to overcome this by dividing networks into blocks that fit into RAM, where networks are then 'stitched' together by searching for single–mutant neighbours between each of the artificial blocks with the same phenotypes Göbel (2000) and Kospach (2003). However, this type of method involves more complex



(a) Depth–first search **(b)** Breadth–first search

Figure 2.3: If one doesn't keep track of sequences that have already been searched, an infinite loop will be encountered when the connections do not form a strict tree structure. Here dotted lines indicate extra connections. **a)** Depth first search: Apon reaching sequence 4 sequence 2 is found again, forming an infinite loop through 3 and 4. **b)** Breadth first search: One looping connection forms a complex backtracking through the whole space. First level = $\{1\}$; second level = $\{2, 3\}$; third level = $\{4, 5, 6, 12\}$; forth level = $\{7, 8, 9, 10, 11, 12, 6\}$.

coding, and is still not computationally optimal (Gallagher, A., 2004; 2005, personal communication)

2.2.2 Array based search

Instead of attempting to implement a complicated routine shuffling information in and out of RAM, I searched for a more simple algorithm to avoid the space constraints. Instead of dividing the set into arbitrary blocks, I divided it into groups based on phenotype (phenotype set). Each phenotype was given a unique phenotype identity (PID) as a more parsimonious recording method than dot–bracket notation. Since the files were generated lexicographically, the first sequence

AAAAAAAAAAAAAAAA

never folded, and was part of the 'open' network with a PID of 0 (PID0). Hereafter PID0 refers to the open phenotype network. At length-10 dividing the space by phenotype means that each set could then be fitted into RAM individually as an array. A neighbourhood was generated for each sequence in the array, and the array was then searched for each neighbour. Because the array only contains sequences mapping to a single phenotype, if the neighbour was present in the array it was a neutral neighbour, and the sequences could be labelled as being part of the same network. This method was not scalable to the length-16 space, since the large size of

some of the phenotype sets, particularly PID0, did not fit into the available RAM. As the size of the phenotype set increases, searching for sequences within an array also becomes far more time costly, even when using an efficient binary search. Because of these constraints, this method was eventually simply used to cross check the networks generated by the other methods outlined below.

2.2.3 Sorting algorithm

The problem with the array–based search outlined above is that it requires large amounts of memory to run. When there is not enough RAM available, the alternative is to use a random access file on the hard–disk. However, this is impracticable to use on a large file, because the time taken to physically jump to different parts of a large file on a disk becomes prohibitive. On the other hand, streaming files to and from a hard disk is significantly faster than randomly accessing them.

The second method then, rearranges the phenotype set files alphabetically to facilitate stream–based computation. Instead of reading the whole file into an array, each pair of adjacent sequences in the phenotype file are compared. Because they were generated alphabetically, if a pair of sequences has the same bases at all the positions bar the last one, they are one-mutant neighbours and can be assigned the same provisional network ID number (pNID). The set can then be re-sorted using the fast and efficient UNIX `sort` command by changing which position counts as the first of the sequence between n and 2. In each different alphabetically sorted file, the most similar sequences for that ordering will occupy the places immediately adjacent in the file (see Table 2.2). If any pair of sequences are identical after discounting the last sorted position, they are neutral neighbours and their pNIDs can be recorded as equivalent.

The pNID relations can be tracked and consolidated using an equivalence tracker, which iterates through from high to low pNID values, replacing higher values with the lowest 'synonymous' pNID value with which it is associated. The result is a single file for each phenotype, each line of which contains the sequence and its NID code. This method relies on the efficiency of the UNIX `sort` command, and provides a feasible running speed when the sorting and streamed comparisons were parallelised on the NGS computer cluster (`http://www.grid-support.ac.uk`). Although the UNIX `sort` command does run faster with more RAM assigned, the major benefit of this algorithm is that the space requirements are limited to hard disk memory, rather than RAM.

Position sort order			
123	231	312	Final file
AAA:1	AAA:1	AAA:1	AAA:1
AAG:1	AUA:2, 1	AAG:1	AAG:1
AUA:2	UUA:4	GGG:3	AUA:1
GGG:3	AAG:1	AUA:2	GGG:3
UUA:4	GGG:3	UUA:4, 2	UUA:1

Table 2.2: An example phenotype set containing 5 sequences. In the initial file (first column), the sequences are sorted alphabetically, so AAA and AAG appear next to each other. They only differ by the last sorted character, so can be given the same ID number. None of the other sequences differ from the neighbours on either side of them by just one letter. In the second column, the first sorted character is now at the last position in the sequence and the last sorted character is at the penultimate position. This means that sequences AAA and AUA now lie adjacent. As they only differ at the last sorted position (here, position 2), they are neighbours and their network IDs can be made equivalent. After the neighbours in each possible sorted order are calculated, the pNIDs are consolidated using an equivalence relation table (4,2 and 1 are all part of the same network). Finally the consolidated NIDs can replace the provisional ones in the original file (last column). This example phenotype set contains one network of 4 sequences, and one 'network' with a single sequence.

The sorting algorithm provides an efficient search for calculating the size and shape of the neutral-networks within each phenotype. However it has a significant shortcoming when calculating which networks neighbour each other. As the networks are split into separate phenotype files, it is very difficult to track down the phenotypic network neighbours of a particular sequence. One can refold all the single–mutant neighbours at the boundary of the network, but this only gives the neighbour's phenotype and not its network. Alternatively one can search through the lists of sequences within other phenotype files on hard disk, which is prohibitively slow, even though the files are sorted. Even when the two are combined (finding the phenotype by folding and then searching through the phenotype file for that phenotype) it is still too costly computationally. This shortcoming becomes particularly important when considering trajectories across more than one network, because different networks of the same phenotype can have different network neighbourhoods. Once the networks had been calculated using the sorting algorithm, a new method allowing more in-depth analysis of the space had to be found.

2.2.4 Sequence addressed array

All the methods outlined up to this point have used inefficient memory–intensive `String` objects to hold the sequence information. However, the whole map can be

fitted into RAM if each string is converted into a unique integer. This way, the integer still holds the sequence coding information but is also used as an array address, where the data held in the array are the PID or NID values for each sequence.

Each text–based sequence string can be represented as a bit–string where each letter in the sequence is two bits (A=00, C=01, G=10, U=11). The bit code is then converted into a unique decimal integer. For example the sequence

CGAGUCAUCC

is represented by the bit code

01100010110100110101

which in turn converts to the decimal number

404789

This integer forms the array address for that sequence, and the datum at that address is first the PID, or once it has been calculated, the NID.

To calculate the local neighbourhood of a sequence, bitwise operations on the integers change the bit code of the integer. This means that it is possible to avoid having to do computationally costly character swaps using `switch` statements on the orriginal character strings. The PID or NID of each neighbour can be easily looked up in the array, and in this way it is possible to calculate the neutral networks using a breadth–first search. In this method the sequences which have already been visited during a breadth–first search are recorded as such by means of a special temporary ID. This means that all the searched neighbours from the previous level of the search can be tracked in the array, and immediately ignored if encountered again, without the need for a separate list of searched sequences.

There are still significant memory constraints on this more efficient method. The PID/NID is stored as a `short` integer, which in Java requires 2 Bytes (16 bits). Therefore the minimum memory requirement for a single native array for length-16 would be $4^{16} \cdot 2$ Bytes or 8GB. However, the Java language has a maximum array size dictated by the maximum value of the integers that make up the addresses (4 bytes or 2^{32} bits). As arrays cannot have negative addresses, and Java does not support unsigned integers, the maximum array index in Java is actually $2^{31} - 1$. Further to this, there is a maximum memory limit to data structures in Java of 2GB, so if short integers are used as the data values in the array, the maximum assignable size is halved again. As this is an *absolute* maximum, and the data structure of the array

itself takes up a small amount of memory, the array for the length-16 dataset actually must be split into 5 parts. The viability of coding the length-16 space in 5 arrays in this way has been successfully tested, but calculations of the networks have not been carried out due to a lack of access to a computer with more than 8Gb of RAM. This method has successfully been used for sequence lengths between 10 and 14, where the full array of short integers requires just under 540MB of RAM. The zero network was not recalculated using this method, as it is by far the most costly to calculate.

The major advantage of this method is that the entire space is stored and accessed from a single array in RAM. This allows phenotypic network neighbours and trajectories across the space to be calculated more easily than when using the sorting algorithm, where phenotypes are accessed one at a time. It means that the sorting algorithm was used to exhaustively map the size and number of the neutral-networks in the larger size-16 space, which would not fit into RAM using any of the methods, but most of the later work in this thesis only considers the shorter sequence length maps using the more versatile integer coding method. The integer array algorithm was not deployed on the NGS because of the set–up time costs – with CPU time not a limiting factor it was more productive to run on a desktop with a simpler interface.

2.2.5 Time comparisons

This section details various time measures taken of the computation times for the two main algorithms. First, the ln(time) taken in seconds to perform neutral network finding of all networks excluding PID0 on the Apple eMac for lengths 10, 12 and 14.

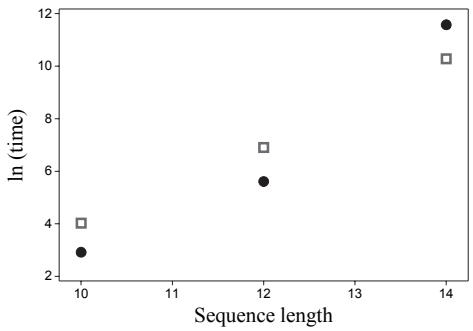
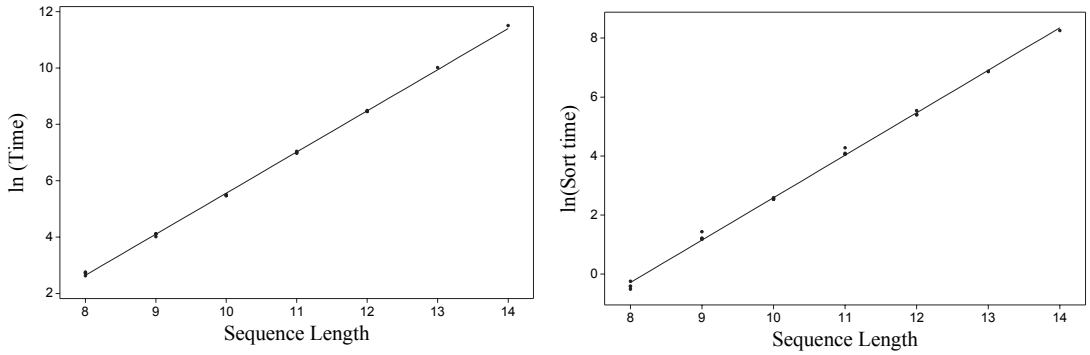


Figure 2.4: Ln (Time) in seconds taken to map the networks at lengths 10, 12 and 14 using the Integer array algorithm and the sorting algorithm on the 1Ghz eMac. □ = Sorting algorithm, • = Integer algorithm.

Interestingly the integer array is outperformed by the sorting algorithm at the length-14 calculations, even though the reverse is true for the smaller datasets. This

is despite the open network not being recalculated using the integer array. The increase in number of calculations in the sorting algorithm makes the increase log-linear, reflected in the more complete set of sequence length calculations shown in figure 2.5.



(a) **(b)**

Figure 2.5: a) Ln (Time (s)) taken to calculate the neutral networks in a map at a given sequence length using the sorting algorithm on the NGS grid. ln (Time (s)) = - 9.042 + 1.460(Sequence length) **b)** The maximum sort time for an iteration of the external UNIX sort for each sequence length (3 repeats for each sequence length). ln (Sort time (s)) = - 11.81 + 1.440(Sequence length)

The time taken for the length-16 space was not directly comparable, because it was manually parallelised to run on more than one node of the NGS grid, each within a seven day maximum execution time, as well as to conform to restrictions on the scratch space available as temporary storage on the local hard drives. The time taken to sort the phenotype files was significantly longer for the PID0 at length-16 than was predicted from the equation given in the legend of figure 2.5. The reason lies in the fact that the PID0 file was the first which was too large to fit into RAM, and therefore required switching to a temporary HDD file, with much slower access. The overall time to complete the length-16 space was approximately four weeks, 25% longer than the time predicted from the relationship in figure 2.5. If the length-16 space or longer were to be investigated using the integer array, the most time–efficient method would involve calculating neutral networks using the sorting algorithm, and the network relations calculated using the integer array method.

2.2.6 Space comparisons

Below is a table detailing the space comparisons for the files at different sequence lengths. The space saving advantage of using integers over strings is clear.

	File size	
Sequence length	String file	Integer file
10	28Mb	2Mb
12	448Mb	32Mb
14	7168Mb	518Mb
16	112Gb	–

Table 2.3: The memory requirements for each of files containing the complete genotype–phenotype map for sequence length-10–16.

2.2.7 Checking procedures

All algorithms were checked against a test dataset small enough to be analysed by hand. In addition, the sorting algorithm yielded identical results to the integer array for the length-10, 12 and 14 spaces and also matched the first string array based search as well as an independent breadth–first search programme written by Alexis Gallagher for the length-10 space (Gallagher, 2004, personal communication).

2.3 Initial network results

This section details the initial network results found by using a combination of the two main methods described above. The first section is necessarily dry, with many measures and numbers used to outline how the networks are structured within the genotype space. The section is therefore concluded with a list of the most important points to be taken away from the complex of results presented before it.

Within the RNA map, most sequences folding into a particular phenotype do form pervasive neutral networks extending across sequence space. The genotype space is therefore significantly different from some other discrete models without neutral mutations (e.g. Gillespie, 1984; Kauffman and Levin, 1987; Kauffman, 1993; Orr, 2002, 2005), but similar to other neutral network explorations using RNA or protein lattice models (e.g. Göbel, 2000; Ebner et al., 2001; Aita et al., 2003; Kospach, 2003).

At the short sequence lengths studied here, the largest group of sequences is the 'open' set (PID0), in which *no* base pairs form. In the length-10 space over 91% of all sequences belong to this group. The percentage falls to just over 44% for the length-16 space. Across all lengths studied, the sequences in the open set form one connected neutral network.

The decrease in size of the open network (from 91% of sequences at length-10 to 44% of sequences at length-16) is set to continue at longer sequence lengths, and

may well eventually result in the non-folding sequences ceasing to remain part of one single network. Forster et al. (2006) found that just 0.00886% of a set of 10^7 random sequences did not fold at a sequence length of 75 bases.

The lack of secondary structure of PID0 can be assumed to have a fatal effect on its efficacy as a functional phenotype (Kikinis et al., 1995; Ke et al., 1998). From this point on I concentrate on phenotypes which do exhibit secondary structure, and so the PID0 network is discounted from the rest of the network results (let us henceforth assume that a non-folding phenotype is inviable). Table 2.4 gives some general statistics about the phenotypes and networks for the even number sequence lengths investigated.

		Total count of		
Temperature	Length	phenotypes	networks	sequences
30	10	19	122	87921
	12	71	716	3736994
	14	269	5334	105783590
	16	1009	37514	2398409233
37	10	19	102	52656
	12	57	645	2451912
	14	228	4603	74731841

Table 2.4: General statistics for the genotype spaces of different sequence lengths (excluding PID0). See text for mapping function parameter details.

When the folding algorithm is set at the higher temperature (37°C), the stability of folded structures is put under pressure. Base–pairs in structures which are only marginally stable at the lower temperature break down, meaning that some sequences change their phenotype and at lengths-12 and 14, the less stable phenotypes disappear altogether.

2.3.1 Number of phenotypes

At 30°C, the length-10 space consisted of 19 folded phenotypes, rising to 1009 at length-16 (see Appendix A). The total number of phenotypes increases exponentially with sequence length, but at a slower rate than the increase in the number of sequences. This means that as sequence length increases, each phenotype set contains more sequences on average.

Within a single sequence–length space, the distribution of sequences folding into different phenotypes is right skewed (Fig 2.6). Many phenotypes contain a relatively

low number of sequences, and a few phenotypes contain a large number of sequences. This becomes more pronounced as sequence length increases, and corroborates the findings of Schuster et al. (1994); Grüner et al. (1996b) and Huynen et al. (1996).

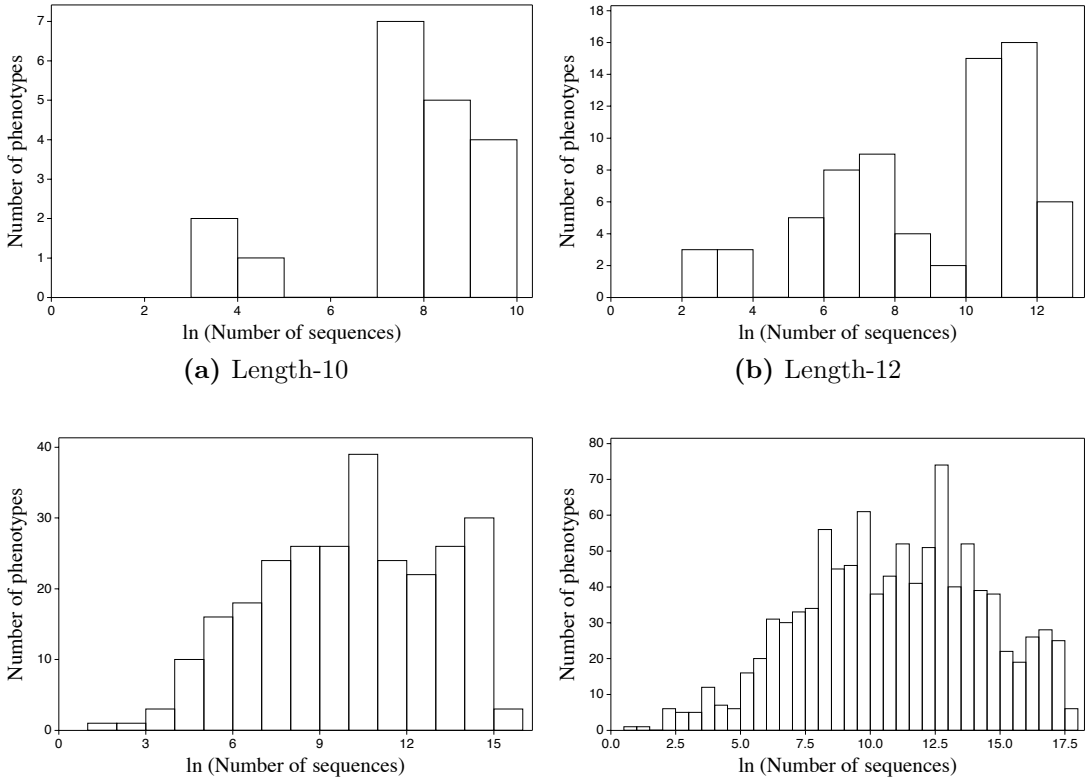


Figure 2.6: Distribution of ln(No. of sequences) per phenotype at sequence lengths 10, 12, 14 and 16, excluding the open phenotype for each

2.3.2 Number of networks

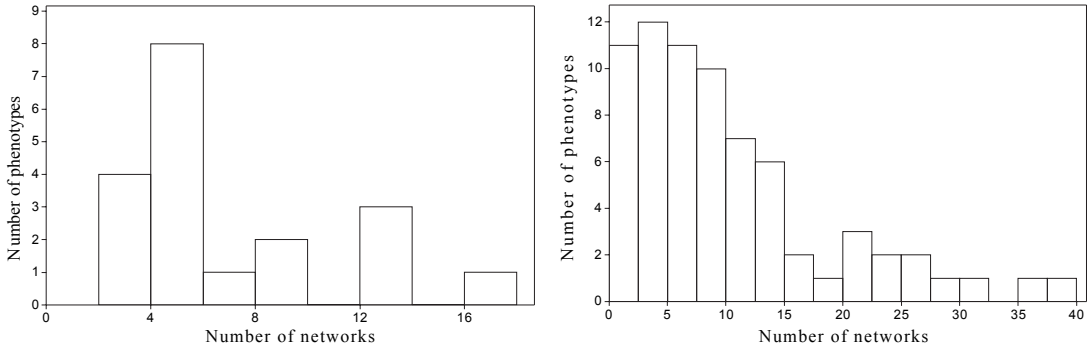
Most phenotypes are mapped to by a set of disjunct neutral networks. In fact, it is unusual for just one network to encompass all sequences mapping to a particular phenotype (except for the open network). In the length-10 space all of the folding phenotypes are mapped to by disjunct networks, while in the length-16 space 39 phenotypes have just one continuous network mapping to them. The largest number of networks in any one phenotype set (length-16) is 265. In the length-10 map the largest number of networks is 17 for one phenotype.

The mean and median number of networks per phenotype increases with increasing sequence length (Tab. 2.5; Fig. 2.7). In the same way as the sequences per phenotype,

Sequence length	mean	median	min	max
10	6.42	4	2	17
12	10.08	8	1	40
14	19.83	13	1	125
16	37.18	20	1	265

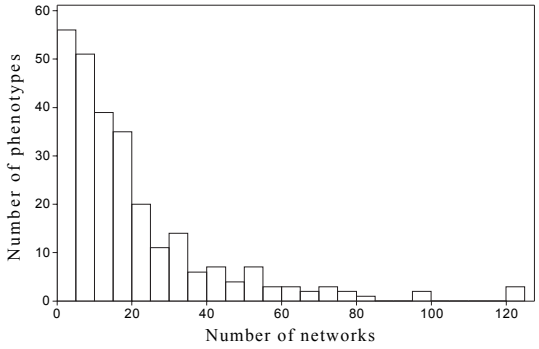
Table 2.5: Networks per phenotype for maps of different sequence lengths, excluding PID0 at 30°C

the distribution shows a right hand skew indicating that a few phenotypes have large numbers of disjunct networks, but most have just a few.

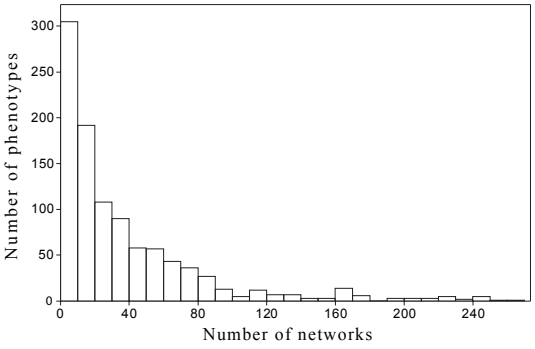


(a) Length-10

(b) Length-12



(c) Length-14



(d) Length-16

Figure 2.7: Distribution of number of networks per phenotype at sequence lengths 10, 12, 14 and 16, excluding the open phenotype for each.

This result can be partially explained by considering the RNA specific shape of a phenotype in relation to the number of networks: generally, the more base–pairs, the more networks. The result is a positive correlation between the number of networks in a phenotype set and the number of base pairs (Fig. 2.8). As highlighted in section 2.1.2, it is not possible in this model to swap paired bases and maintain

a continuous network. The largest numbers of networks are found within a phenotype which is unable to maintain structural stability across all the different base–pair combinations possible. This network break–up happens particularly when two neutral networks are formed because the G-U intermediate does not map to the same phenotype. The lower energy of a G-U bond means that it is more likely not to form than either and A-U or a G-C, meaning that there is no way of linking the latter two into a single network.

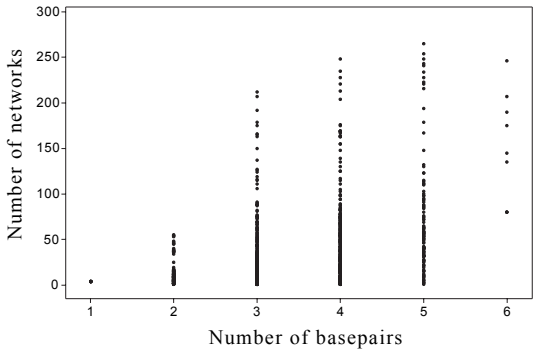


Figure 2.8: Number of networks per phenotype against number of base pairs in that secondary structure for length-16. The distribution of number of networks with 5 base pairs is larger than those with 6 because the only phenotypes with 6 base pairs are stable chains with no breaks or bulges, whereas those with 5 base pairs include less structurally stable phenotypes, with a higher possibility of the genotypes with a less stable G-U intermediate folding into another phenotype.

2.3.3 Sequences per network

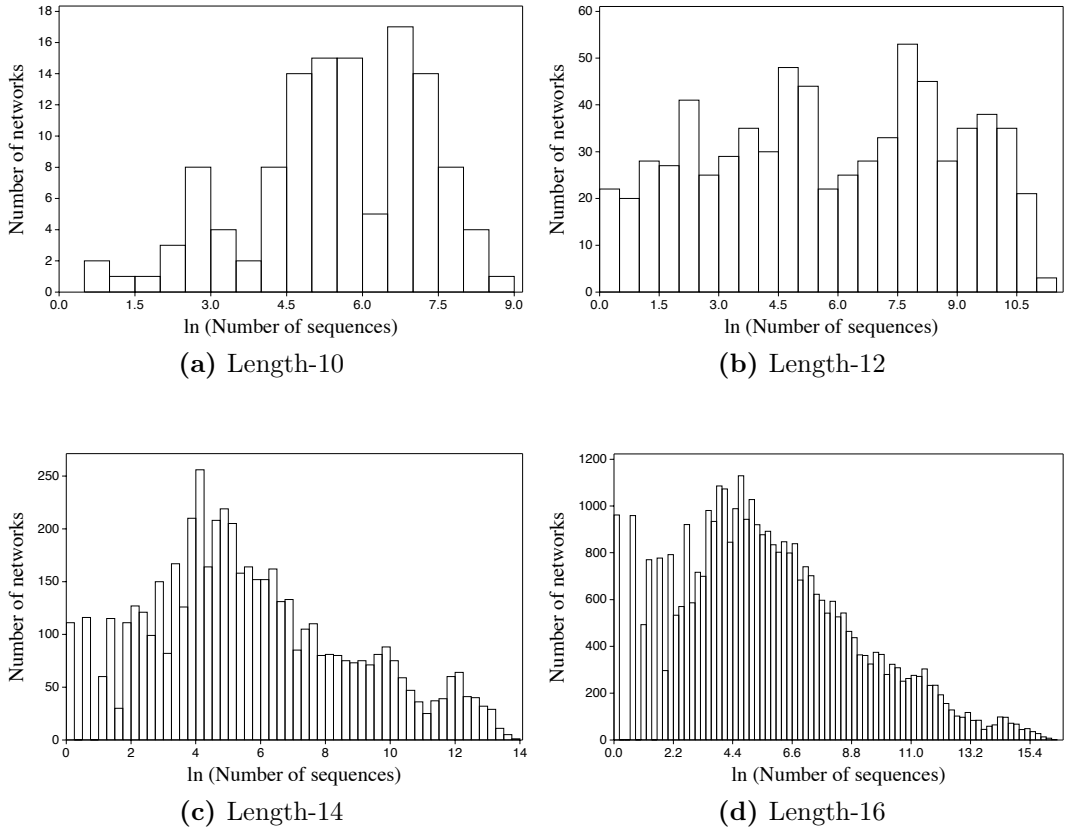
Within a given phenotype the distribution of sequences between networks is not equal: one or two of the networks usually retain the majority of the sequences that fold into that phenotype, Grüner et al. and Reidys et al.'s so called 'giant component' (Grüner et al., 1996a; Reidys et al., 1997). This unequal distribution means that even highly disjunct phenotypes containing a small total number of sequences can have one or a few networks which are larger and percolate further through the space than might be expected.

So, as sequence length increases, the number of foldable sequences shows an exponential increase. This increase is larger than the increase in the number of networks, which is in turn larger than the increase in number of phenotypes. However, the distribution of sequences between networks becomes more skewed at longer sequence lengths (Fig. 2.9). This means that the largest networks take a greater proportion of the folding sequences at longer lengths. In fact, the median number of sequences

Sequence length	mean	median	min	max
10	720.7	247	2	5183
12	5219	261	1	72774
14	19832	187	1	969740
16	63934	210	1	12027112

Table 2.6: Number of sequences per network, excluding PID0 at 30°C

per network does not rise with sequence length, despite the mean number of sequences increasing massively (Table 2.6). Many small networks exist at the shorter sequence lengths, but small networks become increasingly common as sequence-length increases.



(a) Length-10 **(b)** Length-12

(c) Length-14 **(d)** Length-16

Figure 2.9: Distribution of ln(No. of sequences) per network at sequence lengths 10, 12, 14 and 16, excluding the open phenotypes for each.

2.3.4 Symmetry

Finding symmetries within the genotype space has the potential to provide an interesting insight into the structure of that space. There are many ways in which symmetry might occur: if, for example, completely inverting a sequence results in a phenotype exactly opposite in its binding pattern to that of its complementary sequence (Fig. 2.10). Alternatively, swapping base pairs from G-C to C-G could result in symmetries between disjunct networks of the same phenotype.

Symmetries within genotype space also have the potential to reduce the total number of calculations needed in a dataset. If there is symmetry across the whole space for example, the size of the space is reduced to the size of the repeated sub-unit. A symmetrical network requires only half the number of folding calculations and/or neighbour relations to be calculated before it is then possible to complete the space by generating a mirror image, which is much less computationally expensive than folding the whole space.

Original sequence
GGCUCGCGUAGAUGGC (.(((.....)).).)
Full inversion of sequence
CGGUAGAUGCGCUCGG ((((.....)).)).
Symmetrical sequence of complementary bases
CCGAGCGCAUCUACCG .....

Figure 2.10: Sequences are not necessarily predictable in their folding even if a complementary sequence is known.

As figure 2.10 shows, symmetries do not arise between networks in any predictable way. In the most extreme asymmetry, a phenotype set contains just one neutral network, where none of the complementary base–pair sequences fold into the same phenotype. In some cases there is limited symmetry within a network, where a network contains all the base combinations at variable positions. The occurrence of this is not predictable, and so is of no help in reducing the number of calculations required. This network structure is investigated further in section 2.5.

Part of the reason that symmetries do not exist lies in the non-planar shape of the bases, meaning that opposite sequences do not necessarily have opposite structure. Furthermore, bases can interact with the positions on either side of them in the chain, as well as those they could pair with, decreasing the chance of exact symmetries existing. Additionally, as in real life, the folding algorithm carries biases depending on the direction that a sequence is considered.

Despite all this, many phenotypes are mapped to by a series of approximately symmetrical networks. They are often divided by inverted base–pairs, and therefore separated by a simple double mutant. Even a small amount of asymmetry can lead to each network having different network neighbourhoods and therefore different evolutionary potentials.

2.4 Connectivity

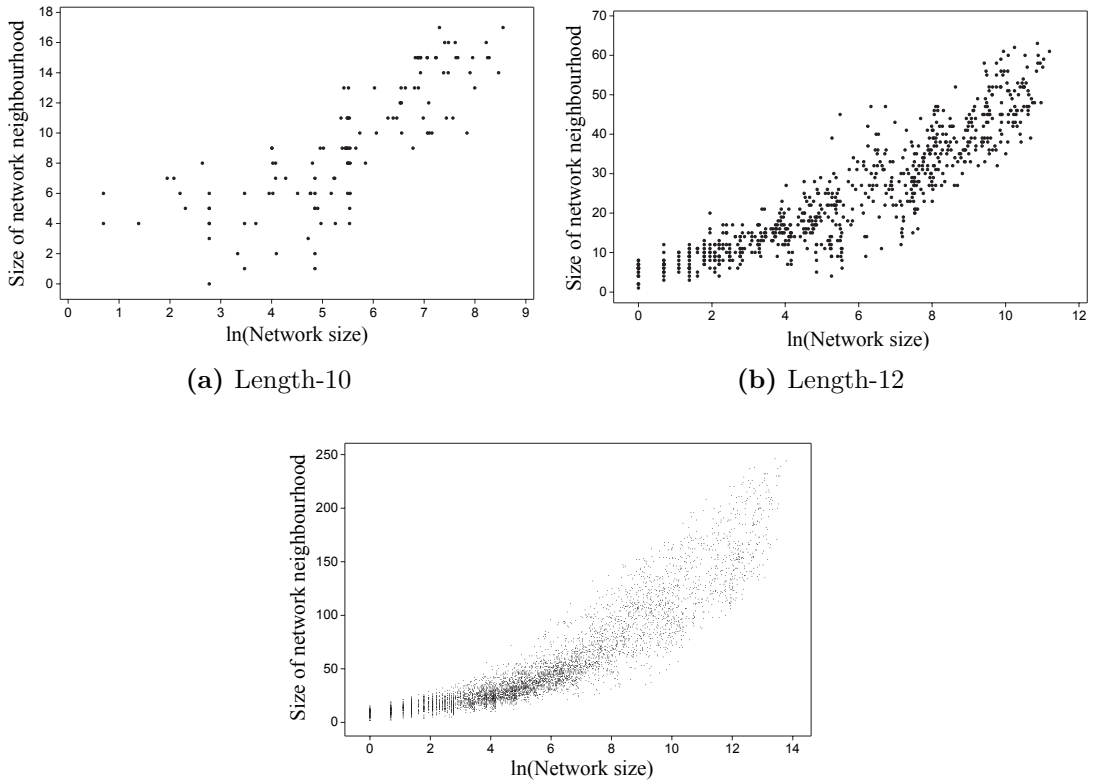
Even though only 8.4% of sequences fold into secondary structure at length-10, these 8.4% form highly inter–connected networks, and the 'landscape' described by the mapping is not one of isolated peaks surrounded by a sea of open structures (Fig. 1.9b). In fact all but three networks in the length-10 space have more than one folded phenotype among their neighbours, and only one of the three is completely surrounded by the open network. At length-12, only one network has just one folded phenotype neighbour, and at lengths 12 and 14, no networks are completely isolated within the open network.

In contrast, at length-10 every network neighbourhood includes a portal to PID0. At length-12, one network (NID-459) does not contain a network neighbour to PID0, while at length-14 there are 141 networks which do not have any portals to the open network at all.

2.4.1 Network neighbourhood

At a sequence length of ten, the mean percentage of other folding phenotypes that a network is connected to is 51.78%, with none connected to all other phenotypes. At length-14 networks connect to only 19.78% of other phenotypes on average, so the length-10 space is actually more widely connected than the length-14. However, excluding the PID0 network, the most widely connected networks at each sequence length remain very well connected at about 90% of all possible phenotypes. At length-14 just over 8% of networks are connected to more than 50% of other phenotypes. At length-10 this increases to 45.9%.

There is a correlation between the size of a network and the number of neighbours it has (Fig. 2.11). This means that the larger networks contain more connections to different phenotypes. However, at the very largest network sizes, most of the phenotypes have already been found at least once before in the neighbourhood. This means that the number of extra sequences required to find each new network neighbour increases with size of network.



(a) Length-10 **(b)** Length-12

(c) Length-14

Figure 2.11: Correlation between the phenotypic neighbours of a network and its size. Total number of phenotypes found in the space: length-10 = 19, length-12 = 71, length-14 = 269.

2.4.2 Portal distribution within networks

Within each network the vast majority of sequences have at least one local neighbour which is part of the single open phenotype (PID0) network. More importantly, *no folded sequences have a neighbourhood consisting solely of neutral neighbours.* This is considerably different from the two and three-dimensional representations, and so is worth stressing here. It means that every sequence is at the boundary of a network, and a random point mutation always has some probability of changing the phenotype. The result is that the lattice type network with sequences completely surrounded by neutral neighbours does not exist in these spaces (Fig. 2.12).

Although no sequence has a neighbourhood consisting solely of neutral neighbours, most do have many. Together with the PID0 neighbours, these two phenotypes make up the majority of local neighbours for any given sequence. At longer lengths, while the percentage of neutral neighbours within the neighbourhood remains approximately the same, the percentage of PID0 neighbours drops as the percentage of

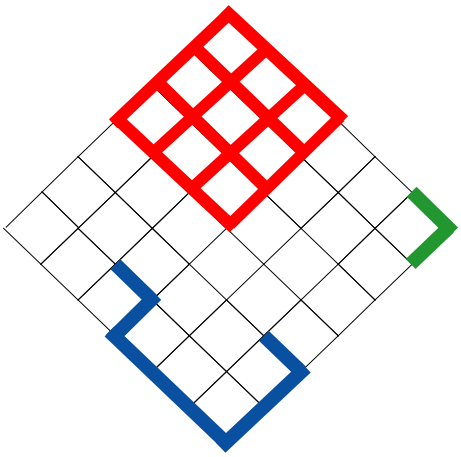


Figure 2.12: The lattice–type network (shown in red), where sequences within the network are completely surrounded by neutral neighbours does not exist within the more complicated multi–dimensional RNA genotype–space.

the total space occupied by PID0 drops (see table 2.7).

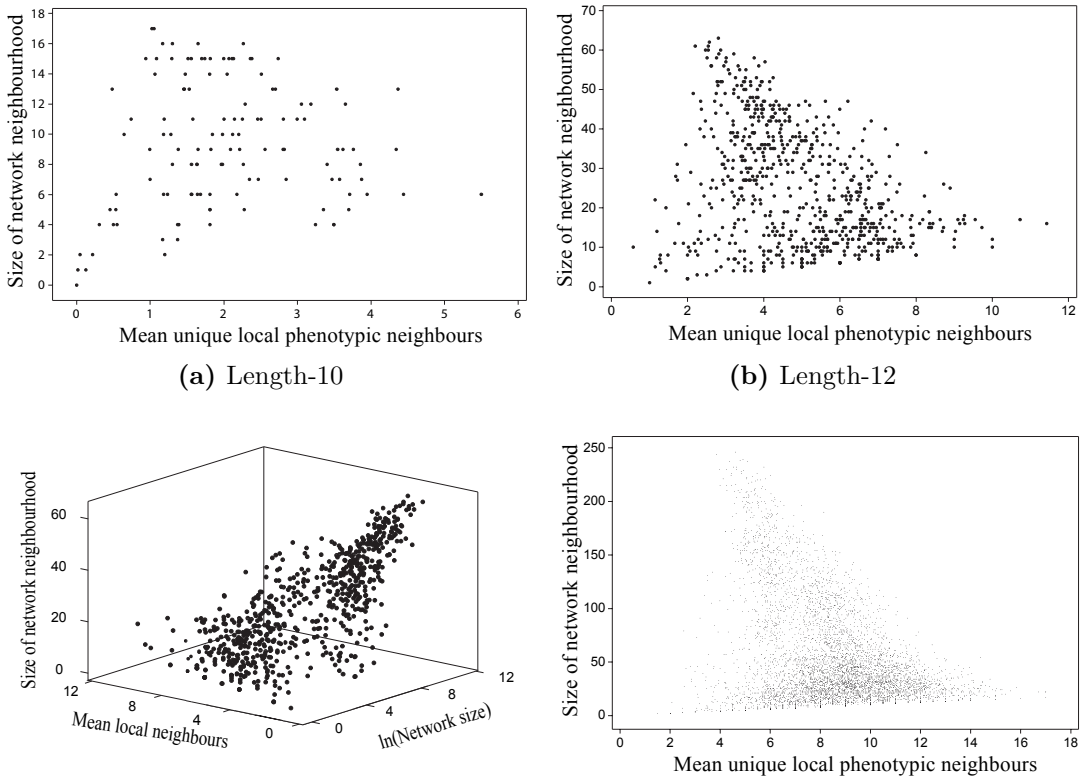
	Local neighbourhood	Mean number of neighbours		
Length	size	Neutral	PID0	Unique PIDs
10	30	13.34	13.16	1.87
12	36	17.09	11.48	3.31
14	42	20.22	9.58	4.97

Table 2.7: Mean number of neighbours per viable sequence. The number of unique PIDs does not include PID0.

The number of unique alternative phenotypes in a given local neighbourhood is on average also far lower than the number found in the network neighbourhood (Fig. 2.13). In this sense 'unique' means that if more than one genotypic local neighbour codes for the same alternate phenotype, that phenotype is still only counted once. This gives us our first suggestion that in this model, neutral networks do facilitate evolutionary innovation.

Figure 2.13 shows that sequences in the networks with the largest network neighbourhoods do not actually have the largest average local connectivity per sequence. In fact it is the smallest well–connected networks whose sequences have the highest average local network connectivity.

Furthermore, there is significant intra-network variation in the number of phenotype neighbours in each local neighbourhood. In some circumstances a small change



(a) Length-10 **(b)** Length-12

(c) Length-12 **(d)** Length-14

Figure 2.13: Graphs showing the number of phenotypic network connections against the mean number of unique phenotypic local neighbours per sequence within that network. Maximum possible number of phenotypic neighbours: length-10 = 18, length-12 = 70, length-14 = 268. **c)** shows the interaction between size of the network, number of local neighbours and the number of network neighbours.

in sequence can have a large effect on the number of accessible phenotypes. For example, if we consider the sequences in table 2.8, changing the 8^{th} base from a Uracil to a Cytosine increases the chances that a mutation will result in a new non-zero phenotype. However, changing the 10^{th} base between an Adenine and a Guanine makes very little difference to the available phenotypes. There are potentially many more different phenotypes accessible from a well connected sequence than are available to the majority of sequences in the network. I shall consider the effect of portal position within a network more thoroughly in chapter 3.

Table 2.8 also shows that many of the local neighbours can belong to the same phenotype. If this is the case, then some networks are better connected to each other and are more likely to be encountered by random drift, because they share more pathways between them. The number of sequences which are directly connected to another phenotype has been used by Fontana and Schuster (1998a,b) to indicate

	Number of phenotype neighbours					
Sequence	PID-0	PID-5	PID-6	PID-10	PID-11	PID-14
GGGUUGCUCA	17	0	1	12	0	0
GGGUUGCUCG	17	0	1	12	0	0
GGGUUUCCCA	5	5	1	12	3	4
GGGUUUCCCG	5	4	2	12	3	4
GGGUUUCCCU	6	2	5	10	5	2
GGGUUUCUCA	17	0	1	12	0	0
GGGUUUCUCG	17	0	1	12	0	0

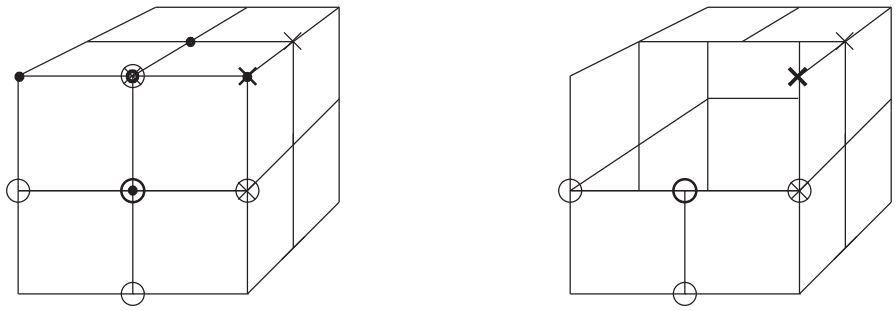
Table 2.8: Subset of the Length-10, PID-10, NID-95, showing the distribution of phenotypes in each sequence's local neighbourhood.

the likelihood of a transition between two different phenotypes. However, these are not the only factors to potentially influence evolutionary trajectories. The internal structure of each network can also have a profound effect on the way in which it is negotiated, as we shall see in chapter 5.

