Principles and Parameters of Molecular Robustness

David C. Krakauer
Joshua B. Plotkin

SFI WORKING PAPER: 2003-02-009

SANTA FE INSTITUTE

Principles and Parameters of Molecular Robustness

David C. Krakauer & Joshua B. Plotkin

(1) Santa Fe Institute, Hyde Park Road, Santa Fe, NM 87501 USA.
krakauer@santafe.edu

(2) Institute for Advanced Study, Princeton, New Jersey 08540, USA

Abstract

A large number of molecular mechanisms exist to ensure the continued function of perturbed cells. Here we explore a range of mechanisms of molecular robustness. A theory of robustness is proposed in which mechanisms are classified into those that purge damage, and those that buffer damage. The strategy upon which a system settles, is a function of the cost of purgation, and thereby a function of population size.

1 Three principles of Robustness

1.1 Canalization, Neutrality and Redundancy

During the course of replication of RNA or DNA, genomes incorporate large numbers of mutations. These mutations can be small, such as modification of a single nucleotide (point mutations), or large involving repeating motifs of nucleotides (micro-satellites), damage to whole chromosomes (genetic instability) or even duplication or loss of whole chromosomes (aneuploidy). The influence of mutation on the evolutionary process is two-fold. On the one hand, mutation leads to phenotypic variance mediated by developmental dynamics, thereby providing the variability required by selection to fix new variants in a population. On the other hand, mutation undermines pre-adapted phenotypes by perturbing development in such a way as to lead to poorly adapted variants. The tension between the advantage of a few novel phenotypes, and the disadvantage of the majority of novel phenotypes is reflected in those mechanisms controlling the rate of mutation, and in those strategies influencing the impact of mutation on phenotypes. Three distinct albeit closely related principles have arisen in an effort to understand the evolutionary response to mutations. The principle of canalization, the principle of neutrality and the principle of redundancy. These are contrasted with the parameters of robustness – the precise mechanisms by which these principles are realized. The principles and parameters metaphor is derived from linguistics [11] where the principles are the invariant properties of universal grammar and the parameters the local rules and practices of language.

The principle of canalization was introduced by Waddington [62] as a means of explaining the constancy of tissues and organ types during development. Canalization refers to those mechanisms that suppress phenotypic variation during development and thereby reduce the cumulative cost of deviations from a locally optimal trajectory. Waddington conceived of deviations as the result of mutations or environmental insults. Mutational canalization and environmental canalization are the terms Waddington applied to the unspecified mechanisms buffering these effects. Evidence for both types of mechanism has been observed [68] [53]. While Waddington conceived of canalization as an ensemble of mechanisms, work bearing directly on the concept has been largely functional, either of a theoretical nature [66], [65], [3], or related to experimental evolution [16] and quantitative genetics [56]. Fewer studies have directly identified the mechanisms of canalization. The term canalization exists as a general evolutionary principle describing adaptive suppression of phenotypic variation during the process of development and as a catch-all phrase referring to those mechanisms buffering deleterious mutations or environmental insults.

The principle of neutrality is best known to biologists in relation to the selective neutrality of alleles in populations. The neutral theory [32] [31] rose to prominence as a means of explaining the higher than expected level of variation in electrophoretic data [42]. Neither the traditional theory of rare wildtypes combined with common deleterious mutations [47], or the balance theory [67] were adequately able to explain the observed diversity. Neutrality refers to the selective equivalence of different phenotypes. The idea does not require that mutations to a wildtype sequence leave the phenotype untouched (although this can be the case), but that the phenotypic differences are beyond the detection limit for selection. As population become smaller, drift effects (random sampling of gametes) amplify [73], implying that selection coefficients must increase in order for selection to dominate drift. Neutrality, unlike canalization, is not assumed to be an adaptive means of suppressing variability. Nor is neutrality concerned with the development of phenotypes. Neutrality is simply a measure of the selective equivalence of phenotypes, and is primarily concerned with finite population effects. Through canalization phenotypes can manifest neutrality, but it is meaningless to say that neutrality is a cause of canalization. More recently interest has turned to the discussion of neutral networks [21] [20]. These networks are sets of selectively equivalent genotypes, connected via single mutational steps. These networks span large volumes of genotype space and provide a natural buffering mechanism. They are often treated as highly epistatic fitness landscapes [60].

The principle of redundancy is more ancient and more widespread. We restrict ourselves to discussing its modern biological interpretation which is derived largely from molecular biology.

A very common means of identifying the function of a gene is to perform a knockout experiment, removing or silencing a gene early in development. By carefully assaying the phenotype, the putative function of the absent gene may be revealed. However, in many such experiments, there is no scoreable phenotype: the knockout leaves the phenotype as it was in the wildtype. Biologists then refer to the gene as redundant. This is taken to mean that this gene is but one of two or more genes contributing to the phenotype, and where removal of one leaves the phenotype unchanged. Of course what we might be observing is something like experimental neutrality, an effect below the experimental detection limit [6]. Assuming that we are able to detect small changes, redundancy describes the degree of correlation among genes contributing to a single function [58] [57]. As with canalization and neutrality the mechanisms (parameters) giving rise to robustness through redundancy are not explicitly stated. Unlike canalization, redundancy does not include a developmental component. Redundancy is merely a measure of the degree to which a set of genes share the burden of function. Moreover, unlike canalization, redundancy need not be adaptive – it can be accidental (although this is unlikely). Once again, one can say redundancy gives rise to neutrality, but not neutrality gives rise to redundancy.

In conclusion all three principles, canalization, neutrality and redundancy are associated with a reduction in selective variance. Canalization is the adaptive suppression of variance during development. Neutrality is the selective equivalence of phenotypes. Redundancy is the overlap of gene function. Only redundancy and canalization assume that phenotypic variance and selective variance are colinear. Very different phenotypes can be selectively equivalent and hence neutral. Neither redundancy nor neutrality assume adaptation, whereas canalization is always an evolved character (in the sense of Waddington).

2 Beyond the redundancy principle: the principle of anti-redundancy

2.1 Redundancy (R) as error buffering, and anti-redundancy (AR) as error elimination

Perhaps the most obvious way in which genes correlated in function promote redundancy is through several copies of a single gene [58]. This situation arises through gene duplication and leads to what are referred to as paralogous copies. Redundancy attributed to paralogues have been found in homeotic genes [45], transcription factors [43] , signal transduction proteins [26], metabolic pathway genes [48], and among the variable genes encoding antibody peptides [69]. It is thought that paralogues promote robustness by 'backing-up' important functions. The idea is that if one copy should sustain damage, then the paralogue will be sufficient to generate the required protein. However, this line of reasoning can be problematic. Because one copy is in principle as fit as two copies, in time one copy is expected to be lost from the population through random mutation [63]. To preserve two or more redundant, paralogous genes, there should be some asymmetry in the contribution of each gene to their shared function [50] [36]. For example, those genes making the larger fitness contribution could experience higher rates, of deleterious mutation than those making smaller fitness contributions. Without such an asymmetry, those genes with the higher mutation rates are lost by random drift. Duplicated genes constitute a redundant mechanism as both genes contribute to an identical function, and in the wildtype condition, only one gene is required. In a later section we provide case studies for mechanisms demonstrated to give rise to redundancy at the phenotypic level. We have discussed this one example early in order to consolidate our intuition of redundancy before describing a further principle of robust design – anti-redundancy.

In many organisms redundancy is rare. In viruses and bacteria, for example, the need for rapid replication and translation leads to small genomes with no or few duplicate genes, a small number of controlling elements, and overlapping reading frames. As a result, a single mutation will often damage several distinct functions simultaneously [33]. Within multicellular eukaryotes checkpoint genes, such as p53, enhances the cell-damage caused by mutations which

might otherwise accumulate in a tissue [40]. The decline in telomerase enzyme during the development of a cell lineage effectively ensures that cells are unable to propagate mutations indefinitely [59]. Similarly, it has been conjectured that the loss of key error repair genes in mitochondria, might reduce the rate of mildly deleterious mutation accumulation [44]. In each of these cases we observe the evolution of mechanisms that promote anti-redundancy – that is, mechanisms which sensitize cells or individuals to single gene damage and thereby eliminate them preemptively from a tissue or from a population of individuals. Unlike redundancy in which genes act together to share the burden of function, anti-redundant mechanisms amplify the damage to other genes. The effect of these mechanisms is to increase the selective cost of each mutation.

3 Redundancy and fitness landscapes

3.1 Fitness landscapes as statistical means of describing redundancy

The concept of a fitness landscape, first articulated by Sewall Wright in 1932 (Wright, S. 1988. Surfaces of adaptive value revisited. Am. Nat. 131:115-123), has proved an enormously useful tool for evolutionary biologists and population geneticists. A fitness landscape is simply an assignment of a fitness value – that is, an intrinsic growth rate in the absence of density limitation – to each possible genotype. In other words, the landscape encodes a mapping between genotypes and their Darwinian fitness.

We will use single-peak fitness landscapes to encode the degree of redundancy or anti-redundancy of an organism. Multiple peak landscapes are discussed in a following section. We shall study landscapes that are symmetric around a central "wild-type" genotype. We define the set of genotypes as the set of all L-bit strings of zeroes and ones. There are thus 2^L possible genotypes, where L is the size (in bits) of the genome. One of these genotypes is distinguished as the "wildtype" whose fitness equals or exceeds all others. We will consider single-peak landscapes: genotypes far from the wildtype have lower fitness, whereas genotypes near the wildtype have higher fitness.

The L "bits" of which a genome is comprised may be interpreted as L genes (each of which is functional or mutated), or alternatively as L nucleotide positions. We do not loose any generality by assuming that each bit assumes two states, zero or one. The gene-wise or nucleotide-wise interpretation of the bits will depend upon context.

In these terms, the amount of redundancy of an organism is described by the steepness of its fitness landscape. Loosely speaking (see below for a formal definition), an organism features a buffered or redundant genome if mutations away from the wildtype do not dramatically lower fitness. Conversely, the genome is characterized by anti-redundancy if mutations precipitously reduce the fitness. In other words, the degree of redundancy reflects the rate of phenotypic penetrance of deleterious mutations, as described by the geometry of the fitness landscape.

3.2 Hamming classes and multiplicative landscapes

The landscapes we consider are symmetric around a fixed wildtype genome. In other words, the fitness of a genotype depends only upon the number of mutations between that genotype and the wildtype. This implicitly assumes that mutations to any part of the genome are equally deleterious.

The Hamming distance between two genomes is defined as the number of positions, or bits, at which the two genomes differ. The Hamming distance between a given genotype and the wildtype is thus a number between zero and L. All genotypes which are exactly k mutations away from the wildtype comprise the kth *hamming class*. By our symmetry assumption, the fitness of a genome depends only the hamming class in which it lies.

We will generally explore multiplicative fitness landscapes, wherein the fitness, w_k of the kth Hamming class is proportional to

$$w_k \propto (1-s)^k \tag{1}$$

The wildtype sequence, $k = 0$, is maximally-fit, and each deleterious mutation reduces fitness by an amount $(1 - s)$, independent of the other loci. The parameter s measures the deleteriousness of each mutation. By varying the magnitude of s we vary the steepness of the landscape and hence the effective degree of redundancy in the genome. A large value of s yields a steep, or anti-redundant, landscapes. A small value of s results in a shallower, redundant landscape.

The fitness loss caused by a random mutation to a genome is generally very small in a wide range of organisms. In Drosophila, for example, measured values of s rarely exceed one percent – despite the fact that several individual mutations are known to be lethal (REFS). Throughout this paper, we will generally assume that the deleteriousness s of each mutation is small. Such an assumption is supported by current evidence [61].

The fitness formulation in Eq. 1 is the canonical example of a *non-epistatic* landscape. In other words, mutations at one loci have the same deleterious effect on fitness independent of the status of other loci. Non-epistatic landscapes are generally easier to analyze than epistatic landscapes, and they will be the focus of our attention. Nevertheless, we we pause to introduce a simple formulation of epistatic landscapes:

$$w_k \propto (1 - s)^{k^\alpha} \tag{2}$$

When $\alpha = 1$, this formulation reduces to the non-epistatic case. When α exceeds one, however, the landscape features *antagonistic epistasis*: each additional mutation has an increasingly deleterious effect on fitness. When α is less than one, the landscape features *synergistic epistasis*: each additional mutation has an less deleterious effect on fitness.

For a single, fixed landscape, geneticists are accustomed to thinking of the degree of redundancy itself in terms of the degree of epistasis in the landscape. Yet in this paper, we have chosen to model redundancy by comparing a *family* of landscapes with varying degrees of steepness. We feel that such a framework – whether the family of landscapes has no epistatis ($\alpha = 1$), antagonistic epistatis ($\alpha > 1$), or synergistic epistatis ($\alpha < 1$) – yields a more intuitive measure of redundancy.

3.3 Landscape normalization

In the previous section, we have introduced a model of redundancy in terms of the steepness, s, of a fitness landscape. Below, we will compare a family of landscapes by varying the steepness s. We will investigate under what conditions an organism will prefer a steep, anti-redundant landscape, and under what conditions it will prefer a shallow landscape. In other words, we will allow organisms to evolve the steepness of the landscape itself.

What are the constraints on the evolution of fitness landscapes, and thereby, on the evolution of developmental programs? If an organism were allowed to evolve its landscape steepness in a fixed environment, it would certainly always prefer the shallowest landscape possible ($s = 0$), so that all genotypes have the maximal fitness. In other words, in a fixed environment, an organism will evolve towards maximum redundancy whereby mutations do not effect fitness. In reality, however, this solution is not allowable in light of physiological constraints – *i.e* intrinsic molecular costs associated with the evolution of redundancy.

Genotypes cannot evolve towards both maximum fitness and maximum redundancy simultaneously. Of the molecular mechanisms of redundancy, discussed below, all incur some cost to the organism – through increased genome size, increased metabolism, or reduced binding specificity. We will model this cost by enforcing a tradeoff between the maximal height of the fitness landscape and its steepness – that is, we normalize all landscapes to have total volume one:

$$w_k = \frac{(1 - s)^k}{\sum_{j=0}^{L} (1 - s)^j} = \frac{s(1 - s)^k}{1 + (1 - s)^L (s - 1)}. \tag{3}$$

Eq. 3 enforces a tradeoff between redundancy and wildtype fitness, constraining the family of landscapes which we consider. Although the precise form of the tradeoff curve is arbitrary

and relatively unimportant (other normalizations yield similar results) it is essential that we impose some tradeoff between maximum fitness and redundancy.