2.5 Network density

The effect that neutral networks have on evolution depends on the structure *within* the network as well as the connections *between* networks. Importantly, the high dimensionality of the space leads to many more 'external' boundaries within the network than is obvious from thinking about a 3-D landscape. Even sequences at the 'centre' of a network have single mutant neighbours coding for other phenotypes. If one considered a network as the skin of a cube, then sequences on the corners of the cube are the 'most external' (they have three neutral neighbours). Sequences on the edges between the corners or on the faces of the cube have 4 external neighbours, but are still on the boundary of the network in some directions (Fig.2.14a). When some sequences are 'missing' from the network, there can be a significant change in the number of neutral neighbours per genotype across a network, which strongly influences the number of paths taking the shortest route between portals (Fig.2.14b). In fact, in this toy example, removing just one sequence means that the minimum path length between the front top right and the front top left of the cube increases from two mutations to four.

The mean number of neutral neighbours is an indicator of the number of shortest route paths across a network, and has been suggested as a measure of connectedness by Reidys et al. (1997). The more neutral neighbours each sequence has, the more



(a) **(b)**

Figure 2.14: a) Every sequence is at the boundary of its network. However, some sequences have more internal (neutral) neighbours (and therefore fewer external neighbours than others). **X** = 3 neutral neighbours, • = 4 neutral neighbours, ∘ = 4 neutral neighbours. **b)** When a sequence is missing from the network, there is a significant effect on the number of possible paths across the network. **X** = 2 neutral neighbours, ∘ = 3 neutral neighbours. NOTE: The diagrams here represent a network as the skin of the cube, there are no internal sequences.

ways of drifting between two points in the network, and the higher the chance of a direct path. However, when the full extent of a network is known, a similar measure can be derived more simply across the network – If we know how variable each position is in the sequence, then we can calculate the maximum possible number of sequences in any given network by working out all the possible combinations of bases at variable positions. The maximum size minus the actual size of the network gives us the number of sequences which are 'missing' and the ratio of the number of genotypes *actually* in the network to the number of sequences *possibly* in the network gives what I shall call the *network density* (Table 2.9).

Example network	All base combinations	Missing
AAA	AAA	
AAC	AAC	
*	ACA	ACA
ACC	ACC	

Table 2.9: In this example network there are three genotypes. Positions two and three are variable in the sequence, with either an A or C at position two and an A or C at position three. Finding all combinations of these variable positions indicates that there is one more combination of bases (ACA), which we might expect to be a member of the example network but which is not. The network density is therefore $3/4 = 0.75$.

It is straightforward to calculate the maximum number of sequences, as long as a record is kept of which positions varied to which bases throughout the network. If

m_k positions can vary between k bases, the number of sequences in the network is:

$$\prod_{k=1}^{k=4} k^{m_k} \tag{2.1}$$

It is also possible to calculate the potential number of neutral neighbours for each sequence as:

$$\sum_{k=1}^{k=4} (k-1)m_k \tag{2.2}$$

For example, the maximum number of sequences in the network in table 2.9 is: $1^1 \cdot 2^2 \cdot 3^0 \cdot 4^0 = 4$, and the potential number of neutral neighbours is: $0 \cdot 1 + 1 \cdot 2 + 2 \cdot 0 + 3 \cdot 0 = 2$.

If all the combinations of bases at all the variable positions are present in the network, the base changes needed to traverse between any two points within the network can occur in any order. We can say that they make a 'face' in the hypercube. Any mutation which is neutral somewhere in the network is neutral everywhere in the network. However, when sequences are missing, in other words the lower the density, then some of those changes can only be made when specific bases are present at different positions in the sequence. The more sequences that are missing, the fewer direct paths there are between sequences. Returning to the example network in Table 2.9, we see that there is only one path between AAA and ACC which remains neutral (via AAC). The other path, via ACA will not remain on the neutral network. As network density decreases, the chance that the only path between two sequences is longer than the (Hamming) distance increases, seen in figure 2.14b between the two front top corners.

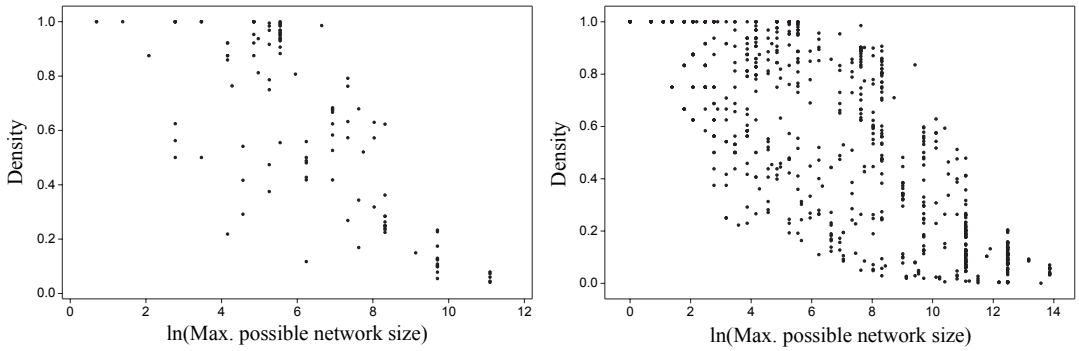
Table 2.10 shows the breakdown of network sizes and densities for PID10 in the length-10 space. We can see that network 61 and network 70 both potentially contain $1^6 \cdot 2^1 \cdot 4^3 = 128$ sequences. However, network 61 actually contains 118 sequences, with 10 sequences folding into different phenotypes, giving a face density of 0.92.

As the maximum possible number of sequences increases, the number of sequences that actually do fold into the network tends to be a much smaller fraction of that maximum (Fig. 2.15). This indicates that the more variable and therefore potentially larger a network can be, the less dense it actually is.

The effect of network shape and structure is further considered in chapter 5. For these short lengths even the least dense networks can still contain many neutral neighbours. However, as we shall see in the next chapter, some pathways between portals do require a number of steps greater than the Hamming distance. This gives a tantalising hint of the complex neutral pathways that might be expected at longer sequence lengths, where the networks are even more convoluted and less dense.

NID	m_k				Net Size		Neutral Mutants		Network
	k:1	2	3	4	Potential	Actual	Potential	Actual	density
39	4	2	-	4	1024	639	14	12.8	0.624
55	2	4	-	4	4096	1078	16	12.9	0.263
61	6	1	-	3	128	118	10	9.41	0.922
69	4	2	-	4	1024	428	14	11.3	0.418
70	6	1	-	3	128	128	10	10	1
76	6	-	1	3	192	151	13	9.74	0.787
78	6	-	-	4	256	248	12	11.7	0.969
87	6	-	-	4	256	248	12	11.7	0.969
89	6	1	-	3	128	122	10	9.66	0.953
90	6	1	-	3	128	128	10	10	1
95	4	2	-	4	1024	539	14	11.5	0.526
103	3	3	-	4	2048	704	15	12.1	0.344
112	6	-	-	4	256	244	12	11.7	0.953
118	6	-	-	4	256	254	12	11.9	0.992
120	6	1	-	3	128	128	10	10	1
121	6	-	-	4	256	256	12	12	1
122	6	-	-	4	256	238	12	11.5	0.930

Table 2.10: Breakdown of network characteristics of PID10 in the length-10 space



(a) Length-10 **(b)** Length-12

Figure 2.15: Graphs showing the relationship between density of the network and maximum possible size for lengths 10 and 12. Length-14 is similar. Temp= 30°C.

2.6 Comparison to networks in other models

When comparing the networks found in this model to those of other genotype–phenotype mapping models, the most striking fact is their qualitative similarity. This extends from comparisons of RNA secondary structure to protein lattice folding models. Furthermore, the networks found in this and other exhaustive models are similar

to those found in models sampling much larger maps (e.g. Fontana and Schuster, 1998a,b; Smith et al., 2003). This echoes Tacker et al.'s and Kospach's work showing that varying the input parameters had little effect on the structure of the space (Tacker et al., 1996; Kospach, 2003). There are, however, quantitative differences in network size and disjointedness between this and the most closely comparable models. At length-16, the number of phenotypes (1010) was higher than those found by Göbel (274 for length-16) or Kospach (741 for length-16). The explanation lies in the fact that any change in biophysical binding energies between different versions of the `RNAfold` software can change the number of phenotypes found, as can changing other parameters such as the exclusion of isolated base pairs. These factors explain the difference between Göbel's work, Kospach's and this study (they both excluded isolated base pairs and used earlier versions of `RNAFold`). Likewise, altering the temperature parameter from 30°C to 37°C changes the quantitative size of networks, but does not have a significant qualitative effect. Most importantly the use of a different local neighbourhood condition results in more disjunct networks within this model, but as we have seen, these smaller disjunct networks are still relatively well connected.

2.7 Summary of results

Basic network facts

- Almost all of the sequences in the genotype space have neutral neighbours with the same phenotype, which together form extensive neutral networks.
- The largest phenotype set forms one large network mapping to the unfolded 'open' phenotype (PID0).
- The other phenotypes are generally mapped to by discontinuous and asymmetric networks.
- The distribution of sequences between networks and between phenotypes is very right-skewed meaning most sequences belong to a few networks, and most networks contain just a few sequences.

Inter-network connectivity

- Networks are highly interconnected. All but 4 networks (3 in length-10, 1 in the length-12 space) have more than one viable phenotypic network neighbour.

- No network contains links to all other phenotypes.
- Larger networks have larger network neighbourhoods.
- Most networks have more connections to certain networks than others.

Intra-network structure

- No folded sequence has a local neighbourhood consisting only of neutral sequences. In other words every sequence is part of the network boundary.
- The average number of phenotypes per local neighbourhood is much smaller than the number of phenotypes in the network neighbourhood.
- Within a network some sequences are far better locally connected to different phenotypes than others.
- Most networks are missing sequences which link parts of the network together, meaning that mutations at some positions are only neutral conditional on certain bases being present at other positions.

2.8 Discussion

The existence of neutral or nearly–neutral mutations in natural systems has become widely accepted over the last 50 years. This means that developing models which include genotypic degeneracy is an important part of evolutionary research. Over the last 10 to 15 years, discrete genotype–phenotype mapping of RNA sequence to secondary structure has proved to be a tractable model system with which to investigate the effect of neutral mutations and genetic accessibility on evolution.

In this chapter I have laid out an exhaustive RNA genotype–phenotype map at sequence lengths between 10 and 16 bases, where the mapping between genotype and phenotype is based on the `RNAfold` secondary structure prediction software. A different approach to calculating the local neighbourhood of a sequence, by considering *only* single point-mutations, rather than base-pair mutations, meant that the networks are more disjunct than in the most similar previous models (Göbel, 2000; Kospach, 2003).

In this kind of small model, the genotype space is dominated by sequences which do not fold into any secondary structure at all. The decrease in size of the open network is set to continue at larger sizes, and may well result in the non-folding

sequences ceasing to remain part of one single network. However, this is unlikely to make a qualitative difference to the space.

It is perhaps enlightening to view the PID0 network as an arbitrary cut-off point like where the shore meets the ocean, below which all other phenotypes were fatal. The ocean is fatal for those who can't swim, so it becomes an arbitrary cut-off point. Drawing a cut-off point for longer sequence lengths at some other reasonable phenotypic character or fitness level should then result in a very similar space. Just because the ocean is a cut-off point, doesn't mean that the landscape stops at the shore line. The landscape may continue further down, but is irrelevant, because the selective disadvantage is almost certain to prove instantaneously fatal. In other words the shorter sequence–length maps with large amounts of PID0 ocean around the networks could be used to model evolution close to the optima in a larger landscape.

If one considers that the number of phenotypes increases exponentially, and far faster than the number of local neighbours, the increase in available phenotypes in the local neighbourhood is outstripped by the number of available phenotypes in the space as a whole. This means that even with fewer PID0 neighbours, the percentage of available phenotypes in the network neighbourhood of any particular network is likely to decrease.

Although the raw number of phenotypes in the average network neighbourhood increases as sequence length increases, the total number of phenotypes increases faster, meaning that the average network neighbourhood contains a lower percentage of all the available phenotypes. This means that there is a higher potential for the space to contain local optima at larger sizes because there are likely to be fewer connections between the networks coding for phenotypes with the highest fitnesses. In fact, as length increases, this property become more important, because the most advantageous step from any one point becomes less and less likely to take an evolving population close to the fittest phenotype. In other words, more intermediate steps are required to negotiate a genotype space where the fittest phenotypes are less accessible.

The existence of neutral networks connected to each other as outlined above fulfils one of the essential requirements under which mutation and drift can become an important force within adaptive evolution. However, the presence of pervasive and connected networks is not enough on its own to confirm that drifting across them really does expand the search space for an evolving population.

The story is complicated by the uneven distribution of portal sequences across networks. Some sequences can be directly connected to many more different local phenotypic neighbours than the average. If these hotspots can act as a link between

different phenotypes, then it may enable adaptive evolution to occur without any need for any or at least much neutral drift, even it the hotspot is part of a large and well connected network.

Imagine the fitness skyscraper again, with its flights of stairs between floors set randomly across the width of the building. Now picture several sets of fire escapes dotted around the building. They have entrances to most if not all of the floors, but just because they have access into or out of a room on a particular floor, there is no need to cross or even enter the room to continue climbing to the top of the building. However, if those fire escapes don't exist i.e. each portal sequence has different phenotypic neighbours, and the distance between each portal is large, then drift can play an important role in the evolutionary dynamics of a population. If that is the case, the size and structure of a network, as well as the distribution of portal connected to it, play a pivotal role in defining what effect networks have. This will be investigated further in the next chapter.

Chapter 3

Minimum path lengths within a single neutral network

3.1 Introduction

Exhaustive searching of RNA genotype–phenotype maps has shown the existence of extensive neutral networks (Grüner et al., 1996a,b; Göbel, 2000; Kospach, 2003). These networks are made up of genotypes which all code for the same phenotype, connected by virtue of differing at only a single position from at least one other sequence in the network. Furthermore, some sequences are connected to viable neighbours in other networks. These so called ‘ portal ’ sequences connect different networks together, by providing a simple point–mutation transition from one to the other. It has been postulated that connected networks can increase the evolutionary search space of a population and allow them to escape from apparent phenotypic stasis via a series of neutral mutations through a network (Huynen et al., 1996; Wilke, 2001b). However, the mere presence of connected neutral networks is no guarantee that they will have any effect on adaptive evolutionary pathways (see Figure 1.9c). In fact we must consider the size and shape of networks and the distribution of portal sequences within them to ascertain the effect of neutral networks on evolution (Fig.3.1).

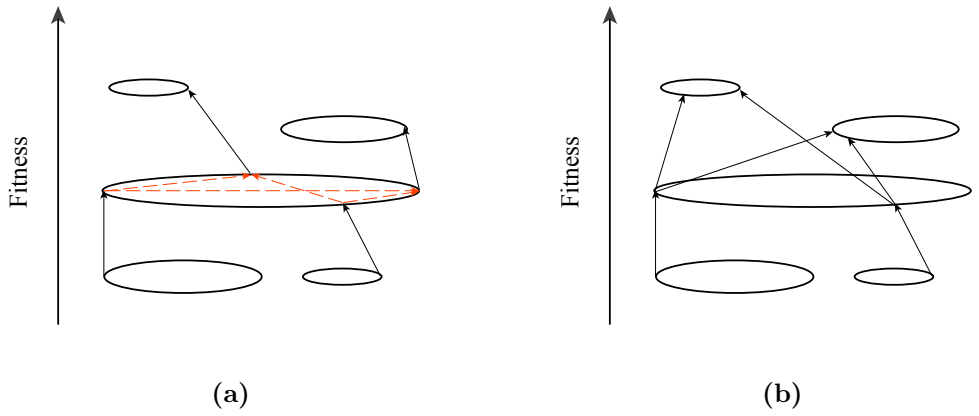


Figure 3.1: Each ellipse represents a neutral network of connected genotypes, and arrows between networks indicate point mutation connections between the portal sequences of different networks. **a)** The size, shape and connectivity of a network become important when the portals to other networks occur at different points within the network (red dashed arrows are potential neutral paths). **b)** The portals in the network all connect to the same neighbours. In this example a drift–based search is not important because all the portals connect to the same set of alternative phenotypes, so neutral drift through the network does not increase the evolutionary options available.

This chapter provides evidence to show that the neutral networks outlined in chapter 2 are structured in such a way that finding a new phenotype often does actually require one or more neutral mutations (Fig. 3.1a rather than Fig. 3.1b). As

might be expected intuitively, the average number of neutral steps required to cross a network increases with the size of the network. Furthermore, the path lengths to rare phenotypes are on average longer than to common ones.

Consider again the analogy of the fitness skyscraper: Each room (neutral network) on a particular floor has staircases leading to some of the other floors in the building. There are however, no doors between rooms on the same floor (networks of the same phenotype cannot be joined at any point, or they simply form one network). Once you arrive on the landing at the top of the stairs (the portal sequence), the number of steps it takes to cross the room to the next staircase depends on the size of the room. A larger room generally takes longer to cross than a smaller one. However, this is complicated by the number of staircases leading out of the room. If there are 10 staircases leading to the floor you are looking for, the chances are good that one set will be reasonably close to your entry point, and on average one of those staircases will be closer than if there were just one staircase at the far side of a much smaller room.

The most interesting discovery presented here is that when a network is not densely populated with sequences, *and* there are rare portal genotypes within it, the shortest path(s) between two portal sequences can be greater than the Hamming distance between them. In fact, very rarely, the minimum length neutral path can be longer than the whole sequence length. Returning to our room in the sky scraper, we could imagine entering a 'C' shaped room from one end, where the only stairs leading to the floor you want is at the other end of the 'C'. It would be less distance as the crow flies to cut across the middle, but there are walls in the way!

With these relatively rare paths longer than the Hamming distance in mind, I argue that using an exhaustive method has an advantage over simulated evolutionary or random walks (for example Reidys et al., 1997; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Smith et al., 2002, 2003), where a sub-sample of the possible paths or trajectories may not capture all the details of the space.

3.2 Inter-portal minimum path calculation

In this chapter an exhaustive calculation was carried out across the length-10 space, to find the minimum distances between all pairs of phenotypic portals within each network. Within a particular network, any sequence with a different phenotype in its local neighbourhood of $3n$ single–mutant neighbours was assessed as a potential 'entry portal' into the network. From that sequence, an expanding breadth–first search

was used to find the minimum number of neutral single–step mutations required to find an 'exit' portal to each of the phenotypes in the network neighbourhood. The neighbourhood of a network is made up of all the phenotypes that exist in the local neighbourhood of at least one sequence in the network (see section 2.1.3). Using a breadth–first search through a network, rather than just counting the number of positions that have changed between two portals, means that any paths longer than the Hamming distance are recorded as such.

In the length-10 space, 119 out of the 122 networks have more than one phenotype in their network neighbourhood. Across these 119 networks, 3,373,430 minimum–length paths were recorded between 53,432 network neighbour pairs, with the majority of paths being across the largest networks (Fig. 3.2). This amount to: For every viable sequence in the space, a separate path was recorded for each viable local neighbour of the sequence, to every network neighbour. Larger networks have larger boundaries, with more portals to more different phenotypes, and therefore more paths between them.

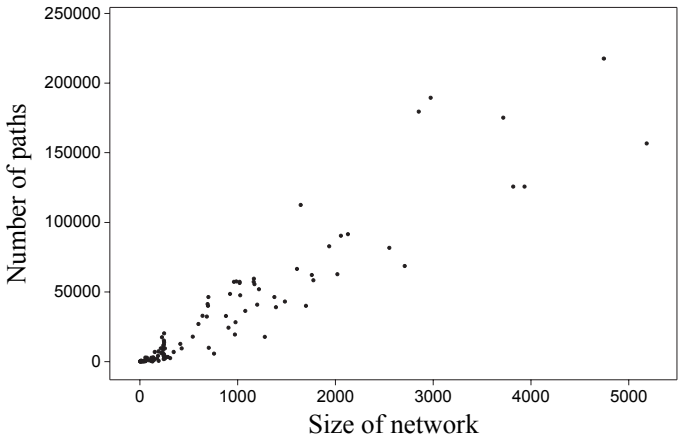
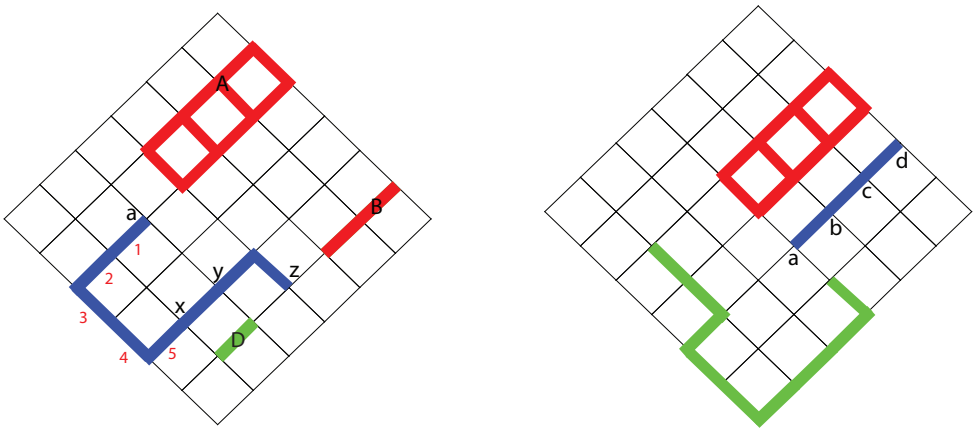


Figure 3.2: The positive correlation between the number of paths across the network, and the size of the network.

For each of the portal pairs within each network, tallies were kept of the number of unique paths, and the minimum, maximum and mean distances (see Fig. 3.3a and table 3.1 for an example). Thus a well connected portal sequence might have several paths recorded from different networks in its local neighbourhood to other phenotypes in the network neighbourhood. If the networks in a portal's local neighbourhood code for different phenotypes, they are given an inter-portal distance of zero neutral mutations: A doorway into the room can have a staircase which leads both up and

down to different floors. (See Fig. 3.1b and portal z in Fig. 3.3a and a in Fig. 3.3b for examples).



(a) **(b)**

Figure 3.3: Two examples of inter-portal calculations. See table 3.1 for the numerical synopsis. **a)** There are 4 portals on the blue network, two to the red phenotype (a & z) and three to the green (x, y and z). We can see that z is a portal to both green and red phenotypes. Traversing across the blue network from red to green there are two possible entry portals: a & z, with minimum distances of 5 (numbered on the network) and 0 respectively (to x & z). From green to red there are three portals: x, y & z. The minimum distances through blue from green to red are 3, 2 and 0 respectively. Portal a acts as an entry portal from network A, but never as an exit, because a path to z always requires fewer steps from any of the green portals than a path to a. **b)** Examining the portal sequences on the blue network: Entry from green must be via portal a, giving a distance of 0 to the red network (no neutral changes are required to move through blue to the red network). However, entry from red can occur at a, b, c, or d. The portal distances to the green network are 0, 1, 2 and 3 respectively.

Network	Portal pair	count	min	max	mean
a	Red to green	2	0	5	2.5
	Green to red	3	0	3	1.67
b	Green to red	1	0	0	0
	Red to green	4	0	3	1.5

Table 3.1: The numerical summary of the portal distances shown across the two blue networks shown in Fig. 3.3.

It is important to note that the paths across a network need not be symmetrical (Fig. 3.3). This asymmetry means that all the entry portals from one phenotype can be close to an exit to a second phenotype, without the converse necessarily being true. For this reason many entry portals will never feature as exits. Where paths longer

than the Hamming distance exist across a network, the minimum path length and the Hamming distance between the sequences were recorded. Because of the combinatorial explosion of paths as the number of sequences, networks and phenotypes increases, the exhaustive calculation of path lengths was carried out for the length-10 space only.

3.3 Results

Across the whole space, the mean inter-portal distance was 1.88 point mutation steps between any two random portals in the same network (see Appendix B for a breakdown of average path lengths by network). However, there is significant variation in the mean inter-portal distance between each pair of network neighbours. This is mainly explained by two factors: The size of the network through which a set of paths is being traced, and the number of exit phenotypes present within a particular network.

In contrast the commonness of the entry–portal within a network has no effect on the average distance to other networks, even when the path is between two rare phenotypes.

3.3.1 The effect of network size

Much of the variation in the inter-portal distance can be explained by the size of the network being traversed. As might be expected, the mean inter-portal distance shows a positive correlation with the number of sequences in the network, i.e. the larger the network, the higher the chance of having to make multiple point mutations to traverse it. However, the relationship is approximately log–linear, meaning that the average number of neutral steps required plateaus at a relatively small network size (Fig. 3.4).

This log–linear relationship is brought about by the relationship between sequence length and network size (Fig.3.5). The increase in network size between a small network and a medium network is often due to a different number of variable positions in the sequence. For each extra position which is variable, the maximum Hamming distance (the *diameter*) across the network increases, meaning that the mean inter-portal path length is likely to increase too.

However, a change in size from a medium sized network to a large network is more often due to additional alleles (bases) being present at positions where there is already some variability in the sequence. This extra variability does not normally increase

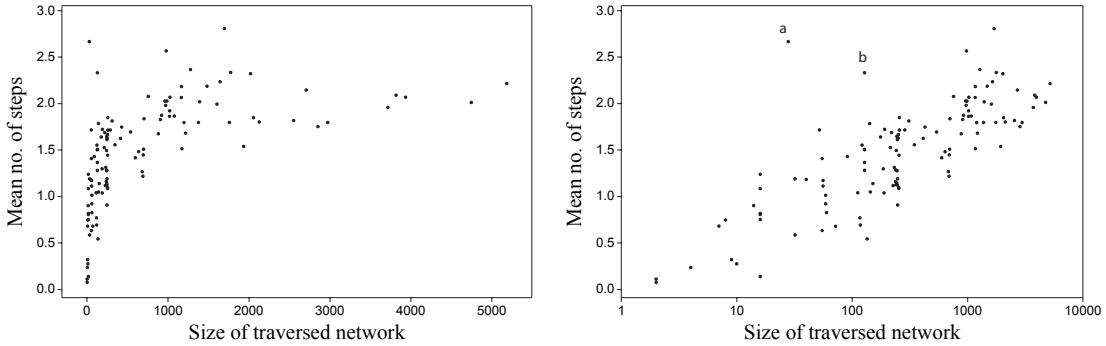


Figure 3.4: **a)** The positive relationship between the mean number of neutral steps in a network against the size of that network. Each datum is the mean across all the portal pairs in one network. **b)** The same data but with a log scale on the x axis indicating a log-linear relationship between inter-portal distance and size of network. The two outliers with higher than expected mean distances for their size both have exceptionally small network neighbourhoods for their size.

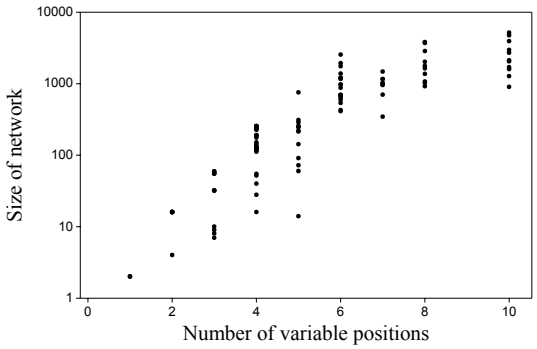


Figure 3.5: Once network size gets above a certain point, the increase in network size is mainly due to increased variability at positions where some variability already exists, and therefore does not increase the diameter of the network.

the diameter of the network, because a change from the current base to any other is equally likely at a particular position. Once this kind of saturation point is reached, any further increase in the mean path length *through the network* is dependent on the existence of path distances longer than the Hamming distance. Because path lengths longer than the Hamming distance make up only a small fraction of the total paths across a network, their contribution to the mean path length is limited. The result is that the change in mean path length per unit size between a medium and large network is smaller than the change in mean path length between a small and medium sized network.

3.3.2 Differences between exit portals

The mean inter-portal path lengths across all the networks in the space show variation due to the number of sequences leading to a particular portal phenotype. The negative correlation with increasing number is only seen when a phenotype is calculated as the exit portal from, rather than the entry portal to a network. The rarer the exit portal phenotype, the longer the path length to it on average (Fig. 3.6a). When the means for each exit portal across all networks are broken down into the means for each exit portal within a particular network, plotted against the number of exit portals in that network (Fig. 3.6b), the effect is slightly obscured by the variation in the size of the network being traversed.

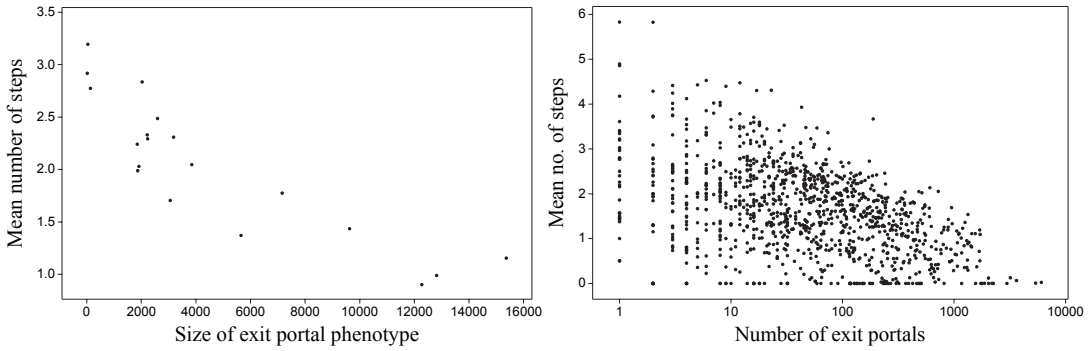


Figure 3.6: a) Mean inter-portal distances from any entry portal to each phenotype across the whole of the length-10 space against the number of sequences belonging to that exit portal phenotype **b)** Within each network, the mean distance to a particular phenotypic neighbour is correlated to the number of exit portals to that neighbour.

When the size of the traversed network and the number of exit portals to a particular phenotype are plotted against the means on the same three-dimensional graph, then almost all of the variation in the mean number of neutral steps is explained (Fig. 3.7).

As it is difficult to see the shape of the 3–D graph in the two–dimensional figure 3.7, I have used a GLM model (Fig. 3.8) to indicate how the data points lie. The small EMS shows that most of the data points lie very close to an imaginary plane drawn on figure 3.7. This model cannot be used for statistical inference due to non independence of the data points.

The fact that the number of exit portals has a negative correlation with the average inter-portal distance indicates that portal sequences are distributed approximately randomly across the network. If they weren't, there would be more variation in

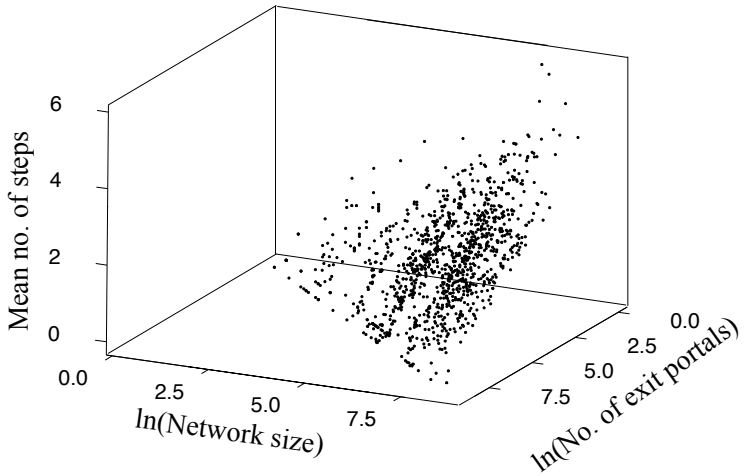


Figure 3.7: The combination of the size of the network being traversed and the number of exit portals to a particular phenotype in that network is an accurate predictor of the mean inter-portal distance between any pair of network neighbours across the length-10 space.

```
Analysis of Variance for 'mean no. of steps', using Adjusted SS for Tests

Source                DF   Seq SS  Adj SS  Adj MS        F
ln(no. of portals)     1   212.42  761.75  761.75  8154.46
ln(network size)       1   756.66  756.66  756.66  8099.98
Error               1132   105.75  105.75    0.09
Total               1134  1074.82


S = 0.305638   R-Sq = 90.16%   R-Sq(adj) = 90.14%
```

Figure 3.8: GLM fitting the model `mean no. of steps = ln(network size) + ln(no. of exit portals)`

inter-portal distances, dependent on where each group of portals was located within a network. Imagine non-randomness in the room analogy: Suppose there are 11 staircases leading out of the far end of the room, 10 leading to one floor and 1 to the other. It doesn't really matter where you start from in the room, the mean inter-portal distances are roughly the same to an exit to either of the two possible floors, despite many more sets of stairs being available to one than the other. If however, the doorways are distributed randomly around the room, the chances of coming across one of the 10 staircases first, is much higher than coming across the single one. Because we see mean inter-portal distances correlated very strongly with the number of exit portals, we can conclude that they are not normally clustered in a particular section of the network. This conclusion is enhanced by the lack of correlation between the

number of entry portals into a network and the mean distance to the other phenotypes across the network (Fig. 3.9a). Each new entry portal that is added at random into

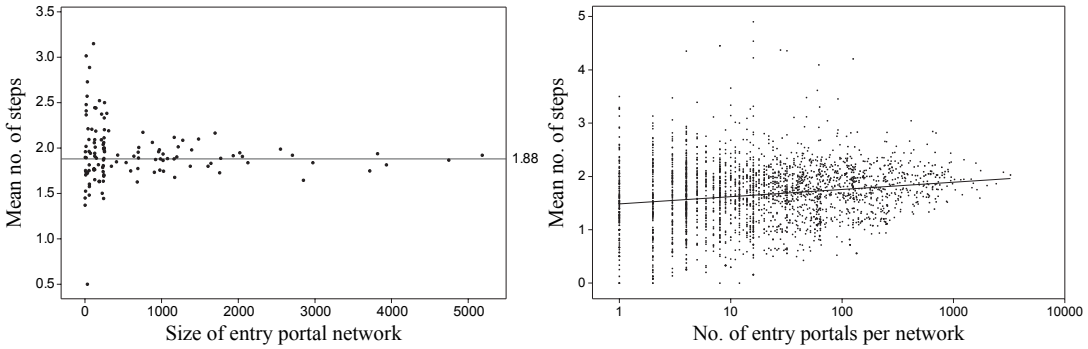


Figure 3.9: a) Mean inter-portal distance from each entry portal phenotype to all exit portals across all networks against the number of sequences in the entry portal network. **b)** Mean inter-portal distances for each entry portal to all exit portals within a particular network against the number of entry portals. There is a slight positive correlation with increasing network size, but this is explained by a positive correlation between the number of portals and the number of sequences in a network (the maximum number of portals is limited by the size of the network). Both measures do not really show any unusual variation from the mean.

a network will on average be the global mean distance away from a particular exit. If entry portals were not distributed randomly, for example collected in one corner of a network, then we would expect to see a difference from the standard distribution seen around the mean 3.9a, especially at the higher numbers of entry portals, where any effect would be exacerbated.

3.3.3 Minimum number of steps

The minimum *possible* number of steps between two portals corresponds to the Hamming distance between them. However, there is no guarantee that a direct Hamming distance path will remain on a neutral network. Any time a minimum path involves more than one change at a particular position, the path becomes longer than the Hamming distance between the entry and exit portals. This occurs when a particular base at one position only becomes neutral once a change or changes at other positions have occurred (Fig. 3.10).

In section 2.5 the *density* of a network was suggested as an indicator of the number of potential routes between segments of the network. Network density can differ significantly, even between the disjunct networks of a single phenotype (See table 2.10). This means that there is marked variation between networks as to how easy it is to

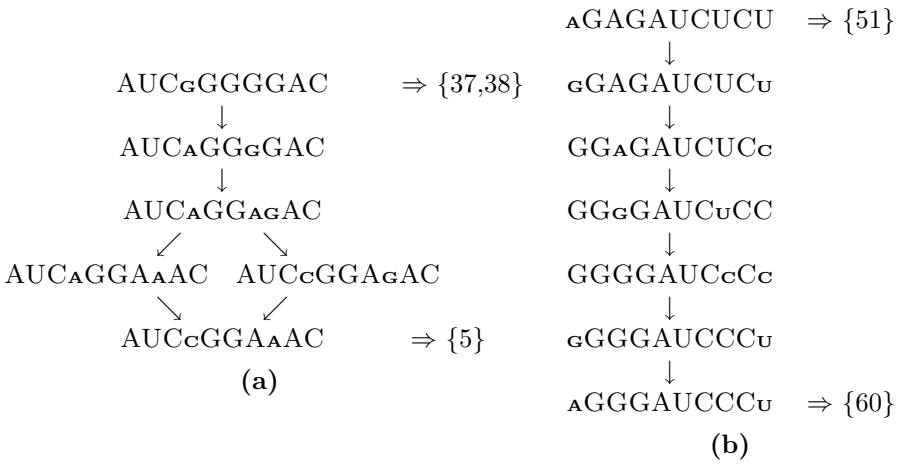


Figure 3.10: a) The only minimum path length longer than the Hamming distance in NID-4. The minimum path between a portal to NID-37 and NID-38 and a portal to NID-5 is 4 steps, while the inter-portal Hamming distance is only 3 base changes. Here a change is required from a 'G' to an 'A' at position 4, before the change at position 7 become neutral. After that change has been made, the two final chances can be made in either order. **b)** NID-26 has many paths longer than the Hamming distance. From the initial portal sequence (AGAGAUCUCU), the phenotype is stable with a mutation to a G-U base-pair at the 1^{st} and 10^{th} positions. However, it is not stable at the 3^{rd} and 8^{th} positions. Once the 1^{st} and 10^{th} positions have the stronger pair of a G-C rather than A-U, then a G-U base-pair *is* stable at the 3^{rd} and 8^{th} positions. Once the A-U at positions 3 and 8 has changed to G-C, the 1^{st} and 10^{th} positions can revert to A-U, to become a portal to NID-60. A total of 6 changes, where the Hamming distance between the portals is just 2.

traverse each one. At high density there are many possible neutral paths between two portals because most of the possible base combinations are present in a network, hence changes at one position are usually independent of the changes at another. When the density of a network equals one, the minimum path length is always the Hamming distance between two sequences, because any change at a given position is independent of changes at every other position. However, as the density drops, routes longer than the minimum hamming distance occur with a higher frequency (Fig. 3.11). A full list of the networks containing minimum inter-portal paths longer than the inter-portal Hamming distance is shown in Appendix C.

Even at a very short sequence length (10 bases), the structure of the space is complicated enough that the Hamming distance path is not always obtainable. Although these paths can number up to 2000 sequences within the length-10 space, they still make up only a very small fraction (usually less than 1%) of all the paths across a network. The highest proportion was the 3.63% in NID-82. It is particularly interesting to note that there are three networks within which there are inter-portal paths as long as the sequence length, and in one case (in network 68) some paths actually

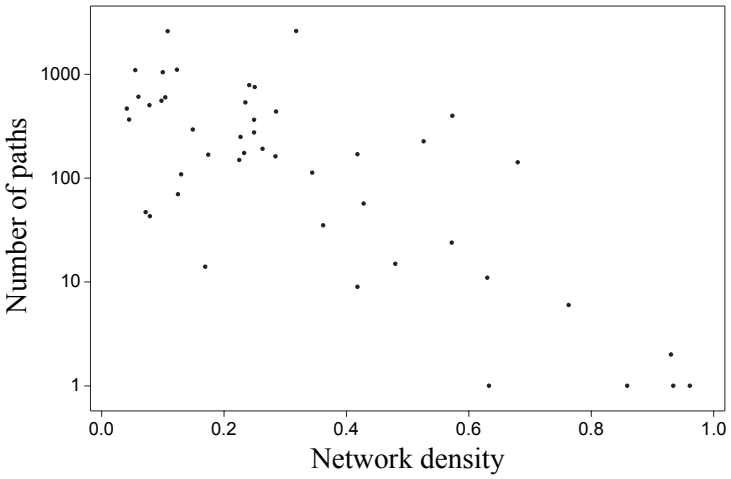


Figure 3.11: The relationship between the number of paths longer than the Hamming distance between individual portal-pairs in each network, and the density of that network.

exceed the sequence length.

3.4 Summary and discussion

Within the RNA genotype–phenotype map, we have already seen in chapter 2 that the neutral networks in genotype space are large, map to a relatively smaller set of phenotypes, and almost all are connected to a subset of the other networks in the space. Networks can be said to be connected when they abut their network neighbours at *portals*, where a simple single-step mutation links the networks coding for different phenotypes.

The portals are spaced across each network in such a way that the mean distance between any two random portals to different neighbours is reasonably large (1.88 neutral steps), indicating that portal sequences do not all share the same phenotypic neighbours in their local neighbourhoods (Fig. 3.1a rather than Fig. 3.1b). However, there are two important factors that strongly influence the expected inter-portal distance required to find a particular phenotypic portal: First, there is a positive non-linear relationship between mean distance and network size. Mean inter-portal distance across a particular network increases rapidly from small to medium sized networks, but less rapidly as network size gets even larger. Second, the rarity of portals to a particular phenotype has a negative log-linear correlation with the mean distance.