4 A quasispecies description of robust populations

In order to investigate evolution of a population on a fixed landscape – and, eventually, the evolution of the landscape itself – we will use the quasispecies formulation introduced by Eigen [14] [15]. The quasispecies equation provides a very general framework for exploring mutation and selection in a heterogeneous, infinite population [7].

Eigen's quasispecies framework considers a large population of L-bit genomes, x_i, reproducing with imperfect fidelity according to their assigned fitnesses, w_i, with fixed total concentration:

$$\dot{x_i} = \sum_{j=1}^{j=2^L} w_j x_j Q_{ij} \ - x_i W. \tag{4}$$

In this equation, $W(t) = \sum w_j x_j(t)$ denotes the mean population fitness. The mutation matrix Q_{ij} denote the probability of genotype i mutating into genotype j during a replication event.

Although Eq. 4 is nonlinear, the change of variables

$$y_i(t) = \frac{x_i(t)}{\exp\left(-\int_0^t W(s)\ ds\right)} \tag{5}$$

produces a linear system whose solution satisfies $x_i(t) = y_i(t)/\sum_j y_j(t)$.

In the present investigation, we assume that the fitness depends only upon the hamming class of a genotype. If we define the kth hamming class as the sum $z_k = \sum_{H(i)=k} y_i$ over the abundances of all sequences i which are k bits away from a fixed wildtype sequence, then Eq. 5 reduces to :

$$\dot{z_k} = \sum_{i=0}^{L} z_i w_i P_{ki}. \tag{6}$$

The value w_i denotes the fitness of a genome in the ith Hamming class, and P_{ki} is the probability of mutation from an i-error genotype to a k-error genotype during replication.

We will assume that each locus has a fixed chance of mutating at each replication event, independent of the other loci. Thus the chance of mutation P_{ki} from Hamming class i to class k is determined by the per-base forward and backward mutation rates, p and b:

$$P_{ki} = \sum_l \binom{i}{l}\binom{L-i}{k-i+l} p^{k-i+l}(1-p)^{L-k-l} b^l (1-b)^{i-l} \tag{7}$$

where we sum from $l = \max(0, i-k)$ to $l = \min(i, L-k)$. In this sum l denotes the number of back-mutations. The backward mutation rate b may be less than or equal to the forward mutation rate p, depending upon the interpretation of the loci as genes or nucleotides.

4.1 Population mean fitness at equilibrium

For a fixed landscape, we are interested in the mean fitness of a population that has reached mutatation-selection equilibrium. For a purely multiplicative landscape ($\alpha = 1$) with equal forward and backward mutation rates ($p = b$) the mean fitness can be solved exactly [25]. The equilibrium mean fitness will depend upon the genome length, L, the mutation rate p, and the steepness of the landscape, s.

The dominant eigenvector of $(w_l P_{kl})$ provides the equilibrium relative abundances of the hamming classes. Moreover, the corresponding dominant eigenvalue equals the equilibrium mean fitness. As suggested in [72], we look for an eigenvector of the binomial form $z_k =$

$\binom{L}{k}a^k(1-a)^{L-k}$, where a must yet be determined. In order to compute a, we solve the discrete-time equivalent of Eq. 4, which reduces to the same eigensystem problem [72].

Consider the random variable V_k defined as the hamming class after one generation of replication (with mutation) of a genotype starting in hamming class k. The generating function of V_k is defined as

$$G(V_k) = \sum_{i=0}^{\infty} \mathbb{P}(V_k = i)X^i.$$

where X is a formal variable.

For a one-bit genome ($L = 1$), we clearly have

$$\begin{aligned} G(V_1) &= p + qX \\ G(V_0) &= q + pX, \end{aligned}$$

where $q = 1 - p$. For $L > 1$, V_k is sum of L independent random variables, one for each bit in the genome. Hence $G(V_k)$ is the product of generation functions:

$$G(V_k) = (p + qX)^k(q + pX)^{L-k}.$$

Given the current abundance of each hamming class, $\mathbf{z} = (z_0, z_1, \ldots z_L)$, then

$$G(V_{\mathbf{z}}) = \sum_k z_k(q + pX)^{L-k}(p + qX)^k$$

is the generating function for the hamming class after mutation of a randomly chosen individual in the population. Similarly,

$$G(V^{\mathrm{s}}_{\mathbf{z}}) = \sum_k (1-s)^k z_k(q + pX)^{L-k}(p + qX)^k$$

is the generating function for the hamming class after mutation of an individual chosen according to its fitness.

In equilibrium, the eigenvector $\hat{\mathbf{z}}$ satisfies $G(V^{\mathrm{s}}_{\hat{\mathbf{z}}}) = \lambda \sum_k \hat{z_k} X^k$, or

$$\sum_k t^k \hat{z_k}(q + pX)^{L-k}(p + qX)^k = \lambda \sum_k \hat{z_k} X^k,$$

where we have defined $t = 1 - s$. Substituting our binomial assumption for the equilibrium eigenvector $\hat{\mathbf{z}}$, we may solve the following system to find the value of a:

$$\begin{aligned} \sum_k \binom{L}{k} a^k(1-a)^{L-k} t^k (q + pX)^{L-k}(p + qX)^k &= \\ \lambda \sum_k \binom{L}{k} a^k (1-a)^{L-k} X^k & \end{aligned}$$

Equivalently, we solve

$$\begin{aligned} [ta(p + qX) + (q + pX)(1-a)]^L &= \lambda[aX + 1 - a]^L \\ ta(p + qX) + (q + pX)(1-a) &= \lambda^{-L}(aX + 1 - a). \end{aligned}$$

Setting coefficients of X^0 and X^1 equal, we find that

$$\begin{aligned} tap + q(1-a) &= \lambda^{-L}(1-a) \\ taq + p(1-a) &= \lambda^{-L} a \\ &= \frac{tap + q(1-a)}{1-a} \cdot a, \end{aligned}$$

and so

$$a = \frac{1}{2}\left(1 - p + \frac{2p}{s} - \sqrt{\left(1 - p + \frac{2p}{s}\right)^2 - \frac{4p}{s}}\right). \tag{8}$$

In equilibrium, the mean hamming distance from the wildtype is $\bar{k} = aL$.

The mean population fitness in equilibrium, $\bar{w}$, is easy to calculate once we know the complete (binomial) distribution of equilibrium hamming classes: $\bar{w} = \sum w_k \hat{z}_k$. Note that $\bar{w}$ depends on genome length L, the mutation rate p, and the landscape steepness, s.

An approximation for mean fitness

In some situations, it may be difficult to find the full distribution of equilibrium hamming classes, $(\hat{z_0}, \hat{z_1}, \ldots, \hat{z_L})$, as we did above for a multiplicative landscape. Nevertheless we may often be able to find the first two moments $\bar{k}$ and $\mathrm{var}(k)$ of the equilibrium hamming distribution. In such cases we can recover a good approximation of the mean fitness according to the Taylor expansion of k around $\bar{k}$:

$$\bar{w} = w(\bar{k}) + \frac{1}{2}\mathrm{var}(k)w''(\bar{k}) + \ldots$$

where we use the notations w_k and $w(k)$ interchangeably. In our case $w(k) \propto (1-s)^k$ and $w''(k) \propto (1-s)^k \log(1-s)^2 \approx (1-s)^k(s^2 + s^3 + O(s^4))$. Since we are assuming a small selective value s throughout, $w''(k)$ is negligible compared to s. Therefore, we may use

$$\bar{w} \approx w(\bar{k}) \tag{9}$$

as good approximation of equilibrium mean fitness in terms of the fitness of the equilibrium mean hamming class.