Although the distances between portals are shorter in smaller networks, having to make neutral changes to find a certain phenotype is not a phenomenon limited to the

largest and therefore most commonly encountered networks. This means that even if fitter phenotypes tend to map from smaller neutral networks, as suggested by van Nimwegen and Crutchfield (2001) in their 'royal staircase' model, neutral drift is still likely to have a part to play – if for no other reason than that the number of steps required to find a portal to a small network is likely to be higher than average in the first place.

The fact that the number of exit portals has a negative relationship with the distance across a network, independent of network size, means that they are likely to be distributed approximately randomly across the network. If they weren't, then we would not expect to see such a clear correlation between distance and the log of the number of portals. Given the random nature of the distribution of portals in a network, it seems unlikely that there are there are areas of the space with many portals in different networks all linked together, and reducing the effect of neutral networks. In other words, the skyscraper we encountered earlier does not have fire escapes running the height of the building – linking all the floors together without having to enter any of the rooms. This will be further investigated in chapter 4.

In this chapter I have established that portals to different phenotypes are not connected directly through the RNA genotype space, which means that neutral drift across a network is likely to play a significant role in the adaptation of a population or species evolving over a series of networks. This is because a population evolving on the map would be liable to quickly reach a point where there were no advantageous mutants within each individual's local neighbourhood. At this stage, by drifting neutrally across the network it found itself inhabiting, a population has the potential to increase its evolutionary options, because some individuals eventually mutate into different phenotypes which were not accessible from the initial genotypes first encountered on the network.

However, the mean inter-portal distances calculated here only apply to the inter-portal distance between two specific phenotypes. In contrast, when a population is not very close to a 'peak' in the 'landscape', the mean neutral step distance to any advantageous neighbour is likely to be significantly less. This is because as the number of advantageous phenotypes increases, the number of advantageous portals increases greatly – and we have already seen in this chapter that the more phenotypic portals off a network the lower the mean distance until one of those portals is found.

As a population becomes better adapted and as advantageous mutations get rarer, discovery of a fitter phenotype is more likely to require a neutral step or more. We

might expect that the mean distance across the penultimate network in an evolutionary trajectory to be the 1.88 neutral mutations required here.

Even within the short sequence–length map studied here, there are a number of paths which are longer than the inter-portal Hamming distance. These paths although not common, are not so rare that we might expect to never come across them if this model is a reasonable approximation of a real genetic space. If this situation arose reasonably often, then simply using the Hamming distance for comparisons of phylogenetic closeness might underestimate the distance between two divergent sequences like those seen in figure 3.10a. This occurs in a similar way to underestimates due to multiple substitutions at a single site when more divergent species are compared (e.g. Guadet et al., 1989). A more interesting scenario is if we consider the example given in figure 3.10b. In this case, the bases at some positions must change along the path, but subsequently revert to their initial values. If we simply compare the Hamming distance between the two sequences at either end, we could not know that positions 3 and 8 had changed at all. In other words it is possible that in networks with very low density, Hamming distance measures become an increasingly unreliable method of measuring phylogenetic closeness as suggested by Novella et al. (2004).

When we consider the network densities of the length-12 and length-14 genotype spaces (see figure 2.15), we can see the larger networks have even larger and less dense networks, and it is therefore likely that as sequence length increases and average density decreases, paths longer than the Hamming distance play an increasingly important role.

Chapter 4

Minimum path lengths across RNA genotype space

4.1 Introduction

In this chapter, I investigate the impact of neutral mutations within an evolutionary trajectory. The hypothesis from chapter 3 is that neutral mutations across a neutral network increase the accessibility of new (and potentially fitter) phenotypes. This 'drift–based search' has the potential to be an extremely important mechanism by which an evolving population can avoid phenotypic stasis from which no further phenotypic innovation is possible.

We saw in the previous chapter that the average neutral path length across a network was 1.88 steps in the length-10 space. However, this figure was calculated by considering every pair of phenotypic network neighbours independently in turn. When *all* the phenotypic connections in a network are considered, the distance to any different phenotype will be far less than 1.88 steps. This means that, at least initially, the inter-portal distances across a network are likely to be far less (Fig. 4.1b). As a walk gets closer to the optimum we might expect the mean number of neutral steps required to increase, with the final step in the length-10 space averaging 1.88 mutations.

Returning to the sky scraper analogy: When we are near the bottom of the building most of the staircases lead upwards, so the first set a blind man comes across is likely to lead up. In contrast near the top of the building most lead down, so a person must search more of the room to find an 'up' staircase. If the building contains fire-escapes leading all the way to the top, a person finding one of them would not need to even enter any of the rooms on the way up.

To test this hypothesis, I first assign ranks to each phenotype, with higher ranked phenotypes designated 'fitter'. I use an adaptive walk simulation where each step must be adaptive or neutral within the exhaustively calculated genotype space from chapter 2. The result is an approximate measurement of the number of steps (adaptive *and* neutral) usually required to find the fittest available phenotype.

This model is far more complex and nuanced than that of the preceding chapters. However, analysis of the detail can provide interesting insights into the particular nature of the space, as well as enabling us to draw some more general conclusions. The main ones being: first, that neutral steps through one or several networks are often a prerequisite to the highest fitness phenotypes becoming accessible. Second, that on any given adaptive walk, neutral steps do make up a significant proportion of the total number of mutations along its path (Fig. 4.1a rather than Fig. 4.1b).

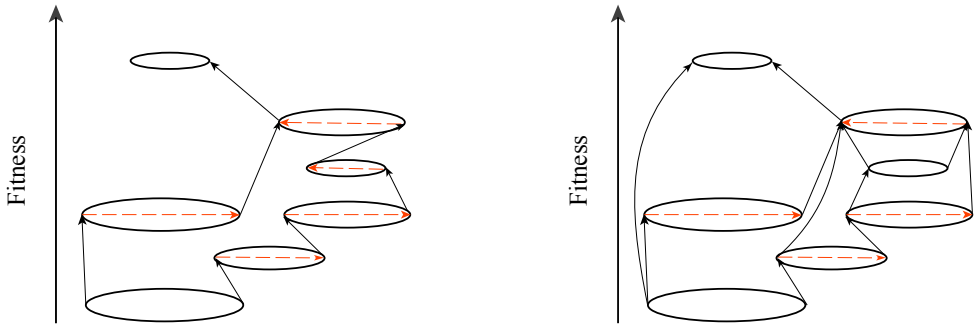


Figure 4.1: a) An example genotype space structure, where long neutral pathways (red dashed lines) are important. The distribution of portals means that many neutral steps are required to get access to fitter phenotypes. **b)** In a better connected genotype space, there is often an adaptive step accessible, even though some long pathways do exist.

The model makes a number of important and potentially unrealistic assumptions about the nature of the genotype space, and how an evolving population might tackle it. However, these assumptions are a necessary step to establish a framework from which work can progress, and analysing the effect of changing them can itself reveal interesting results, as will be demonstrated in sections 4.4-4.6.

I shall now briefly review other models which have considered the availability of adaptive mutations within a discrete genotype space in the next two sections (4.1.1 & 4.1.2), before introducing my own and presenting its results. These models fall into one of two main categories, depending on whether fitness is assigned to the phenotype, or directly to the genotype. However, both are linked by the assumption that available mutants are limited to single point substitutions from a given genetic sequence. At the molecular level, the accessible mutants of any sequence are the $3n$ local neighbours which differ from it by one base at just one position.

In those based around a set of genotypes with fitnesses directly assigned, neutral mutations are generally assumed not to exist, leading to a one–to–one mapping between genotypes and fitnesses, though genotypic fitnesses can be correlated (e.g. Gillespie (1984); Kauffman and Levin (1987); Orr (2003); Rosenberg (2005); Orr (2006b)).

Where fitness is assigned to a phenotype, it can also be randomly assigned or correlated with some notion of phenotypic similarity. Either way, it is normal to use an explicit, biologically grounded genotype–phenotype function to map genotypes to phenotypes. This usually results in some degeneracy between genotype and phenotype

(a many–to–one mapping) and hence the existence of extended neutral networks between genotypes coding for the same phenotype (e.g. Huynen et al., 1996; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Smith et al., 2003).

4.1.1 Genotype–fitness models

In 1984, Gillespie presented a genotype space model which predicted the mean number of steps taken over an evolutionary walk. He showed that after a shift in the adaptive landscape displaces the population from the optimum genotype, evolution should respond by favouring a series or 'burst' of local adaptive mutations which return the population to the highest attainable fitness.

The model assumes that only local mutants are accessible, and that no neighbours are neutral, but instead have randomly distributed fitnesses. This means that the landscape does not have a degenerate many–to–one mapping, and is very rugged, with many local optima (Kauffman and Levin, 1987; Kauffman, 1993; Orr, 2003), and thus the path lengths are short, before the supply of accessible adaptive mutations are exhausted.

Orr (2003) proposed that a *greedy* fitness algorithm (one which always selects the fittest of the available mutants Kauffman and Levin, 1987) produces the minimum average path length for adaptive walks starting at a random sequence and continuing to a local optimum. He argued that there is no stronger way in which selection can act because the fittest option is always being taken and on average that route will take fewer steps to reach the optimum. He went on to calculate that the mean number of steps under this fitness algorithm was e-1, where $e \approx 2.72$. I shall return to this result later in this chapter to use as a null hypothesis, when considering the effect of a many–to–one mapping has on path length.

Kauffman and Levin (1987) and Kauffman (1993) also used a genotype to fitness model based around NK boolean networks to produce an adaptive landscape with a tunable level of epistatic interactions, making the fitnesses correlated. The inclusion of epistatic interactions adds enough complexity that the walks in the landscape become too complicated to resolve analytically, especially when neutral networks are also included (e.g. Barnett, 1998; Newman and Engelhardt, 1998; Smith et al., 2002). The model essentially becomes as complex as a genotype–phenotype model and requires simulations of adaptive walks through the space to calculate path lengths, making them more similar to the models outlined below, than the analytical models outlined above.

4.1.2 Genotype–phenotype models

This type of model explicitly maps genotypes to phenotypes, with fitnesses assigned to each phenotype. They can be broken down into two sub-categories. The first involves simulating a population undergoing adaptive walks in a large genotype space (Huynen et al., 1996; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Smith et al., 2003). The second, illustrated by the work of Schuster et al. (1994), Grüner et al. (1996a), Grüner et al. (1996b), Reidys et al. (1997), Reidys et al. (2001), Deeds et al. (2003) and Sumedha et al. (2007b), and to a lesser extent in chapter 2, relies on a topological or graph–based approach to mapping out sequences, networks and phenotypes.

Many of the simulation models show that phenotypic innovation often only occurs after a population has undergone a period of genetic drift. Van Nimwegen and Crutchfield (2000) used a very simple mapping, similar to traversing a single network, to show that population size and mutation rate can have an important effect on the number of generations it takes to find an adaptive portal on the far side of a network. However, in the more complex models which include multiple networks and connections, most of the work has focused on 'observing' populations undergoing genotypic drift between phenotypic changes, rather than examining the contribution that the constituent factors such as the underlying structure of the space and the stochastic dynamics of the population make to it. In general, attempts to consider the effect that the underlying structure has on mutational accessibility have focused on measuring its geometry or topology directly.

These geometric or topological approaches generally reduce the space to a set of graph–type properties and statistics, and one of the original results was to show the existence of extended neutral networks in the first place. Schuster et al. (1994) and Sumedha et al. (2007b), among others, have shown that any sequence of a 'common' phenotype is only a few mutations (a short Hamming distance) away from all the other common phenotypes in the space, so called 'shape space covering'.

This kind of method places much emphasis on the geometric or topological relationships within the space (or graph), but as Sumedha et al. (2007b) point out, when the aim is to investigate the *biologically plausible* evolutionary trajectories implied by the network structure, using measurements which do not consider the restrictions placed on mutations by phenotypic constraints will often identify a shorter *geometrically possible* path than is *biologically* possible. A short Hamming distance between two sequences does not necessarily mean that there is a short accessible path between those two points (Fig.4.2).

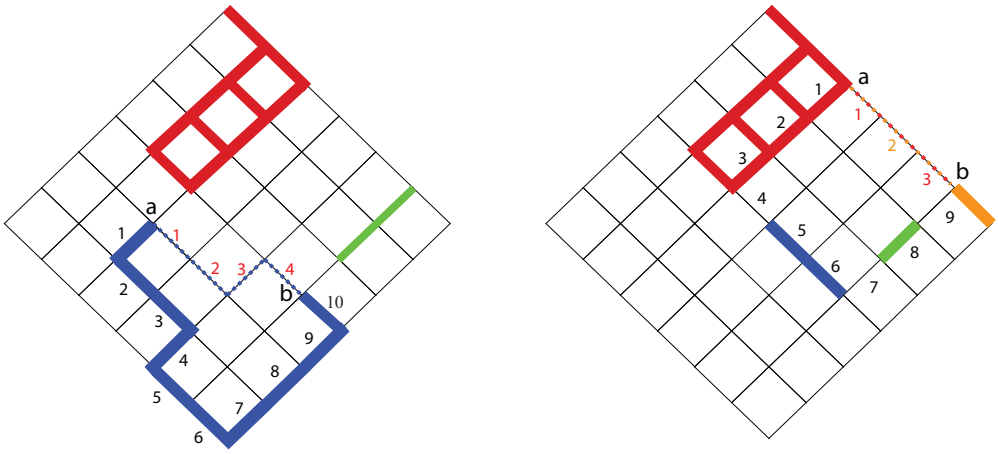


Figure 4.2: The arrangement of the neutral networks can impose extra restrictions on the availability of particular sequences, not captured by simply considering the Hamming distance between them. **a)** Within a single network. **b)** The effect can be amplified over multiple networks. In this case no single neutral path is longer than the Hamming distance, but the adaptive walk through all the networks is substantially longer than Hamming distance between the start and end sequences.

In summary, the direct genotype–fitness models use a simple adaptive landscape lacking neutral mutations. The genotype–phenotype models tend to either observe populations evolving through genotype space, without distinguishing between the effects of population dynamics and the underlying structure of the space, or examine the geometric and/or topological structure of the space, without considering how that might affect a population evolving through it.

4.2 Model

The model presented here, aims to bridge that gap by using a simple adaptive walk method similar to that used by Orr (2003) within a space closer to those used in Schuster et al. (1994), Grüner et al. (1996a) and Sumedha et al. (2007b).

To this end, each phenotype was ranked randomly with no ties, except for the unfolded phenotype (PID0) which was always considered fatal and given a rank of 0. An increase in rank was assumed to confer such a large fitness advantage that a phenotype was always selected over one ranked lower, and always beaten by one of higher rank. At any one step, the adaptive walk always takes the fittest accessible phenotypic option (the greedy fitness algorithm). If no adaptive mutant is directly accessible, an expanding breadth–first search through the network is conducted until

one is found, or the optimum reached. These strict assumptions imply that stochastic population–level effects do not play a role in defining the trajectory at any stage. They are in place to calculate an estimate of the *minimum* number of steps in a walk, from which other factors such as mutation rate and population size might increase, but are unlikely to reduce the number of steps. In certain circumstances these assumptions may be close to the truth for some fast evolving organisms. For example, Wahl and Krakauer (2000) calculated that a virus can produce thousands of copies of every point substitution neighbour in every generation, in which case the fittest mutants may well usually be fixed.

Though the calculation of *average* path length produces a minimum, each walk has no 'foresight' over which option to take at any particular stage, and therefore cannot be guaranteed to find the absolute minimum path through the space. Figure 4.3 shows an example of where selecting an adaptive mutant which was not the fittest would open up a shorter path to the optimum.

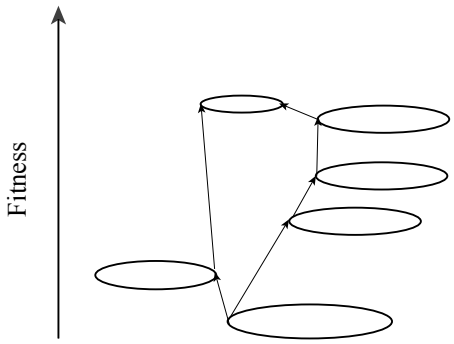


Figure 4.3: With the lack of foresight of the adaptive walk method, any given path may not actually be the shortest way of crossing the space. In this example selecting the fittest mutation at the first step leads to a longer path than if the less fit mutation had been selected, despite both paths ending with the same phenotype.

Because of the added complexity of navigating neutral networks when compared to Gillespie or Orr's one–to–one landscape, the model includes three different adaptive walk methods of calculating the minimum paths. Each serves a different purpose in terms of elucidating the affect that the underlying structure has on an evolutionary trajectory. They share certain properties: each continues until no adaptive mutant is accessible; if there is more than one accessible mutant to the fittest phenotype, then one of them is chosen at random.

The three ways of calculating adaptive walks are laid out in figure 4.4, and explained here: In the first, each step must be advantageous. The walk finishes when

there are no adaptive mutants to be found within the local neighbourhood of $3n$ genotypes surrounding the last step of the walk (see Fig. 4.4a). This mirrors Gillespie and Orr's mutational landscape models in that the availability of adaptive steps is limited to a sequence's local neighbourhood. It differs in the fitness distribution of the local neighbours, which are assigned according to the underlying RNA genotype space, not randomly. This shall be referred to as the Local Neighbourhood (LN) walk.

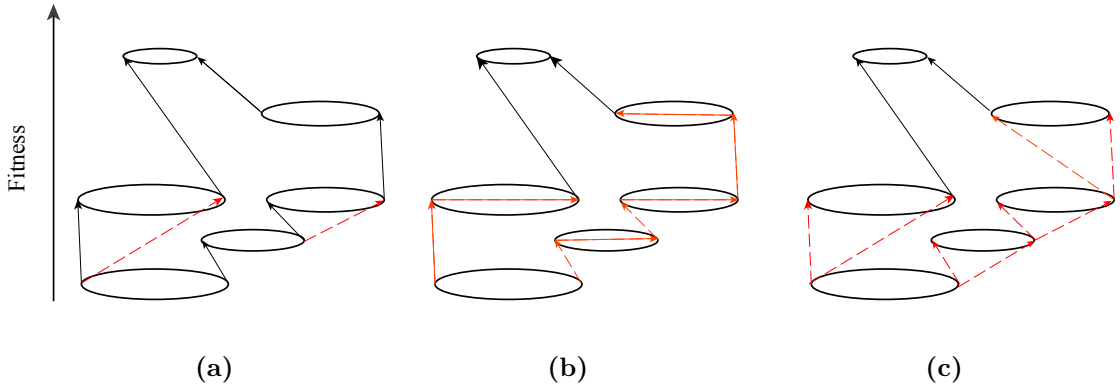


Figure 4.4: Red dashed arrows indicate paths with the potential for further advantageous mutations. Black arrows indicate paths with no further possible advantageous mutations. **a)** The Local Neighbourhood (LN) process: All the local neighbours of a sequence are available, but the network neighbourhood is not. **b)** The Neutral Step (NS) process: If an adaptive mutant is not available, a breadth–first neutral network search is used to find the closest adaptive portal. **c)** The Network Mapping (NM) process: The entire network neighbourhood is immediately accessible to any sequence in the network. The fittest network neighbour is chosen irrespective of which is closest to the current sequence.

The second, Neutral Step (NS) method, like the LN method, always takes the most advantageous option available locally. If no adaptive step is possible, but neutral steps are, a breadth-first search through the network is carried out to find the nearest advantageous portal genotype (Fig. 4.4b).

The third method, Network Mapping (NM), does not search for the fittest adaptive phenotype in the *local* neighbourhood of a sequence, but instead considers every phenotype accessible from somewhere in the whole neutral network instantaneously (Fig. 4.4c). This effectively reduces the genotype space from the 4^n unique sequences to the number of neutral networks, because all the sequences in one network have the same accessibility. In this situation, the size of the network neighbourhood is variable, unlike Gillespie's and Orr's adaptive landscape models. However, the fitness dynamics are more similar to the mutational landscape model than during an LN walk, because none of the neighbours are neutral and each phenotype's fitness is randomly

assigned, though the existence of disjunct networks mean that the mapping is still not one–to–one.

The network mapping method is partly included to establish if there is a negative effect of taking the minimum neutral path across a network (in the NS walk). If the distribution of portals in a network means that the fittest genotypes are rarely the closest to the entry portal, then being restricted to a breadth–first search across a network might lower the chance of the walk reaching the global optimum (Fig. 4.5a). In contrast, the network mapping method can also be used to ascertain how well the networks are connected to each other. The more well connected the space is, the fewer the number of steps are needed to reach the optimum using network mapping (Fig. 4.5b).

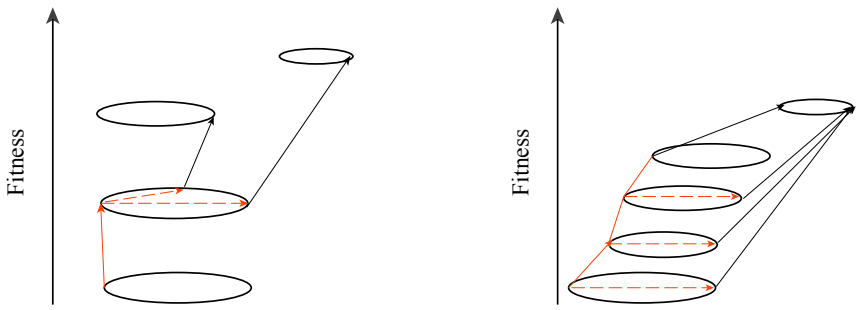


Figure 4.5: a) In this set of networks, the NS walk through the neutral network gets stuck in the lower local optimum because its portal lies closer to the entry portal. By contrast, the network mapping method ignores distance across the network and therefore reaches the fitter (global optimum) phenotype. **b)** In this set of networks the optimum is found by all three walk methods. The network mapping method requires only one step, because there is a portal from the starting network directly to the optimum, whereas the other two each require 4 steps.

4.2.1 Simulation parameters

The adaptive walk approach taken in this chapter is the first in this thesis not to involve exhaustive searching of the space. In the case of the length-10 network, an exhaustive search testing all sequences through every possible ordering of phenotypes amounts to at least $4^{10} \times 19! = 1.28 \times 10^{23}$ paths, an infeasibly large number.

Instead, I sample the paths across the exhaustively calculated genotype spaces from chapter 2. Using an exhaustively calculated genotype space confers certain advantages compared to performing simulations over a genotype space which is not.

The key one is that any neutral steps can be calculated by using a breadth–first search rather than simulating the random drift of a population. A breadth-first search guarantees to find the minimum number of steps across each network, where as a drifting population does not. Furthermore, it is simple to calculate if fitter phenotypes will ever become accessible in this model, saving computational time on a futile search. As the total set of phenotypes are also exhaustively calculated it is simple to assign fitness ranks to them. This all means that the focus is on the accessibility of mutants due to network structure, rather than understanding how the dynamics of an evolving population are influenced by that structure.

The paths for the three different adaptive walk methods were traced through the space starting from 1000 randomly chosen non-PID0 sequences. The average path lengths for each method were calculated for the same 1000 sequences over a number of different phenotype rankings: 100 at length-10, 25 for length-12, and 5 for length-14.

Walks in the length-10 space were repeated using a much larger sample of 10,000 initial starting genotypes for 20 different random fitness orderings, and for 500 initial sequences through 200 fitness orderings, and neither showed any difference in the results.

The other assumptions of the model are similar to those used by Orr: the population is effectively reduced to a single point in genotype space; mutation is weak enough that we can ignore double and triple mutants as being very rare (Maynard Smith, 1970; Gillespie, 1984; Orr, 2003), and also weak enough that we can assume selection to be much stronger than mutation, allowing phenotypic changes to occur instantly.

4.2.2 Minimum path length predictions

Orr calculated analytically that the minimum average number of steps in a greedy adaptive walk was ≈ 1.72 steps across an uncorrelated but discrete landscape with a one–to–one mapping between genotype and fitness. I shall use this calculation as the basis of a null hypothesis with which to test the effect of imposing a more complex many–to–one mapping on the space.

As Orr assumed a long sequence length in his model, he could make certain approximations when calculating the figure of ≈ 1.72. At the short sequence lengths used in this chapter, I cannot. Therefore I shall now outline how I have adapted his calculations, influenced by a further paper by Rosenberg (2005), to achieve the modified approximations in table 4.1.

Orr's assumption of long sequence length meant that he could ignore the chance of a walk starting at a local optimum. The same assumption meant he could also

conveniently ignore the slight reduction of unique local neighbours at each step (the second step in a walk does not introduce another $3n$ neighbours, but actually another $3(n-1)$, where n is sequence length). The mean path length was calculated by Orr (2003, eq. 4) as:

$$\bar{L} = \sum_{K=1}^{\infty} K P_K$$

where $\bar{L}$ is the mean number of steps, K is the exact number of steps and P_K is the probability of the path containing exactly K steps. At a short sequence length, and therefore relatively small neighbourhood, we must include the probability that the initial sequence is a local optimum, first pointed out by Kauffman and Levin (1987) as being:

$$P_0 = \frac{1}{(3n+1)}$$

It is then possible to substitute Rosenberg's equation for P_K for an exact sequence length (n) (Rosenberg, 2005, eq. 2) in place of Orr's:

$$P_K(n) = [1 - \alpha_K(n)] \prod_{k=1}^{K-1} \alpha_K(n)$$

where n is the sequence length, and $\alpha_K(n)$ is the probability that at least one of the neighbours of the current sequence has a higher fitness. The product term is the probability that the adaptive walk has a length of K steps or more. α_K can be calculated exactly for low K from table 1 in Rosenberg (2005), and was calculated up to K=20 to give the approximations in table 4.1. The resultant estimates of path lengths are slightly lower than the ≈ 1.72 given by Orr.

Sequence length	$\bar{L}$
10	1.60
12	1.62
14	1.63

Table 4.1: The average path length calculated for different sequence lengths according to Orr's mutational landscape model

I propose two opposite ways in which the existence of a many–to–one genotype–phenotype function might influence the path length. First, when a sequence is part of a neutral network, the number of neighbours around the boundary of the network reachable by neutral drift can be much larger than the $3n$ local neighbours (where n

is sequence length). This larger number increases the probability that at least one neighbour is positive, and thus can increase the mean number of steps in an adaptive walk.

Conversely, an explicit genotype–phenotype function imposes a strict upper limit on the number of possible phenotypes. This means that especially at shorter lengths, there is a high chance of reaching the global optimum after just a few adaptive steps, or even starting at a local optimum, reducing the mean number of adaptive steps.

4.3 Results

4.3.1 Path length

The mean path length of the neutral step adaptive walk is far longer than predicted by Orr, and on average includes a number of neutral steps (Table. 4.2). This indicates that network structure within the RNA space is arranged so that higher fitness phenotypes often only become accessible after a period of neutral drift.

In fact, table 4.2 shows that the mean total number of steps in a neutral step walk is approximately double the number of advantageous steps, meaning that on average *each adaptive step requires a neutral one.*

4.3.1.1 Neutral step walk

The number of *adaptive* steps in the neutral step (NS) walk is larger than the number of adaptive steps predicted in table 4.1. The large network neighbourhood increases the chance of finding another adaptive mutant, and therefore increases not just the mean path length, but the mean number of adaptive steps. In fact, as well shall see in section 4.3.2, the end point of many path lengths is reaching the global optimum, rather than getting stuck in a local one.

The mean total path lengths show that when neutral network structure is taken into account, shape–space covering becomes a bad predictor of the *evolutionary* distance between phenotypes across the genotype space. Sumedha et al. (2007b) suggested that even though most phenotypes could be found by changing the bases at just 20% of the positions of a genotype, the conditions imposed by network structure mean that many phenotypes may not be accessible at that distance. In the RNA genotype space, we see this to be true. When network structure is considered, the average path length is more than 40% of the sequence length and increases for longer sequence lengths.

(a) Advantageous mutations

Length	Type	mean	med	min	max
10	LN	0.69	1	0	6
	NS	1.83	2	0	7
	NM	1.04	1	0	4
12	LN	0.97	1	0	6
	NS	2.47	2	0	9
	NM	1.07	1	0	4
14	LN	1.15	1	0	6
	NS	3.11	3	0	9
	NM	1.12	1	0	3

(b) Total mutations

Length	Type	mean	med	min	max
10	NS	3.72	3	0	17
12	NS	5.19	5	0	32*
14	NS	6.90	6	0	23

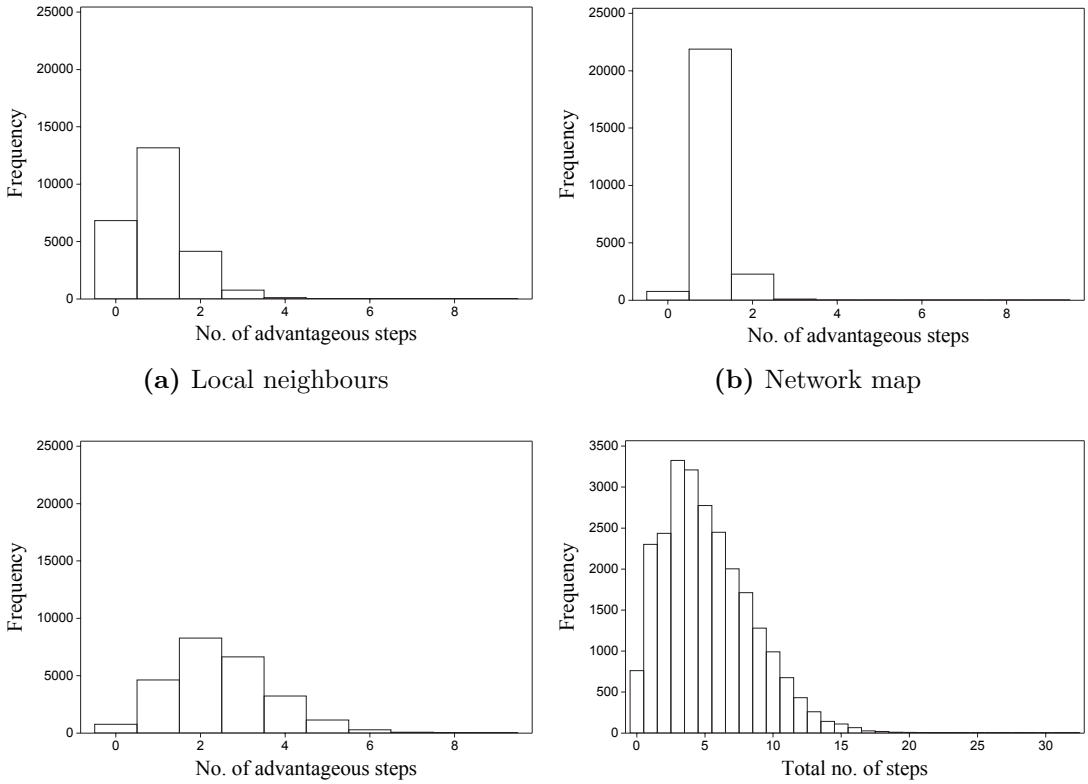
Table 4.2: **a)** Summary of the number of adaptive steps required to reach the maximum attainable fitness for lengths 10-14. The number of different sequences tested was 1000, the number of different generations tested was 100 for the length-10 space, 25 for the length-12 space and 5 for the length-14. **b)** Summary of the total number of mutations to reach the final fitness for the NS walk for lengths 10-14. *The longest path was found in the length-12 space. The larger sample size (25 fitness orderings) led to a more extreme value being recorded than in the length-14 space.

Indeed, the longest paths can be as much as two and a half times the sequence length, however these are at the extreme right hand tail of the distribution of path lengths, and relatively few paths are longer than the sequence length (Fig. 4.6).

Though relatively few paths are longer than the sequence length, the path length *is* often longer than the Hamming distance between a walk's start and end genotypes (Fig. 4.7). The proportion of paths which are longer than the Hamming distance increases with increasing sequence length, with over 50% of all paths longer than the Hamming distance in the length-14 space.

Table 4.3 gives details of the longest neutral step path in the length-12 simulation. There are 7 phenotype transitions and 25 neutral mutants required to reach the optimum.

In this walk, the final phenotypic transition involve a neutral path longer than the Hamming distance between the portal sequences. The 17 neutral steps within NID-



(a) Local neighbours **(b)** Network map

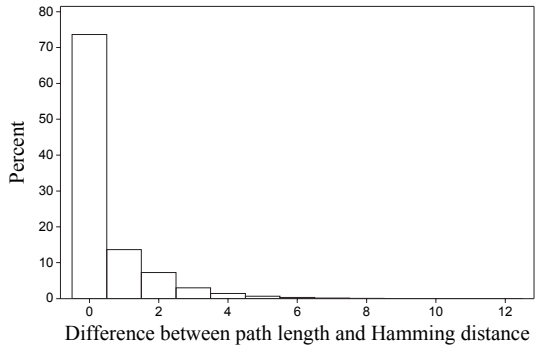
(c) Neutral step

Figure 4.6: Distribution of paths lengths within the length-12 network according to the different adaptive walk methods. NOTE: Because neutral mutations are not counted in the LN or NM walks, the number of adaptive steps and the total number of steps are the same.

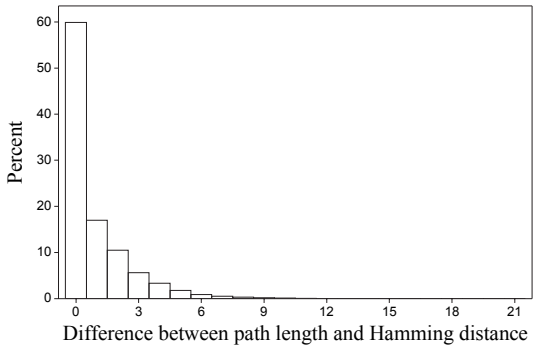
196 is two steps short of the longest path recorded within a single network (also NID-196). The 19 neutral steps between UCUCUACGGUGG and CCUCGAAAGUGA cover a direct distance of just 5 mutant steps 'as the crow flies'. NID-196 has a particularly low density of 6.59×10^{-4} making it an excellent candidate for long and winding neutral paths.

4.3.1.2 Local neighbourhood walk

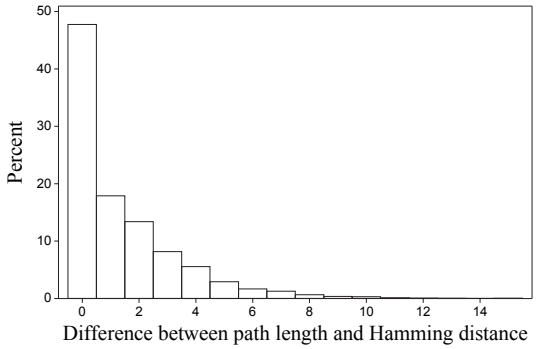
In comparison to the neutral step walk, there are significantly fewer steps in the average local neighbourhood walk (Table 4.2, Wilcoxon signed ranks test statistic length-10= 70092.5, N for test= 74167, p< 0.0005; length-12= 9468.0, N for test= 20845, p< 0.0005; length-14= 23.0, N for test= 4506, p< 0.0005, for all sequence lengths). However, the average path length does increase as the sequence length increases. In fact, by length-14 a local neighbourhood (LN) path contains more steps on average than in the network mapping walk. The path length increases because



(a) length-10



(b) length-12



(c) length-14

Figure 4.7: Distribution of the difference between neutral step path length and the Hamming distance between the start and end genotype for each path at lengths 10-14.

Sequence	NID	fitness	rank increase	neutral steps
GGUAUGCCUGAC	320	8	8	0
+++C++++++++	94	16	-	1
C++++++++++	94	16	15	0
++G+++++++++	380	31	2	0
++++++++A+++	321	33	-	1
++++A+++++++	321	33	26	0
+++++A++++++	319	59	-	2
+++++++++CG+	319	59	6	0
++U+++++++++	389	65	-	4
++++U++GG++A	389	65	5	0
+C++++++++++	196	70	-	17
++++G+AA++++	196	70	1	0
CCUGGAAAGCGA	439	optimum		

Table 4.3: The longest path recorded through the length-12 genotype space. '+' indicates the base has not changed from the *previous* genotype. Every position in the sequence has changed at least once in the 32 changes across the path, but the Hamming distance between the starting and final sequences is only 11, and the Hamming distance across NID-196 is 3.

with the increase in sequence length comes an increase in the number of phenotypes as well as the number of viable neighbours (the proportion PID0 sequences decreases), meaning that there are more potentially advantageous options at each step.

This does not stop the mean LN path lengths being below the minimums predicted using the equations of Orr (2003) and Rosenberg (2005) (see table 4.1). The reason is that neutral network structure of the space influences the number of *potentially adaptive* local neighbours of each sequence so that it is below the $3n$ used in Orr's calculation. In the case of the length-10 space, on average 13.34 of the 30 possible local mutants code for a neutral neighbour, and a further 13.16 code for PID0 (and are fatal), leaving only 3.5 potentially adaptive viable neighbours. Furthermore, if similar sequences code for similar structures many of those potentially adaptive mutants are likely to code for the same phenotype, reducing the number still further. If we consider the number of phenotypic neighbours for each sequence as the effective neighbourhood size (E), which is the number of observed unique viable phenotypic neighbours per sequence from table 2.7. Using this value of neighbourhood size instead of $3n$, we can recalculate the predicted path lengths using the equations from section 4.2.2 (Table. 4.4).

Sequence length	E	$\bar{L}_{Orr}$	$\bar{L}_{Simulation}$
10	1.87	0.65	0.69
12	3.31	0.82	0.97
14	4.97	1.06	1.15

Table 4.4: Estimates of mean path length ($\bar{L}_{Orr}$) calculated from the revised mean number of unique phenotypes in the neighbourhood of each sequence (E), compared with the mean path lengths of the LN method taken from table 4.2a

The estimate of the mean number of steps now slightly underestimates the simulation data. The most likely cause is a low estimate of the effective neighbourhood size. The fewer phenotypic neighbours a sequence has, the less likely it is to be encountered during an adaptive walk. The local neighbourhood walks are therefore more likely to move through relatively well connected areas of the genotype space where each sequence has an effective neighbourhood size above the average recorded across the whole genotype space.

We saw in figure 4.7 that many of the NS paths involve a greater number of steps than the Hamming distance between the two end sequences of a path. Much rarer are cases in LN walks where the number of *adaptive* steps in a path is greater than the Hamming distance between the end sequences (for example in the length-12 space

there were just 89 across 25,000 walks). Furthermore, these paths were never more than one step longer than the Hamming distance. The important point is that no matter how rare these paths are, their very existence indicates that a local mutation based step-wise path between two sequences longer than the Hamming distance is not limited to walks involving neutral networks+.

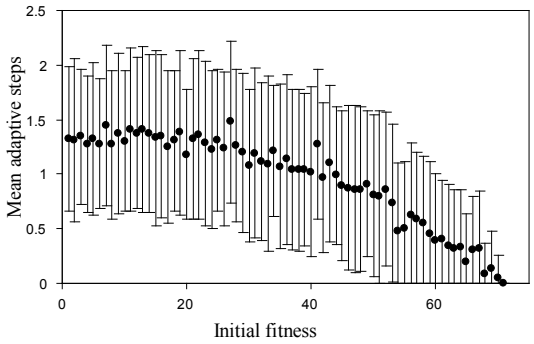
4.3.1.3 Network mapping

The network mapping (NM) walk calculates the number of adaptive steps that are required to reach an optimum, considering the whole network neighbourhood rather than just the local neighbourhood of a sequence. For this reason, a change in network potentially brings a much larger variety of different phenotypes into the neighbourhood than a single point–mutation change, increasing the chance of the NM walk finding the global optimum after just a few steps. In fact the networks are so well connected that the average number of adaptive steps in the network mapping walk starts just above one step for length-10, and only rises slightly over the sequence lengths investigated here. This is highlighted by comparing the distribution of path lengths shown in figure 4.6 with that of the LN walk. The distribution of LN path lengths is wider, and actually contains *more* paths above 1 step than the NM method, despite its much more restrictive neighbourhood assumptions. The reason is that many NM walks only take one step to reach the optimum.

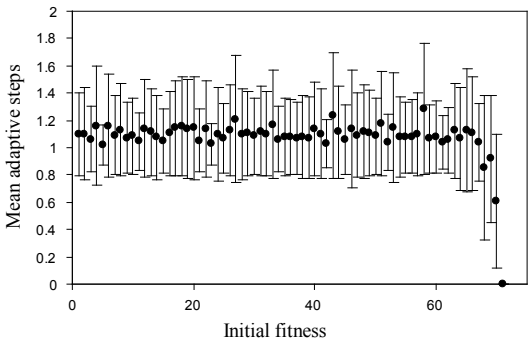
A given network mapping walk therefore requires fewer *adaptive* changes per path than the comparable neutral step walk. This indicates that the distance restriction placed on the neutral step method restricts the options available to it at any one step, and an NS walk often has to make more adaptive changes to reach the end of a walk instead (see figure 4.5b).

4.3.1.4 Initial fitness vs Path length

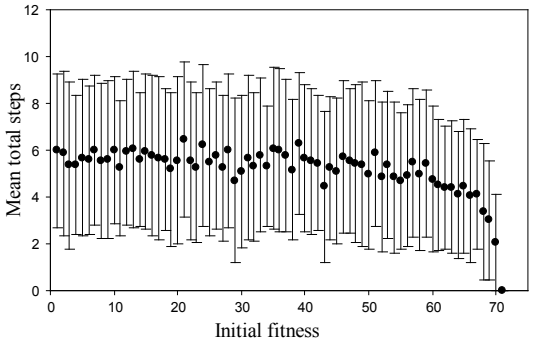
When paths are started at random points within the space, one might suppose that a lot of the variation in path length would be due to differences in rank of the initial fitness (Fig. 4.8). In the LN walk, we do see a decrease in the mean number of steps with increased initial fitness rank, but do not see a reduction in the variation in path length as the mean decreases. For the two network based paths, the variation in path length is surprisingly independent of initial rank. The mean and standard deviation only decrease when the initial fitness rank is quite close to the optimum. This indicates that the last steps of a walk are the most difficult, and make up the defining part of path length.



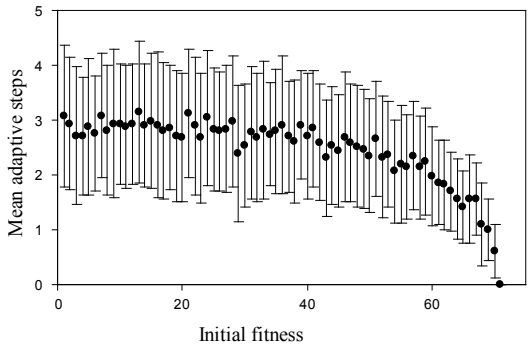
(a) Local neighbours

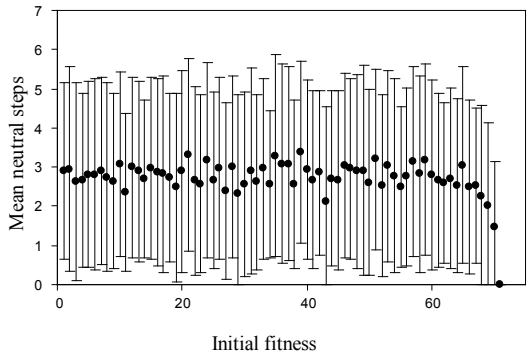


(b) Network mapping



(c) Neutral step walk (total)





(d) Neutral step walk

Figure 4.8: Average number of steps per path within the length-12 space against the initial fitness they started from. Error bars are 1 std.dev

It appears that the 'shape space covering' property of the genotype space means that most networks are relatively easily accessible until an adaptive walk gets close to the optimum. At this point the network structure imposes restrictions which mean that to reach the optimum, adaptive walks require many more steps than we might expect from geometric considerations alone. Given that evolution often occurs among populations which are relatively well adapted, this is an important finding when

so much weight is attached to 'fine–tuning' of the phenotype to become absolutely optimal.

4.3.2 End fitness

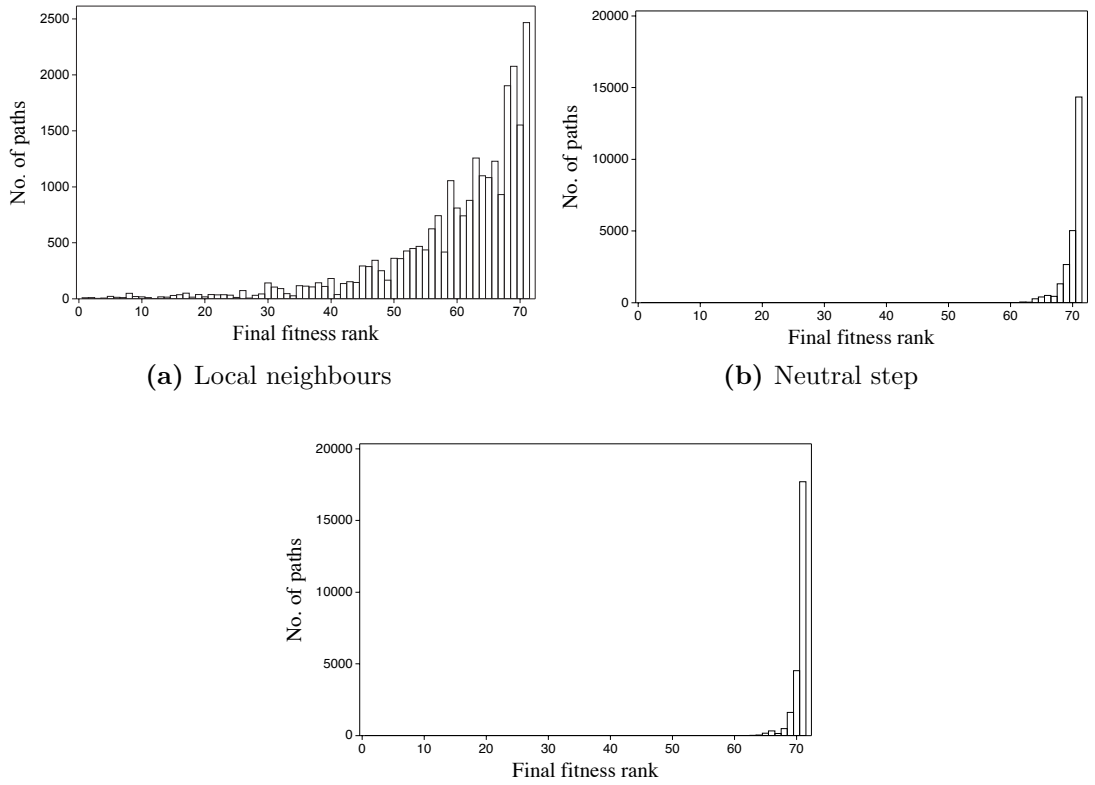
The final fitness rank of each path through the space is usually in the top 20% of phenotypes, irrespective of the walk method, suggesting that the space is quite well-connected even when only the local neighbourhood is available. However, using neutral networks significantly increases the chance of reaching the optimum phenotype (Table 4.5, Fig. 4.9, LN v NS, Wilcoxon signed ranks test length-10= 1591377.5, N for test= 73591, p<0.0005; length-12= 98970.0, N for test= 20714, p<0.0005; length-14= 1366.5, N for test= 4488, p<0.0005).

Maximum fitness

Length	Type	mean	med	min	$\%_{opt}$
10	local	14.74	16	1	18.95
	neutral	18.65	19	2	76.40
	network	18.80	19	2	84.60
12	local	59.96	63	1	9.89
	neutral	70.06	71	50	57.36
	network	70.49	71	59	70.82
14	local	241.50	249	22	5.04
	neutral	268.21	269	240	61.44
	network	268.81	269	261	85.22

Table 4.5: Summary of the final fitness ranks for lengths 10-14. Max fitness: length-10 = 19, length-12 = 71, length-14 = 269.

Although the difference between the average final fitness of the two network based methods is not a very obvious when we compare the two histograms in figure 4.9, the mean final fitness rank of the network mapping method is significantly higher than that of the neutral step method (Table 4.5, Fig. 4.9, NM v NS, Wilcoxon signed ranks test, length-10= 118688556.0, N for test= 17291, p< 0.0005; length-12= 3823721.0, N for test= 7070, p< 0.0005; length-14= 119483.5, N for test= 1540, p< 0.0005). The percentage of NS walks reaching the global optimum is also significantly lower than for the network mapping. Given that the NS walk has to traverse through more networks to get to the optimum (table 4.2a), and often has to take several neutral steps (table 4.2b), the NM method does not derive a huge advantage from having instant access to all the phenotypes in a network's neighbourhood. Occasionally,



(a) Local neighbours **(b)** Neutral step

(c) Network mapping

Figure 4.9: Distribution of final fitnesses at length-12 for the three different adaptive walk methods. The local neighbourhood method provides a much lower chance of reaching the global optimum than either of the other two methods.

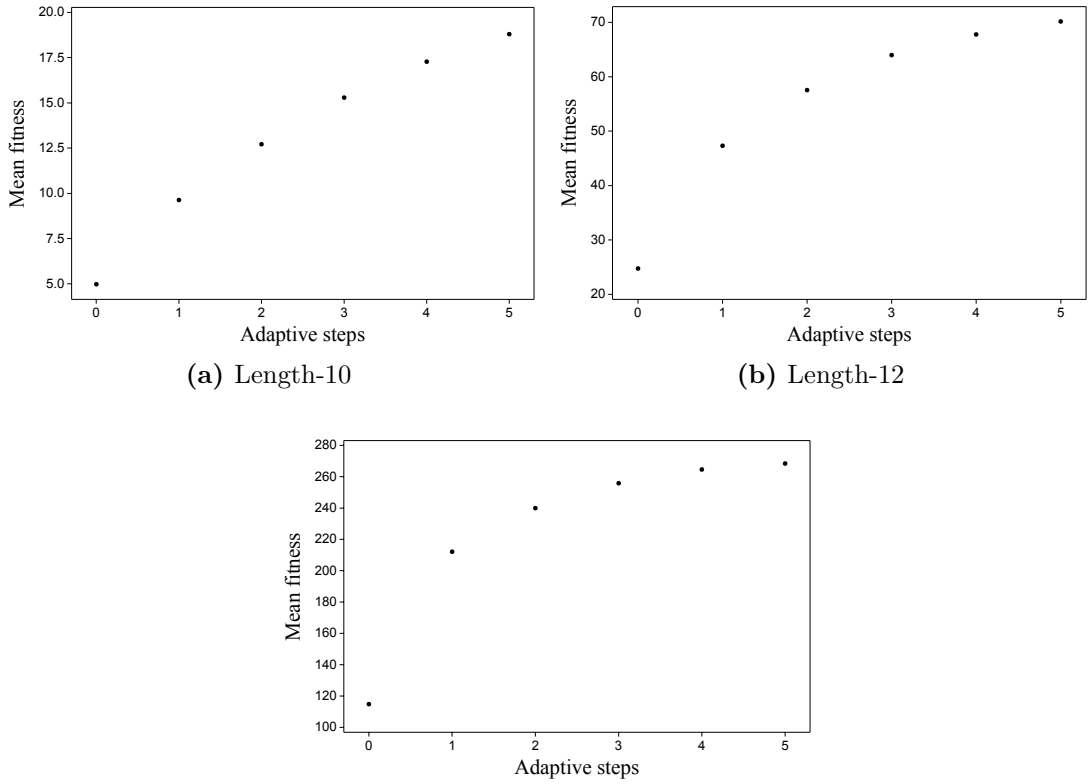
the neutral step method does fall just short of the global optimum, but it is rarely many ranks lower than the NM walk. The percentage of LN and NS walks reaching the optimum fall as sequence length increases, indicating that larger genotype spaces might be more rugged than the one examined here.

4.3.3 Path trajectories

We saw in the last section that the initial fitness rank had a surprisingly small effect on mean path length. This is because the networks are so well connected together that the accessibility of adaptive mutants often only become limiting when there are relatively few fitter phenotypic options left. In this section, I break down the neutral step adaptive walks to assess how the evolutionary trajectory changes over the course of a path. In this section I also attempt to reconcile the predictions made at the start of the introduction, with the data found in the NS simulations: that the average number of steps required to cross each network rises with fitness rank, and

that crossing the penultimate network in the length-10 space requires 1.88 neutral mutations on average (as found in chapter 3).

Holder and Bull (2001), Imhof and Schlotterer (2001), Orr (2002), Elena and Lenski (2003) and Barrett et al. (2006) have all shown that when a population is far from the genotypic or phenotypic optimum, large fitness changes are more likely. This phenomenon is also seen in the NS walks (Fig. 4.10).



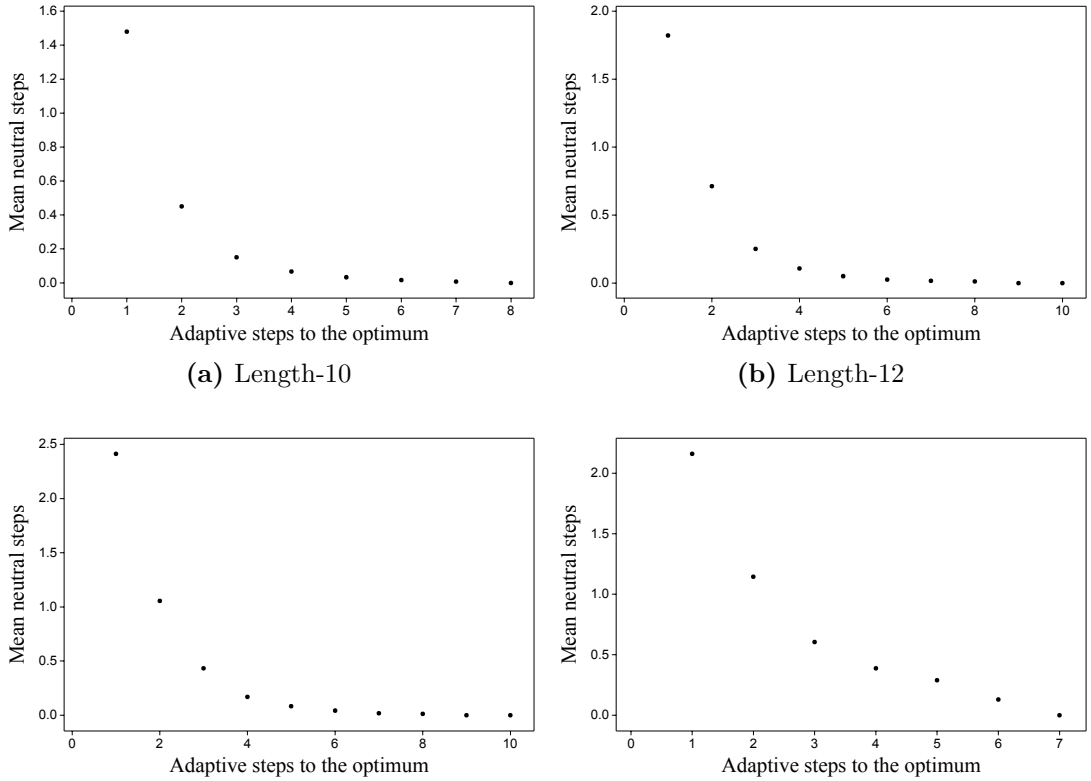
(a) Length-10 **(b)** Length-12

(c) Length-14

Figure 4.10: The mean fitnesses for each step in all the walks made up of 5 adaptive steps. The increase in fitness decreases over the walk. At shorter path lengths the rise in fitness is even sharper, as generally the final step of any path is very close to the optimum.

A similar but inverse principle applies to neutral steps, where the first few steps on an adaptive path involve the fewest neutral steps. Early in a walk the most accessible adaptive mutants require few or no neutral steps to reach. It is only the towards the end of an evolutionary trajectory that phenotypes requiring several neutral steps become the most accessible option (Fig. 4.11).

In the next three paragraphs, I shall attempt to reconcile the results from the NS simulations (Fig. 4.11) with the predicted 1.88 neutral steps per network.



(a) Length-10 **(b)** Length-12

(c) Length-14 **(d)** Length-10,single fitness gain

Figure 4.11: a), **b)** and **c)** Show the mean number of neutral steps per network (y axis) required for each adaptive change rises sharply as the path nears the optimum. Whatever the path length, traversing the penultimate network normally requires at least one neutral step. **d)** The average number of neutral steps to cross a network when the adaptive phenotype is just one rank higher than the current one.

The closer a path gets to the optimum, the fewer advantageous phenotypes there are, and the higher the chance of needing a neutral step to find an adaptive mutant. In the length-10 space, the mean number of neutral steps required to cross the penultimate network is 1.48. This is below the 1.88 steps predicted in the previous chapter by calculating the average number of steps between each pair of network neighbours across every network in the space.

This is not the whole picture however. An analysis of the final step includes those walks which reached the optimum with a large jump in fitness. This means that more than one phenotype could be available as an adaptive portal, potentially increasing the number of available exits, and so is not strictly comparable to the mean distance across a network between two fixed phenotypes.

We can account for this by restricting our walk analysis to those steps where the increase in fitness rank is just one. The data in figure 4.11d is restricted in this way.

In this case, the mean length for the final step is 2.16, larger than the 1.88 predicted in chapter 3. If the data are restricted still further so that only steps increasing the fitness rank from 18 to 19 are considered (perhaps the truest comparison with the results from chapter 3), the mean number of neutral steps required rises to 2.25.

How can we explain this discrepancy? A possible cause is that phenotypes are assigned a fitness rank randomly, and irrespective of their size. Because of the distribution of sequences within phenotypes, the highest ranked phenotype(s) will often only contain relatively few sequences. We saw in section 3.3.2, that inter-portal distance was negatively correlated with the number of portals to a particular phenotype. Thus, because small networks have low numbers of portals, we might expect the mean distance to reach a small network to be higher than the global average for the space. This could be tested with a further analysis including the sizes of the penultimate network and the final network. As well as testing the reason whether the number of neutral steps is higher in the simulations because of the final network size, an analysis of this sort could also test if having engaged on a previous walk influences the number of steps taken across the penultimate network.

In summary, we have seen the importance that neutral networks can have across the whole RNA genotype space. They increase the accessibility of adaptive mutants when compared to an uncorrelated landscape, especially at the end of an adaptive walk. The increase in accessibility allows the majority of neutral step walks to reach the global optimum phenotype from any random starting point in the genotype space, but the shortest path is often longer than the minimum topological distance from the starting point to the global optimum phenotype.

In the next three sections, I consider how resistant the model is to changing or relaxing some of the assumptions that have been imposed until now. While delving into the details of each individual assumption uncovers interesting properties of the space, this invariably ends up posing as many new questions as giving answers. Most importantly, these sections highlight the robustness of the main conclusions that neutral mutations are very important for accessing areas of higher fitness, and that those neutral mutations make up a significant part of a neutral step adaptive walk.

4.4 Fitness algorithm

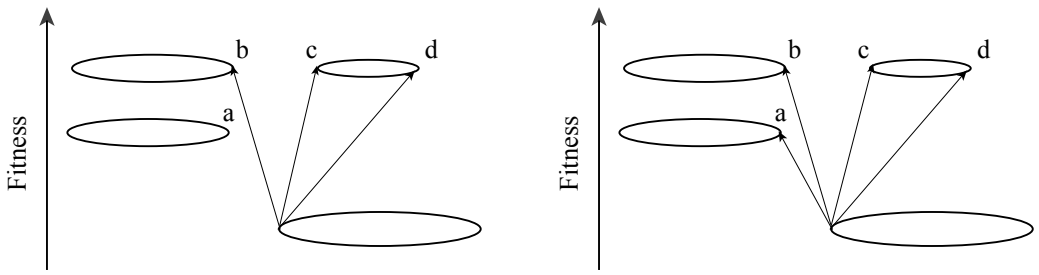
The greedy fitness algorithm used in the first half of the chapter presents an extremely powerful selective force. It has the property of minimising the average number of adaptive mutations required to reach an optimum, and therefore enables calculation

of a 'minimum path length'. Thus it facilitates comparison with Orr's (2003) minimum path length when considering how the many–to–one nature of neutral network structure in genotype space compares with a one–to–one genotype–fitness mapping. This kind of selection of the fittest mutant may occur in some large populations of fast evolving organisms i.e. those which can test their local neighbourhood in a single generation (Wahl and Krakauer, 2000). However, for many populations it is probably quite unrealistic. Here, I relax the fitness algorithm to an alternative (weaker) selection algorithm adapted from Gillespie (1984), which takes into account the probability of fixation, biased in favour of fitter phenotypes (Gillespie, 1984, Eq.20). The probability of mutating through a particular portal (q) is

$$P(q_{Y_i}) = (Y_j - Y_i) / \sum_{k=1}^{j-1} (Y_j - Y_k) n_k$$

$$i = 1, 2, \ldots, j-1.$$

where Y_j is the rank of the current phenotype and Y_i is the rank of the i^{th} (fitter) phenotype. n_k is the number of portals to phenotypes with fitness rank k in this model (Fig 4.12).



(a) Greedy algorithm **(b)** Gillespie algorithm

Figure 4.12: The two methods of calculating portal selection: **a)** Greedy algorithm: $P(a) = 0$, $P(b, c, d) = 1/3$. **b)** Fitter mutants given more weight. $P(a) = 1/7$, $P(b, c, d) = 2/7$

Simulations were performed as before, for the length-10 and length-12 spaces, by calculating path lengths for 1000 new randomly selected sequences across sets of randomly ranked phenotypes. The same sequences were repeatedly used for each different fitness ordering with each of the two fitness algorithms, but each algorithm used different sets of sequences. This means that a relevant statistical test including all the data is difficult to perform, due to nesting of sequences within each algorithm. Repeating the work using the same sequences for each algorithm would allow better

statistical comparison between the two algorithms, if a precise comparison of the effect of changing to this particular fitness algorithm or any other were required.

Using Gillespie's algorithm increases the path lengths of all three walk methods, especially for the network mapping walk (Table 4.6). The NM walk is most affected because it is the most likely to have simultaneous access to a number of different fitter phenotypes. The greater the number of adaptive options at any one step, the slimmer the chance of choosing the fittest option, and the more subsequent steps required.

(a) Advantageous mutations

		Greedy algorithm			Gillespie algorithm		
Length	Type	mean	med	max	mean	med	max
10	LS	0.69	1	6	0.75	1	6
	NS	1.83	2	7	2.10	2	8
	NM	1.04	1	4	1.78	2	7
12	LS	0.97	1	6	1.17	1	6
	NS	2.47	2	9	3.05	3	11
	NM	1.07	1	4	2.48	2	8

(b) Total steps

		Greedy algorithm			Gillespie algorithm		
Length	Type	mean	med	max	mean	med	max
10	NS	3.72	3	17	4.17	4	18
12	NS	5.19	5	32	6.30	6	24

Table 4.6: Comparison of the path lengths of the two different fitness algorithms tested. The number of different sequences tested was 1000. For Gillespie's algorithm, the number of different fitness orderings was 25 for length-10 and 12.

In contrast to path length, there is no overall pattern to the changes in final fitness rank. The neutral step method actually has a slightly higher mean fitness than under the greedy fitness algorithm, and is also higher than the network mapping method under Gillespie's algorithm.

In the length-10 space, the final fitnesses are remarkably unaffected by the change in algorithm. This highlights the infrequency of coming across different adaptive phenotypes at a particular step, when the local neighbourhood is so limited. As sequence length increases, the number of adaptive phenotypes at a certain distance is likely to increase, especially at the start of a walk, when the population is further from the optimum.

End fitness

		Greedy algorithm				Gillespie algorithm			
Length	Type	mean	med	min	% at opt.	mean	med	min	% at opt.
10	LN	14.74	16	1	18.95	14.79	16	1	19.52
	NS	18.65	19	2	76.40	18.65	19	3	75.7
	NM	18.80	19	2	84.60	18.63	19	3	84.37
12	LN	59.96	63	1	9.89	59.69	63	1	11.06
	NS	70.06	71	50	57.36	70.51	71	48	76.05
	NM	70.49	71	59	80.82	70.23	71	48	69.40

Table 4.7: Comparison of the final fitness ranks for lengths 10 and 12. Max fitness: length-10 = 19, length-12 = 71.

Using Gillespie's algorithm has a larger affect on final fitness in the length-12 space, where there is more often range of adaptive steps in any one neighbourhood. In the NM walks, the chance of getting to the global optimum is actually 10% less than when the greedy algorithm is used. It seems that when many high ranked phenotypes are available, including those just below the optimum, the chance is higher of getting caught in a local optimum. This indicates that many of the highest fitness networks are not connected to each other.

The neutral step method does not suffer this decrease in sequences reaching the global optimum. This can be explained by considering the differences in fitness rank: when there is a large difference in fitness rank between the current phenotype and the global optimum, the bias means there is a high probability of jumping to a much higher fitness rank. When those sub-optimal networks are all available to a NM walk, they have almost the same probability of being selected as the optimum. Once at a sub-optimal network, the chance of finding another step to the optimum is reduced.

In contrast, because fewer options are available to the NS walk at any one step, the final steps towards the optimum tend to be from a higher current fitness rank. This means that the difference in probability between selecting the global optimum step is far more likely than selecting the local optimum (Fig. 4.13). The effect could be tested by using relative fitness ranks, rather than absolute fitness ranks.

4.5 Non-random initialisation

Gillespie's original premise was that if the landscape shifted, the population would return to the optimum in a burst of mutations. The purpose of this section is to

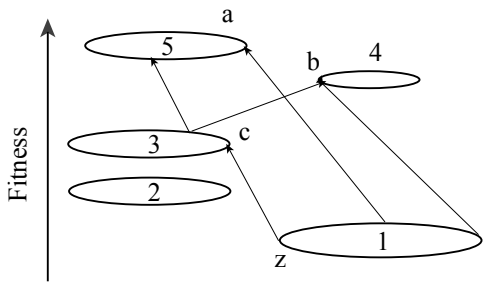
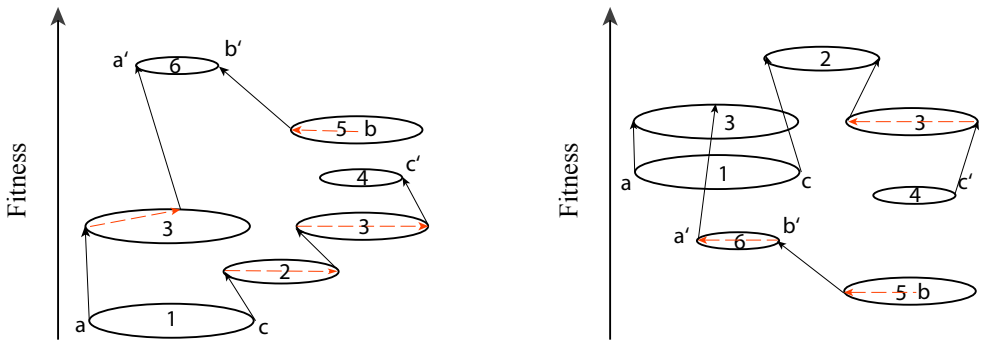


Figure 4.13: The difference between a NM and NS method in reaching the global optimum when both sequence start from z. Fitness ranks are given as numbers, possible portals are labelled a, b and c. First consider the NM method: all network neighbours are equally accessible from z, so $P(a) = 4/9$, $P(b) = 3/9$, $P(c) = 2/9$. In contrast, in the NS method, the neutral path distance makes a difference. In the first step: $P(a) = 0$, $P(b) = 0$, $P(c) = 1$, because the portal to c is closest. In the second step, $P(a) = 2/3$, $P(b) = 1/3$. So, chance of reaching the global optimum is higher under the NS method ($2/3 > 4/9$). The smaller difference in ranks after the first step means that the probability of taking reaching the optimal network is actually higher with the NS method, than with the NM method.

investigate whether simply having *previously evolved* confers an advantage over starting from a randomly chosen point in the space. We can model this by reshuffling the order of the phenotypic fitness ranks, and then comparing each of the three walk methods from the random starting sequence and 'pre-evolved' end sequence of the previous neutral step (NS) walk over the new phenotypic ranking (Fig. 4.14).



(a) **(b)**

Figure 4.14: a) Under the first random phenotype ranking, the initial genotypes are tracked to the end of their respective neutral step walks. **b)** In each subsequence phenotype ranking the start and end genotypes from each neutral step walk are used as the start sequences of new adaptive walks. In this step we can see that a′ and b′ converge on the same genotype, and in the next generation will start from the same position.

One might predict that pre-evolved sequences have an advantage if the space has

some regions where connectivity between networks or sequences is higher than the average. Over time, pre-evolved sequences would converge on these areas. The basis of this expectation lies in a (possibly weak) selective pressure towards areas of the network which have many phenotypic neighbours. If a path happens to come across a well connected area of the space it is likely to remain within that area over future alternative rankings, because of the increased probability that any given optimum lies nearby. Pre-evolved sequences which start in a region of higher connectivity are likely to require fewer steps and reach higher fitnesses that randomly chosen sequences, which do not.

Table 4.8 indicates that with the raw data, the adaptive path lengths of pre-evolved sequences show an increase in the number of steps for the LN and NM walks, and a decrease for the NS walk. This is what we might have predicted if the area occupied by pre-evolved sequences were better connected. Longer paths for the LN walk indicate more potential for adaptive change, while shorter paths for the two network based methods indicate that portals lie closer together, and less neutral steps are required to reach the optimum.

(a) Adaptive steps

		Random start			Pre-evolved start		
Length	Type	mean	med	max	mean	med	max
10	LN	0.69	1	6	0.91	1	5
	NS	1.82	2	7	1.73	2	7
	NM	1.04	1	4	1.11	1	4
12	LN	0.97	1	6	1.07	1	6
	NS	2.47	2	9	2.31	2	9
	NM	1.07	1	4	1.14	1	4

(b) Total steps

		Random start			Pre-evolved start		
Length	Type	mean	med	max	mean	med	max
10	NS	3.72	3	17	3.14	3	18
12	NS	5.19	5	32	4.64	4	22

Table 4.8: a) Comparison between random initial starts and 'pre-evolved' sequences. Sample size = 1000 walks, Fitness orderings = 100 for length-10, 25 for length-12. **b)** Summary of the total number of steps in the NS adaptive walk.

However, it turns out that to test this hypothesis is far more problematic than

might be thought, as fair comparisons are difficult to draw. Before presenting the results for final fitnesses, I shall consider the two most important factors which confound our ability to test the prediction about the connectivity of the space.

The first confounding factor is that any sequence which starts from the final step of a previous run must have at least one phenotypic neighbour. Once the phenotypes are re-ranked, each starting genotype has at least a 1/2 chance of immediately undergoing an adaptive step in both the local neighbourhood and neutral step walks. A randomly chosen starting sequence does not have this guarantee, especially at short sequence lengths, where sequences can often have nothing but neutral and PID0 neighbours. We can counter this factor by only including those paths whose first step is an adaptive one. The chance of both random and pre-evolved sequences taking a second step should then be equal, with the only differences due to connectivity. The mean initial fitness of walks starting with an adaptive step is likely to be lower than the average, because the chance of taking at least one step rises the further one starts from the optimum (see figure 4.8).

When we take into account the increased likelihood of a pre-evolved genotype taking an immediate adaptive step, the mean path length for the LN method is actually shorter for the pre-evolved sequences (Table 4.9). So, the difference in path length for the LN walk appears to be due to the increased chance of taking an initially adaptive step from a pre-evolved starting point, rather than because those pre-evolved starting points are in a more well connected part of the genotype space. However, the decrease in the the number of steps in the NS walk remains, and the network mapping walk produces slightly longer paths when the initial genotypes are pre-evolved. Both these effects are explained by the second confounding factor.

The second confounding factor is that pre-evolved sequences have a tendency to start from smaller networks than the randomly chosen ones. Because sequences are initially randomly chosen from the whole genotype space, larger networks are likely to have more sequences in the sample of 1000 than smaller ones. But because phenotypes are then ranked randomly, and because there are more rare phenotypes than common ones (Schuster et al., 1994; Göbel, 2000), there is a good chance of a phenotype with not many sequences being the global optimum. The pre-evolved genotypes are generated from the end of the previous NS walk, and so are therefore as likely to be in a small network as a large one. They are in fact only *almost* as likely, because network connectivity is correlated with network size, and therefore there exists a loose correlation between the number of pre-evolved sequences and network size (Fig. 4.15).

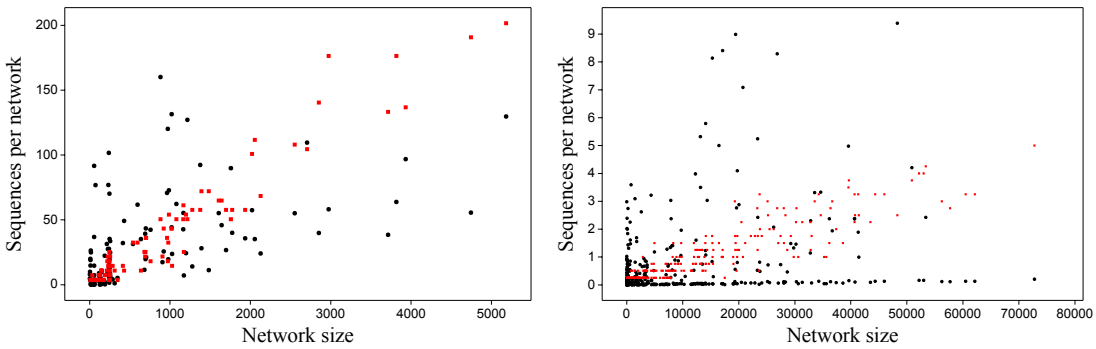
Adaptive steps

		Random start			Pre-evolved start		
Length	Type	mean	med	max	mean	med	max
10	LN	1.24	1	6	1.21	1	5
	NS	2.19	2	7	2.00	2	6
	NM	1.13	1	4	1.22	1	3
12	LN	1.33	1	6	1.31	1	6
	NS	2.74	3	9	2.53	2	9
	NM	1.10	1	4	1.12	1	4

Total steps

		Random start			Pre-evolved start		
Length	Type	mean	med	max	mean	med	max
10	NS	3.88	4	17	3.41	3	16
12	NS	5.35	5	32	4.73	4	22

Table 4.9: Comparison of walks whose first step was an adaptive one, and where each starting sequence is unique within that fitness ordering.

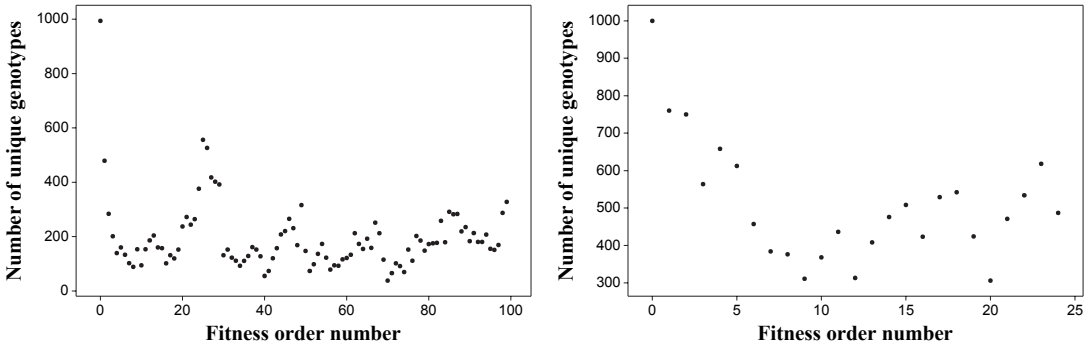


(a) length-10 **(b)** length-12

Figure 4.15: Initial sequences are chosen randomly from within the space. This means that there is strong correlation between the frequency with which a particular network appears and its size Pearson correlation co-efficients: length-10 = 0.965, length-12 = 0.892 (■). Because phenotypes are ranked randomly, any network has an equal probability of being the global optimum or close to it. This means that the correlation between network size and frequency of sequences is much less strong for pre-evolved genotypes Pearson correlation co-efficients: length-10 = 0.518, length-12 = 0.276 (●).

The result is that as the simulation proceeds over different fitness orderings, the distribution of sequences shifts from sequences mainly starting in large networks, to many of them starting in small networks. This in itself is not a problem. However, it has several implications for calculating the subsequent adaptive walks. One is that smaller networks generally have fewer connections and shorter neutral inter-portal distances, so we might expect smaller networks to have an negative effect on both the path length and final fitnesses of each adaptive walk. Another is that within small networks initially random sequences can converge on the same few genotypes, potentially reducing the independent sample size (Fig. 4.16, and see Fig. 4.14 for an example of convergence). This in turn is likely to lead to higher variation between the means, as many paths are repeated by sequences which have converged.

We can counter this factor by comparing walks starting in similar sized networks, and only considering unique starting sequences. However, care must be taken not to hide the magnitude of any differences due to the connectivity of the space – If all the sequences that end up converging on highly connected areas also converge on a single sequence within it and are thus not considered, the effect will be underestimated.

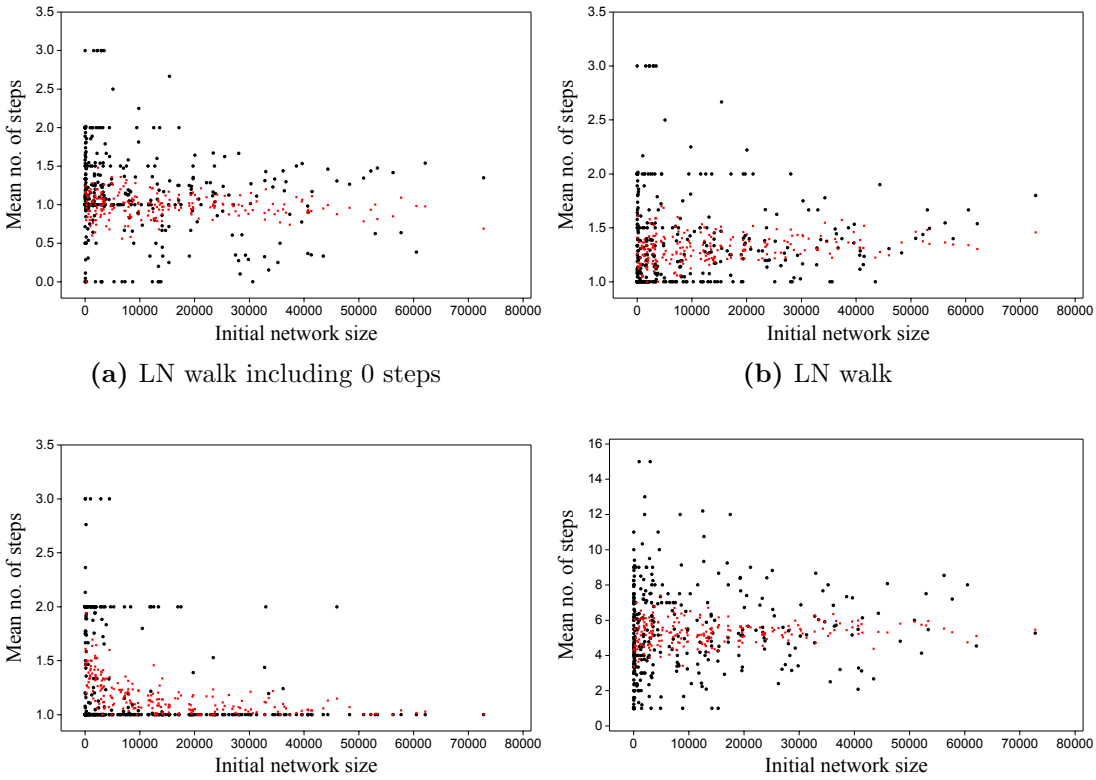


(a) length-10 **(b)** length-12

Figure 4.16: Of the initial 1000 random sequences, 3 were duplicated in the length-10 space, and none in the length-12 space. However, the count of unique sequences drops very quickly as walks converge over subsequent phenotype orderings. After the initial reduction, there is substantial variation from one ordering to the next, depending on the size of the fittest phenotype(s). At each different fitness ordering, there is the potential for divergence as well as convergence depending on the available paths through the space.

The network mapping walk produces slightly longer paths when the initial genotypes are pre-evolved. This is due to smaller networks lacking the connectivity to make an initial network–based step to the optimum (Fig. 4.17).

In the NS walks, table 4.8 and table 4.9 indicate a reduced mean path length for the pre-evolved genotypes. However, the effect is difficult to see in figure 4.17d. In



(a) LN walk including 0 steps

(b) LN walk

(c) NM walk

(d) NS walk

Figure 4.17: The number of steps per walk against initial network size at length-12. In **(b)**, **(c)** and **(d)**, the data is filtered so that only walks where the first step is adaptive are included, and no walk can start from the same sequence within a particular phenotype ordering. ●=pre-evolved genotypes, ■=random genotypes. **a)** LN walk. The pre-evolved genotypes have a higher mean across all initial network sizes when all the paths are considered (Table 4.8). **b)** This difference disappears if we consider only the sequences which take at least one adaptive step (Table 4.9). **c)** NM walk. For random and pre-evolved sequences, starting from small networks require more steps to get to the optimum because those networks are less well connected. **d)** NS walk. See main text for details. Across all the walk methods, pre-evolved starting sequences tend to have increased variation in path length across all initial network sizes.

fact the distribution of sequences in the small network sizes is skewed, with many walks starting in smaller networks having shorter path lengths and a few starting in small networks having longer path lengths. The small networks with many sequences are also the most connected ones, and therefore are the ones most commonly seen as pre-evolved starting sequences. It is these well connected small networks which lead to the overall reduction in NS path length. In other words, pre-evolved sequences have a higher chance of ending up in small well connected networks, from which it is easier (requires less steps) to find a new optimum. Though the size is influenced by the way in which networks were assigned, the fact that more paths end up in well connected networks means that there is some increase in evolvability to be derived

from having previously evolved within the space.

Table 4.10 shows that despite all the potentially confounding factors, there is little difference between the final fitness of random or pre-evolved starting sequences for the two network based methods. This is despite a small difference in the mean initial starting fitnesses especially at length-12.

Raw data

		Random start Mean fitness			Pre-evolved start mean fitness		
Length	Type	start	end	% at opt.	start	end	% at opt.
10	LN	10.06	14.74	18.95	9.94	16.7	29.62
	NS	10.06	18.65	76.40	9.94	18.71	78.47
	NM	10.06	18.80	84.60	9.94	18.80	83.43
12	LN	35.50	59.96	9.89	38.91	63.83	11.71
	NS	35.50	70.06	57.36	38.91	70.01	57.5
	NM	35.50	70.49	70.82	38.91	70.48	67.4

Unique initial sequences taking at least one step

		Random start Mean fitness			Pre-evolved start mean fitness		
Length	Type	start	end	% at opt.	start	end	% at opt.
10	LN	10.06	16.08	24.22	10.25	16.78	31.10
	NS	10.06	18.66	76.29	10.25	18.71	78.64
	NM	10.06	18.83	85.53	10.25	18.83	85.57
12	LN	28.59	62.21	11.08	34.61	64.22	16.10
	NS	28.59	70.07	57.70	34.61	70.09	55.19
	NM	28.59	70.50	71.08	34.61	70.47	65.20

Table 4.10: Summary of the final fitness ranks for lengths 10 and 12. Max fitness: length-10 = 19, length-12 = 71.

The local neighbourhood walk becomes slightly more likely to reach the optimum when initialised from a pre-evolved sequence, both in the raw data set, and in the modified one. In the modified data set, this difference comes despite taking no more steps than the random starting sequences. This results indicates that although the NS walks are not likely to take more adaptive steps across a walk, they are more likely to end at the optimum.

In summary there is some evidence to suggest that neutral paths tend to terminate in areas of the space which are more locally connected than is generally the case.

The resulting pre-evolved sequences tend to be more evolvable, even when that pre-evolution does not confer any initial fitness advantage. As a result, we see NS walks requiring slightly fewer steps per walk, and LN walks advancing to higher fitness than random sequences. Perhaps most importantly, by considering the confounding factors, the differences between path lengths seen in the raw data were reduced, and the most important conclusion drawn from this section is that the effect of being pre-evolved is *not* very large – There are no regions of the space which act like fire-escapes through a building, linking floors without the need to cross the rooms in between. This section has also highlighted that even within a small genotype–phenotype map, the nature of the space is so complex that it is sometimes difficult to examine and extract useful results and conclusions.