4.2 Stochastic dynamics and the influence of population size

The quasispecies framework introduced by Eigen applies only to infinite populations of replicating individuals. But we are primarily interested in the effects of redundancy and anti-redundancy in finite and even very small populations. For a constant, finite population size, the population mean fitness does not steadily approach a fixed equilibrium value. Instead, the stochastic process of mutation and selection produces variation in the mean population fitness over time. Nevertheless, the stochastic process approaches a steady state – that is, the expected population fitness (where expectation here denotes ensemble average) assumes a fixed value for large times.

Assuming a small mutation rate p, moment equations [72] allow us to compute the steady state population mean hamming class in terms of the population size N:

$$\langle\bar{k}\rangle = \frac{L}{2}\left(1 + \frac{2p}{s} + \frac{1}{2sN} - \sqrt{\left(1 + \frac{2p}{s} + \frac{1}{2sN}\right)^2 - \frac{4p}{s} - \frac{1}{sN} - 2p}\right). \tag{10}$$

Note that the ensemble average, $\langle\cdot\rangle$, is taken after the population average. Substitution into Eqs. 3 and 9 yields the expected equilibrium mean fitness of a finite population in steady state. These equations determine the relationship between mean population fitness, the strength of selection, the rate of mutation, the genome length, and the size of the population.

Figure 1 shows the relationship between the level of redundancy, s, and the expected mean population fitness for several different population sizes. Both in theory (Fig 1a) and in individual-based stochastic simulations (Fig 1b) we see that redundancy increases the mean fitness in small populations, while it decreases fitness in large populations. This result has an intuitive explanation. In small populations, mutational drift contributes disproportionately to the population fitness. There is a large temporal variance in the mean hamming class, and redundancy can effectively mask these mutations. Small populations are thus better served by shallow landscapes – *i.e.* by slightly decreasing the fitness of the wildtype, but increasing the

fitness of its nearby neighbors. Large populations, however, are not at risk of being "swept off" the fitness peak by the stochastic fluctuations that afflict small populations; the temporal variance in the mean hamming class is small. It is better, therefore, for large populations to amplify the phenotypic penetrance of deleterious genes via sharp landscapes.

4.3 Evolutionary accessibility of robust landscapes

Our results on equilibrium mean population fitness (Figure 1) constitute a population-based argument for the evolution of redundancy in small populations and anti-redundancy in large populations. These results do not, in themselves, demonstrate that such strategies are evolutionarily stable or achievable. In other words, we must yet demonstrate that individual replicators subject to individual-level selection evolve degrees of redundancy consistent with the optimal population mean fitness. If we allow individuals to modify the heritable steepness of their own individual landscapes through mutation, however, we have previously shown that small populations do, indeed, evolve towards redundancy, and large populations towards anti-redundancy via individual-level selection (Figure 2).

The evolutionary stability of these two strategies – sensitivity in large populations and redundancy in small populations – has an intuitive explanation. The stability rests on the fact that flatter landscapes have lower fitness peaks. A large population on a steep landscapes is highly localized near the wildtype (low $\bar{k}$). Mutants with different s-values are thus most often generated near the wildtype – precisely where a more shallow landscape would be disadvantageous to them. Conversely, small populations with shallow landscapes are de-localized (high $\bar{k}$). In this case, landscape mutants tend to arise far from the wildtype – precisely where a steeper landscape would decrease their fitness. Thus the landscape itself acts as a

mechanism for ensuring the robustness of the incumbent strategy, in each population size.

4.4 The importance of back-mutation

Unlike many treatments of the quasispecies-equation, we allow for back mutations. According to two well known principles of population genetics, the neglect of back mutations introduces pathologies into the equilibrium state of both infinite and finite populations. If back-mutations are neglected then, according the Haldane-Muller principle, the mean equilibrium fitness of an infinite haploid population is independent of the landscape's steepness (provided the wildtype is maintained). Hence, without some rate of back-mutation, we cannot detect a preference for one landscape over another in an infinite population.

Similar problems apply if we ignore back mutations in small, finite populations. In this case, the dynamics will proceed by the gradual accumulation of mutations. Once a mutation is shared by all the members of the finite population, then (ignoring back-mutations) the mutation is fixed for all time thereafter. This phenomenon, called Muller's ratchet [47] [22] [17], implies that (for a finite genome length L) the equilibrium mean fitness will equal the minimum fitness, regardless of the steepness of the landscape. Therefore, in order to detect adaptive benefits or costs of redundancy – in finite and infinite populations – we cannot ignore back mutations.

Despite the importance of not ignoring back-mutation, it is important to allow for differences between the forward mutation rate p and the back-mutation rate b. If we interpret each bit of the genome as a base-pair, then certainly $p \approx b$; if we interpret each bit as indicating whether or not a given gene is functional, then the forward mutation rate will exceed the back-ward rate. Fortunately, our qualitative results (Figs 1b, 2) remain essentially unchanged, for forward mutations occurring at twice the rate of backward mutations; small populations still favor redundancy and large populations anti-redundancy. However, as the the backward mutation rate becomes proportionately smaller than the forward rate, larger populations are required to favor steep landscapes. This makes intuitive sense as, when back-mutations are rare, drift away from the wildtype is more problematic and requires more buffering.

4.5 Epistatic effects

The analytical results on mean equilibrium fitness derived in Sections 4.1 and 4.2 apply to families of non-epistatic landscapes: $w_k \propto (1-s)^k$, normalized to unit volume. We used such landscapes because they are analytically tractable and because epistatic effects can often be confounding (although more so for diploid models). In this section we briefly discuss the consequences of epistasis, $w_k \propto (1-s)^{k^\alpha}$, on the tendency to evolve redundancy or anti-redundancy.

For small to moderate degrees of synergistic or antagonistic epistasis, $0.8 < \alpha < 1.2$, all of our qualitative results remain essentially unchanged: redundancy is preferred in small populations and anti-redundancy in large populations. However, it is interesting to note that antagonistic epistasis ($\alpha > 1$) accentuates this general trend. When comparing a family of antagonistic landscapes with varying degrees of steepness, small populations have an even stronger preference for redundancy and large populations for anti-redundancy. Similarly, synergistic epistatis ($\alpha < 1$) mitigates these preferences (See Figures 3 and 4).

4.6 Selective values

We must emphasize that our results on redundancy, anti-redundancy, and population size assume (i) haploid asexual reproduction, (ii) constant population size, (iii) roughly equal forward and backward mutation rates, (iv) finite genome length, and (v) small selective values s. These are all reasonable assumptions for a broad range of biological circumstances. In particular, the assumption that a random mutation in the genome has a small deleterious effect, s [61]. There are few extant organisms known for which, on average, a single random mutation decreases replicative ability more than one percent.