4.6 Phenotype-fitness correlation

Until this point I have made the assumption that fitness has no correlation with phenotype. This is unlikely to hold true in many real life situations, and so in this section I examine the effect of a simple correlation between phenotype and fitness.

Unlike the genotype–phenotype function for RNA secondary structure, there is no obvious natural phenotype–fitness function. Here I correlate a simple count of bases pairs with fitness. We can consider this as a simple proxy for structural stability, often an important molecular characteristic. The fitnesses are ranked according to

$$Y_i = p_i + r$$

where p_i is the number of base pairs in the phenotype, and r is a random number drawn from a uniform distribution between 0 and 1. This means that the random number simply decides ties between base-pairs and that more base pairs are always fitter.

If similar sequences code for similar phenotypes, any correlation between phenotypes will mean that the genotype space is also correlated. Perelson and Macken (1995) and Orr (2006b) both studied correlated landscapes within an adaptive landscape and found that path length increased with the degree of correlation. The results of my correlated landscape model are shown in tables 4.11 and 4.12.

The path lengths for the network mapping method are very similar between correlated and random landscapes, as shown in table 4.11, and are not discussed further here.

(a) Adaptive steps

		Random ranking			Correlated by number of BPs		
Length	Type	mean	med	max	mean	med	max
10	LN	0.69	1	6	0.78	1	5
	NS	1.83	2	7	1.60	1	6
	NM	1.04	1	4	1.02	1	4
12	LN	0.97	1	6	1.15	1	5
	NS	2.47	2	9	2.22	2	8
	NM	1.07	1	4	1.11	1	5

(b) Total steps

		Random ranking			Correlated by number of BPs		
Length	Type	mean	med	max	mean	med	max
10	NS	3.72	3	17	2.68	2	13
12	NS	5.19	5	32	4.28	4	20

Table 4.11: Comparison of the path lengths in a random and correlated landscape. The number of different sequences tested was 1000, the number of different fitness orderings was 25 for length-10 and 12 in the correlated landscape. For comparison the original data from the random ranking simulation are repeated from table 4.2.

The path lengths for the LN walk are longer in the correlated landscape. This fits with the theoretical work of Perelson and Macken (1995) and Orr (2006b). The increase in length indicates that a walk in a correlated landscape is more likely to find an advantageous neighbour within its local neighbourhood at each step.

By contrast the number of adaptive steps in the NS model is reduced, and the total number of steps even more so. With a higher chance of adaptive steps in the local neighbourhood of a given step on a walk, less neutral searching is required on each walk. However, this does not explain the drop in the mean number of adaptive steps. I suggest that the reason for the drop in the number of adaptive steps has less to do with the correlation of the landscape, and instead the positions of the particular phenotypes which are correlated. In the correlated landscape, the phenotypes with the most sequences and highest connectivity are those with the intermediate number of base pairs. If an adaptive walk is not already at high fitness, it is likely to be able to access an intermediate network which has a wide degree of connectivity, and facilitate access to the global optimum, or close to it.

The correlation within the space, including the well connected intermediates,

means that the average final fitness and the percentage of walks reaching the optimum increases for all the walk methods (Table 4.12).

End fitness

		Random ranking			Correlated by number of BPs		
Length	Type	start	end	% at opt.	start	end	% at opt.
10	LN	10.06	14.74	18.95	11.70	16.08	35.13
	NS	10.06	18.65	76.40	11.70	18.80	86.24
	NM	10.06	18.80	84.60	11.70	18.88	90.51
12	LN	35.50	59.96	9.89	38.47	64.59	18.40
	NS	35.50	70.06	57.36	38.47	70.46	65.88
	NM	35.50	70.49	70.82	38.47	70.79	83.42

Table 4.12: Comparison of the final fitness ranks for lengths 10 and 12 between the random and correlated ranking of phenotypes. Max fitness: length-10 = 19, length-12 = 71.

4.6.1 Phenotype correlation and continuing from an optimum

In the correlated landscape, when starting genotypes are seeded from the end of the previous neutral step walk ('pre-evolved'), there is a marked reduction in the mean path length for all walking methods (Table 4.13). This is because the probability of a pre-evolved genotype initially coding for a phenotype of high fitness rank is high. In fact the mean path length for the network mapping method is well under 1 step. In the neutral step (NS) walk, the ratio of neutral steps to adaptive steps is reduced still further compared to random starting sequences. Importantly, even when an NS adaptive walk starts close to the optimum of a correlated landscape, it still requires an average of 0.43 neutral mutations for every adaptive one it takes.

If we consider the percentage of all the walks reaching the optimum from pre-evolved sequences, we can see that many of them get caught in local optima. This is another indication that when using a greedy algorithm, starting from lower fitness results in a better chance of reaching the global optimum. In other words the highest ranking phenotypes, those with the most base pairs, are often not connected to each other directly. We can make sense of this in terms of RNA folding, if one considers that it is very difficult to jump between different phenotypic shapes, but retain the same number of base pairs.

(a) Advantageous steps

		Random			Pre-evolved		
Length	Type	mean	med	max	mean	med	max
12	LN	1.15	1	5	0.63	1	4
	NS	2.22	2	8	0.82	1	6
	NM	1.11	1	5	0.71	1	3

(b) Total steps

		Random			Pre-evolved		
Length	Type	mean	med	max	mean	med	max
12	NS	4.28	4	20	1.18	1	15

Table 4.13: Comparison of random and pre-evolved genotypes in a correlated landscape for length-12. The number of different sequences tested was 1000, the number of different fitness orderings was 25 for each of the ordered phenotypes.

End fitness

		Random			Pre-evolved		
Length	Type	start	end	% at opt.	start end	% at opt.	
12	LN	38.47	64.59	18.40	67.64	70.02	55.91
	NS	38.47	70.46	65.88	67.64	70.52	67.22
	NM	38.47	70.79	83.42	67.64	70.55	69.13

Table 4.14: Comparison of the final fitness ranks for length-12 between random and pre-evolved genotypes in the correlated landscape. Max fitness: length-12 = 71.

4.7 Summary of results

In the first half of this chapter I considered in some detail the trajectories of three different adaptive walk methods across the RNA genotype space. In the second half, I altered some of the assumptions of the initial model to ascertain under what conditions neutral steps remain an important part of adaptive walks, as well as further probing the ways in which genotype space structure affects mutational accessibility. The main findings can be summarised as follows:

Accessibility of the space

- Neutral network structure enables a population to reach adaptive phenotypes which would otherwise not have been accessible.

- When phenotypes are not correlated to fitness on average *every adaptive step requires a neutral one.*
- Even when the landscape is correlated, and when initial fitnesses are very high every adaptive step requires 0.43 neutral ones.
- The RNA genotype space is very well connected at short sequence lengths. The majority of sequences starting from a random point in the genotype space can reach the global optimum.
- No areas of the space are so well connected that neutral networks are not required.
- Some areas of the space are slightly better connected than others, and sequences starting in those areas require fewer steps/and or have a better chance of reaching the global optimum
- Local optima do exist, but usually with relatively high fitness rank

Walk trajectories

- Path length is not strongly correlated with initial fitness.
- The number of steps in a path is often longer than the Hamming distance between the two end sequences of the path.
- The largest fitness gains in an adaptive walk tend to be early in the trajectory. Conversely the longest neutral paths across a single network tend to be late in the trajectory.

4.8 Discussion

We have seen in this chapter that neutral steps allow access to a greater range of phenotypes than are available in a local neighbourhood, and in doing so increase the final fitness rank achieved by an adaptive walk. In fact within the RNA genotype space with randomly assigned phenotype fitnesses, on average *each adaptive step requires a neutral step.* Even when phenotypes are correlated, neutral mutations are still required within an adaptive walk. This result highlights the importance of considering neutral mutations in a genotype-phenotype mapping where they are known to exist. The mean length of walks involving neutral steps in this kind of

landscape is significantly longer than those found in simple one–to–one genotype–fitness maps, and any walk has less chance of getting caught in a local optimum.

In fact, the results indicate that the majority of paths have the potential to reach the global optimum, and if they do not, can usually climb to a very high fitness rank. However there is a downward trend within the sequence lengths tested here. Reaching the global optimum may therefore become less and less likely at longer sequence lengths. The impact of this depends on the level at which epistatic interactions work in the genome.

If there are further epistatic interactions between all parts of a longer sequence, the landscape will become increasingly rugged, and the chance of a particular random genotype being able to evolve to the global optimum is reduced. However, various authors have suggested that the genome is made up of more discrete modules (e.g. Wagner and Altenberg, 1996; Ancel and Fontana, 2000; Mezey et al., 2000; Carroll, 2001; West-Eberhard, 2003). If the genome is compartmentalised in this way, so that only a few genes (or base positions) interact epistatically, then this model will hold for each block, unless the blocks themselves interact epistatically.

Even when modelling a relatively small genotype space, as in this chapter, there is huge potential for extremely complex interactions and confounding factors to affect our ability to untangle the most important effects from each other, as typified by section 4.5. Careful effort is required to negotiate these successfully, especially in even more complex spaces.

At longer sequence lengths with larger local neighbourhoods, the potential for locally accessible adaptive mutants grows. This means that the trajectory of a path is is likely to become initially steeper. With a greater variety of adaptive neutral neighbours, the difference between the greedy fitness algorithm and Gillespie's algorithm is also likely to grow. These two methods of selecting adaptive steps are just two of many possibilities. Others examples include changing the ranks in Gillespie's algorithm from absolute ranks to comparative ranks or assigning fitness values rather than ranks. The affect on final fitness and path length could vary hugely depending on the algorithm chosen, and deserves further consideration.

In section 4.5 we saw that when walks are initiated from the end sequences of a previous walk, the number of unique starting sequences varied between iterations of the fitness ordering. The reduction is explained by convergence on small numbers of very fit sequences. However, after every genetic convergence the number of unique sequences usually bounced back. This indicates that many adaptive walks have the potential to take one of several paths, and can end up diverging significantly. A

further study on the repeatability of evolutionary trajectories within the genotype map could yield interesting insights into genetic divergence as well as convergence, perhaps providing a method by which speciation could easily occur (Gavrilets, 1997, 2003).

The final assumption I tested considered the effect of correlating fitnesses between phenotypes. This has the potential to impose an extra level of structure onto the genotype space. Even with a high degree of correlation between initial and final fitnesses, and therefore very short mean path lengths, a reasonable proportion of all steps on a path were neutral. As with the selection algorithms, there are an infinite number of ways of applying different phenotype correlations, because there is no natural biological model to follow. Finding a natural further correlation perhaps based around selection for particular unbound bases at a position rather than purely phenotypic structure could yield interesting insights into a more realistic genotype–phenotype mapping (Hall and Williams, 2004, showed that binding affinity of IREs depends on the particular bases at certain unbound positions).

Two of the most important assumptions of this model are the most difficult to test, because they are assumptions about the nature of the space or the dynamics of a population rather than the parameters of the model. The first is the lack of environmental change, the second is the strength of selection. The impact of changing these assumptions is discussed now.

4.8.1 The lack of environmental change

In this model, a change in the fitness ordering was applied only after each adaptive walk had been traced to its conclusion. The effect that neutral networks have on an evolutionary trajectory can itself be greatly affected by the stability of the environment. Let us first consider the case where the environment remains static for a longer period of time. A population will reach the optimal sequence, and then spread out across the optimal network. If the environment shifts the optimum at this point, the standing variation in the population means that many of the portals in the network's neighbourhood will already be accessible to members of the population. As we saw in section 4.3.1.3, when the whole network neighbourhood is available, the chances are good that one of the network neighbours is very highly adaptive, if not the global optimum. Under these circumstances the variation between individuals within the neutral network could lead to those already close to the most adaptive portals taking a short and direct path to the optimum as seen under the network mapping walks.

With an intermediate level of environmental restructuring, we saw that a population is likely to end up inhabiting the more *evolvable* areas of the space. In these areas there is a higher concentration of phenotypic neighbours and therefore any adaptive neighbours are likely to be more accessible, conferring an evolutionary advantage.

If the environment was constantly in flux, and a population never reached anywhere near the optimal sequence, then neutral steps are likely to play a less important role in evolutionary trajectories. The pattern across an evolutionary trajectory is for very few neutral steps to be required at the beginning of an adaptive walk. Therefore where a population is constantly battling to follow environmental shifts, the chances are higher than local adaptive mutants will be accessible.

The size of an environmental shift also has the potential to affect a population's subsequent final fitness. A large change can eventually benefit a population stuck in a local optimum, because the less fit it becomes after the shift, the more pathways are available to find its way to the global optimum, and therefore, the lower the chance of getting trapped sub-optimally again.

4.8.2 Relaxing selection

Throughout this thesis selection has been considered a very black and white issue. Neutral mutations are strictly neutral, and any difference in fitness large enough to allow selection to act instantaneously. If selection is weak, and a phenotypic transfer requires a slow shift of gene frequencies over many generations, a population may continue to drift past the adaptive portal, and find a more advantageous option further away from the original entry point. The effect may be exacerbated at long sequences lengths, where the chance of getting any particular mutation decreases, and so the chance of drifting past an adaptive one does too. This effect may change the trajectory of an adaptive walk, with more neutral steps occurring earlier in a trajectory, but resulting in larger fitness gains.

The alternative is that if a mutant phenotype is only weakly deleterious, the mutants may not be purged quickly from the population. Further mutations might then be capable of crossing the valley and potentially reduce the path length of an adaptive walk, especially considering the results in figure 4.7 show that many of the walks through networks take more steps than the direct path between the path's end points. However, van Nimwegen and Crutchfield (2000) showed that crossing even a weakly deleterious valley was orders of magnitude less likely than drifting across a single neutral network to find a fitter phenotype, and that valley width was actually more important than depth.

Van Nimwegen and Crutchfield's result indicates that adaptive walks are more likely to follow the adaptive–neutral step trajectories described here than cross a fitness valley, even if the intermediates suffer a relatively small fitness disadvantage. This is especially true when we consider that each path usually involves a number of adaptive steps as well as neutral ones, and the periods of neutral drift are often punctuated by adaptive changes which can reduce the time available for a given lineage to cross a valley. We saw in figure 4.7 that many of the adaptive walks involving neutral mutations took paths longer than the Hamming distance between the end sequences.

The result leads to an important inference about the nature of phylogenetic differences, and the molecular clock (Zuckerandl and Pauling, 1962). When the end products of two divergent sequences are assessed using a pair-wise comparison, even what appears to be a highly conserved gene has the potential to have undergone larger numbers of essential genotypic changes, which have subsequently converged on the original sequence. This could lead to an unexpected difference in the ratio of changes between synonymous and non-synonymous mutations as suggested by Novella et al. (2004). In this case, rather than parallel evolution of synonymous sites, some adaptive positions would appear more constrained than they actually are. The effect is to underestimate the rate of non-synonymous mutants, and therefore to underestimate the phylogenetic distance between two species by assuming that the transition between genotypes was a direct one.

In conclusion, considering and understanding the underlying structure of the genotype–phenotype map can lead to important conclusions about the evolutionary potential of a population. Most importantly within this model, it has been shown in every model variant that neutral mutants play an important role in making higher fitness genotypes more accessible, and that they make up a significant portion of the number of steps in an adaptive walk.

Chapter 5

The effect of shape on drift through networks

5.1 Introduction

Neutral networks can provide a way of making advantageous mutations more accessible via a series of single step neutral mutations (Chapter 3; Chapter 4; Huynen et al., 1996; Fontana and Schuster, 1998a; Smith et al., 2002). Until now a distance based approach has served well to calculate this accessibility (i.e. where a number of neutral steps is required, the shortest is the most accessible). While doing so, it was convenient to assume that the population moved as a single point, which is common within population genetic theory (Wright, 1982; Gillespie, 1984; Lande, 1985; Orr, 2003). However, when the population is considered as a quasi-species–type cloud of mutating individuals diffusing out across a network, then it is possible that the shortest path may not necessarily be the easiest, and therefore not the most likely. In this kind of situation, a direct measure of the number of steps becomes ineffective.

Van Nimwegen and Crutchfield (2000) and Sumedha et al. (2007a) have both explored the effect of population size and mutation rate on the way a cloud of individuals diffuses across a neutral network. Van Nimwegen and Crutchfield also calculated that the time taken to find a distant advantageous portal was orders of magnitude more likely across a neutral network, their 'entropic barrier', than waiting for a short–lived lineage to negotiate a 'fitness barrier' of deleterious intermediates. In this chapter I use a simulation model similar to that of van Nimwegen and Crutchfield's to show that the time taken to drift across a network is strongly influenced by a network's size and structure as well as population size and mutation rate.

Much of the other work quantifying drift over networks including that of Sumedha et al. (2007a), has focused on the selection for and evolution of mutational *robustness*, testing van Nimwegen et al.'s 1999 prediction that at higher mutation rates, a larger, denser neutral network confers an advantage to a population, because the chances of a mutated offspring retaining high fitness is greater (e.g. van Nimwegen et al., 1999; Wilke, 2001b; Krakauer and Plotkin, 2002; Lenski et al., 2006; Elena et al., 2007).

In contrast, in this thesis, we are interested in drift across a network as a means of accessing further adaptive mutants. Assessing the effect of network size and shape is particularly important in RNA genotype space, because the neutral networks can vary hugely. They range all the way from the dense, symmetrical 'face' structures highlighted in section 2.5 (Fig. 5.1a), to much less dense networks which form elongated paths through the genotype space (Fig.5.1b). Intermediate networks of all shapes and sizes lie between these two, indicating the potential for a rich variety of network shape effects on the time taken to find an advantageous phenotype.

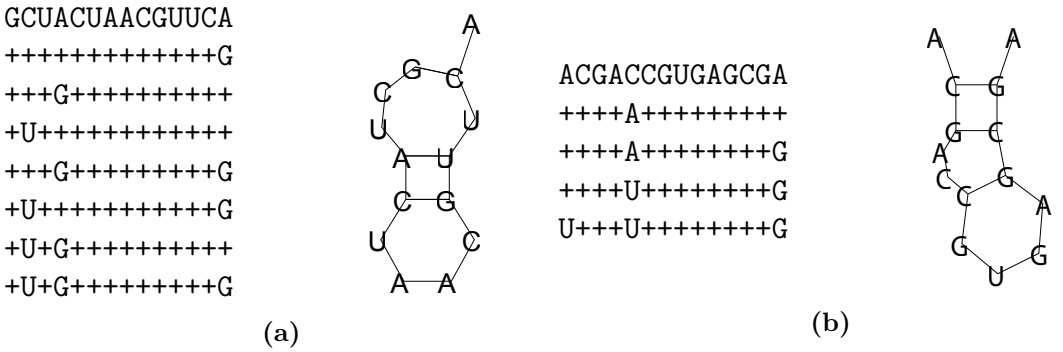


Figure 5.1: Two example networks from the length-14 space. In this figure, an '+' indicates that the base is the same as that in the first sequence. **a)** Network ID-4405: A neutral mutation can occur at position 2, 4 or 14, while still coding for the same phenotype (to a U, G and G, respectively). These mutations are always neutral, no matter which of the 2 bases are present at the other variable positions, meaning that each sequence has three point mutant neighbours in the network, and there are 6 paths the first and last sequence which all require three mutations. **b)** Network ID-1286 contains 3 fewer sequences, with variable bases at the 1^{st}, 5^{th} and 14^{th} positions. The 'network' forms a single path of sequences, where a mutation at any given position is only neutral when a specific combination of bases are present at the other variable positions. The intermediate sequences in the table each have two neutral neighbours and the end sequences have just one. There is single path requiring 4 steps in order between the first and last sequences.

In the rest of this chapter I characterise the effect of various network parameters in idealised toy network structures based on the alternative network topology highlighted in figure 5.1. These basic network shapes provide the building blocks from which it is possible to build more complex networks explored in sections 5.3.1 and 5.3.2.

The results are based on simulations of a finite population drifting over a range of different network sizes and shapes. The final section looks briefly at the time taken to drift between 'real' portals in an example network taken from the length-10 RNA space.

5.1.1 Simple network layout

In this section I introduce the two simplest network structures. The first is based on the 'dense' face-type network structure and we can call a *lattice*, because paths criss-cross the network. The second is based on the conditional networks found in the RNA space which consists of a single path (called a *string*), where each mutation is only neutral after a previous change has occurred[1]. They are at the opposite limits of possible network structure, and can contain vastly different numbers of sequences.

[1]This kind of conditional neutrality has also been observed in vesicular stomatitis virus by Quer et al. (2001)

Despite this, they can be directly compared using two distance measures. The first is the inter-portal distance, which here is calculated as the minimum distance between the starting genotype of the simulation (the *entry portal* and the closest advantageous genotype or *exit*). This means that the distance measured includes the adaptive step. i.e. the population has gone through the portal, rather than just finding the begining of the portal, which was the definition used in chapters 3 and 4. The second measurement is the diameter of the network. This is defined as the maximum distance across a network and is to a large extent independent of the number of sequences contained in the network (Fig. 5.2). For this first section, the entry portal and exit are always situated on opposite sides of the network, and therefore the inter-portal distance is equal to the diameter.

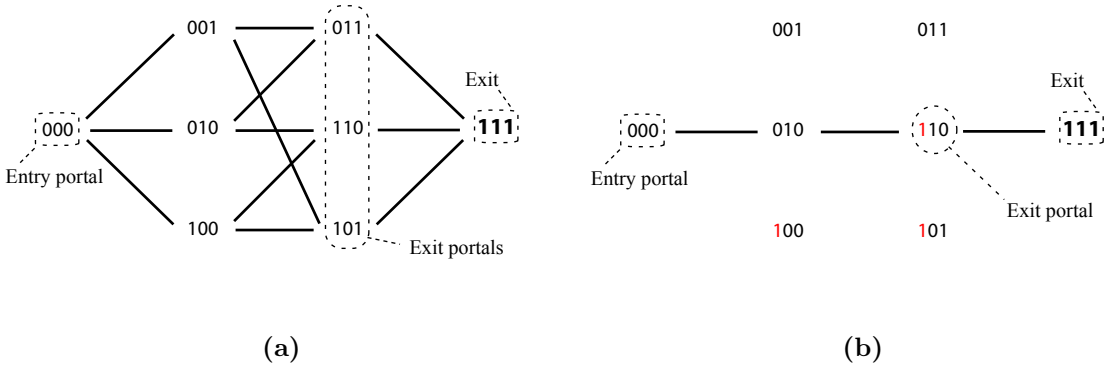


Figure 5.2: Examples of the lattice and string networks consisting of three binary loci. Because the entry portal and exit are on opposite sides in each network, the inter-portal distance is the same as the diameter – three steps. **a)** The lattice network. Each change can occur in any order, but mutations at all three loci are required to move from entry portal to exit. The first neutral mutation must reduce the Hamming distance between the current genotype and the exit, the second has a 2/3 chance of reducing it and a mutation from one of the exit portals has a 1/3 chance of reaching the exit. **b)** The string network contains fewer sequences for the same inter-portal distance, and has only one path. The number 1 allele at the first position (shown in red) is contingent on having a number 1 at the second position. Neither 100 or 101 is part of the network. After the first neutral mutation, a back mutation is as likely as a forward mutation at each Hamming distance from the exit.

In a lattice structure network all the combinations of alleles at the variable positions are neutral (Fig. 5.2a), except for one advantageous genotype (the *exit*). The particular combination of alleles in the exit genotype interact epistatically, because in any other combination none confer a selective advantage. When the entry portal and exit are on opposite sides of a lattice, the first viable mutation at any locus will reduce the inter-portal distance. However, the closer a sequence sequence is to the

advantageous genotype, the lower the chance that a further mutation will reduce the distance, and the higher the chance of a back mutation.

By contrast, in the string network only a subset of all the possible allelic combinations are neutral (Fig. 5.2b). Each neutral genotype can only be reached via a point mutation once a specific combination of alleles are present at other loci. For example, in the network in figure 5.1b a neutral mutation from an A to U at the first position is conditional on the 5^{th} position being a U and not an A (or any other base).

The epistatic interactions between loci are more complicated in the string network than in the lattice. A particular allele is deleterious unless a specific combination of alleles occur at other loci (Fig. 5.2b). Once this condition is met the allele becomes neutral. The allele at the final locus switches between being deleterious when the rest of the combination are not in place, to advantageous when they are. In both networks, a genotype which differs at one locus (position) from the advantageous phenotype (i.e. a Hamming distance of 1 away) is an exit portal (Fig. 5.2).

I shall now go on to explain the details of the simulation model used to calculate the average network crossing times.

5.2 Simulation model

When they were assessing how a population negotiated a fitness barrier, van Nimwegen and Crutchfield (2000) used an analytical model to calculate the expected time until the first individual with a portal genotype occurred in the population. They found that once the difference between the local peak fitness and the valley fitness got below a critical level (the error threshold), the valley essentially became a neutral network, where the population was free to drift across all the genotypes. Under this regime, accurate predictions from their analytical models broke down, and could only provide order of magnitude estimates. Instead, they used simulations to predict entropic barrier crossing times. As this chapter focuses on how populations drift across different shaped networks, I follow van Nimwegen and Crutchfield (2000) in simulating a finite population drifting across the network. The simulation ends when the first individual with the exit genotype appears. Again following van Nimwegen and Crutchfield, τ can be defined as the average number of generations it takes for the advantageous phenotype to occur. A further advantage of the simulation model is that it can be easily extended to examine even more complex networks taken from the RNA network space, as shown in section 5.4.

Throughout this chapter I continue to assume that selection is far stronger than mutation. This means I can assume that movement through a portal occurs 'instantaneously'. Thus the simulation is initialised with every individual having the entry portal genotype. Any genotype which is neither advantageous nor part of the network is assumed to be fatal. The standard assumption that all mutations have an equal probability of occurring is also retained, so for the simple networks there are two neutral mutants and two fatal ones at each variable position. Generations are non-overlapping. If μ is the mutation probability per position per generation, the probability of mutating between two genotypes can be calculated for each pair in the network.

$$P(i) = \left(\frac{\mu}{3}\right)^{m_{ij}} (1-\mu)^{n-m_{ij}}$$

where n is the length of the sequence and m_{ij} is the Hamming distance between the genotypes i and j. Figure. 5.3 and table 5.1 outline a simple two position lattice example.

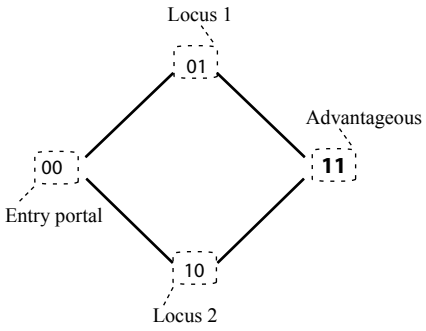


Figure 5.3: A simple two locus binary lattice network example.

	Entry	Locus 1	Locus 2	Advantageous
Entry	$(1-\mu)^2$	$\frac{\mu(1-\mu)}{3}$	$\frac{\mu(1-\mu)}{3}$	$(\frac{\mu}{3})^2$
Locus 1	$\frac{\mu(1-\mu)}{3}$	$(1-\mu)^2$	$(\frac{\mu}{3})^2$	$\frac{\mu(1-\mu)}{3}$
Locus 2	$\frac{\mu(1-\mu)}{3}$	$(\frac{\mu}{3})^2$	$(1-\mu)^2$	$\frac{\mu(1-\mu)}{3}$

Table 5.1: Table of mutation probabilities within a small two locus network. Note that back mutation is included and so the probability of mutation between two genotypes is symmetrical, except for the advantageous genotype. There is never any back mutations from the advantageous genotype, because the simulation stops at the first occurrence of and individual with the advantageous genotype.

In any given generation the expected proportion of the population with a particular genotype is calculated by summing the probability of mutating to it, from

each of the individuals in the previous generation. New individuals are then drawn randomly according to the calculated proportions to replace those of the previous generation. The pool of potential offspring is considered large enough that sampling occurs with replacement i.e. that those proportions do not change. These assumptions imply that the offspring population is drawn from a multinomial distribution, and is therefore subject to drift. The contrasting situation would be sampling without replacement, which would effectively lead to a set of independent random walkers (See van Nimwegen et al. (1999); Sumedha et al. (2007a)).

To calculate the probability of a particular genotype occurring in any given generation, we need to take into account the number of individuals with each different genotype from the previous generation. Using the example in table 5.1, let N be the total population size and $N_{entry,k}$ be the number of individuals with the entry genotype at generation k, and $N_{b,k}$ and $N_{c,k}$ be the number at loci 1 and 2 respectively. At the start of the simulation, all the individuals are of the entry phenotype so $N_{entry,0} = N$. The probability of drawing an individual with an entry genotype at generation k is therefore

$$P\left(Entry, k\right) = \frac{N_{a,k-1}\left(1-\mu\right)^2 + N_{b,k-1}\frac{\mu}{3}\left(1-\mu\right) + N_{c,k-1}\frac{\mu}{3}\left(1-\mu\right)}{\left(1-\frac{2\mu}{3}\right)^2\left(N_{a,k-1}+N_{b,k-1}+N_{c,k-1}\right)}$$

The first term of the numerator is the expected number of offspring of entry genotype individuals that do not mutate and the other two are the expected number of individuals with back mutations from loci 1 and 2. The number of advantageous genotype individuals is always zero in generation $k-1$, so the chance of getting a back mutation from it is also zero. The denominator is made up of the expected number of all non-fatal offspring. More generally one can say that the expected proportion of any genotype a in generation k is

$$P(a,k) = \frac{\sum_{i\in A} N_{i,k-1}\left(\frac{\mu}{3}\right)^{m_{a,i}}\left(1-\mu\right)^{n-m_{a,i}}}{\sum_{j\in A}\sum_{i\in A} N_{i,k-1}\left(\frac{\mu}{3}\right)^{m_{i,j}}\left(1-\mu\right)^{n-m_{i,j}}}$$

where A is the set of all viable genotypes and $N_{i,k}$ is the number of individuals with the i^{th} genotype in generation k. The number of individuals $N_{a,k}$ is drawn from a multinomial distribution:

$$\left(N_{a,k}\right)_{a\in A} \sim Multinomial(N, \left(P_{a,k}\right)_{a\in A})$$

$$\begin{aligned} \text{where } N_{entry,0} &= N, \\ N_{a,k} &\in \{0, \ldots, N\}, \\ N_k &= \sum_{a \in A} N_{a,k}. \end{aligned}$$

The simulation was run over a range of inter-portal distances from 2-8 steps for 100 trials of each. Sequence length was fixed at 10, and 4 combinations of mutation rate (0.001 & 0.0001) and population size (1000 & 10,000) were tested. Forster et al. (2006) calculated that if the product of genotypic mutation rate and population size was above 30 ($n\mu N > 30$), then selection for mutational robustness can become important (see section 5.3.2). The parameter values chosen therefore reflect product values of 1, 10, 10 and 100, to cover the range across which selection for mutational robustness might come into play.

Where computationally feasible, I also calculate the number of generations taken to find the advantageous genotype when *no* network exists (i.e. jumping from the entry portal directly to the fitter genotype in a single generation).

The method of calculating proportions, and then drawing from the distribution, is subtly different from van Nimwegen and Crutchfield (2000) and Sumedha et al. (2007a), who mutated each individual after it was drawn randomly from the parental generation. The method used here reduced the simulation running time. When fatal mutations occur in the offspring generation, then extra individuals need to be drawn to replace them. Because of the binary lattice nature of their model, van Nimwegen and Crutchfield (2000) did not include fatal mutants in their model.

We can test if this difference has an effect by reproducing the conditions used by van Nimwegen and Crutchfield, and briefly considering binary loci with no fatal mutations.

Figure 5.4 shows the relationship between τ, mutation rate and population size as calculated by my model. Across a fixed size network, the correlation between τ and μ and τ and N show good agreement with those of van Nimwegen and Crutchfield (2000, Fig. 4). The exponents of the power law relationships suggested by van Nimwegen and Crutchfield

$$\tau \propto \frac{1}{N^{\alpha}}$$

and

$$\tau \propto \frac{1}{\mu^{\beta}}$$

are given in table 5.2.

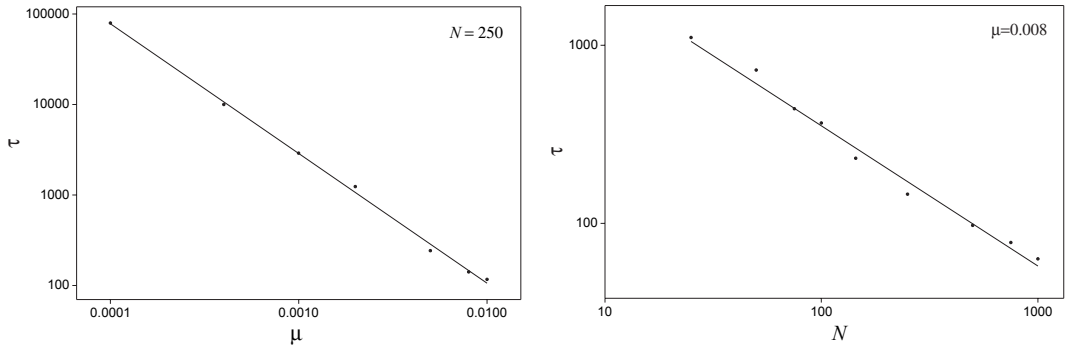


Figure 5.4: The effect of N and μ on portal discovery times for a fixed binary lattice network with different values of μ and N respectively. As with van Nimwegen and Crutchfield's data, τ is more variable in relation to N than μ. Inter-portal distance is fixed at 5, and network diameter (L) is fixed at 10. The lines on each graph are lines of best fit. See table 5.2 for details of the co-efficients

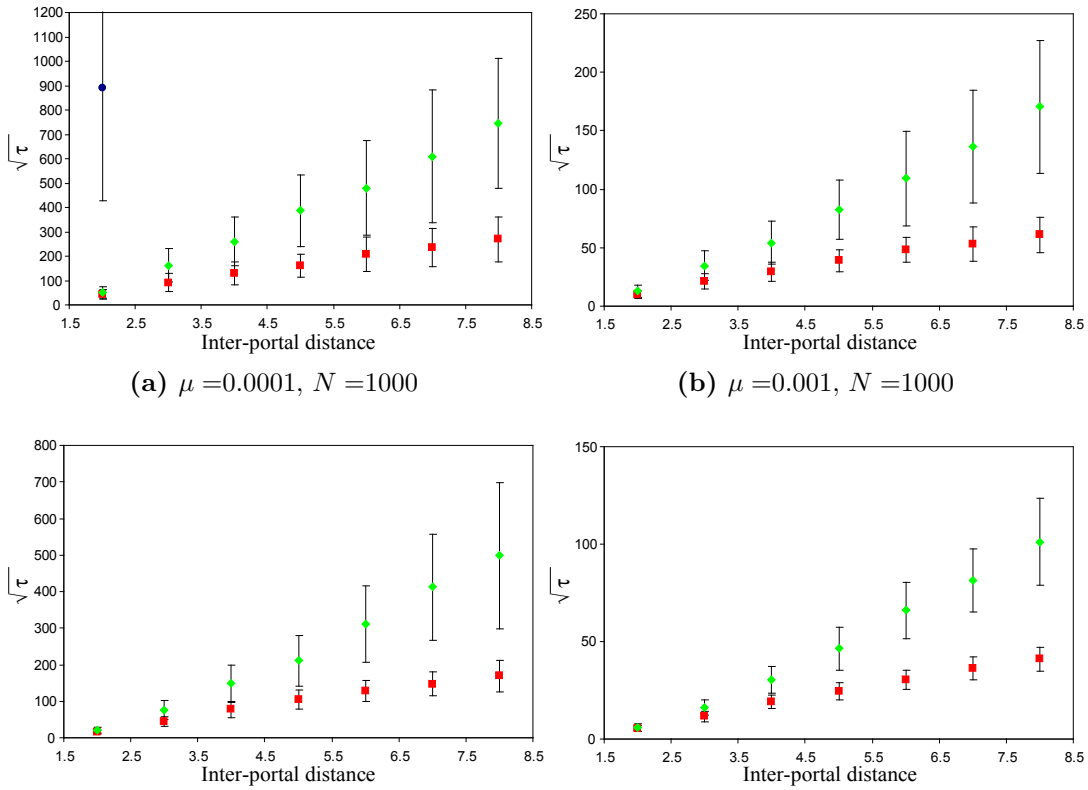
Model	exponent	
	α(when $\mu = 0.008$)	β(when $N = 250$)
van Nimwegen	0.761 ± 0.03	1.365 ± 0.014
This model	0.757 ± 0.02	1.395 ± 0.03

Table 5.2: The exponents for this model and their confidence intervals, compared with those calculated by van Nimwegen and Crutchfield (2000, table 1).

5.3 Results

For the string and lattice networks τ increases approximately quadratically with inter-portal distance (Fig. 5.5). However, for each combination of mutation rate and population size the string network takes substantially more generations to drift across than the lattice network. In fact under each of the $N\mu$ combinations, a string network with an inter-portal distance of 4 steps takes longer to drift across than a lattice network of 8 steps.

Where the intermediate genotype is fatal (i.e. there is a fitness not an entropic barrier), the average number of generations required to make a number of simultaneous base changes is very large (Fig. 5.5a, ●). Making two simultaneous changes (crossing a steep fitness barrier of width two), takes more time on average than drifting more than 8 steps on a single path when $\mu = 0.0001, N = 1000$. When three base changes are required simultaneously, the number of generations required becomes infeasibly large in this strict selection environment.



(a) μ =0.0001, N =1000 **(b)** μ =0.001, N =1000

(c) μ =0.0001, N =10000 **(d)** μ =0.001, N =10000

Figure 5.5: $\sqrt{\tau}$ against inter-portal distance for two different values of μ and N. Figure 5.5a includes the average time taken for a single double mutation to occur (the 'no network' case at an inter-portal distance of 2). ● = no network, ■ = lattice network and ◆ = string network. Error bar = 1 St.dev.

5.3.1 Entry portal position

In reality, and certainly in the RNA genotype space, the inter-portal distance is likely to be smaller than the diameter. With this in mind, the simple networks can be extended to increase the diameter without increasing the inter-portal distance (Fig. 5.6).

When the diameter is larger than the inter-portal distance, the average time taken to find the advantageous genotype can be longer than when the diameter equals the inter-portal distance. This is because some lineages in the population are likely to start by drifting in a direction away from the exit portal. In other words purging selection becomes less effective at driving the population in the 'correct' direction (see Fig. 5.6b).

The effect is most pronounced where the network diameter is significantly larger than the distance between portals, giving the population space to get lost, but where

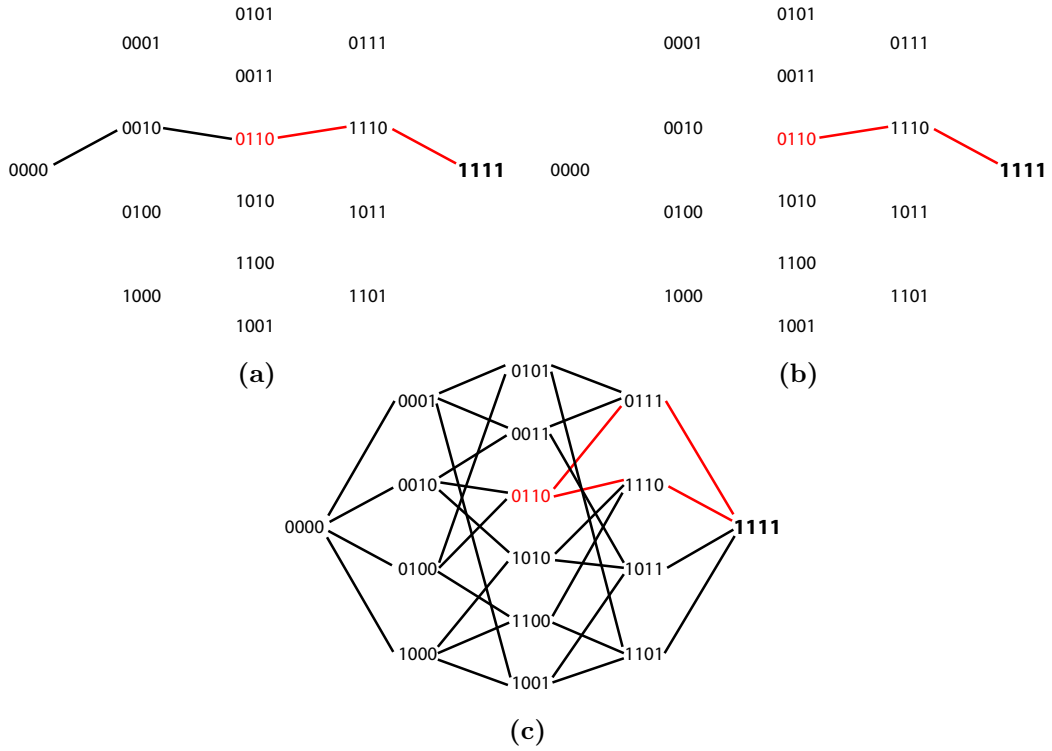
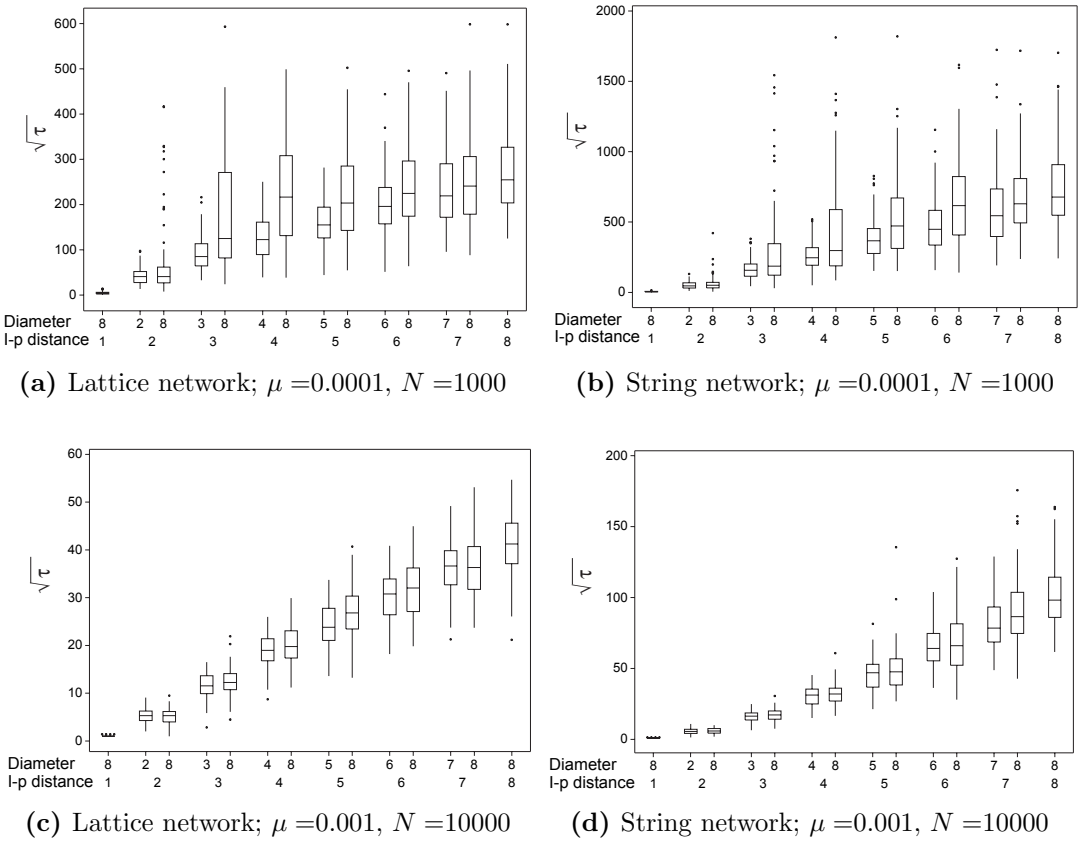


Figure 5.6: **a)** The entry portal (shown in red) is in the centre of a string network. This means that the inter-portal distance (2 steps) is shorter than the network diameter (4 steps). On average, it takes longer to find the exit in this case than in Fig. 5.6b, even though the number of mutations required are the same. Some lineages of the population are likely to drift in a direction which takes them further away from the exit portal. Here only the red lines reduce the distance to the exit. **b)** In a network which does not extend beyond the entry portal, any mutation away from the exit is deleterious and purged. The population is concentrated between the entry and exit portals, increasing the probability of an advantageous mutation occurring after fewer generations. Here both the inter-portal distance and network diameter are 2 steps. **c)** The same as **a**, but for a lattice network (inter-portal distance = 2, network diameter = 4).

the inter-portal distance is itself not insignificant, giving the population time to get lost.