However, it is important to be aware of the equilibrium behavior of landscape families which allow very large s values, $s > 0.1$. In these cases, even small populations which start, comprising mostly wiltypes, can preserve the wiltype and fail to evolve flatter landscapes. In other words, if the landscape is *sufficiently* steep, then even a small, constant-sized population of wiltypes evolves to keep the landscape steep. Mutants are rapidly removed before they have a chance to evolve towards shallow landscapes. This phenomenon is perhaps not so significant in multicellular organisms, but becomes important for replicating RNA molecules, under prebiotic conditions [37].

4.7 Multipeak landscapes

In the preceding analysis and discussion we have assumed a symmetrical single peak landscape model. Our interest has been the preservation or maintenance of an optimal genotype configuration - designated the wildtype. In biology landscapes are rarely single peaked, and it frequently is the case, that different genomes map onto identical fitness values. In other words we have neglected neutral networks. We have done so for several reasons: (1) multipeak landscape models are largely concerned with optimization, (2) neutral network models are frequently single peaked, (3) and model tractability.

A recurrent question in the study of multipeak landscapes is the fixation probability of reaching a maximum peak or the number of mutations required to reach from one peak to another [77]. In landscapes with multiple equivalent maxima, optimization leads to a symmetry breaking event in which a single peak is selected according to the initial system configuration. In rugged landscapes with irregular local optima, the problem is not so much a question of symmetry breaking, but the preferred mechanism for hill climbing when numerous suboptimal solutions exist. In other words combinatorial optimization problems. Neither of these questions is concerned with the robustness of the final solution - the stability of the optimal system configuration - the wildtype. A robustness problem in a multipeak landscape would address the question as to the mean population fitness as a function of the spatial separation of iso-fit maximum peaks. Thus, in relation to our present concerns, transitions between alternative functional wildtypes for different population sizes and mutation rates. This has not yet been attempted and is beyond the scope of the present paper.

Existing analytical studies of neutral networks assume single peak landscapes. These are frequently plateau landscapes in which one neutral network occupies the high ground and another the low ground ([77], [78]). While many different genotype configurations map onto a single fitness class, there are only two fitness classes. Our single peak landscape model with high positive epistasis reduces to this case.

One problem with robustness of rugged landscapes involves adopting a model formalism that remains tractable. A variety of models have been used to study optimization on rugged landscapes including spin glasses, genetic algorithms [77], NK-models ([81]) and stochastic additive scale population genetics models ([80]). These are all powerful formalisms, but they result in few transparant results when assuming many heterogeneous peaks.

5 The parameters of redundant and anti-redundant design: Case studies

As a result of our modeling efforts we have discovered that population size influences the degree of redundancy we expect to be expressed by a genome. In large populations of microorganisms, such as viruses and bacteria, and in large populations of rapidly dividing cells within multicellular organisms, we predict an evolution towards antiredundant mechanisms. For small populations, on the other hand, we expect a tendency towards redundancy. In biological systems we find a large variety of molecular mechanisms capable of producing redundancy and antiredundancy. While our quasispecies formulation does not incorporate the explicit details of these mechanisms, it does provide a statistical treatment of the parameter s, that is assumed to be the developmental end-point of all these processes. In this section we provide brief case studies for six adaptive mechanisms of redundancy and the same number for antiredundancy. The principal purpose of this section is to demonstrate the utility of dividing molecular mechanisms into two groups according to the classification suggested by our model.

Those mechanisms described as redundant or antiredunt all influence development and somatic processes by modifying the effective degree of deleteriousness, s, of mutations. Redundant mechanisms (low s values) preserve mutations by masking their influence. These mechanisms are termed redundant as a consequence of their neutrality in the wildtype and their buffering capacity in the mutant. Antiredundant mechanisms (high s values) remove mutant genomes from populations (either of individuals or cells). The term antiredundancy derives from a capacity to amplify mutational damage.

When available, both sets of strategies can be exploited, even simultaneously, by a single organism according to population size constraints. Where data on incidence of mechanism in relation to population size are available we report results. However these are usually fairly qualitative as population size estimates are hard to come by, and the incidence of mechanisms often reflects sampling bias rather than genuine absence. It is also the case that a plurality of selection pressures impinge on each of these mechanisms making evaluation of data from literature sources particularly difficult and urging caution in interpretation!

5.1 Redundancy through dominance modifiers

Perhaps the best known example of the buffering of mutation occurs upon the mutation of a single allele in a diploid organism. Fisher [19] noted that a great number of these mutations leave the phenotype unchanged. In other words, the wildtype is almost completely dominant over the deleterious mutation. Fisher proposed that this observation was the outcome of a protracted selective process, in which modifier genes evolved to increase the recessivity of mutations. Wright [73] stood in opposition to this view, proposing that dominance was a non-selected (neutral) consequence of the kinetic structure of metabolic pathways. Wright's idea can be reinterpreted as stating that enzymes have little influence on the flux through a pathway unless rate limiting. Thereby the reduction in enzyme concentration by a half upon mutation into the hemizygote, will have little effect on the pathway. This interpretation has been verified through metabolic

control theory [28]. However it is now thought that the robustness of kinetic pathways can themselves be evolved properties. Rather than thinking in terms of modifiers dampening the expression of a deleterious allele (sensu Fisher), modifiers are now thought to act on the kinetic parameters of enzymatic pathways [5].

By definition dominance is a property of diploid genomes, and thus dominance is rarely observed in short-lived populous microorganisms. However there are populous diploid organisms. In drosophila there are estimates of the average coefficients of dominance (h) for deleterious spontaneous mutations. A parameter value of h =1 signifies complete dominance, h = 0.5 signifies co-dominance. In Drosophila this is estimated at h = 0.1 [74]. In Daphnia the estimated average value of h is 0.3 [75]. Thus a range from approximately co-dominant to dominant.

5.2 Redundancy through epigenetics & imprinting

Epigenetics describes heritable changes in gene expression without changes in underlying nucleic acid sequences. Imprinting is a special case of epigenetic inheritance, and is often thought of as the expression of only one allele at a locus, dependent on the parental origin of the allele. Mutations to DNA, giving rise to repeated runs of nucleotides containing nonsense and missense mutations, arise relatively frequently through recombinational slippage. DNA-DNA pairing can detect these repeats and induce MIP (methylation induced premeiotically), by which duplicated sequences are extensively methylated leading to transcriptional gene silencing (TGS) [71]. Defective genes are no longer expressed. Imprinting plays an important role in guarding against the transformation of healthy cells into cancerous cells. Loss of imprinting is often an early step in cancer progression. Once imprinting is lost, it becomes very difficult to distinguish homologous from non-homologous chromosomes. In an attempt to repair large mutant sequences arising through 'microsattelite instability' during homologous recombination, there is inappropriate recognition, and this increases the incidence of mutation [12]. Epigenetic silencing can therefore promote redundancy by hiding the effects of mutation.

Imprinting has been reported in a number of non-mammalian groups, and by implication, in groups that tend to live in large population sizes. These groups include the yeasts, the dipterans and even in the plants: rye and maize [27]. In most of these cases imprinting is as crude as the elimination of chromosomes from one parent. However imprinting is far more common among mammals (smaller population sizes) and also more elaborate.

5.3 Redundancy through autophagy

Autophagy is a cell membrane trafficking process that occurs in response to changes in cell nutrient concentrations or specific kinds of mutation. During autophagy, cytoplasmic material is sequestered into double membrane compartments or vesicles, known as autophagosomes. These vesicles then fuse with the lysosome, which releases hydrolases breaking down vesicle contents, allowing them to be recycled [1]. The importance of autophagy in relation to redundancy becomes apparent in cancer. Transformed cells dependent on hormones for growth, are killed through autophagous processes upon the removal of hormone. Moreover, overexpression of the autophagy related gene beclin is capable of reversing the transformed state of cancerous cells and inhibiting their ability to grow [54]. In other words, autophagy is capable of breaking down and recycling the translated protein products of oncogenes.

5.4 Redundancy through mRNA surveillance

A sizeable fraction of mRNAs of eukaryotes contain premature termination codons. These arise through misincorporation errors during transcription, or derive from mutations within the DNA template. The result is the production of mRNAs encoding nonsense. These mRNAs are found to be less stable than the wildtypes as a result of 'nonsense-mediated mRNA decay' (NMD) or mRNA surveillance [51]. NMD is thought to protect cells from the deleterious effects of high concentrations of truncated proteins, reducing the number of defective mRNA transcripts prior

to translation. To date at least seven different genes have been discovered involved in NMD [9]. NMD is thus a mechanism of redundancy as it discovers and eliminates errors in the mRNA but is unable to remove defectve DNA.

5.5 Redundancy through tRNA suppresors

Whereas mRNA surveillance intercepts error prior to translation into protein, tRNA suppressors intercept errors during translation. This is achieved through a special class of modified tRNA, in which the anticodon is modified to be able to recognize a nonsense codon [13]. In the absense of tRNA suppresor molecules, termination codons within the mRNA are recognized by proteins known as releasing factors, terminating translation. In the presence of tRNA suppressors, the termination codon is bound by a tRNA suppressor and an amino acid is inserted into the growing polypeptide. In this way a nonsense mutation is transformed into a missense mutation. There exist tRNA suppressors for each type of termination codon. An interesting consequence of suppression is that the suppressor tRNAs must compete with the protein release factors for the termination codons. If suppression is too effective, then there will be extensive readthrough of the true termination codon, producing an excess of C-terminal product. This problem is overcome by making sure that the suppressors are much less than 100% effective (from 50% for amber to 10% for ochre). Redundancy is promoted through suppression by masking many of the nonsense mutations to the DNA sequence.

5.6 Redundancy through chaperones

Chaperones are proteins that facilitate the folding of nascent polypeptides. The majority of chaperones reside within the endoplasmic reticulum (ER), through which most polypeptides pass after translation in order to be folded for export from the cell or for recirculation in the cytoplasm. Within the the ER, chaperones play an essential role in protein quality control, both retaining misfolded or misassembled proteins, eliminating proteins through ER-associated protein degradation, or preventing the accumulation of unfolded proteins through the unfolded protein reponse (UPR) [18]. Retention is facilitated by an excess of glycosylation on slowly folding defective proteins acting as a signal for retention by chaperones. Degradation requires firstly the recognition of aberrant polypepetides, secondly retrotranslocation (export of the proteins from the ER back to the cytoplasm), and finally degradation of the polypeptide by proteosomes. The yeast chaperones, BiP and calnexin have been implicated in each of these pathways. Chaperones can promote redundancy by reducing the impact of mutations on protein structure. Since chaperones operate at the level of translation upwards, they can only buffer the effect of mutation, and are unable to purge mutations.

5.7 Anti-redundancy through overlapping reading frames

Single sequences of DNA or RNA encoding parts of more than one polypeptide are said to possess overlapping reading frames. In principle, three different amino acid sequences, can be obtained by initiating transcription from each of the three nucleotides constituting a single codon. This gives rise to three different readings of a genetic message, all of them out of phase with one another. Alternatively, transcription might begin in phase, but from a codon further downstream than the traditional initiation codon. Overlapping reading frames are a preferred strategy of genomic compression found among viruses, bacteria, and even some eukaryotic genes [49] [33]. Overlapping genes are of interest as they can increase mutational load: a single mutation can result in damage to more than one protein. In section 2 we introduced the multiplicative fitness landscape of the form, $w_k = (1-s_1)^k$. If we now consider a single sequence of length $2N-M$ containing two genes of length N and with an overlapping region of length M, the multiplicative fitness landscape for this sequence is rendered as, $w_k^{(o)} = \sum_i \binom{k}{i} p^i q^{k-i} (1-s_1)^i (1-s_2)^{k-i}$, where $p = \frac{M}{2N-M}$, $q = 1-p$ and s_1 and s_2 are the selection coefficients for mutation to one gene and two genes respectively, where $s_2 > s_1$. Note that $w_k^{(o)} < w_k$ and hence, on average, mutations

to genomes with overlapping reading frames will tend to be more deleterious than mutations to genomes without overlapping reading frames. This gives rise to an enhancement of point mutations consistent with anti-redundancy.

Overlapping reading frames are primarily a mechanism of genomic compression, and hence are rarely, if ever observed in eukaryotes [33]. Thus this is a mechanism that produces anti-redundancy incidentally and always in large population sizes.

5.8 Anti-redundancy through non-conservative codon bias

The genetic code gives rise to high levels of synonym redundancy. There are four nuclotides and a triplet code, whereas there are only 20 amino acids. This produces a ratio of 16:5 codons to amino acids. Assuming an equal abundance of each of the codons, and a selective equivalence or 'neutrality' of each codon, then we would expect equal frequencies of nucleotides in the genome. This is not observed, different species often have consistent and characteristic codon biases [2]. The possible causes of codon biases are numerous including translation selection for increased gene expression [38] [29], translation selection for parasite immunity [34]), and structural stability [30] weak selection and drift. One potential consequence of codon bias is to increase the rate of amino acid substitution in proteins. As an example we can consider a GC rich genome in which for each of the amino acids G or C nucleotides are used preferentially. If we consider the four codons for Serine we have TCT, TCC, TCA and TCG. In a GC rich genome TCG will be most common. In this genome random mutations are most likely to introduce Gs or Cs at each site. Given this assumption, around 100% of mutations to the first site, and around 50% of mutations to the second and third sites, will lead to a different amino acid. Thus around two thirds of mutations to the serine codon are deleterious. With equal frequencies of G, C A and T, around one half of mutations are deleterious. Thus codon bias can lead to a greater chance of a non-synonymous amino acid substitution following a point mutation, promoting anti-redundancy in the genome.

Variation in codon usage is greater among microorganisms than in mammals, birds, amphibians and reptiles (Codon usage database: http://www.kazusa.or.jp/codon/). This could reflect any number of different independent variables, including population size. In large populations greater variation is expected simply as a consequence of random mutation. This indirect effect could still effect variation in gene expression rates, and thus the level of redundancy.

5.9 Anti-redundancy through apoptotic check point genes

Apoptosis or programmed cell death describes a series of adaptive phases cells undergo, including mitochondrial breakdown, blebbing, degradation of chromatin, and membrane fragmentation, before being engulfed by phagocytic cells. The genetic and cellular cues initiating the apoptotic pathway are numerous. These include, infection, developmental signals, the removal of trophic factors, heat stress and mutation. Without apoptosis, deleterious mutations that leave cells capable of proliferation or increase the rate of proliferation, can lead to an increase in the frequency of mutant genes contained within body tissues [24]. Tumor suppressor genes or checkpint genes, such as the transcription factor P53, respond to mutation by inducing apoptosis. The range of mutations P53 is capable of responding to, includes double strand breaks in DNA, chemical damage to DNA, and DNA repair intermediates [40]. Apoptotic check point genes enhance the deleterious effects of mutation so as to increase their likelihood of being purged through local selection pressures. These are therefore mechanisms for increasing the selective cost of mutation, and are mechanisms of anti-redundancy as they remove defective cells and genomes.