The effect of increased network diameter on τ is not as pronounced as that of increasing the inter-portal distance or the overall shape of the network. However, we can see in figure 5.7a and figure 5.7b that $\sqrt{\tau}$ is significantly larger when the inter-portal distance is substantial, and network diameter is even larger. The variance can also increase and become more right skewed when the diameter is larger than the inter-portal distance. As might be expected, the lower limit does not decrease for a given inter-portal distance, but the upper limit can increase, and in some intermediates the upper limit of the distribution edges towards that where the inter-portal distance is as large as the network diameter (e.g. Fig. 5.7a, Inter-portal distances$>$ 4, network



(a) Lattice network; μ =0.0001, N =1000

(b) String network; μ =0.0001, N =1000

(c) Lattice network; μ =0.001, N =10000

(d) String network; μ =0.001, N =10000

Figure 5.7: Box-plots showing $\sqrt{\tau}$ at different start points in a network with a maximum size of 8 neutral loci. These are compared to a smaller network where the inter-portal distance is equal to the network diameter. Other combinations of μN produced intermediate results.

diameter= 8).

The effect that entry portal position has on τ is influenced strongly by the mutation rate and population size. The largest effect is seen at small population size and high mutation rate (Fig. 5.7a & Fig. 5.7b). Random drift on a small population is more likely to carry the whole population away from the exit, and therefore increase τ. In contrast, when there is a high mutation rate and large population size (Fig. 5.7c & Fig. 5.7c) individuals diffuse out in all directions across the network, allowing the exit to be found almost as quickly as when the network does not extend further.

The structure of the network also interacts with the network diameter to influence where the maximum effect is seen. In a string network, there is weak selection pressure towards the centre of the network, because that is the area of maximum robustness (the two end sequences have only got one neutral neighbour). This means that the maximum effect of having a larger network diameter is seen when the start point is on the exit side of the midpoint of the string. In contrast, in a lattice every genotype

has the same number of neutral neighbours, and so the largest effect is seen when the start point is in the middle of the network.

5.3.2 Combination networks

The completely regular structures laid out above do not reflect those found in most neutral networks. For instance, a base–paired position in RNA may be more constrained by its pairing requirements than a position which has no pairing requirements.

While variable unbound positions can be likened to a lattice, a string in RNA can occur when changes at a pair of bound positions switch from A-U to G-C or vice versa. This can only be achieved neutrally via a G-U intermediate, and not via an A-C. Furthermore, entry into a lattice may be constrained by the strength of the bonds between base pairs. For example, genotypic combinations involving large purines at unbound positions may only be stable if strong G-C bonds are found at the base-paired positions, meaning that changing within a lattice may only be possible when a string is in a certain configuration. These interactions can lead to a multitude of ways in which networks can be structured.

A simple start in attempting to model this kind of phenomenon is to combine the lattice and string structures in a single network. I have called the result a 'frying pan' network. One section (the 'pan') of the network has a lattice structure, allowing changes in any order, and the other is restricted to single base changes occurring in a linear fashion (the 'handle') (Fig. 5.8).

In the case of a composite network such as this, it is not just the network structure which is asymmetrical. The time it takes for a population to traverse the network is also different in different directions. When the entry portal is in the handle, and the exit is in the pan, the average number of generations required to traverse the network is lower than in the reverse situation. The reason is as follows – once the two types of network are joined, sequences in each of the two sections can have completely different numbers of neutral neighbours, and hence the probability of moving in one direction or the other is completely different (Fig. 5.8b). Furthermore, Forster et al. (2006) calculated that in the absence of any other selective criteria, when $n\mu N > 30$ selection for mutational robustness, a phenomenon first calculated analytically by van Nimwegen et al. (1999), means that the section of the network with a higher degree of neutrality is favoured, because individuals in that section produce less fatally mutated offspring, raising their reproductive rate slightly.

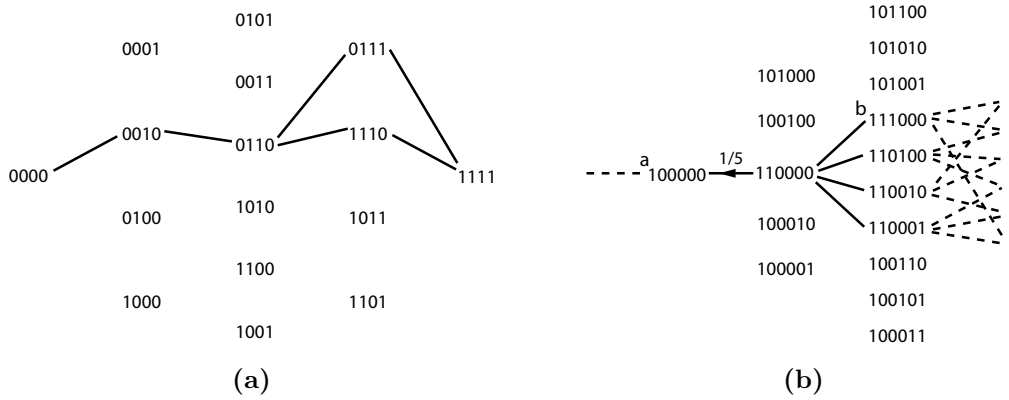


Figure 5.8: a) The connections in a 'frying pan' network, which is a combination of a lattice structure and a string structure. In this example, the network diameter and inter-portal distance are 4 steps, and the lattice diameter is 2 steps. **b)** If we consider the joining point of a larger frying pan network, there is a differential probability of moving from one half to the other based on the size of the lattice. The larger the lattice, the lower the probability of any given mutation being into the handle. In this example, the lattice size of 4 binary alleles means that there are 4 neighbours leading into the pan, and one leading into the handle, thus the chance of a mutation leading into the handle is 1/5. Compare this with **(a)**, where the lattice size is 2, and therefore the chance of a mutation leading into the handle is 1/3. Furthermore, genotype a in the handle has only got 2 neutral neighbours, whereas genotype b has 4, this means that on average b will leave slightly more offspring than a because fewer of its offspring will be fatal mutants.

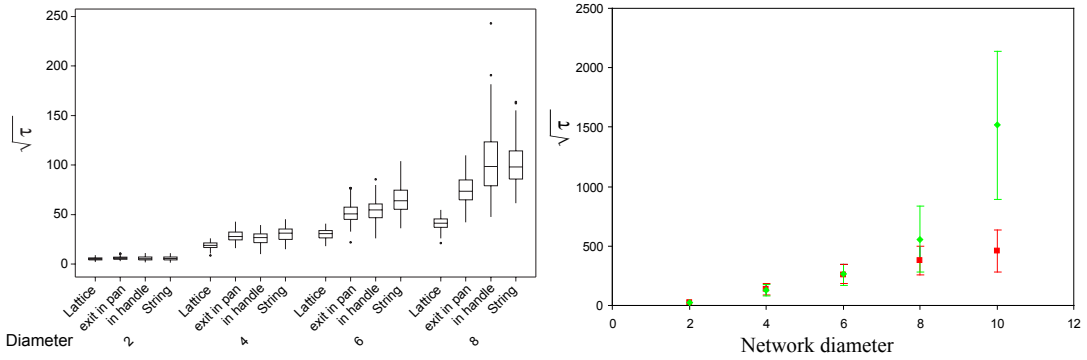
Entering the pan section from the handle presents no noticeable barrier to a drifting population. Figure 5.9a shows that the average number of generations from handle to pan is approximately half-way between the time taken for the pure lattice and pure string networks.

In contrast, drifting towards the handle becomes less and less easy as the lattice diameter grows. In figure 5.9a, by the time the network diameter has increased to 8 steps, the average number of generations required to traverse the network from pan to handle is equal to that of a pure string network. Figure 5.9b makes clear that when traversing from pan to handle, the increase in τ increases much faster with increasing network diameter quadratic relationship observed from handle to pan or across the uniform networks from section 5.3.

5.3.2.1 Entry start point within combined networks

The effect of the pan–handle junction, as opposed to just an increase in lattice diameter, can be seen more clearly if we alter the start point over a fixed diameter frying pan network (Fig. 5.10).

When the inter-portal distance is low, the entry portal is still in the handle and $\sqrt{\tau}$ increases approximately in line with the distance between entry and exit. Once



(a) μ =0.001, N =10000 **(b)** μ =0.0001, N =10000

Figure 5.9: The effect of asymmetry on $\sqrt{\tau}$ across combination networks of different diameters. In each network the lattice diameter equals the string diameter. **a)** The lattice and string network crossing times are shown for comparison with the frying pan network in both directions. **b)** $\sqrt{\tau}$ increases faster than quadratically with increasing network diameter, when heading from pan to handle (◆). It remains approximately quadratic when drifting from handle to pan(■). Error bars=1 St.Dev.

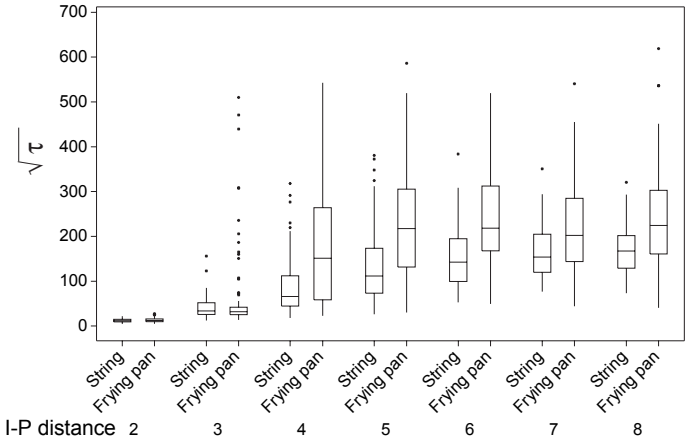


Figure 5.10: Box-plot showing the effect of different starting positions in a fixed diameter frying pan network (lattice size = 4, string size = 4) on $\sqrt{\tau}$. The comparable (size = 8) pure string network for different starting points is included for comparison. $\mu = 0.001$ and $N = 1000$.

the starting sequence is close to the pan section of the network, any increase in the inter-portal distance has little effect on τ (Fig. 5.10, distances 4-8). In fact it is the low probability of moving from the pan to the handle which becomes the constraining factor, rather than the inter-portal distance. The barrier between handle and pan is so difficult to overcome that once the entry portal is in the pan, $\sqrt{\tau}$ is larger than across a pure string network with an inter-portal distance of 8.

This result is especially interesting when we consider selection for mutational robustness or lack of it. Forster et al. (2006) observed selection for mutational robustness

when $n\mu N > 30$. In figure 5.10 $n\mu N = 10$, below that threshold, and yet we see that the time taken to cross the frying pan is longer than the time taken to cross the comparable pure string network. In fact it takes longer to navigate across the frying pan network from anywhere in the lattice than it does to drift the whole way across the pure string. The difference must be due to the asymmetric probability of entering each section, rather than the different number of neutral neighbours leading to selection for mutational robustness. This conclusion is enhanced if we reconsider the results shown in figure 5.9a. $n\mu N = 100$ in that case, above Forster et al.'s threshold, but here the frying pan network was no more difficult to cross at a diameter of 8 steps than the pure string. The network acts like a one-way check–valve or diode, only allowing flow in one direction and so I call this the 'diode effect'.

In summary, the chance of any individual in a population mutating into a handle from the pan decreases as pan (lattice) size increases, and thus the time taken to find an exit in the handle also increases. This effect is not due to selection for mutational robustness, because it occurs at all the values of $n\mu N$ tested here, and the effect is actually greatest at low population size, when drift should negate selection for mutational robustness.

5.4 'Real' RNA networks

The simulation model based around idealised networks can be simply extended to consider networks taken from the RNA genotype space. In this section I simulate the time it takes to cross a more realistic network than those examined above. In a 'real' network, each variable position can potentially contain 4 neutral alleles, rather than two neutral and two fatal, as used earlier in the chapter.

Network ID 12 (PID3) was chosen from the length-10 space, because it is well connected to other phenotypes, and of a computationally manageable size (containing 2020 sequences). This size is far larger than the simple networks examined earlier, but the maximum inter-portal distance across the network is 9 steps (See Appendix B) – 10 to the furthest new phenotype, which is the same as the maximum used in section 5.3.2. As before, the inter-portal distance *through a network* can be longer than the Hamming distance between the two sequences.

Within NID-12, the sequence GCCUGAAAAG was selected as an entry portal because it has at least one different phenotypic neighbour at every inter-portal distance between 1 and 9 mutations away (Table 5.3).

PID	Number of genotypes at each distance										
	1	2	3	4	5	6	7	8	9	10	11
1	0	2	8	6	0	0	0	1	7	3	0
2	0	0	0	12	106	282	248	105	68	20	0
4	0	0	0	0	16	78	193	350	362	125	0
5	0	0	1	1	7	26	52	119	158	49	0
6	0	0	0	0	5	21	75	156	231	149	0
7	0	0	0	0	1	9	22	23	10	1	0
9	1	3	17	12	57	81	157	163	110	36	0
10	0	0	0	0	6	6	35	118	111	49	0
11	0	0	1	2	2	23	75	108	72	37	0
12	0	0	0	0	3	4	2	1	2	0	0
13	0	0	0	0	1	1	14	58	58	12	0
14	0	0	0	0	0	2	7	19	16	0	0
15	0	0	0	0	0	4	14	16	15	9	0
16	0	0	0	0	0	0	0	0	1	0	0
17	0	0	0	0	0	0	0	7	9	5	0
18	0	0	0	0	0	0	1	2	2	1	0

Table 5.3: The number of unique genotypes coding for each phenotype by their minimum path length through the network from the entry portal sequence GCCUGAAAAG.

For each simulation, one phenotype was chosen from the network neighbours as being 'advantageous'. The set of all viable genotypes then consists of the 2020 sequences of network ID-12 plus all the advantageous genotypes which are in the neighbourhood of the network (Fig. 5.11). As the network structure is intermediate between a lattice and string, we might expect intermediate values of τ, depending on the location of the portals, and therefore the shape of the network between the two sequences.

Figure 5.12 shows a clear correlation between $\sqrt{\tau}$ and minimum inter-portal distance. The mean time does always lie in between the values of the pure string and pure lattice network for all the portal distances except the largest one (to PID16). Interestingly, the large size of the network, which we might have expected to increase τ because populations might find it very easy to get lost, does not have a large influence on the length of time it takes to find a particular exit. At no stage is τ larger than the string network, indicating again that at higher values of μ and N, network structure and size *towards* the exit is always more important than structure and size *away* from it.

The one exception is the single unique portal to PID16. It lies 9 steps across the network from the entry portal (Hamming distance of 7). This genotype was not once discovered in several hundred hours of CPU time (in comparison 50 trials to find a

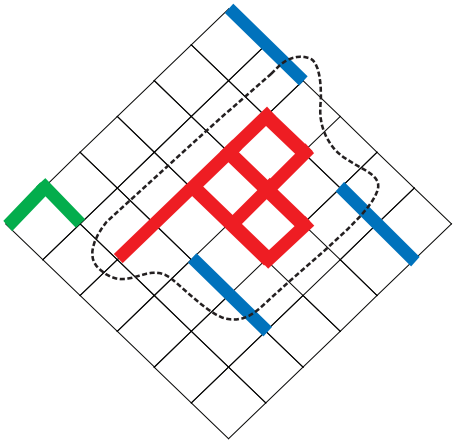


Figure 5.11: All the genotypes which belong to the neutral network (red) together with all the 'advantageous' (blue) network neighbours constitutes the set of all viable genotypes (included within the dotted line).

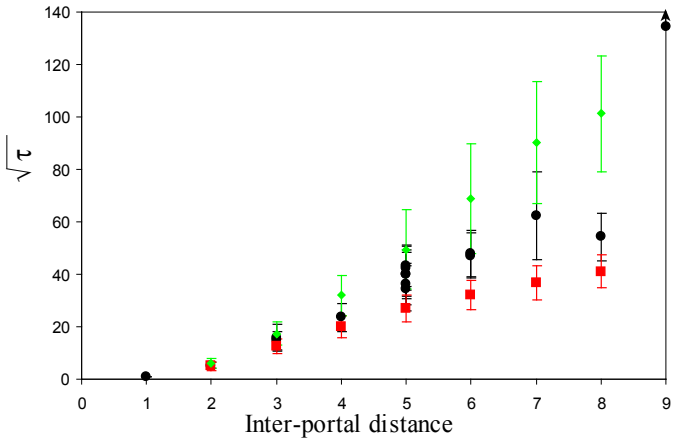


Figure 5.12: Average traversal times to cross network ID-12 (●) and the corresponding lattice (■) and string (◆) networks at different inter-portal distances. $\mu = 0.001$ and $N = 10000$, number of trials = 100, except NID-12 distances 7 & 8, where No. of trials = 50. Error bars = 1 Sd.Dev.

genotype at distance 7 or 8 took 20 hours CPU time) and repeated over a range of combinations of μ and N, including those where $n\mu N \ll 30$.

The conclusion is that the chance of reaching such a far flung and unlikely genotype is so small as to be virtually impossible. When the shape of the network is taken into account, the vast majority of individuals remain so far removed from the area of the network which is close to the advantageous genotype that it is extremely unlikely that it will ever be discovered.

Though it is not possible to say from these results exactly what the structure

of the network is, we can get an indication from the calculations from section 2.5. The network density is 0.123. The variability of the positions within the sequence is this: 6 positions vary between four bases, 2 between two, and 2 are completely constrained. This means that the network diameter is at least 8 steps, indicating there is the potential for reasonably long neutral paths.

The relatively low density means that there are likely to be at least some paths which are relatively convoluted within the space, and indeed the path to PID16 is longer than the Hamming distance between the entry portal and the adaptive exit. It is perhaps instructive to consider that though the network contains just 2020 points, it defies simple structural examination, and any future work will have to consider ways of accurately grasping the complexity of network structure.

5.4.1 Number of exits

Though all the other values of $\sqrt{\tau}$ for the NID-12 sit in-between those of the corresponding lattice and string networks, there is some variation between the different PID targets. $\sqrt{\tau}$ is generally lower where there are more advantageous genotypes at the minimum inter-portal distance (Fig. 5.13). The more exit genotypes that border a network, the greater the number of paths that lead to an exit, and therefore there is an increased probability of finding one of them. In fact the average time taken to find the 7 genotypes that code for PID-17 is lower than to find the one genotype coding for PID-18 even though PID-18 has a shorter minimum path length (7 compared to 8).

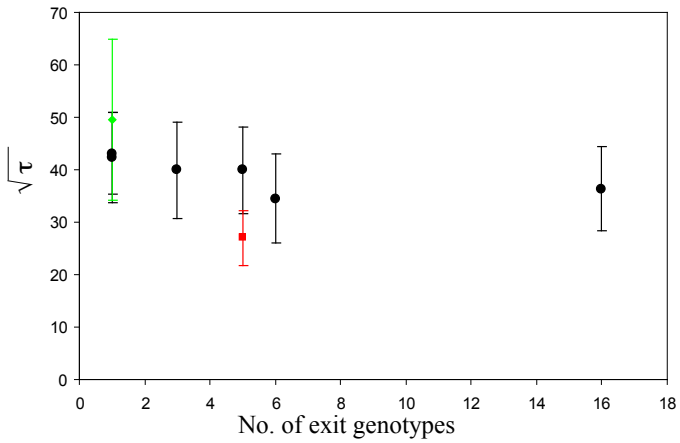


Figure 5.13: Average traversal times to drift an inter-portal distance of 5 steps against the number of exit genotypes at that distance. NID-12 = ●, lattice = ■and string = ◆. $\mu = 0.001$ and $N = 10000$. Error bars = 1 Sd.Dev.

However, the effect may be compounded by factors other than the number of neighbours. It is quite possible that the difference in $\sqrt{\tau}$ could be due to differences in shape of the particular part of the network leading to each different portal, rather than the number of genotypic exits once the population gets there. This is perhaps the reason why in the example shown in figure 5.13, an increase from 6 to 16 neighbours does not continue to decrease the average time any further than it had already dropped from 1–6. The interplay between shape and number of advantageous genotypes is complex, and requires further consideration.

5.5 Summary of results

In this chapter we have seen that network size and shape has an important impact on the time it takes a population to cross an *entropic barrier* to find an advantageous portal on the far side of a network. There is a complex interplay of size and shape with population size and mutation rate, all of which combine to influence barrier crossing time. The results can be summarised as follows:

Distance

- For fixed N and μ the average generation time to cross a network is related approximately quadratically to number of steps required to cross it.

Shape

- Networks with a larger number of possible paths (lattices) take less long to drift across.
- Shape interacts with distance, so that the larger the distance between the portals, the more effect network shape has on the time taken.
- Under the strict selection criterion used here, where no network is present the number of generations required to find an advantageous phenotype by simultaneous mutations is almost always infeasibly long.

Inter-portal distance v Diameter

- An increase in the size of the network beyond the inter-portal distance can increase the average number of generations required.

Directional bias in complex networks

- A population is much less likely to find a portal at the end of a rare and difficult to reach part of a network if there are many other more common options in other directions – the diode effect.
- The position of the portals within an irregular network is of huge importance to the length of time it takes to negotiate an entropic barrier. The effect is not simply caused by selection for mutational robustness.

5.6 Discussion

Van Nimwegen and Crutchfield (2000) calculated that a population is more likely to find a distant advantageous genotype by navigating an *entropic barrier* involving neutral drift through a lattice network than by traversing a fitness barrier. In this study I have shown that changing the shape of the network has a very strong effect on the time it takes a population to negotiate that entropic barrier. If the shape is limited to a single conditional path (a *string* network), the average number of generations required to cross the network is significantly higher than if all the allelic combinations at all the variable loci are neutral (a lattice). This is an important finding, because there are examples of lattice and string type networks in the RNA genotype space, and the shape of networks is likely to vary across any genotype space they exist in.

5.6.1 Implications for RNA networks

Most network structures in the RNA space lie somewhere in between the lattice and string. They generally have diameters larger than the average inter-portal distance, and more irregular shapes. This is because the physical structure of a phenotype exerts different pressures on different positions in the genotypic sequence. Positions which form one half of a base pair are particularly likely to be more constricted in their neutral neighbours compared with unbound base positions. Furthermore, their neutrality can be contingent on which bases are present at other positions in the sequence. This can be particularly true when changes include less stable base pairs, especially G-U intermediates. Figure 5.14 repeats the example from section 3.3.3, where a more stable intermediate is required to allow a base pair to neutrally mutate.

We saw in section 5.3.1 that the size of a network in a direction away from the exit can also have an effect on the time it takes a population to cross an entropic barrier. If the diameter of a network is significantly larger than the inter-portal distance, the

```
{51}⇒  AGAGAUCUCU
       G+++++++++
       G++++++++U
       G++++++++C
       G+G++++++C
       G+G++++C+C
       G+G++++C+U
       AGGGAUCCCU  ⇒ {60}
```

Figure 5.14: A neutral path across network-26. Entering from NID-51, to change an A-U pair to a G-C pair at the 3^{rd} and 8^{th} positions it is necessary to first change the base pair at positions 1 and 10 from A-U to G-C. This is because the third hydrogen bond between a G-C base pair provides extra stability to hold the structure together through the weaker G-U intermediate at positions 3 and 8. The two sequences on either end of the string are only a Hamming distance of 2 away from each other, despite requiring 7 steps.

population can drift off in the wrong direction (especially at low μ and small N). The effect is relatively weak when the network is regularly shaped with binary loci at each position. In other words, though the size and shape of the network *away* from the exit can have an effect, the size and shape from the entry portal *towards* the exit is far more important. However, there is potential for the effect of size away from the exit to be amplified in the RNA space. Each position has can have up to four neutral bases, rather than the two used in the toy networks. The more neutral genotypes there are that lie in the opposite direction to the exit, the greater the chance of a population getting lost in the wrong part of the network. Most importantly, the location of the entry and exit points within an irregular network can have a very strong effect on the time taken to find the exit. At times an adaptive portal may even be almost impossible to find, because the path to reach it involves making the difficult transition from the pan to the handle of a frying pan type network, a 'diode effect'.

5.6.2 Robustness and Asymmetry

When a network is not of a uniform shape, sequences differ in the number of neutral neighbours in their neighbourhoods (i.e. they vary in their *robustness* to mutations). In this case, the respective positions of entry and exit portal can have a large impact on the time it takes to find the exit (see section 5.3.2). Depending on the rate of mutation and the size of the population, drifting towards an advantageous portal can become quite difficult (For example finding the single PID16 genotype in the real network example in section 5.4). However, the effect does not appear to be due to

selection for mutational robustness *per se*, but because the probability of taking a step towards a less robust area is lower than the reverse. The results from section 5.3.2 indicate that finding an exit in a remote and difficult to reach part of a network is *more* difficult in a regime where drift dominates over selection for mutational robustness (when μN is low).

5.6.3 Robustness and evolvability

Portal genotypes do not have to lie away from the robust 'centre' of a network like the genotype coding for PID16 does. In fact, the very notion of a 'centre' as laid out by Forster et al. (2006), where "there is selective pressure that keeps the population away from the fringes of the neutral network, and pushes it towards the more densely connected areas in the centre", can be quite conceptually misleading in this context, as it implies a small, concentrated area. In a frying pan–type network there are *more* sequences in the 'centre' (the pan), which share the same higher degree of robustness, than there are in the handle, on 'the fringes'.

Because there are more boundary sequences surrounding the pan section than the handle in total, drift towards a more robust area might even go hand in hand with evolvability, though on average an individual genotype in the pan is less likely to produce mutant offspring with a different phenotypic character. If we apply the formula for the number of neutral neighbours per sequence (Eq. 2.2), in a binary lattice example with a diameter of four steps. We get $(n-4)\times 0+4\times 1$. In a binary lattice the local neighbourhood is not $3n$ sequences but only n (because a base at any position can only vary to the alternate base). This means that each binary lattice sequence has $n-4$ non-neutral neighbours. There are 4^2 genotypes in a binary lattice (diameter=4) so for a sequence length of 10, the total boundary of the network is

$$B = (10-4)4^2 = 96$$

In contrast a string network of diameter=4 contains five genotypes. Each end sequence has one neutral neighbour, and the other three have two. So the total number of network neighbours is:

$$B = 2\times 9+3\times 8 = 42$$

The results hold for 4 bases, but the calculations are not shown here as they are significantly more complicated, because of overlapping neighbourhoods among positions where bases can only vary between two options neutrally.

No part of a lattice is more robust than any other, so after a period of equilibration, an individual has an equal probability of inhabiting any point in the pan section (Derrida and Peliti, 1991; Sumedha et al., 2007a). Increasing the diameter of a lattice increases the number of genotypes, and hence decreases the proportion of a population at any particular point. This could potentially slow the discovery of a particular adaptive portal. However, it also increases the number of genotypes around the boundary of the lattice. As each lattice sequence lies equally on the boundary, there are more alternative genotypes (and potentially phenotypes) available to a population inhabiting the lattice than the handle (Fig 5.15).

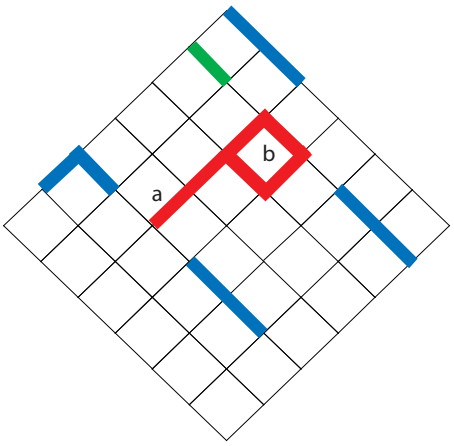


Figure 5.15: In this (inaccurate) representation of a frying pan network, the handle section (a) contains two genotypes and 2 of its 5 boundary sequences code of viable phenotypes in this example. This means that there is an average of one phenotypic mutant per genotype. The pan section (b) contains 4 genotypes, and 3 of its 7 boundary sequences code for viable phenotypes. In this example it has an average of 0.75 local phenotypic mutants per genotype, lower than the handle section. However the total number is higher, increasing the chance than one of them will be advantageous.

The impact of this effect is most strongly felt where there a large difference between the number of neutral neighbours in two or more areas of the network; so when there is more than one neutral allele at each locus the effect is magnified. In the RNA space, with 4 potentially neutral bases at each position, there is the potential for a large degree of disparity between the robustness of different areas of a network. This may mean that adaptive walks across the whole of the genotype space by a population modelled as a cloud of individuals, rather than tied to a single point, would result in *longer* path lengths than those recorded in chapter 4. Any time the shortest path lay towards the end of a handle, the walk taken by a drifting population cloud might well involve a number of more likely steps to a portal further away but easier to reach. As

the majority of paths only involve a few neutral steps, it is possible that many will be unaffected by whether the population is modelled as a single point or a cloud, but for the rare longer paths, irregular network structure is likely to have a profound effect.

5.6.4 Network shape and valley escapes

Throughout this chapter I have assumed that the level of selection (and of neutrality) is very strong i.e. that all non-neutral or non-advantageous neighbours are fatal. It was necessary to make this assumption to simplify the calculations and make clear the effects of changing the network shape and size. Even then, we have seen that the complex interplay between mutation rate, population size and network size and shape can be difficult to untangle.

If the assumption is not met, then it is possible that shorter paths through a 'valley' of slightly less fit phenotypes could replace a long and winding neutral path. Valleys of different fitness significantly complicate the calculations with real networks and have not been considered here. If they were, van Nimwegen and Crutchfield (2000) suggest that width is more important than depth, and that a population crosses at the narrowest rather than the shallowest point. From the results in this chapter, it appears that the shape of the network would also have a very strong effect on where and when a population would cross a valley. If the narrowest valley is at the end of a long handle, or even if a network contains a large lattice in a direction away from the narrowest valley, a population may eventually be more likely to cross at another point. The majority of the population will not be focused at the point from which the valley was easiest to cross. If many more maladaptive and therefore short–lived valley walks set off from a more robust part of the network, a different longer path might become the most likely route.

5.6.5 Population size and mutation rate

The mutation rates used in this chapter are at the higher end of the spectrum recorded in the natural world (e.g. those found in HIV Hahn et al., 1986), but are significantly lower than those used by most other studies (e.g. van Nimwegen and Crutchfield, 2000, used per position mutation rates between 0.002 and 0.05, and Sumedha et al., 2007a, used genomic mutation rates between 0.01 and 0.25). They were chosen so that the product of the genotypic mutation rate and population size ($n\mu N$) bracketed the value at which Forster et al. (2006) observed selection for mutational robustness.

$n\mu N$ is a very natural value to consider, as it is the average number of new mutants occurring in a population per generation. In this respect, a small population with a large mutation rate, and a large population with a small mutation rate can create the same number of new mutants per generation. The difference between the two comes due to random sampling effects (Sumedha et al., 2007a).

The way that drift scales with N while $n\mu N$ remains constant may play an important role in real world populations. It is likely to be particularly important in relation to the size of the network (Forster et al., 2006). When N is far larger than the network, as was usual in this study, for reasonable μ we might expect individuals to spread out across the whole network, and find any available phenotype, even those less accessible parts. In contrast, if the network is much larger than population size, a small μN might allow the whole population to drift across the network, rather than being trapped statically in a quasispecies-like cloud at the most robust point(s) in the network. The situation is complicated further still if we consider that mutation rate, at least, can be under selection pressure itself (Bedau and Packard, 2003; Andr and Godelle, 2006).

The interaction between μN, network size and network shape warrants further study, with the RNA genotype space model being ideally placed to tackle it relatively simply.

5.6.6 Distance versus neighbours

When the population size and mutation rate are such that at least the local neighbourhood of a population is likely to be well explored, the distance to a portal and the shape of the network are likely to be the overriding factor in dictating which portal is discovered first. Fontana and Schuster (1998b) based their notion of phenotypic accessibility from a particular network as the proportion of the network's shared boundary. However, we have seen in this chapter that the accessibility of an advantageous phenotype varies hugely depending on the positioning of the portals into and out of a network. A blanket figure for the whole network is likely to be an inaccurate indicator of accessibility.

Consider a population starting from a single genotypic origin. If it has an advantageous mutant in its local neighbourhood ($3n$ sequences), any point mutation has a $1/3n$ chance of resulting in that mutant. If a second adaptive phenotype requires a single neutral step to become accessible, the larger number of total neighbours to

search means that the number of adaptive mutants must be higher to keep the proportion of adaptive mutants to non adaptive mutants the same, even without considering the reduced chance of making two steps as opposed to just one.

5.6.7 Time scale and punctuations

With the mutation rate at the high end of the scale and the population size at the low end, the number of generations taken to cross networks in these simulations are unlikely to be an accurate reflection of the times it might take real real populations to navigate neutral networks if they were to require it. However, they do show that it takes relatively few generations for a population to drift a small number of neutral steps, so small inter-portal distances are likely to be reasonably easy for a population to navigate. In this situation the neutral mutants which lead towards a portal may never build up to a detectable frequency in the population before being replaced by an adaptive mutant, and so may appear to have never occurred. Because the average number of generations taken to find a portal rises steeply with increasing distance, we can see from the graphs in figure 5.5 that, depending on the network shape, the time scale for a rare longer path of approximately 4 or more mutants (400-1,000,000 generations) puts it in the order of magnitude of the phenotypic punctuations seen in *E. coli* (Elena and Lenski, 1997) over 3,000 generations, or in the fossil record (Eldredge and Gould, 1972; Gould and Eldredge, 1977), over longer time periods. Thus shorter neutral paths present little barrier to evolution, and may occur with a relatively high frequency, without necessarily even being noticed, while longer paths fit the empirical data on phenotypic punctuated transitions, as pointed out by Fontana and Schuster (1998a), Ebner et al. (2001), Crutchfield (2002), Smith et al. (2003) and Wolf et al. (2006).

5.6.8 Conclusion

In conclusion, the underlying structure of each neutral network within the genotype–phenotype map has the potential to play an important role in defining the accessibility of adaptive mutations, not just in terms of how easy it is to reach an adaptive mutant, but whether it is possible at all within a reasonable time scale. This chapter has made good progress in helping us understand the role of that structure in relation to population size and mutation rate. Even when the model makes simplifications as to the neutrality of networks, or the strength of selection, the interactions between population size, mutation rate and the size and structure of less regular networks make it

more difficult to predict precisely the time taken to find an adaptive solution by drift. However the advances made in this chapter make it possible that potential further work, modelling the effects of more complex networks on adaptive potential, could have a profound effect on the way we consider apparent punctuations in phenotype, particularly important in assessing the evolutionary potentials and vaccine escapes routes of viruses (Koelle et al., 2006).

Chapter 6

Discussion and conclusions

6.1 Discussion

The existence of neutral or nearly–neutral mutations in natural systems has become widely accepted over the last 50 years. Developing models which include genotypic degeneracy is therefore an important part of evolutionary research. It is only over the last 10 to 15 years that discrete genotype–phenotype map models have filled this role within an adaptive landscape framework. If neutral mutations are not included in models of adaptive evolutionary change, we risk excluding selection for genetic robustness (van Nimwegen et al., 1999; Wilke, 2001a; Wilke and Adami, 2003), and also an important property of the map's degenerate nature – to facilitate adaptive change by drift across neutral networks. That is, in the search for fitness optima, neutral mutations can provide a way of exploring a far larger fraction of the fittest phenotypes than would otherwise be mutationally accessible (Lipman and Wilbur, 1991; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Smith et al., 2003; Wroe et al., 2007).

This kind of model, based around whole genotypes rather than studying independent loci, has increased the emphasis on the importance of epistasis. The facilitating force that neutral mutations provide can be thought of as the eventual expression of an epistatic interaction between a particular set of mutations at different loci.

Poelwijk et al. (2007) point out that while traditional evolutionary studies have compared the end points of evolutionary trajectories, 'unseen intermediates' define evolutionary outcomes. Tracking a series of mutations across genotype space allows us to study these unseen intermediates. Furthermore, Whitlock et al. (1995) argue that evidence of epistasis is difficult to find even if it is integral to the evolutionary process (but see Elena and Lenski, 2001, for an example of epistatis in *E. coli* mutants). By considering the epistatic interactions of unseen intermediates, a genotype–phenotype map model can lead to important insights into the evolutionary process even when neutral networks are not a large feature of the space. The best example from this thesis are the adaptive walks traced through the degenerate genotype–phenotype map. The number of *adaptive* genotypic mutations undergone by a population evolving across the map can be larger than the direct genetic distances calculated between sequences (see section 4.3.1.1). Although this result was found within a map where neutral networks are very important, the presence of paths consisting of purely adaptive steps that are longer than the Hamming distance between the end sequences indicates that adaptive mutants can interact in a complex epistatic way, where occasionally a back mutation can increase fitness.

The existence of this kind of path leads to an important inference about the nature of phylogenetic differences. When the end products of two divergent sequences are assessed using a pair-wise comparison, even what appears to be a highly conserved gene has the potential to have undergone a larger number of genotypic changes, which have subsequently converged on the original sequence. This effect could lead to underestimation of phylogenetic distance. In other words, using Hamming distance as a measure of accessibility or of genetic closeness can be confounded by long path lengths through genotype space, especially if they involve several neutral mutations.

Within the rest of this discussion, I shall first consider the computational methods used to model the spaces from this thesis, and argue that their design makes them ideal for future work studying the RNA (and potentially protein) genotype space(s). I then consider the assumptions of the RNA space model, and argue why it is a plausible and valuable tool for investigating evolutionary dynamics. This in turn leads to further consideration of some of the results presented in chapters 2–5, and their assumptions, including some speculation as to the impact of these findings on other evolutionary phenomena.

6.1.1 The RNA model

The method of mapping genotype spaces used in this thesis provides a framework within which further simple models can be used to directly answer questions about how the underlying structure of the genotype–phenotype map affects the potential trajectories of populations evolving through it. This is a development from previous exhaustive searches of the RNA genotype space, in which the emphasis was placed more firmly on the underlying statistical properties of the networks themselves.

The nature of the particular genotype space, i.e. an exhaustive map of short sequences, has advantages and disadvantages. Using short sequence lengths reduces the biological applicability of the results as they are not long enough to exactly mirror the RNA sequence lengths used in various signalling functions within the cell (e.g. tRNAs and mRNA regulatory elements). However, I assert that this is more than compensated by the increase in tractability and the ability to perform precise calculations and targeted simulations, which are only possible because the space is exhaustively mapped. In particular an exhaustive mapping allows us to be certain of capturing the affects of rare networks and events.

As the very existence of life is highly improbable, taking samples or sub-sections of the space might miss the rare or improbable results which are in fact among the most important in evolution, precisely because they do only happen very rarely (for

example long and winding paths through a network). A small map also allows clearer analysis of the elements within it, meaning that it is easier to *explain* a particular phenomenon, rather than simply observe it. Even within this kind of model, the space is still so complicated that there may be many alternative explanations for a particular phenomenon (e.g. those seen in section 3.3.2 and section 4.5).

The methodology of the model has made two important steps forwards in relation to other exhaustive RNA models. The first is the sorting algorithm, which allows (relatively) fast and efficient calculation of the connectedness of sequences with the same phenotype. This algorithm is of particular interest to future neutral network modellers, because it will work for sequences of any length, and up to phenotype sets of significant size (for example the PID0 set from the length-16 space contains 1,896,558,063 sequences). Computationally, the use of the UNIX sort algorithm means that the algorithm has a much lower total RAM requirement than the other methods using a breadth-first search.

The second important development is the use of a memory efficient integer array to store the entire genotype map in instant access RAM. Though it is slower to calculate network connectedness than the sorting algorithm for longer sequence lengths, this instant access allows exploration of the boundaries of each network, as well as the actual networks themselves, and it is thus simple to track populations across the whole genotype space. For example, the integer array could be used as the basis of a model which extended and combined the work in chapters 4 and 5, to trace the way in which a population cloud drifts across the whole genotype space.

With the rapid increase in hard disk storage space, it would now be feasible to map the length-18 networks using the sorting algorithm, especially if the sequences which did not fold into any secondary structure were not considered. However, given the similarities of the preceding sequence lengths, this is of questionable value without the computational resources available to apply further models using the integer array method over the whole resultant genotype space.

These two computational methods in combination with the Vienna RNA secondary structure prediction software provide a basic toolset from which it is possible to launch further experiments into the nature of the RNA genotype–phenotype map and its effect on populations evolving through it. However, the sorting algorithm and the integer array could both equally well be applied to other genotype–phenotype maps, where a different mapping function is substituted for the `RNAfold` programme.

The way in which the discrete genotype–phenotype model tackles epistatic interactions means that it could easily be modified to model co-evolution. At the risk of

a combinatorial explosion of sequence space, one could model every combination of genotypes from all the co-evolving entities in the genotype space, and map the fitness differences between epistatically interacting genotypes to them.

I shall now consider some of the assumptions made within these models, and the impact of those assumptions not being met have in more realistic situations.

6.1.2 Assumptions of the genotype–phenotype map

The profile of how genotypes map to phenotypes in an accessible degenerative manner remains remarkably constant across a wide range of parameters within the RNA model (Tacker et al., 1996; Kospach, 2003). Changing the sequence length, the parameters for the `RNAfold` secondary structure prediction algorithm, and even changing the connectedness of the local neighbourhood do not fundamentally change how genotypes map to phenotypes (Tacker et al., 1996; chapter 2). I shall now consider the implications of changing or relaxing these three assumptions in more depth.

6.1.2.1 Longer sequence lengths

The sequence lengths used in the genotype space in this thesis are shorter than any of the RNA molecules found in real life systems. In this case it is necessary to ask how plausible the map is, and whether the structure of the genotype space approximates that of sequence space at longer lengths?

Fortunately there has been a considerable amount of work modelling longer molecules of RNA, all of which shows a very similar pattern to the smaller genotype space in the characteristic distributions of phenotypes and sequences (Huynen et al., 1996; Fontana and Schuster, 1998b; Schuster and Fontana, 1999; Wilke, 2001a; Sumedha et al., 2007b). Most importantly those models tracing evolutionary trajectories through the space have observed similar properties of sustained genotypic drift followed by sudden phenotypic transitions indicating that networks are also connected in a relatively similar way to the models presented here (Huynen et al., 1996; Fontana and Schuster, 1998a; Fontana, 2002).

The adaptive walk simulations from Chapter 4 suggest that as sequence length increases a smaller percentage of walks get to the optimum. As the number of phenotypes increases, the landscape becomes more rugged, and therefore the number of local optima increases. We might have expected this to reduce path length; however, the mean path length also increased with sequence length, indicating that evolutionary walks at longer sequences lengths may have to go through more steps, but will be less capable of reaching the global optimum.

All of this speculation about longer sequence lengths supposes that the level of epistatic interactions within a short sequence of RNA are likely to extend to longer sequences. One way in which the genotypic map might be accurately modelled by short sequence lengths in RNA is if genotypes are broken up into modules, as predicted by Wagner and Altenberg (1996) and Ancel and Fontana (2000). A reduction in epistatic effects between different stacks in a tRNA shape for instance could lead to semi-independent optimisation of each particular region of the molecule across a map very close to that laid out in chapter 2. This kind of modularity might mean that there was increased ruggedness within the map, because optimisation of sub-units are restricted once a particular basic structure has been evolved early in a trajectory (Collins et al., 2007), leading to an increase in populations getting stuck in local optima. Its existence in the RNA genotype space would increase the direct applicability of models such as those presented here, which correspond to sub-units of the space.

Either way, modelling shorter sequence lengths does not result in a qualitative difference in the structure of the genotype space compared to models using longer sequences, and has the advantages of completeness and increased tractability.

6.1.2.2 Mutations

The second major assumption of the model is that the only mutational force is single point substitution. However, deletions, insertions and inversions are all well known as potential mutational pathways. From a methodological point of view it is possible to include these mutations into a model of a population evolving across the genotype space, but their inclusion changes one of the basic tenets of the connectedness of the space – That the only accessible genotypes are those that differ at just one position. Neutral networks could become more connected if new mutational pathways link previously disjunct networks together. In fact other models of RNA genotype space including 'base-pair' mutations, result in more connected networks than the one presented in chapter 2 (Grüner et al., 1996a; Göbel, 2000; Kospach, 2003, c.f. chapter 2). As we saw in section 2.1.2, this kind of increase of mutational pathways can increase the accessibility of adaptive mutants outside the boundary of the point-mutant neighbourhood. However, the large increase in possible mutants with distance means that there is a much lower chance of finding an adaptive mutant by a larger mutation (Maynard Smith, 1970). The result is that larger mutations are unlikely to replace drift across neutral networks as the primary method of accessing adaptive mutants at distance.

So although expanding the range of types of mutation included in the model will change its quantitative results, comparison with a model using an expanded mutational relationship (that of 'base–pair mutants') shows that the changes are only likely to affect the results in a quantitative manner.

6.1.2.3 Genotype–phenotype function

Finally I consider perhaps the most important set of assumptions within this model. Namely, those that revolve around the nature of the phenotype. Its simplicity and its hard and fast link with fitness could potentially lead to distortions in the way that we view the space.

Over the considerable literature using an RNA based map, no study has found that genotypes map to phenotypes in a way which does not result in the existence of neutral networks. However, this does not mean that no matter what function maps genotypes to phenotypes, neutral networks will always exist. Considering this potential for distortion, it is perhaps surprising that almost every study on the nature of the genotype–phenotype map in RNA has assumed that fitness is tied to secondary structure and little else (but see Collins et al. (2007) for an example using secondary structure and molecular stability, and Ancel and Fontana (2000) for an example using molecular plasticity of secondary structure). The reasons why are twofold. First, in such a highly dimensional and unintuitive space, anything more complicated becomes so difficult and computationally expensive to model and also so difficult to follow that to use a more complex relationship for the sake of realism could potentially obscure our understanding of the processes involved. Second, that secondary structure is a very obvious and natural function to map between genotype and phenotype in a way that not many other features are. It is grounded in biophysical reality, and it is therefore possible to avoid making any further assumptions about the nature of fitness, except that each phenotype codes for a distinct fitness. This leads us to the following question: Is using secondary structure alone a valid simplification?

The nature of the phenotype leads to an exceptionally degenerate mapping between genotype and phenotype. This in turn leads to large neutral networks, and thereby increases the potential for neutral drift to have adaptive consequences. In practice, secondary structure is unlikely to define fitness precisely. For example, within the ferritin mRNA regulatory region know as the Iron Responsive Element (IRE), the hair pin loop secondary structure has a conserved bulged base within the base-pair stack which plays an important role in binding affinity (Addess et al., 1997; Hall and Williams, 2004). In this case, it is clearly not just the structure of the

phenotype that matters, but also the particular bases at certain positions within the genotype.

Including an extra constraint such as this has the effect of increasing the total number of phenotypes and therefore decreasing the number of genotypes within each network. Furthermore it is now not just the interactions of bases at different positions which influence the secondary structure, but the interaction of bases with different structures which altogether influence fitness. The result is that there are more possible fitness values, and potentially an extra layer of epistatic interactions.

Even if the primary function of a phenotype remains simple, and maps directly to fitness, it is unlikely that all the genotypes will code for exactly the same fitness. For instance, in RNA the molecular stability of the secondary structure (Collins et al., 2007), its ability to fold into more than one shape (Ancel and Fontana, 2000), or its robustness to mutation (Wilke, 2001a) could each influence fitness and reduce the number of genotypes which map to the same fitness.

The more complicated the phenotype, and the more interactions between the different elements of phenotype which make up fitness, the more complicated the mapping between genotype, phenotype and fitness. As fewer sequences share a fitness value in common, we might expect neutral networks to become more and more disjunct within the genotype space. With this break up, the pattern of adaptive evolution through the space falls back towards the one–to–one mappings characterised by the models of Gillespie (1984) and Orr (2003).

So, is using such a simple phenotype justifiable given the effect may be to produce unrealistically large networks which connect the space to a higher degree than in most real genotype–phenotype maps?

In fact the results indicate that most of the neutral paths across a single network are limited to just a few steps. These short paths are likely to be relatively unaffected by the break up of larger networks. It is also worth considering that even if genotypes are not strictly neutral in relation to each other, many of the genotypic effects on fitness may actually be very small. If this is the case, then nearly neutral mutants might form networks in much the same way as strictly neutral ones.

6.1.2.4 Strict neutrality and near neutrality

After the initial controversy over the neutral theory, Ohta (1992) refined it by considering a finite population, within which a small selective pressure may not be enough to significantly alter the probability of a slightly deleterious allele drifting to fixation.

In a different approach, Eigen's quasispecies theory puts forward an error threshold, which is the level of selective advantage below which a population can no longer be maintained at a peak due to mutation pressure (Eigen, 1971). Van Nimwegen and Crutchfield calculated that due to a finite population size, a population was lost from a single adaptive peak genotype just above the error threshold predicted by Eigen, and was then free to drift across a slightly deleterious neutral network of fixed depth below the peak (van Nimwegen and Crutchfield, 2000). However, at present it is less clear what influence a slight 'slope' of fitness might have over the diameter of a nearly–neutral network (Fig. 6.1).

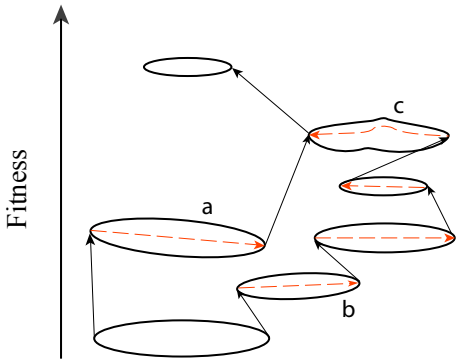


Figure 6.1: A visualisation of nearly neutral networks. Each network may no longer strictly neutral with relation to fitness. Networks can change slightly in fitness uniformly across a whole network (a and b), or can vary slightly in fitness between different regions of the network (c).

An indication into what happens over a gradual incline in fitness is given by Wroe et al.'s protein lattice model, which showed phenotypic transitions that were facilitated by the gradual increase in pleiotropic function for the new phenotype, while still crossing the nearly-neutral network where the old phenotype dominated function (Wroe et al., 2007). Whether a slight decline in fitness across a network would be navigable by a drifting population is still an open question.

Within the RNA genotype space framework, a future study on the effects of increasing phenotypic complexity and/or simulating near neutrality could involve any number of different options. Assigning different fitnesses to changes in the unbound bases mimicking those in the IRE hairpin would result in a discrete break up of the original structure based networks. Instead, including a notion of secondary structure stability could create a continuous fitness gradient on nearly–neutral networks that retained the same secondary structure. Yet another option is to include the propensity of a sequence to fold into a sub-optimal free energy structure. All these options

have the potential to decrease the size of individual networks and therefore increase the ruggedness of the genotype space.

With any increase in phenotypic complexity comes the risk of using speculative assumptions about the nature and magnitude of fitness affects, and thus defeating the aim of making the model more biologically plausible. A good way of studying nearly-neutral networks which does not use speculation on fitness criteria is by using a finite sized population, drifting across an irregularly shaped network as in chapter 5. This inherently introduces nearly-neutral topology. The different numbers of neutral neighbours affects the number of viable mutated offspring a particular genotype produces, and so presents a ready made model for testing the effects of slightly varying fitnesses across a network on the accessibility of adaptive mutants.

6.1.2.5 Summary of model assumptions

In summary, the model makes many assumptions and simplifications which increase its simplicity and tractability. Many of these necessary assumptions are likely to break down at least somewhat in more realistic circumstances, but the result is unlikely to be a complete reversal of the results presented in the preceding chapters, far more likely is that the space becomes more complicated, with many other factors, such as near neutrality or different types of mutation playing a role in dictating genetic acessibility. I therefore argue that the model provides a valuable insight into the the particular effect that neutral networks have within RNA, and that that effect is likely to be seen, if moderated, under more realistic conditions. In particular modelling evolution using a genotype-phenotype map can lead to a greater understanding of the evolutionary relationships between genotype and phenotype, even if neutral mutations play a less important role than they appear to here. I now consider whether neutral networks are seen more widely outside the evolution of RNA sequences.

6.1.3 Neutral networks: a widely encountered phenomenon?

Perhaps the strongest argument in favour of the existence of neutral networks within molecular biological systems comes from the presence of large amounts of genetic variation within naturally occurring populations, the observation of which led to Kimura and King and Jukes's neutral theory of molecular evolution (Kimura, 1968a; King and Jukes, 1969). To generate large amounts of variation, genotypic sequences must be (nearly) neutral with respect to each other. They must also be accessible, else that variation would never have arisen in the first place. As neutral networks can

be defined as a group of accessible neutral mutations, it is therefore likely that they exist at least in some form within the gentoype–phenotype map of most populations showing a degree of variability at the molecular level.

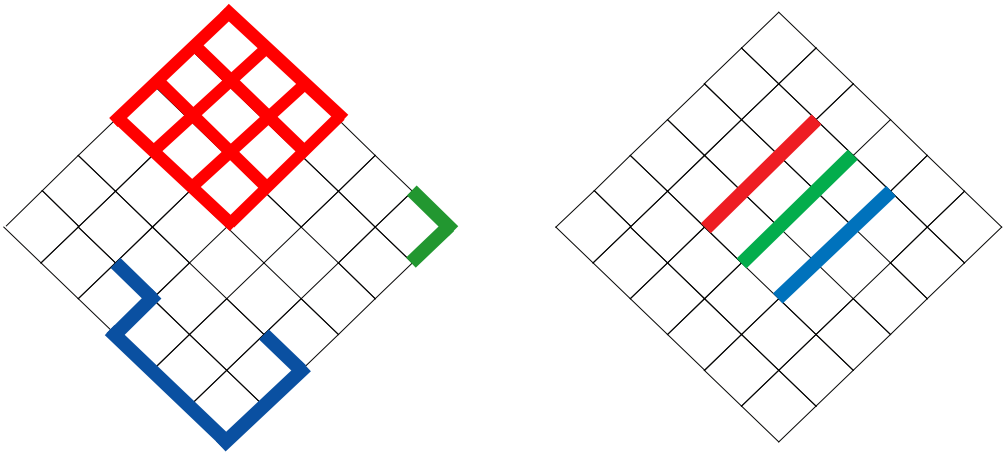
Protein lattice models of the genotype–phenotype map show very similar properties and characteristics to those of RNA secondary structure models, including the existence of pervasive neutral networks (Lipman and Wilbur, 1991; Govindarajan and Goldstein, 1997b; Bastolla et al., 1999; Bornberg-Bauer and Chan, 1999; Babajide et al., 2001; Aita et al., 2003; Wroe et al., 2005, 2007). In fact, neutral networks in proteins have the potential to be more neutral than in RNA, leading to the smoother landscapes seen by Wroe et al. (2005). Some of the degeneracy between protein genotype and phenotype lies in the degeneracy of the triplet codons mapping to amino acids. This means that unlike RNA, a mutation at a particular base can change the genotype without having any effect on the phenotypic expression of the protein. In addition to this highly degenerative relationship, some of the positions in most polypeptide chains can be extremely variable without having a significant effect on the phenotype, further increasing the degenerate nature of the map. While there is an extensive literature on the other factors that can affect the fitness of a *genotype*, e.g. bias in favour of particular codons due to selection for translational efficiency (Ikemura, 1985), selection for mutational robustness or anti-robustness (Plotkin et al., 2004; Archetti, 2006), or selection for translational accuracy (Stoletzki and Eyre-Walker, 2007), these factors are perhaps most likely to have a minor effect on fitness compared to changes in phenotype, creating nearly-neutral networks from which the probability of escaping is similar to the effects of network shape seen in chapter 5.

There is no reason to suppose that neutral networks would not be found on an even broader scale than at a molecular level. A recent model by Koelle et al. (2006) showed a good fit between the population dynamics observed in Human Influenza A (subtype H3N2) and a neutral network model of its evolution. Within the literature, there is speculation that neutral networks provide a parsimonious explanation for the phenotypic punctuations seen in the fossil record (Fontana and Schuster, 1998a; Crutchfield, 2002; Smith et al., 2003), where the genotype–phenotype map consists of the whole genome and its phenotypic expression, the organism.

6.1.3.1 The effect of neutral networks on adaptive evolution

While (nearly) neutral networks may well appear often in molecular biological systems and beyond, their existence alone is not enough to guarantee that they increase the accessibility of adaptive mutants. The nature of the accessibility between networks

is also crucial to whether neutral steps are ever used (Fig. 6.2). Calculating this is a far more difficult proposition than simply confirming the existence of networks.



(a) **(b)**

Figure 6.2: Two examples of genotype–phenotype maps where networks exist, but have no impact on the accessibility of adaptive mutations: **a)** Networks exist in the map, but the boundaries do not contact each other, making transitions very unlikely. **b)** Networks exist and contact, but each point is connected to the same set of neighbours, so if an adaptive step is possible it is immediately available.

Given the range of phenotypic mutations that are known to be available in genetic systems (e.g. Lenski and Travisano, 1994; Papadopoulos et al., 1999; Buckling et al., 2003; Burch and Chao, 1999, 2000; Alipaz et al., 2005; Rundle et al., 2006), neutral networks must allow phenotypic transitions at some level, even if they are only very rarely adaptive, networks for viable phenotypes are unlikely to all be completely isolated within genotype space (Fig. 6.2a)

The key factor concerning the influence of neutral networks is the existence of extensive epistatic interactions between different elements of the genotype. If epistasis is not common between genetic elements, adaptive accessibility is not increased by neutral networks, because a mutation always has the same effect regardless of its genotypic environment (Fig. 6.2b).

Consider again the fitness skyscraper analogy where neutral networks are the floors of the tower, and adaptive mutations the stairs between them (see section 1.2.2.1). When there are neutral networks but no epistasis, at each neutral step across any floor we find another set of identical stairs, each leading to exactly the same set of floors. In this case it doesn't matter which set of stairs you find first, or how many

neutral steps you take on a particular floor, the probability of reaching the penthouse does not change.

The epistasis required for neutral networks to have a significant impact can be shown at the molecular level, but is difficult to prove at any higher level (Whitlock et al. (1995), but see Burch and Chao (1999), Elena and Lenski (2001) and Poelwijk et al. (2007) for examples). This means that firm evidence that neutral networks increase adaptive mutational accessibility at the whole organism scale and therefore provides an explanation for the punctations seen in the fossil record remains a distant prospect.

So far in this discussion I have considered the assumptions made in relation to the underlying nature of the map. I shall now go on to consider some other factors which affect evolutionary dynamics, and assess the possible outcomes of including them within a genotype space model.

6.1.4 Population modelling considerations

Whether the population is modelled as a point or as a cloud of mutating individuals makes a large difference to the dynamics of evolutionary change. When modelled as a point, large population size makes genetic drift a very slow process (Wright, 1982; Lande, 1985), but if a large population is modelled as a cloud, then it can spread and diffuse across the space, leading to the conclusion that large population size decreases the time it takes to discover new mutants across the genetic landscape (Weinreich and Chao, 2005).

In fact, $n\mu N$ is the number of new mutants produced in each generation (where n is sequence length, μ is mutation rate per position per generation and N is the population size), and so has an important influence on the rate at which evolution proceeds. Though $n\mu N$ can remain constant, the amount of random drift due to having a high mutation rate and small population size, or vice versa can have a complex effect on the average time it takes to cross an 'entropic barrier', and can interact with the size and shape of neutral networks in a complex way.

The first steps in unravelling these interactions were taken under simplified conditions in chapter 5. However, from the initial analysis of drift across realistic networks it is still unclear under what circumstances selection for mutational robustness combined with the difficulty of reaching the fringes of a network makes a long path become unlikely. It is also unclear under what conditions the population starts to drift around at random – potentially drifting further than the closest portal, because no lineages went in the right direction.

To provide further insight into the effect that more complex shapes have on the patterns of drift–based evolution, a model must be found which expresses the complexity of the shape in a way which can be used to explain the results rather than just observe them.

A further consideration when a population cloud reaches a selective portal, is that the strength of the selective pressure is unlikely to be so strong that the shift in genotypes is instantaneous, as was assumed in chapter 4. This means that there is a non-zero chance of some members drifting past the nearest adaptive portal to a more adaptive or more frequently encountered portal further away. Modelling the chance of different or multiple portals being found requires estimates of the absolute fitness values to calculate the selective advantage. For this reason it was not considered in this thesis. However, how often the interplay between selective advantage, number and distance of portals, and population drift results in a more distant portal being used by a population is an interesting question, and could be modelled over a set of customised toy networks similar to those used in chapter 5, as well as directly within the RNA genotype space. Under these kind of conditions selection for genotypes on the far side of different portals can lead to population divergence, especially if the portals lead to disjunct networks of the same fitness.

In chapter 4, convergence and subsequent divergence of lineages was seen within the RNA genotype space (section 4.5). The number of divergent paths taken by different lineages which had started from the same sequence indicates that having the choice of two or more identically fit adaptive options at a particular stage is not uncommon. When these options lead to disjunct networks with the same phenotype, the subsequent population divergence can be swift and drastic on both a genotypic and phenotypic scale. That there is no initial difference in the fitnesses of the divergent populations implies that both have a good chance of initial survival, and perhaps proves an excellent way for speciation to occur (Gavrilets and Gravner, 1997).

6.1.5 The effects of environmental change

Whether the environment is more static or dynamic in nature has been a point of contention dating back to Fisher and Wright (reviewed in Skipper, 2002). However, in many cases within genotype–phenotype map models the environment is assumed to be static. The level of environmental change certainly has the potential to have a defining effect on the trajectory of any evolving population (Collins et al., 2007).

Holder and Bull (2001), Imhof and Schlotterer (2001) and Elena and Lenski (2003) have all shown empirical evidence that when a population has been subject to a

sudden environmental shift, mutations with large effect are initially favoured, and as time passes, any further adaptive change is likely to have a smaller affect on fitness.

Within the RNA genotype space, we saw in chapter 4 that mutations do follow the pattern of initially large adaptive steps. We also saw that the reverse is true for neutral pathways. The effect that the genotype–phenotype map has on evolution is then totally dependent on the level of environmental change. With a high frequency of large sudden changes, any population is frequently likely to have local access to mutations that are highly adaptive and increase in accessibility due to neutral network structure is rarely required or seen.

The rate of an environmental shift also has as much potential to affect a population's subsequent final fitness as the magnitude. If environmental change is continuous rather than stochastic, a population will equilibrate at a particular point on the curve of the size of fitness change against time (Fig. 4.10). If the environmental change initially degrades fitness faster than mutation and selection can increase it, then fitness will decrease over time. At lower fitness, the size of the adaptive benefit increases, eventually stabilising the population at a balanced point where the slope of benefit per adaptive change matches the rate of negative environmental change. If environmental change is too large, then selection cannot keep up, and a population's fitness becomes uncorrelated with that of the landscape. At very low rates of change, a population will remain very close to the peak, and is likely to undergo a large amount of drift for each fitness increase.

Finally, environmental stochasticity or heterogeneity, including the effects of frequency dependent selection, can potentially act to reverse some of the near–neutrality brought about by the factors discussed in section 6.1.2.4.

When the fitness criterion is not constant over time and/or space, exact fitness definitions can become blurred or overlap, so that different individual genotypes or phenotypes can be equally fit within the larger environment. More complex developmental pathways involving phenotypic plasticity (e.g. West-Eberhard, 2003) or individuals matching their phenotype with the particular section of a heterogenous environment that maximises their fitness (e.g. Todd et al., 2006), are just two possible ways in which nearly neutral fitnesses could be re-neutralised.

6.1.6 Time scale and punctuations

It is extremely difficult to predict absolute time scales from models where the initial parameters have a very large quantitative affect on the outcome, and where a number

of simplifying assumptions have been made. However, it is possible to make some general statements about the pattern of time scales involved.

In chapter 5, drift across a simple model neutral network took just a few generations to drift one or two neutral steps. If we consider Wahl and Krakauer's calculation that a virus can produce thousands of copies of each of its local neighbourhood in a single generation, it seems likely that many small, rapidly reproducing organisms would find traversing short distances extremely easy, and indeed practically instantaneous on an evolutionary time scale (Wahl and Krakauer, 2000). As population size and mutation rate correlate inversely with body size, we might expect larger organisms to require more generations to perform the same kind of drift based search. However, the relatively small number of generations required means that even with a reasonable amount of environmental change, short neutral paths might well feature in evolutionary trajectories when required.

Because an adaptive mutant will be selected once found, the steps immediately preceding it may never build up to a detectable frequency in the population before being replaced by an adaptive mutant, and so may appear to have never occurred. When the total neutral path consists of a small number of steps, this may mean that the neutral intermediates are extremely difficult to observe empirically within a population.

As the average number of generations taken to find a portal rises quadratically with inter-portal distance, we can see from figure 5.5, that the time scale for a neutral path of approximately 4 or more neutral steps is in the order of magnitude of the phenotypic punctuations seen in the fossil record (Gould and Eldredge, 1977) and empirically in *E. coli* (Elena et al., 1996). Fontana and Schuster (1998a), Ebner et al. (2001), Crutchfield (2002), Smith et al. (2003) and Wolf et al. (2006) have all postulated that this phenomenon can be explained by neutral drift across a network increasing the accessibility of a adaptive mutant, whose sudden appearance causes the phenotypic shift. Section 5.4 has cast doubt on the accessibility of the longest paths recorded within the space, so it is reassuring for the proponents of neutral drift hypothesis of phenotypic punctuations that intermediate length neutral paths require a number of generations in the right order of magnitude to appear punctuated.

Thus shorter neutral paths present little barrier to evolution, and could occur with a relatively high frequency, but be extremely difficult to spot. Longer paths on the other hand fit the empirical data on phenotypic punctuated transitions.

6.1.7 Recombination and ploidy

All the work in this thesis has been with haploid asexual organisms. As with including more complicated mutations, recombination changes the number and arrangement of the neighbours which are accessible from any one point. There are three regimes under which recombination might occur between two sequences. Within a single neutral network, across two neutral networks of the same fitness, or across two networks of unequal fitness where phenotypic transitions are not instantaneous.

Within a single network, the effect of recombination depends to a large extent on the structure of the underlying genotype space. If the fitness of a particular point mutation is conditional on other positions within the sequence (e.g. the *string* network from chapter 5), then recombination between sequences does not increase the speed at which a population can cross the network. This is because each advance along the network only occurs via a new mutation, not currently present in the population. However, recombination can allow exploration of different phenotypes within a network neighbourhood larger than the point-mutant neighbourhood around the boundary of the network (Fig. 6.3a).

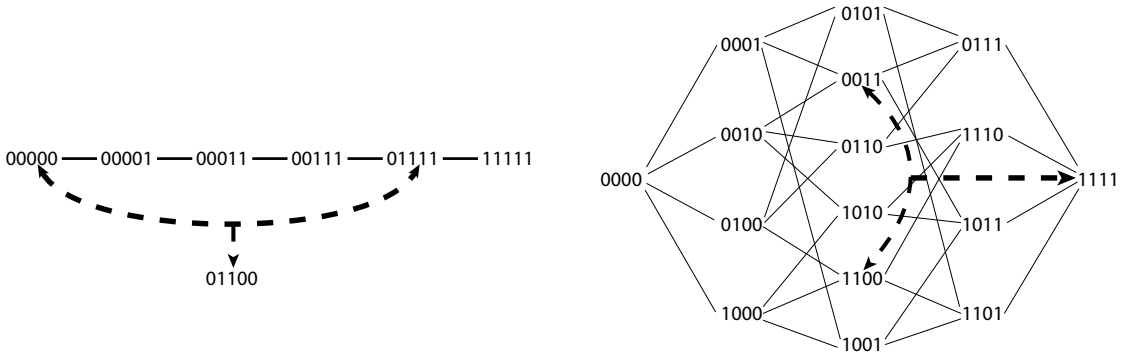


Figure 6.3: a) Recombination in the string network. Because each new step on the network has to arise from a point mutation from a neighbouring genotype there is no possibility of recombination generating novel genotypes within the network. Novel genotypes can lie outside the point-mutant network neighbourhood however. **b)** Recombination in the lattice network allows generation of novel mutants within the network which can be more than a single point mutation from any other genotype so far created. This could allow a population to cross the network in fewer generations than by mutation alone.

By contrast, if the fitness of a particular point mutation is never deleterious within a particular network (e.g. the *lattice* network from chapter 5), then recombination of the sequences allows exploration across the network to occur more quickly, because different neutral genotypes can recombine to incorporate more than one point mutation

in a single step (Fig.6.3b). As network diameter increases, the effects of recombination are likely to become more obvious, with recombination capable of making large leaps forward across the network in a single event. Because all of the combinations are part of the network, any recombination will never explore sequences outside the network itself.

Recombination between two neutral networks of the same fitness will still be influenced by the shape of each, but also depends on the way the networks are arranged with respect to each other in the space. This makes predicting the outcomes virtually impossible, though the affect would be observable in a future study within the genotype space model.

Recombination between sequences on either side of a portal might deliver the most important impact of all. If a population has spent many generations drifting across a large neutral network before finding an accessible portal, there is likely to be a large amount of genetic variation in the population. However, this will be lost in a severe genetic bottleneck if just one or a few individuals discover the portal and recombination does not occur. Recombination between individuals on either side of the portal could allow an evolving population to re-diversify into the new network, even while the individuals on the old network are dying out. This would allow much faster exploration of a new network, or even test new more distant genotypes irrespective of the shape of the network.

The effects of recombination are likely to be highly dependent on the balance between mutation rate and recombination rate (Cui et al., 2002; Xia and Levitt, 2002). However, one possibility if recombination is relatively high, is that string-type networks become even less easy to traverse. Recombination of rarer sequences which have drifted towards the boundary of a network will tend return them towards the centre of the network. Again, the RNA genotype–phenotype space provides an ideal testing ground for this hypothesis.

Mapping diploid combinations of all the haploid sequences at this stage is computationally unfeasible at the lengths tested here. There is also no obvious natural mapping from diploid genotypes to phenotypes. However, the potential to reduce selective pressure in a diploid organism by retaining one functional allele could increase the chance that a lineage crosses a fitness valley using the other allele, which in turn could reduce the impact of neutral networks.

6.1.8 Evolution of the genetic code

Much interest and effort has been focused on establishing whether the genetic code as we see it today is a frozen accident or whether it is itself the product of evolution (Woese (1965), Crick (1968), Jukes (1973), reviewed by Knight et al. (1999) and subsequent work by Knight et al. (2001), Berger (2003), Archetti (2006) and Zhu and Freeland (2006)). The number of bases which make up DNA and RNA (Szathmry, 2003), and how they code for amino acids are both factors on which selection could have acted, with most research focusing on the latter. For example the existence of codon biases indicates that some kind of selective pressure is likely to be being currently exerted on the genetic code though what that pressure is is still debated (Ikemura, 1985; Plotkin et al., 2004; Archetti, 2006; Stoletzki and Eyre-Walker, 2007).

There has also been substantial discussion on whether evolvability – loosely defined as 'evolutionary potential' is also under selective pressure (Kirschner and Gerhart, 1998; Ebner et al., 2001; Knight et al., 2001; Carter et al., 2005; Pigliucci, 2008). In this thesis I have shown that degeneracy in the genetic code increases the evolvability of a population by increasing the chance of future genetic mutations being adaptive, even if no adaptive changes are currently accessible. This begs the question: was a degenerate code selected for because of the increase in evolvability that comes with it? Any advantage that degenerate code eventually brings by increasing the chances of future fitter mutants is unlikely to be selected for, when there is likely to be more direct factors such as efficiency of translation or error minimisation. Those directly selected codes would be likely to out-compete a degenerate one, even if drift within a degenerate encoding were more likely to achieve a higher fitness in the longer term.

Direct selective advantages of degenerate code could be favoured by more direct selective pressures such as selection for genetic robustness (van Nimwegen et al., 1999; Wilke, 2001b; Zhu and Freeland, 2006; Bloom et al., 2007), or as an adaptation to a rapidly changing environment or stochastic environment, where a degree of genetic degeneracy allows the existence of many different genotypes (standing variation), and hence confers an advantage by giving a head start to a population responding to a shifted fitness optimum (Earl and Deem, 2004). In the balance of probability it is unlikely that we will ever be certain exactly why the genetic code is degenerate, but the fact that it is means that the degeneracy and its consequences should be included in models of evolution including those apparently restricted to only considering adaptive change.

6.2 Summary

In this thesis I have presented a biologically based model of RNA genotypes and the secondary structure phenotypes that they code for. Within this framework I have shown that the underlying relationship between genotype and phenotype can play a significant role in determining the accessibility of adaptive mutants. In particular, neutral mutations play a special role in this degenerative map, because they have the potential to substantially increase the accessibility of future adaptive mutants, as has been suggested by previous work (e.g. Lipman and Wilbur, 1991; Schuster et al., 1994; Fontana and Schuster, 1998a; van Nimwegen and Crutchfield, 2000; Ebner et al., 2001; Smith et al., 2003; Aita et al., 2003; Gavrilets, 2003; Wagner, 2005; Koelle et al., 2006; Wroe et al., 2007; Sumedha et al., 2007b),

I have argued in this discussion that a model involving the use of short genotypic sequences provide a degree of precision about the phenomena we observe which is not possible when modelling a sub-section of a larger space, but still remains a valid model of larger spaces. Previous similar exhaustive models have emphasised the geometric or topological properties of the space (Grüner et al., 1996a,b; Göbel, 2000; Kospach, 2003). In this work, the emphasis is placed more firmly on the impact those geometric and topological properties have on a population evolving through the space. This exhaustive map framework is open to extension for higher level investigations of evolutionary phenomena (as it was used in Chapter 4), and provides a valuable alternative perspective to models based around allele frequencies, by implicitly considering the epistatic interactions between loci. The main results of the thesis are summarised here, but the reader is referred to the chapter summaries and discussions for a more in-depth consideration of each set of results.

Neutral networks

- Large neutral networks of sequences form in the genotype space. Every sequence in a particular network has the same phenotype, and shares $(n-1)$ bases in common with at least one other sequence in the network (where n is sequence length) (Chap. 2).

- These networks vary widely in their shape and size, as well as the way in which they are connected externally to other networks in the space (Chap. 2).

- Some networks have such a convoluted shape, that the number of point mutation changes between two sequences when each step must remain on the network,

can be significantly longer than the direct number of point mutation changes between them (the Hamming distance) (Chap. 3).

Local accessibility of adaptive neighbours

- The local neighbourhood of any sequence (the $3n$ point mutation neighbours) always contains sequences coding for other phenotypes i.e. no network has an interior (Chap. 2).
- When a genotype codes for a low fitness phenotype, the local neighbourhood is diverse enough that there is usually a directly accessible adaptive mutant (Chap. 4).

Neutral network accessibility of adaptive neighbours

- In an adaptive walk traced across the genotype space and restricted to only taking adaptive or neutral steps, the majority of walks started at random will find the global optimum (Chap. 4).
- The trajectory of these walks usually involves a mixture of adaptive and neutral steps (Chap. 4).
- When phenotypic fitnesses are uncorrelated, on average every neutral step requires an adaptive one (Chap. 4).
- The more well adapted a genotype is, the more neutral steps are required to access a further adaptive change (Chap. 4).
- The number of steps in a path is often longer than the Hamming distance between the start and end sequences (Chap. 4).
- The number of adaptive *and* the number of neutral steps in any given path increase with sequence length (Chap. 4).

Network restrictions on accessibility

- Network size and shape strongly influences the accessibility of adaptive mutants when neutral steps are required. The more paths there are towards the adaptive mutant, the more accessible that mutant (Chap. 5).

- Some networks are so asymmetric that drift can occur between two sequences in one direction, but not in the other – The network analogue of a check-valve or diode (Chap. 5).
- The longest paths that are possible in the space may often be inaccessible, because they lie on the wrong side of a check-valve type network shape (Chap. 5).

6.3 Concluding remarks

It is likely that the stochastic processes and complex interactions within biological evolution mean that it will be impossible to completely explain every facet of it. However, many of the salient points can be picked out and elucidated using a variety of models. Using a genotype–phenotype map model presents a different way of looking at various evolutionary phenomena compared to a more traditional allele frequency approach, in particular by allowing the inclusion of complex epistatic interactions. This interactions have the potential to have an important role in adaptive evolution, particularly when epistasis causes a change in the sign of the interaction, causing a collection of independently neutral mutations to become advantageous.

The results presented in the preceding chapters give a tantalising glimpse of what is possible with this framework, though they have generated as many questions as answers – providing many avenues for further study. Future models testing the assumptions of this work, in particular the large and strict nature of the neutral networks may well result in the moderation of some of the results, particularly those involving walks across the entire genotype space. However, I assert that those assumptions were all valid and necessary ones to enable the progress that has been made, and that they do not in fact overlay a significant artefact on the fundamental findings that *degeneracy does increase the accessibility of adaptive mutants*.

Within this particular genotype–phenotype map, the results indicate that relatively few generations are required for a population to assimilate neutral mutations into an evolutionary trajectory, and that in turn, those neutral mutations are necessary for onward evolution. This means that on occasion the only way to reach the peak of an adaptive landscape is via a long and winding road, though not necessarily going up-hill all the way.

Glossary

dot–bracket notation A system of coding a phenotypic secondary structure, where periods code for unbound bases, and parentheses for base pairs. 39, 41

entropic barrier The term used by van Nimwegen and Crutchfield (2000) to suggest the barrier to adaptive change caused by having to drift across a network of neutral intermediates. 123, 162

entry portal A portal genotype which is the first genotype found in a new network. It must have at least one viable local neighbour from which a mutation could have occurred. 70, 73, 75

epistasis The property of genetic interactions where the fitness outcome is different from that predicted by the sum fitness contributions of the constituent parts. 18, 21, 151, 161

exit portal A portal genotype which is the last genotype in a network, with at least one viable local neighbour with a different phenotype. 73, 78

genotype space The set of sequences made from considering every combination of bases at every position in the sequence i.e. For RNA, 4^n sequences, where n is sequence length. 25, 77–79, 151–155, 157, 158, 162–164, 166, 167, 169–171

genotype–phenotype map The genotype–phenotype map defines a discrete set of genotypes made up of unique genetic sequences, and the phenotypes that each sequence codes for by means of a mapping between genotype and phenotype. 13, 14, 17, 22, 23, 25, 28, 30, 32, 33, 63, 67, 151–153, 156, 160, 164, 171

greedy fitness algorithm A fitness algorithm in which the population always jumps to the highest ranking mutant at each step in an adaptive walk. 83, 85, 89, 102, 104, 115, 118

Hamming distance The number of character changes between two strings of equal length. 10, 22, 40, 60, 68, 69, 71, 75, 76, 79, 84

indirectly connected Genotypes are indirectly connected if they are more than one simple mutation away from each other, but there is an unbroken chain of simple mutants coding for either of their phenotypes between them. 38

inter-portal distance The distance between two portal genotypes across a single neutral network. NOTE: In chapter 5, this refers to the distance between the entry portal and the exit genotype on the far side of a portal, one step further than the earlier definition. 69, 71, 73, 77, 78, 130–132, 136–138, 140

lattice A network shape where all the combinations of bases at variable positions result in genotypes which code for the same phenotype. 124–127, 129, 130, 134–138, 140, 142, 145, 146, 166

local neighbourhood The set of genotypes that are accessible from a single genotype. When accessibility is limited to single point–mutations in RNA the local neighbourhood consists of $3n$ sequences.. 8, 9, 13, 14, 17, 38, 68, 69, 96, 144

Local Neighbourhood adaptive walk An adaptive walk method where each step must be advantageous and locally accessible. When there is more than one choice, the fittest adaptive neighbour is chosen. 87, 93, 96, 107

network density The density of a network is calculated by dividing the number of sequences found in the network, by potential total number of sequences, calculated by assessing the variability of bases at each position in the sequence. 59, 60

network diameter The maximum distance across a network. 132, 134, 135

Network Mapping adaptive walk An adaptive walk method where the most adaptive network neighbour is taken at each step. 87, 96

network neighbourhood The set of genotypes which are accessible from at least one genotype within the neutral network whose neighbourhood this refers to. 38, 69, 96

neutral network A set of genotypes each of which coding for the same phenotype and sharing the same code at $(n-1)$ positions in its sequence, with at least one other genotype in the network. 4, 32, 33, 45, 47, 68

Neutral Step adaptive walk An adaptive walk method where the fittest local neighbour is chosen. If one is not available a breadth–first search is performed to find the closest adaptive genotype. 87, 88, 91, 95, 99, 100, 106, 108, 114, 115

NID An integer code given to each unique network standing for Network Identity. 76

PID An integer code given to each unique phenotype to provide a more parsimonious way of tracking phenotypes than dot-bracket notation. 41

PID0 The phenotype identity number given to the open structure, in which no base pairs formed. 41, 42, 45, 47, 48, 54

portal A simple link between two genotypes coding for different phenotypes. 67

portal genotype A genotype which includes at least one genotype coding for a viable phenotype within its local neighbourhood. 68

quasi-species A theory first proposed by Eigen (1971), whereby selection acts on a population of closely related individuals, the 'quasi-species', rather than individuals. 10, 11, 123

robustness The ability of a particular genotype to undergo point mutations which do not change the phenotype. One can also consider an area of a network where the neutral connectedness between point mutant neighbours is high to be robust. 123, 129, 134, 135, 137, 142, 144, 147, 157, 160, 162

shape space covering The property of a genotype space where every genotype coding for a common phenotype requires a relatively small number of changes to its sequence to becoming a genotype coding for any other common phenotype. 13, 14, 84, 97

sign epistasis Epistatic interactions which differ in the direction of their effect on fitness than the sum of their constituent parts. 17

string A shape of network where each genotype has a maximum of two neutral neighbours and the phenotypic result of a mutation at a particular position is dependent on the bases present at other variable positions. 124, 126, 130, 134, 135, 137–140, 142, 166

References

Christoph Adami. On modelling life. *Artificial Life*, 1:429–438, 1995.