Most of the apoptosis genes are confined to multicellular eukaryotes. Moreove,r within the metazoa, apoptosis is more frequently observed in large populations of cells. In populations of cells with effective population sizes approacinh zero (such as oocytes and neurons, apoptosis is almost completely inhibited)

5.10 Anti-redundancy through genetic bottlenecks

Genetic bottlenecks arise when the effective population size of a gene, or set of genes, experiences a dramatic reduction. The transmission of mitochondria through the germ line, the transmission of bacterial or viral pathogens between hosts, and the alternation of diploid and haploid generations, all lead to severe bottlenecks in the genetic variation of founder populations. The consequence of this reduction of variation is the exposure of formerly masked mutations [4]. One of the clearest examples is provided by mitochondria. Cells containing in excess of 10% wildtype mitochondrial genomes are capable of almost perfect aerobic metabolism [10]. Without a bottleneck, the binomial sampling of mitochondria to provision daughter cells, leads to few daughters with less than 10% wildtypes. Hence most daughter cells are equally fit. By imposing a bottleneck, heterogeneity in daughter cells is increased, increasing variation in metabolism, and allowing mitochondrial genomes to compete for survival [35]. Genetic bottlenecks are a mechanism for anti-redundancy as they expose and purge deleterious mutations from a population.

The extent of the genetic bottleneck is very nicely correlated with effective population sizes. In those species producing many offspring the bottleneck is minimal, whereas in species producing few offspring the bottleneck is most severe. Thus species producing few offspring make use of the abundance of their gametes to increase the efficiency of local selection pressures [35].

5.11 Anti-redundancy through inactivation of telomerase

The ends of chromosomes are capped by protective, nucleoprotein structures known as telomeres. At each cell division there is a loss of part of the non-coding repeat sequence constituting the telomeres. When telomeres are allowed to erode beyond a certain critical threshold, this leads to the proliferative arrest of mitotic cells. The erosion of the telomeres can be reversed through the action of the telomerase enzyme. In early embryonic development telomerase is active, favoring the steady proliferation of cells and the growth of tissues and organs systems. At maturity, telomerase is inactivated, imposing an upper limit on the life time of somatic cells [59]. In cancer cells, telomerase is very often over-expressed, allowing transformed cells to propagate chromosomes for an almost indefinite number of generations. The expression of telomerase effectively allows mutant lineages to increase in frequency. The loss of telomerase leads to the purging of mutant cells, and is therefore a mechanism of antiredundancy.

The repression of telomerase appears to be confined to humans and other long lived mammals. Telomerase repression is therefore a feature of long lived individuals comprising large populations of actively replicating cells. Rodents do not possess the same stringent controls on telomerase inactivation [75].

5.12 Anti-redundancy through loss of DNA error repair

Mutational damage to DNA is minimized through the actions of mechanisms that recognize changes to the genome and repair them. The mechanisms of DNA repair include excision-repair (removal of damaged regions and replacement), mismatch repair (replacing non-complementary bases in opposite strands of a double helix), and direct repair (the reversal of damage to nucleotides). The loss of one or more of these classes of repair mechanism can lead to the lethal accumulation of mutations and genetic instability of the genome [41]. In finite populations, the accumulation of mildly deleterious mutations, can lead to the eventual extinction of a lineage through Muller's ratchet [17]. The rate of the ratchet can be reduced either by increasing the efficacy of repair, or paradoxically, by eliminating some forms of repair altogether. This latter strategy is a mechanism of anti-redundancy, as it increases the deleterious effects of mutations. It has been suggested that the absence of direct repair mechanism in mitochondria are mechanism that enhance the efficiency of selection.

6 Levels of selection and the robust evolutionary individual

Natural selection operates on any entity capable of replication, assuming a reasonably stable pattern of heredity. Selected entities exist at many levels of biological organization, from short ribonucleotide sequences, through to genes, genomes, cells, organisms, and populations. These levels are called the levels of selection, and they are of interest to biology, as selection at lower levels is often incompatible with stable heredity at more inclusive levels of organization. The ways in which selection acting at lower levels creates higher levels, and the ways in which higher levels feed back to influence lower levels, have been called the 'fundamental problem of biology' [39].

The dynamics of redundancy and anti-redundancy reveal reciprocal relations among the 'levels of selection'. In multicellular organisms, rapidly dividing cells experience selection as members of a large quasispecies – much like viruses or bacteria. Each cell, bacterium or virus is in immediate competition with its neighbours for survival factors and nutrients. There is a premium on fast replication, forsaking the loss of many defective daughters. In smaller populations comprising more slowly dividing cells, robust replication is more important, and fewer cells can be sacrificed in favor of haste.

There is a potential conflict between the organismal and cellular levels of selection. Multicellular organisms living in small population sizes would benefit from a redundant (flatter) fitness landscape, whereas the abundant cells from which the organism is composed would benefit from an anti-redundant (steeper) landscape. In some cases this conflict has a synergistic resolution: anti-redundancy at the cellular level is an effective means of ensuring redundancy and robustness at the organismal level. Anti-redundant mechanisms activated in mutant or damaged cells cause their removal thereby ensuring stability (redundancy) against mutation in tissues. This coordination of interest will not always be observed. Considering only two levels of selection, there are four combinations of strategies available: (1) Redundancy at the cellular level promoting redundancy at the organismal level (for example, polyploidy), (2) Redundancy at the cellular level promoting anti-redundancy at the organismal level (loss of molecular checkpoints), (3) Anti-redundancy at the cellular level promoting redundancy at the organismal level (checkpoint genes inducing apoptosis), and (4) Anti-redundancy at the cellular level promoting anti-redundancy at the organismal level (bottlenecks in organelle transmission within and between generations).

The preferred strategy will depend upon the local population size experienced by the cell and by the organism, and any further constraints placed on their ability to replicate. Whereas bacteria and viruses replicate their genomes over a potentially indefinite number of generations, the somatic cells of many animals, are only able to replicate over a small number of generations (this is often the result of the loss of telomerase). The division of cell lines into somatic and germ line was an innovation of enormous importance for the evolution of organization [27]. It effectively handicapped selection acting at the level of cells in favor of selection acting on the germ line and somatic cell aggregate. This aggregate has come to be known as the evolutionary individual [8]. Individuals are characterized by an evolved common interest among levels (for example cell and organism). The emergence of individuals at more inclusive levels of organization has come to be known as the 'major transitions' [46].

Because our model does not separate germ-line from soma, we cannot directly address the evolutionary conflicts of interest between cells and organism. In other words, we do not consider the developmental dynamics of the individual. A thorough treatment of cancer progression would require such an approach. In cancer, mutant cells strive to increase cellular redundancy in order to mitigate the deleterious effects their mutations. These mutations impose a cost on cells by damaging the individual organism. The 'parliament of genes', coming under selection from the whole organism, seeks to promote cellular anti-redundancy so as to remove mutant cells and increase organismal redundancy (case 3). The fact that so many anti-redundant mechanisms are found to respond to cancer at the cellular level, should be viewed as a victory for the muticellular

individuals.