Kenneth J. Addess, James P. Basilion, Richard D. Klausner, Tracey A. Rouault, and Arthur Pardi. Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins. *J Mol Biol*, 274(1):72–83, 1997.

Takuyo Aita, Motonori Ota, and Yuzuru Husimi. An in silico exploration of the neutral network in protein sequence space. *J Theor Biol*, 221(4):599–613, 2003. doi: 10.1006/jtbi.2003.3209.

Julie A Alipaz, Shu Fang, Naoki Osada, and Chung-I. Wu. Evolution of sexual isolation during secondary contact: genotypic versus phenotypic changes in laboratory populations. *Am Nat*, 165(4):420–428, Apr 2005. doi: 10.1086/428388.

Charles R. Allerson, Alan Martinez, Emine Yikilmaz, and Tracey A. Rouault. A high-capacity RNA affinity column for the purification of human IRP1 and IRP2 overexpressed in *pichia pastoris*. *RNA*, 9(3):364–374, 2003. URL `http://www.rnajournal.org/cgi/content/abstract/9/3/364`.

Lauren W. Ancel and Walter Fontana. Plasticity, evolvability, and modularity in RNA. *J Exp Zool*, 288(3):242–283, Oct 2000.

Paul E. Anderson and Henrik J. Jensen. Network properties, species abundance and evolution in a model of evolutionary ecology. *J Theor Biol*, 232(4):551–558, 2005. doi: 10.1016/j.jtbi.2004.03.029.

Jean-Baptiste Andr and Bernard Godelle. The evolution of mutation rate in finite asexual populations. *Genetics*, 172(1):611–626, Jan 2006. doi: 10.1534/genetics.105.046680.

Marco Archetti. Genetic robustness and selection at the protein level for synonymous codons. *J Evol Biol*, 19(2):353–365, Mar 2006. doi: 10.1111/j.1420-9101.2005.01029.x.

Aderonke Babajide, Robert Farber, Ivo L. Hofacker, Jeff Inman, Alan S. Lapedes, and Peter F. Stadler. Exploring protein sequence space using knowledge-based potentials. *J Theor Biol*, 212(1):35–46, 2001. doi: 10.1006/jtbi.2001.2343.

Francois Bardou and Luc Jaeger. Large phenotype jumps in biomolecular evolution. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(3 Pt 1):031908, 2004. doi: 10.1103/PhysRevE.69.031908.

Lionel Barnett. Ruggedness and neutrality - the NKp family of fitness landscapes. In C. Adami, R. K. Belew, H. Kitano, and C. Taylor, editors, *Alife VI, Proceedings of the Sixth International Conference on Artificial Life*, pages 18–27. MIT press, 1998.

Rowan D H Barrett, Leithen K M'gonigle, and Sarah P Otto. The distribution of beneficial mutant effects under strong selection. *Genetics*, 174(4):2071–2079, Dec 2006. doi: 10.1534/genetics.106.062406.

Nick Barton and Willem Zuidema. Evolution: the erratic path towards complexity. *Curr Biol*, 13(16):R649–R651, Aug 2003.

Ugo Bastolla, Markus H. Eduardo Roman, and Michele H. Vendruscolo. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J Theor Biol*, 200(1):49–64., 1999. doi: 10.1006/jtbi.1999.0975.

Mark A Bedau and Norman H Packard. Evolution of evolvability via adaptation of mutation rates. *Biosystems*, 69(2-3):143–162, May 2003. doi: doi:10.1016/S0303-2647(02)00137-5.

Gerard Berger. Deterministic hypotheses on the origin of life and of its reproduction. *Medical Hypotheses*, 61(5-6 SU -):586–592, 2003. doi: doi:10.1016/S0306-9877(03)00237-8.

Jesse Bloom, Zhongyi Lu, David Chen, Alpan Raval, Ophelia Venturelli, and Frances Arnold. Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol*, 5(1):29, Jul 2007. doi: 10.1186/1741-7007-5-29.

Erich Bornberg-Bauer. Structure formation of biopolymers is complex, their evolution may be simple. In *Pac Symp Biocomput 1996*, pages 97–108, 1996.

Erich Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophys J*, 73(5):2393–2403., 1997.

Erich Bornberg-Bauer and Hue Sun Chan. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A*, 96(19):10689–10694, 1999.

Angus Buckling, Matthew A. Wills, and Nick Colegrave. Adaptation limits diversification of experimental bacterial populations. *Science*, 302(5653):2107–2109, Dec 2003. doi: 10.1126/science.1088848.

Christina L. Burch and L. Chao. Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics*, 151(3):921–927, Mar 1999. URL `http://www.genetics.org/cgi/content/full/151/3/921`.

Christina. L. Burch and Lin Chao. Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature*, 406(6796):625–628, Aug 2000. doi: 10.1038/35020564.

Sean B. Carroll. Chance and necessity: the evolution of morphological complexity and diversity. *Nature*, 409(6823):1102–1109, 2001. doi: 10.1038/35059227.

Sean B. Carroll, Jennifer K. Grenier, and Scott D. Weatherbee. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Blackwell, second edition, 2004.

Ashley J R Carter, Joachim Hermisson, and Thomas F Hansen. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theor Popul Biol*, 68(3):179–196, Nov 2005. doi: 10.1016/j.tpb.2005.05.002. URL `http://dx.doi.org/10.1016/j.tpb.2005.05.002`.

Sinad Collins, Juliette de Meaux, and Claudia Acquisti. Adaptive walks toward a moving optimum. *Genetics*, 176(2):1089–1099, Jun 2007. doi: 10.1534/genetics.107.072926.

Iñaki Comas, Andrés Moya, and Fernando González-Candelas. Validating viral quasispecies with digital organisms: a re-examination of the critical mutation rate. *BMC Evol Biol*, 5(1):5, Jan 2005. doi: 10.1186/1471-2148-5-5.

Michael Conrad. The geometry of evolution. *Biosystems*, 24(1):61–81, 1990.

Michael Conrad, Carl Friedlander, and Morris Goodman. Evidence that natural selection acts on silent mutation. *Biosystems*, 16(2):101–111, 1983. doi: 10.1016/0303-2647(83)90031-X.

Matthew C Cowperthwaite, James J. Bull, and Lauren Ancel Meyers. Distributions of beneficial fitness effects in RNA. *Genetics*, 170(4):1449–1457, Aug 2005. doi: 10.1534/genetics.104.039248.

Jerry A. Coyne, Nicholas H. Barton, and Michael Turelli. Perspective: A critique of sewall wright's shifting balance theory of evolution. *Evolution*, 51(3):643–671, jun 1997. ISSN 0014-3820. URL `http://www.jstor.org/stable/2411143`.

Francis H. Crick. The origin of the genetic code. *J Mol Biol*, 38(3):367–379, Dec 1968.

James P. Crutchfield. When evolution is revolution:origins of innovation. In James P. Crutchfield and Peter Schuster, editors, *Evolutionary Dynamics:Exploring the Interplay of Selection, Neutrality, Accident, and Function*, Santa Fe Institute Series in the Science of Complexity. Oxford University Press, New York, 2002.

James P. Crutchfield and Erik van Nimwegen. The evolutionary unfolding of complexity. In L. F. Landweber, E. Winfree, R. Lipton, and S. Freeland, editors, *Evolution as Computation, Lecture Notes in Computer Science*, New York, 1999. Springer-Verlag.

Yan Cui, Wing H. Wong, Erich Bornberg-Bauer, and Hue S. Chan. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci U S A*, 99(2):809–814, 2002. URL `http://www.pnas.org/cgi/content/abstract/99/2/809`.

Jan Cupal, Stephan Kopp, and Peter F. Stadler. RNA shape space topology. *Artif Life*, 6(1):3–23, 2000.

Charles Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, 1859.

Richard Dawkins. *The Selfish Gene.* Oxford University Press, Oxford, UK., 1976.

Richard Dawkins. *The Blind Watchmaker*. Harlow : Longman Scientific & Technical, 1986.

Eric J. Deeds, Nikolay V. Dokholyan, and Eugene I. Shakhnovich. Protein evolution within a structural space. *Biophys J*, 85(5):2962–2972, 2003. URL `http://www.biophysj.org/cgi/content/abstract/85/5/2962`.

Bernard Derrida and Luca Peliti. Evolution in a flat fitness landscape. *Bulletin of Mathematical Biology*, 53:355–382, 1991.

David J Earl and Michael W Deem. Evolvability is a selectable trait. *Proc Natl Acad Sci U S A*, 101(32):11531–11536, Aug 2004. doi: 10.1073/pnas.0404656101.

Marc Ebner, Mark Shackleton, and Rob Shipman. How neutral networks influence evolvability. *Complexity*, 7(2):19–33, 2001.

Jeffrey A Edlund and Christoph Adami. Evolution of robustness in digital organisms. *Artif Life*, 10(2):167–179, 2004. doi: 10.1162/106454604773563595.

Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971. doi: 10.1007/BF00623322.

Manfred Eigen and Peter Schuster. The hypercycle. a principle of natural self-organization. part a: Emergence of the hypercycle. *Naturwissenschaften*, 64(11): 541–565, 1977.

Niles Eldredge and Stephen J. Gould. Punctuated equilibria: an alternative to phyletic gradualism. In T. J. M. Schopf, editor, *Models in Paleobiology*, pages 82–115. Freeman, Cooper and Company, San Francisco, 1972.

Santiago F. Elena and Richard E. Lenski. Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, 390(6658):395–398, Nov 1997. doi: 10.1038/37108.

Santiago F. Elena and Richard E. Lenski. Epistasis between new mutations and genetic background and a test of genetic canalization. *Evolution Int J Org Evolution*, 55(9):1746–1752, Sep 2001. doi: doi:10.1111/j.0014-3820.2001.tb00824.x.

Santiago F. Elena and Richard E. Lenski. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet*, 4(6):457–469, Jun 2003. doi: 10.1038/nrg1088.

Santiago F. Elena, Vaughan S. Cooper, and Richard E. Lenski. Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272(5269):1802–1804, 1996. doi: 10.1126/science.272.5269.1802.

Santiago F. Elena, Claus O. Wilke, Charles Ofria, and Richard E. Lenski. Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution Int J Org Evolution*, 61(3):666–674, Mar 2007. doi: 10.1111/j.1558-5646.2007.00064.x.

Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

Ronald A Fisher. *The genetical theory of natural selection*. Clarendon Press, Oxford, 1930.

Walter Fontana. Modelling 'evo-devo' with RNA. *Bioessays*, 24(12):1164–1177, Dec 2002. doi: 10.1002/bies.10190.

Walter Fontana and Peter Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280(5368):1451–1455, 1998a. doi: 10.1126/science.280.5368.1451.

Walter Fontana and Peter Schuster. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *Journal of Theoretical Biology*, 194(4):491–515, 1998b.

Robert Forster, Christoph Adami, and Claus O. Wilke. Selection for mutational robustness in finite populations. *J Theor Biol*, 243(2):181–190, Nov 2006. doi: 10.1016/j.jtbi.2006.06.020.

Sergey Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology & Evolution*, 12(8):307–312, August 1997. URL `http://www.sciencedirect.com/science/article/B6VJ1-3X2B5DX-68/2/6634c84b639e201b7e81e3b4eb02d0bf`.

Sergey Gavrilets. Perspective: models of speciation: what have we learned in 40 years? *Evolution Int J Org Evolution*, 57(10):2197–2215, Oct 2003. URL `http://www.jstor.org/stable/3448772`.

Sergey. Gavrilets and Janko Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *J Theor Biol*, 184(1):51–64, Jan 1997. doi: 10.1006/jtbi.1996.0242.

Nicholas Geard, Janet Wiles, Jennifer Hallinan, Bradley Tonkes, and Ben Skellett. A comparison of neutral landscapes -NK, NKp and NKq. In D. B. Fogel, M. A. El-Sharkawi, X. Yao, G. Greenwood, H. Iba, P. Marrow, and M. Shackleton, editors, *Proceedings of the Congress of Evolutionary Computation (CEC2002)*, pages 205–210, Honolulu, Hawaii, 2002.

John H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, 38 (5):1116–1129, Sep. 1984. ISSN 00143820. URL `http://www.jstor.org/stable/2408444`.

John H. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, Oxford, 1991.

Ulrike Göbel. *Neutral Networks of Minimum Free Energy RNA Secondary Structures*. Ph.d. thesis, University of Vienna, 2000.

Stephen J. Gould and Niles Eldredge. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, 3:115–151, 1977.

Stephen J. Gould and Richard C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci*, 205(1161):581–598, 1979.

Sridhar Govindarajan and Richard A. Goldstein. The foldability landscape of model proteins. *Biopolymers*, 42(4):427–438, 1997a. doi: 10.1002/(SICI)1097-0282(19971005)42:4⟨427::AID-BIP6⟩3.0.CO;2-S.

Sridhar Govindarajan and Richard A. Goldstein. Evolution of model proteins on a foldability landscape. *Proteins*, 29(4):461–466, 1997b.

Alan Grafen. On the uses of data on lifetime reproductive success. In T. H. Clutton-Brock, editor, *Reproductive success : studies of individual variation in contrasting breeding systems*, pages 454–471. University of Chicago Press, Chicago, 1988.

Alan Grafen. The formal Darwinism project: a mid-term report. *J Evol Biol*, 20(4): 1243–1254, Jul 2007. doi: 10.1111/j.1420-9101.2007.01321.x.

Walter Grüner, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jacqueline Weber, Ivo L. Hofacker, Peter F. Stadler, and Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monatshefte f. Chemie*, 127:355–374, 1996a.

Walter Grüner, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jacqueline Weber, Ivo L. Hofacker, Peter F. Stadler, and Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monatshefte f. Chemie*, 127:375–389, 1996b.

Jacques Guadet, Jacqueline Julien, Jean F. Lafay, and Yves Brygoo. Phylogeny of some fusarium species, as determined by large-subunit rRNA sequence comparison. *Mol Biol Evol*, 6(3):227–242, May 1989.

Beatrice H. Hahn, George M. Shaw, Maria E. Taylor, Robert R. Redfield, Phil D. Markham, S. Zaki Salahuddin, Flossie Wong-Staal, Robert C. Gallo, Elizabeth S. Parks, and Wade P. Parks. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science*, 232(4757):1548–1553, Jun 1986. doi: 10.1126/science.3012778.

Kathleen B. Hall and D. Jeremy Williams. Dynamics of the IRE RNA hairpin loop probed by 2-aminopurine fluorescence and stochastic dynamics simulations. *RNA*, 10(1):34–47, 2004. URL `http://www.rnajournal.org/cgi/content/full/10/1/34`.

Richard W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160, 1950.

C. Meacham Harrell, Andrew R. McKenzie, Maria M. Patino, William E. Walden, and Elizabeth C. Theil. Ferritin mRNA: Interactions of iron regulatory element with translational regulator protein P-90 and the effect on base-paired flanking regions. *Proc Natl Acad Sci U S A*, 88(10):4166–4170, 1991. URL `http://www.pnas.org/content/88/10/4166.abstract`.

Daniel M. Held, S. Travis Greathouse, Amit Agrawal, and Donald H. Burke. Evolutionary landscapes for the acquisition of new ligand recognition by rna aptamers. *J Mol Evol*, 57(3):299–308, 2003. doi: 10.1007/s00239-003-2481-y.

Jonathan D. Hirst. The evolutionary landscape of functional model proteins. *Protein Eng*, 12(9):721–726, 1999. URL `http://peds.oxfordjournals.org/cgi/content/abstract/12/9/721`.

Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian. Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125(2):167–188, 1994.

K. Kichler Holder and James J. Bull. Profiles of adaptation in two similar viruses. *Genetics*, 159(4):1393–1404, Dec 2001. URL `http://www.genetics.org/cgi/content/full/159/4/1393`.

David Houle and Locke Rowe. Natural selection in a bottle. *Am Nat*, 161(1):50–67, Jan 2003. doi: 10.1086/345480.

Martijn A. Huynen, Peter F. Stadler, and Walter Fontana. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc Natl Acad Sci U S A*, 93(1):397–401, 1996.

Toshimichi Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 2(1):13–34, Jan 1985.

Marianne Imhof and Christian Schlotterer. Fitness effects of advantageous mutations in evolving *escherichia coli* populations. *Proc Natl Acad Sci U S A*, 98(3):1113–1117, Jan 2001. doi: 10.1073/pnas.98.3.1113.

Thomas H. Jukes. Possibilities for the evolution of the genetic code from a preceding form. *Nature*, 246(5427):22–26, Nov 1973.

Thomas H. Jukes and Motoo Kimura. Evolutionary constraints and the neutral theory. *J Mol Evol*, 21(1):90–92, 1984.

Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol*, 128(1):11–45, Sep 1987.

Stuart A. Kauffman. *The origins of order : self-organization and selection in evolution*. Oxford University Press, New York ; Oxford, 1993.

Yaohuang Ke, Jingyang Wu, Elizabeth A. Leibold, William E. Walden, and Elizabeth C. Theil. Loops and bulge/loops in iron-responsive element isoforms influence iron regulatory protein binding. fine-tuning of mrna regulation? *J Biol Chem*, 273 (37):23637–23640, 1998. URL `http://www.jbc.org/cgi/content/full/273/37/23637`.

Zora Kikinis, Richard S. Eisenstein, Andrew J. Bettany, and Hamish N. Munro. Role of RNA secondary structure of the iron-responsive element in translational regulation of ferritin synthesis. *Nucleic Acids Res*, 23(20):4190–4195, 1995.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(129):624–626, 1968a.

Motoo Kimura. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res*, 11(3):247–269, Jun 1968b.

Motoo Kimura. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. *Proc Natl Acad Sci U S A*, 78(9):5773–5777, Sep 1981.

Motoo Kimura and Takeo Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6):1337–1351, 1966.

Jack L. King and Thomas H. Jukes. Non-Darwinian evolution. *Science*, 164(881): 788–798, May 1969.

David A. Kirby, Spencer V. Muse, and Wolfgang Stephan. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A*, 92(20):9047–9051, Sep 1995. URL `http://www.pnas.org/content/92/20/9047.abstract`.

Marc Kirschner and John Gerhart. Evolvability. *Proc Natl Acad Sci U S A*, 95 (15):8420–8427, Jul 1998. URL `http://www.pnas.org/cgi/content/full/95/15/8420`.

Robin D. Knight, Stephen J. Freeland, and Laura F. Landweber. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci*, 24(6): 241–247, Jun 1999.

Robin D. Knight, Stephen J. Freeland, and Luara F. Landweber. Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet*, 2(1):49–58, Jan 2001. doi: 10.1038/35047500.

Bjarne Knudsen and Jotun Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, Jun 1999.

Katia Koelle, Sarah Cobey, Bryan Grenfell, and Mercedes Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, 314(5807):1898–1903, Dec 2006. doi: 10.1126/science.1132745. URL `http://dx.doi.org/10.1126/science.1132745`.

Michael Kospach. *Molecular Evolution of Short RNA Molecules - Neutral Nets in Sequence Spaces and Kinetic Properties of RNA*. Ph.d., University of Vienna, 2003.

David C. Krakauer and Joshua B. Plotkin. Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci U S A*, 99(3):1405–1409, 2002. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=11818563`.

Ulrich Kutschera and Karl J. Niklas. The modern theory of biological evolution: an expanded synthesis. *Naturwissenschaften*, 91(6):255–276, 2004. doi: 10.1007/s00114-004-0515-y.

Jean-Baptiste Lamarck. *Philosophie Zoologique*. Chez Dentu, Paris, 1809.

Russell Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334, jun 1976. ISSN 0014-3820. URL `http://www.jstor.org/stable/2407703`.

Russell Lande. Expected time for random genetic drift of a population between stable phenotypic states. *Proc Natl Acad Sci U S A*, 82(22):7641–7645, Nov 1985.

Russell Lande. The dynamics of peak shifts and the pattern of morphological evolution. *Paleobiology*, 12(4):343–354, Autumn 1986. ISSN 00948373. URL `http://www.jstor.org/stable/2400510`.

Richard E. Lenski and Michael Travisano. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci U S A*, 91(15):6808–6814, Jul 1994.

Richard E. Lenski, Charles Ofria, Travis C. Collier, and Christoph Adami. Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400 (6745):661–664, Aug 1999. doi: 10.1038/23245.

Richard E Lenski, Jeffrey E Barrick, and Charles Ofria. Balancing robustness and evolvability. *PLoS Biol*, 4(12):e428, Dec 2006. doi: 10.1371/journal.pbio.0040428.

Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996. doi: 10.1126/science.273.5275.666.

David J. Lipman and W. John Wilbur. Modelling neutral and selective evolution of protein folding. *Proc R Soc Lond B Biol Sci*, 245(1312):7–11, 1991. doi: 10.1098/rspb.1991.0081.

Eugene V Makeyev and Dennis H Bamford. Evolutionary potential of an RNA virus. *J Virol*, 78(4):2114–2120, Feb 2004.

John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564, February 1970. URL `http://dx.doi.org/10.1038/225563a0`.

Jason G. Mezey, James M. Cheverud, and Günter P. Wagner. Is the genotype-phenotype map modular? A statistical approach using mouse quantitative trait loci data. *Genetics*, 156(1):305–311, Sep 2000. URL `http://www.genetics.org/cgi/content/full/156/1/305`.

Takashi Miyata, Hidenori Hayashida, Teruo Yasunaga, and Masami Hasegawa. The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other. *Nucleic Acids Res*, 7(8):2431–2438, Dec 1979.

Guido Modiano, Gianantonio Battistuzzi, and Arno G. Motulsky. Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc Natl Acad Sci U S A*, 78(2):1110–1114, Feb 1981.

Chrystopher L Nehaniv. Evolvability. *Biosystems*, 69(2-3):77–81, May 2003. doi: 10.1016/S0303-2647(02)00130-2.

Mark E. J. Newman and Robin Engelhardt. Effects of selective neutrality on the evolution of molecular species. *Proc. R. Soc. London B*, 265,:1333–1338, 1998. doi: 10.1098/rspb.1998.0438.

Isabel S. Novella, Selene Zrate, David Metzgar, and Bonnie E. Ebendick-Corpus. Positive selection of synonymous mutations in vesicular stomatitis virus. *J Mol Biol*, 342(5):1415–1421, Oct 2004. doi: 10.1016/j.jmb.2004.08.003.

Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23:263–286, 1992.

H. Allen Orr. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution Int J Org Evolution*, 56(7):1317–1330, Jul 2002. doi: 10.1111/j.0014-3820.2002.tb01446.x.

H. Allen Orr. A minimum on the mean number of steps taken in adaptive walks. *Journal of Theoretical Biology*, 220(2):241–247, 2003.

H. Allen Orr. The probability of parallel evolution. *Evolution Int J Org Evolution*, 59(1):216–220, Jan 2005. doi: 10.1111/j.0014-3820.2005.tb00907.x.

H. Allen Orr. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J Theor Biol*, 238(2):279–285, Jan 2006a. doi: 10.1016/j.jtbi.2005.05.001.

H. Allen Orr. The population genetics of adaptation on correlated fitness landscapes: the block model. *Evolution Int J Org Evolution*, 60(6):1113–1124, Jun 2006b. doi: doi:10.1554/05-701.1.

Dimitri Papadopoulos, Dominique Schneider, Jessica Meier-Eiss, Werner Arber, Richard E. Lenski, and Michel Blot. Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci U S A*, 96(7):3807–3812, Mar 1999.

Alan S. Perelson and Catherine A. Macken. Protein evolution on partially correlated landscapes. *Proc Natl Acad Sci U S A*, 92(21):9657–9661, 1995. URL `http://www.jstor.org/stable/2368535`.

Massimo Pigliucci. Is evolvability evolvable? *Nat Rev Genet*, 9(1):75–82, January 2008. ISSN 1471-0056. URL `http://dx.doi.org/10.1038/nrg2278`.

Joshua B Plotkin, Jonathan Dushoff, and Hunter B Fraser. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature*, 428(6986):942–945, Apr 2004. doi: 10.1038/nature02458. URL `http://dx.doi.org/10.1038/nature02458`.

Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, Jan 2007. doi: 10.1038/nature05451. URL `http://dx.doi.org/10.1038/nature05451`.

William B. Provine. *Sewall Wright and Evolutionary Biology*. University of Chicago press, 1986.

Josep Quer, Christine L. Hershey, Esteban Domingo, John J. Holland, and Isabel. S. Novella. Contingent neutrality in competing viral populations. *J Virol*, 75(16):7315–7320, 2001. doi: 10.1128/JVI.75.16.7315-7320.2001.

Thomas. S. Ray. An approach to the synthesis of life. In C. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen, editors, *Artificial Life II*, volume XI of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 371–408. Addison-Wesley, Redwood City, CA, 1991.

Christian Reidys, Peter F. Stadler, and Peter Schuster. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull Math Biol*, 59 (2):339–397, 1997.

Christian Reidys, Christian V. Forst, and Peter. Schuster. Replication and mutation on neutral networks. *Bull Math Biol*, 63(1):57–94, 2001.

Mark Ridley. *Evolution*. Oxford : Blackwell Science, 3rd edition, 2004.

Merry S. Riley, Vaughn S. Cooper, Richard E. Lenski, Larry J. Forney, and Terence L. Marsh. Rapid phenotypic change and diversification of a soil bacterium during 1000 generations of experimental evolution. *Microbiology*, 147(4):995–1006, 2001.

Noah A. Rosenberg. A sharp minimum on the mean number of steps taken in adaptive walks. *Journal of Theoretical Biology*, 237(1):17–22, November 2005. doi: 10.1016/j.jtbi.2005.03.026.

Howard D Rundle, Stephen F Chenoweth, and Mark W Blows. The roles of natural and sexual selection during adaptation to a novel environment. *Evolution Int J Org Evolution*, 60(11):2218–2225, Nov 2006.

Erik A. Schultes and David P. Bartel. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 289(5478):448–452, Jul 2000. doi: 10.1126/science.289.5478.448.

Peter Schuster and Walter Fontana. Chance and necessity in evolution: lessons from RNA. *Physica D: Nonlinear Phenomena*, 133(1-4):427–452, 1999.

Peter Schuster, Walter Fontana, Peter F. Stadler, and Ivo L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings: Biological Sciences*, 255(1344):279–284, mar 1994. ISSN 0962-8452. URL `http://www.jstor.org/stable/49949`.

Robert A. Skipper, Jr. The persistence of the R. A. Fisher-Sewall Wright controversy. *Biology and Philosophy*, 17:341–367, 2002.

Robert A. Skipper, Jr. The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. *Philosophy of Science*, 71(5):1176–1188, 2004. doi: 10.1086/425240.

Tom. Smith, Phil Husbands, Paul Layzell, and Michael O'Shea. Fitness landscapes and evolvability. *Evol Comput*, 10(1):1–34, 2002. doi: 10.1162/106365602317301750.

Tom Smith, Phil Husbands, and Michael O'Shea. Local evolvability of statistically neutral GasNet robot controllers. *Biosystems*, 69(2-3):223–243, 2003.

Bärbel M. Stadler, Peter. F. Stadler, Günter P. Wagner, and Walter Fontana. The topology of the possible: formal spaces underlying patterns of evolutionary change. *J Theor Biol*, 213(2):241–274, 2001.

Nina Stoletzki and Adam Eyre-Walker. Synonymous codon usage in *escherichia coli*: selection for translational accuracy. *Mol Biol Evol*, 24(2):374–381, Feb 2007. doi: 10.1093/molbev/msl166.

Monroe W. Strickberger. *Evolution*. Sudbury, Mass. ; London : Jones and Bartlett,, 3rd ed. edition, 2000.

Sumedha, Olivier C Martin, and Luca Peliti. Population size effects in evolutionary dynamics on neutral networks and toy landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(05):P05011, 2007a. doi: 10.1088/1742-5468/2007/05/P05011.

Sumedha, Olivier C Martin, and Andreas Wagner. New structural variation in evolutionary searches of RNA neutral networks. *Biosystems*, 90(2):475–485, 2007b. doi: 10.1016/j.biosystems.2006.11.007.

Ers Szathmry. Why are there four letters in the genetic alphabet? *Nat Rev Genet*, 4 (12):995–1001, Dec 2003. doi: 10.1038/nrg1231.

Manfred Tacker, Peter Stadler, Erich Bornberg-Bauer, Ivo Hofacker, and Peter Schuster. Algorithm independent properties of RNA secondary structure prediction. *European Biophys J*, 25(2):115–130, December 1996. doi: 10.1007/s002490050023.

Nobuto Takeuchi and Paulien Hogeweg. Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evol Biol*, 7:15; author reply 15, 2007. doi: 10.1186/1471-2148-7-15.

Paul A. Todd, Richard A. Briers, Richard J. Ladle, and F. Middleton. Phenotype-environment matching in the shore crab (*Carcinus maenas*). *Marine Biology*, 148: 1357–1367, 2006. doi: 10.1007/s00227-005-0159-2.

Erik van Nimwegen and James P. Crutchfield. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? *Bulletin of Mathematical Biology*, 62(5):799–848, 2000. doi: 10.1006/bulm.2000.0180.

Erik van Nimwegen and James P. Crutchfield. Optimizing epochal evolutionary search: Population-size dependent theory. *Machine Learning*, 45(1):77–114, October 2001. URL `http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1023/A:1012497308906`.

Erik van Nimwegen, James P. Crutchfield, and Martijn Huynen. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A*, 96(17):9716–9720, 1999.

Gregor von Mendel. Versuche uber pflanzen-hybriden. *J Hered*, 42(1):3–4, 1951. URL `http://jhered.oxfordjournals.org`.

Michael J. Wade and Charles J. Goodnight. Perspective: The theories of Fisher and Wright in the context of metapopulatinos: When nature does many small experiments. *Evolution*, 52(6):1537–1553, December 1998. doi: doi:10.2307/2411328.

Andreas Wagner. Robustness, evolvability, and neutrality. *FEBS Lett*, 579(8):1772–1778, 2005.

Günter P. Wagner and Lee Altenberg. Perspective: Complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967–976, Jun. 1996. ISSN 00143820.

Lindi M. Wahl and David C. Krakauer. Models of experimental evolution: the role of genetic chance and selective necessity. *Genetics*, 156(3):1437–1448, Nov 2000. URL `http://www.genetics.org/cgi/content/full/156/3/1437`.

Daniel M Weinreich and Lin Chao. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution Int J Org Evolution*, 59(6): 1175–1182, Jun 2005. doi: 10.1111/j.0014-3820.2005.tb01769.x.

Daniel M Weinreich, Richard A Watson, and Lin Chao. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution Int J Org Evolution*, 59(6):1165–1174, Jun 2005. URL `http://www.bioone.org/perlserv/?request=get-abstract&doi=10.1554%2F04-272`.

Kenneth M. Weiss and Stephanie M. Fullerton. Phenogenetic drift and the evolution of genotype-phenotype relationships. *Theor Popul Biol*, 57(3):187–195, May 2000. doi: 10.1006/tpbi.2000.1460.

Mary Jane West-Eberhard. *Developmental Plasticity and Evolution*. Oxford University Press, 2003.

Mary Jane West-Eberhard. Developmental plasticity and the origin of species differences. *Proc Natl Acad Sci U S A*, 102 Suppl 1:6543–6549, May 2005. doi: 10.1073/pnas.0501844102.

Michael C. Whitlock, Patrick C. Phillips, Francisco B.-G. Moore, and Stephen J. Tonsor. Multiple fitness peaks and epistasis. *Annual Review of Ecology and Systematics*, 26:601–629, 1995. ISSN 0066-4162. URL `http://links.jstor.org/sici?sici=0066-4162%281995%2926%3C601%3AMFPAE%3E2.0.CO%3B2-L`.

Claus O. Wilke. Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution Int J Org Evolution*, 55(12):2412–2420, Dec 2001a. doi: 10.1111/j.0014-3820.2001.tb00756.x.

Claus O. Wilke. Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology*, 63(4 SU -):715–730, 2001b.

Claus O. Wilke. Quasispecies theory in the context of population genetics. *BMC Evol Biol*, 5:44, Aug 2005. doi: 10.1186/1471-2148-5-44.

Claus O. Wilke and Christoph Adami. Interaction between directional epistasis and average mutational effects. *Proc R Soc Lond B Biol Sci*, 268(1475):1469–1474, 2001.

Claus O. Wilke and Christoph Adami. Evolution of mutational robustness. *Mutat Res*, 522(1-2):3–11, 2003.

Claus O. Wilke, Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001. doi: 10.1038/35085569.

Claus O. Wilke, Richard E. Lenski, and Christoph Adami. Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. *BMC Evol Biol*, 3(1):3, Feb 2003. doi: doi:10.1186/1471-2148-3-3.

Carl R. Woese. Order in the genetic code. *Proc Natl Acad Sci U S A*, 54(1):71–75, Jul 1965. URL http://www.jstor.org/stable/72993.

Yuri Wolf, Cecile Viboud, Edward Holmes, Eugene Koonin, and David Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol Direct*, 1(1):34, Oct 2006. doi: 10.1186/1745-6150-1-34.

Sewall Wright. The roles of mutation, inbreeding, crossing and selection in evolution. In *Proceedings of the VI International Congress of Genetrics*, volume 1, pages 356–366, 1932.

Sewall Wright. Random drift and the shifting balance theory of evolution. In Ken-ichi Kojima, editor, *Mathematical topics in population genetics*, volume 1 of *Biomathmatics*, pages 1–31. Springer-Verlag, 1970.

Sewall Wright. The shifting balance theory and macroevolution. *Annu Rev Genet*, 16:1–19, 1982.

Richard Wroe, Eric Bornberg-Bauer, and Hue Sun Chan. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys J*, 88(1):118–131, 2005. doi: doi:10.1529/biophysj.104.050369.

Richard Wroe, Hue Sun Chan, and Erich Bornberg-Bauer. A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP Journal*, 1(1):79–87, 2007. doi: 10.2976/1.2739116.

Yu Xia and Michael Levitt. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci U S A*, 99(16):10382–10387, 2002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12149452.

Gabriel Yedid and Graham Bell. Macroevolution simulated with autonomously replicating computer programs. *Nature*, 420(6917):810–812, 2002. doi: 10.1038/nature01151. URL http://dx.doi.org/10.1038/nature01151.

Wen Zhu and Stephen Freeland. The standard genetic code enhances adaptive evolution of proteins. *Journal of Theoretical Biology*, 239(1):63–70, March 2006. URL http://www.sciencedirect.com/science/article/B6WMD-4HPD5F5-1/2/a4c07ca10b9a8811c4d9f1e6886eb05b.

George K. Zipf. *The Psycho-biology of Language*. Houghton-Mifflin, 1935.

Emile Zuckerandl and Linus Pauling. Molecular disease, evolution and genic heterogeneity. In M. Kasha and B. Pullman, editors, *Horizons in biochemistry*, pages 189–225. New York: Academic Press, 1962.

Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9 (1):133–148, Jan 1981. URL `http://www.pubmedcentral.gov/articlerender.fcgi?tool=pubmed&pubmedid=6163133`.

Appendix A

Genotype space broken down by sequence length and phenotype

A.1 Length-10

Breakdown of phenotypes at 30^oC			
PID	**Phenotype**	**No. of seqs**	**No. of nets**
0	.....	960655	1
1	....(....)	2036	4
2	...(....).	2592	4
3	..((....))	7162	12
4	.(((...)))	3062	12
5	.((....)).	12819	9
6	(((....)))	12276	13
7	..(....)..	2235	4
8	..((...)).	1855	2
9	.((.....))	3853	5
10	(((...))).	5651	17
11	((....))..	15363	8
12	.(....)...	2217	4
13	.((...))..	1917	2
14	((.....)).	9626	7
15	((......))	1873	5
16	((....).).	41	4
17	(....)....	3185	3
18	(.(....).)	21	4
19	(.(....)).	137	3

For the full list of the networks at longer sequence lengths please download the electronic version of this thesis, available from the Oxford University Research Archive:

http://ora.ox.ac.uk

Appendix B

Average portal distance

NID	Shortest path details		
	Count	Maximum length	Average length
1	5708	5	2.08
2	2623	4	1.67
3	1198	3	1.41
4	19306	5	1.98
5	59517	7	2.18
6	68657	8	2.15
7	282	2	1.09
8	12033	5	1.61
9	57270	6	2.07
10	46257	6	1.80
11	6	4	2.67
12	62755	9	2.32
13	39905	5	1.45
14	217685	6	2.01
15	90323	7	1.85
16	57419	7	2.03
17	920	3	1.17
18	12558	5	1.63
19	298	2	0.75
20	9226	6	1.53
21	76	3	1.18
22	55560	5	1.51
23	41213	5	1.22
24	57198	6	1.86
25	179586	9	1.75
26	66477	8	1.99
27	4774	3	1.49
28	32228	5	1.27
29	2288	5	1.79
30	25	2	1.24
31	326	2	1.19
32	864	5	1.43
33	360	2	0.59
34	39899	9	2.81
35	252	5	0.83
36	40751	6	1.80
37	189570	8	1.80
38	91555	6	1.80
39	32763	5	1.48
40	156719	7	2.22

Continued on next page

Table B.1 – continued from previous page

NID	Shortest path details		
	Count	Maximum length	Average length
41	2940	3	0.54
42	11406	4	1.13
43	112367	8	2.24
44	57237	6	2.03
45	4085	4	1.30
46	663	3	1.11
47	17322	4	1.12
48	242	2	0.81
49	5996	5	1.69
50	1019	3	1.05
51	3877	5	1.71
52	439	3	1.04
53	62064	5	1.80
54	51818	6	1.68
55	36242	8	1.87
56	6464	4	1.31
57	3878	3	1.14
58	56374	6	1.92
59	125687	7	2.09
60	46266	6	1.51
61	2190	3	0.69
62	26859	5	1.42
63	2707	4	0.77
64	5231	4	1.28
65	2087	4	1.64
66	1207	3	1.01
67	24213	9	1.83
68	17530	11	2.37
69	9238	6	1.75
70	362	3	1.50
71	204	4	1.72
72	82776	4	1.54
73	48470	7	1.87
74	125623	8	2.07
75	47539	9	2.07
76	6750	4	1.14
77	14876	4	1.13
78	13703	4	0.91
79	621	5	0.90
80	9229	4	1.44

Continued on next page

Table B.1 – continued from previous page

NID	Shortest path details		
	Count	Maximum length	Average length
81	245	3	0.68
82	28105	8	2.57
83	58271	10	2.34
84	42961	6	2.19
85	273	3	0.75
86	36	1	0.11
87	20159	4	1.19
88	2412	4	1.81
89	1319	4	1.55
90	972	3	1.37
91	385	3	1.72
92	2707	3	0.68
93	300	1	0.27
94	1124	3	1.28
95	17733	6	1.69
96	3293	5	1.72
97	2843	2	0.63
98	1502	3	0.92
99	0	0	0.00
100	6	3	2.33
101	32677	5	1.67
102	175292	6	1.96
103	9734	6	1.84
104	288	2	0.14
105	2444	3	1.09
106	68	2	0.24
107	81684	6	1.82
108	39006	5	2.02
109	440	3	0.32
110	78	1	0.08
111	9991	4	1.28
112	1674	3	1.65
113	7202	4	1.04
114	44	2	0.82
115	0	0	0.00
116	0	0	0.00
117	6763	5	1.56
118	2693	3	1.09
119	3270	5	1.63
120	450	3	1.51

Continued on next page

Table B.1 – continued from previous page

NID	Shortest path details		
	Count	Maximum length	Average length
121	2102	3	1.85
122	8653	4	1.16
Total:	3373430	Average:	1.88

Table B.1: The average number of neutral substitutions needed to get between any two portals across the network studied

Appendix C

Path length differences

NID	Network	Paths		Max network	Max
	density	number	%age of total	distance	difference
3	0.859	1	5.93E-02	3	1
4	0.633	1	2.61E-03	4	1
5	0.285	439	0.34	7	1
6	0.041	469	0.32	7	2
8	0.48	15	7.73E-02	5	1
9	0.284	162	0.14	5	1
10	0.149	294	0.35	5	2
12	0.123	1105	0.76	9	3
14	0.072	47	1.07E-02	5	1
15	0.125	70	4.20E-02	6	2
16	0.241	787	0.68	7	2
20	0.418	170	1.21	6	2
22	0.763	6	7.13E-03	3	1
24	0.249	366	0.34	6	1
25	0.174	168	5.34E-02	9	3
26	0.098	558	0.42	8	4
34	0.104	601	0.54	9	6
37	0.045	367	0.11	8	4
38	0.13	109	6.61E-02	6	2
40	0.079	43	1.24E-02	5	2
43	0.1	1048	0.42	8	2
44	0.235	536	0.46	6	2
49	0.428	57	0.56	5	2
53	0.572	24	2.15E-02	5	1
55	0.263	192	0.28	8	2
58	0.249	276	0.26	6	1
59	0.233	175	6.66E-02	7	2
67	0.055	1095	2.47	10	5
68	0.078	506	1.22	11	6
69	0.418	9	5.58E-02	5	1
72	0.63	11	8.64E-03	4	1
73	0.225	150	0.17	7	4
74	0.06	609	0.23	8	4
75	0.25	753	0.77	9	2
77	0.961	1	5.97E-03	3	1
82	0.318	2620	3.63	8	4
83	0.108	2599	1.91	10	3
84	0.362	35	3.72E-02	5	1
95	0.526	226	0.75	6	2
101	0.573	399	0.73	5	1

Continued on next page

Table C.1 – continued from previous page

NID	Network density	Paths		Max network distance	Max difference
		number	%age of total		
102	0.227	250	7.28E-02	6	1
103	0.344	113	0.63	6	1
108	0.68	142	0.18	5	1
111	0.934	1	7.80E-03	3	1
117	0.169	14	0.13	4	1
122	0.93	2	1.99E-02	4	1

Table C.1: Networks in the length-10 space with routes between portals longer than the Hamming distance. Column 3 shows the total number of paths for which this is true; column 4 gives the percentage of all paths through that network which are longer than the Hamming distance. Column 5 shows the largest minimum inter-portal path length, and column 6 the maximum difference between the minimum inter-portal path and the Hamming distance