7 Evolving landsape parameters

7.1 Modifier models

We have spoken of redundancy and anti-redundancy in very general terms as principles of robustness. We have also spoken of the parameters of robustness in terms of a diverse set of molecular adaptations. The steepness parameter has been assumed to represent the net contribution of these mechanisms to the final plasticity of the phenotype. Selection acting on the individual must be able to modify the degree of redundancy, through modification of these, or similar mechanisms. Mutations do not only alter fitness, they also alter these mechanisms, and thereby alter the genotype to phenotype map [65]. In order to arrive at an approximately continuous change in landscape steepness, we should assume an approximately continuous degree of variation in the efficacy of mechanisms. From the list of mechanisms that we have provided this is not hard to imagine. Dominance is commonly described as varying between complete dominace through incomplete dominance to recessivity; autophagy, mRNA surveillance and tRNA suppression are stochastic phenomena, working on a variable proportion of products; reading frames can overlap to varying degrees; codon bias can be more or less extreme; and bottlenecks are of varying severity. These mechanisms are all compatible with an approximately continuous distribution of variation. We can therefore think of redundancy as the outcome of a multilocus system in which we have multiple modifiers of redundancy (chaperone genes, methylation genes, tRNA suppressors etc), and where at each locus, there are several alleles. Working out the population genetics of such a system represents an open challenge. An even greater problem is ascertaining why there should be so many different redundancy modifiers.

7.2 Evidence for distributed control

Recent work on the robustness of yeast development, and on the mechanisms of redundancy, have highlighted the distributed, multi-locus, multi-allele nature of redundancy. In yeast, the ability to buffer against the effects of gene knockout, are largely independent from the genetic distance between paralogues [64]. This suggests that gene duplication is unable to produce significant functional redundancy in yeast, and is suggestive of more distributed mechanisms for buffering. Knockout of the chaperone gene HSP90, leads to the formerly undetectable expression (cryptic variation) of polymorphisms at multiple loci [52]. Hence a single gene can conceal variation in multiple different genetic pathways. A review of synthetic lethal mutations in yeast (genes for which a double knockout is lethal), finds that any given gene has a synthetic lethal relationship with at most 26 other genes in the genome[23]. While this would suggest that genes are buffered from the activity of the majority of genes in the genome, it also shows that genes are buffered at multiple loci. The accumulating evidence paints of picture of connected modules within which there is a great deal of dependence, and among which, activity is fairly independent. Above these modules, there are shared processes playing essential buffering roles. These processes give rise to the principles of redundancy and antiredundancy. They also give us a crucial insight into biological complexity.

8 Robust overdesign and biological complexity

The study of macroscopic organization, has lead to an evolutionary worldview in which evolution produces carefully crafted, painstakingly parsimonious, and reliably robust structures. The role of evolutionary theory has been traditionally to explain these engineered properties of biological systems. A standard perspective on adaptation is the degree of agreement between a biological trait and a comparable engineered system. Bird wings and those of aircraft; eyes and the lenses of cameras. While this approach has been extremely fruitful - not least because it enables us to

understand how natural devices operate - it has lead us to neglect the baroque extravagance of natural designs. In other words, we have often ignored the important fact that so much of the natural world seems to be over-designed. When we look at the cell we do not so much think of a John Harrison clock [55], as a Rube Goldberg cartoon [70], with device mounted upon device to accomplish the simplest of tasks.

What is over-design ? In a deep sense this has been the subtext of most of our evolutionary questions: Why proteins when nucleic acids seem cable of so much function? Why diploidy when this requires twice as much resource as haploidy? Why multicellularity when unicellularity seem so efficient? Why cellular differentiation when coordinated control is so uncertain? Why sexuality when asexuality has a two-fold advantage? Why cooperation when selfishness provides higher payoffs? These are all questions in which a reasonable design solution is discarded in favor of an apparently unreasonable solution. Evolutionary biology seems to be the science of unreasonable solutions, whereas engineering is the science of reasonable solutions. This does not mean that biological systems can not be studied in terms of engineering, but that biological problems have different properties to classical engineering problems, and the issues of stability and robustness play a very central role in this difference.

Considering the list of dichotomous solutions listed in the previous paragraph and reviewing some standard explanations for the observed solutions, (1) Proteins over nucleic acids - amino acids act as cofactors increasing binding specificity ; (2) diploidy over haploidy - diploidy facilitates DNA repair; (3) Multicellularity over unicellularity - multicellularity increases control over selfish cytoplasmic elements; (4) Sexuality over asexuality - sex reduces mutational load or helps evade parasitism; (5) cooperation over selfishness - cooperation increases the fitness of aggregates. In each of these examples one factor is repeatedly in evidence: the need for stability or robustness. In other words, our canonical theories for biological organization seem to be implicitly formulated in terms of robust designs. While this comes as little surprise upon reflection, it highlights some very important notions of adaptation that have been neglected by the classical engineering schools of life.

1. Biology is deeply stochastic - no absolute zero
2. Biology is deeply historical - no tabula rasa
3. Biology is deeply conflictual - no garden of Eden
4. Biology is deeply connected - no free agents

Robustness provides a unifying thread running through all of these ideas. Consider only genetics as an example. Stochasticity (1) presenting itself as mutation, leads to the evolution of DNA repair enzymes, mRNA surveillance, tRNA suppression and checkpoints. Historicity (2) present in the canonical genetic code, leads to diverse translational strategies, preferred codon frequencies and biased amino acid usage. Conflict (3) arising from selfish, parasitic elements, leads to diploidy and parliaments of genes. Finally, the fact of networks of inter-dependence (4), leads to modularity and distributed control. In each of these cases, and having considered only the genetic level, we have observed how the notion of robustness is deeply related to the uniquely biological property of over-design, and how over-design reflects a need to incorporate redundancy and canalization. In other words, the remarkable diversity and complexity of living things, arise in part as robust and redundant solutions to instabilities that evolve alongside and above primary design goals.

9 Figure Legends

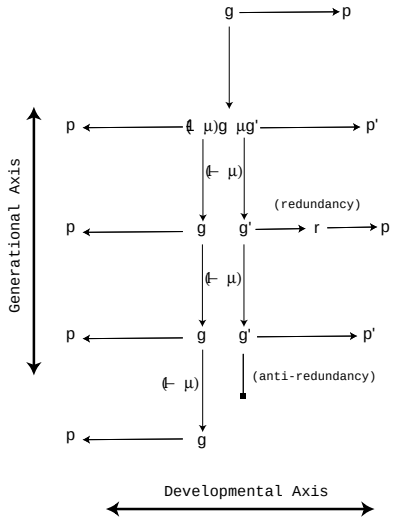

Figure 1: The principles of redundancy and anti-redundancy. A genome (g) develops into a phenotype (p). Genomes are transmitted across generations with perfect fidelity with a probability $1 - \mu$. Genomes experience mutations with a probability μ causing the wildtype g to transform into a mutant g'. The mutant genome g' develops into the mutant phenotype p'. Mechanisms of redundancy occur during development buffering the effect of heritable mutations to produce wildtype phenotypes. However redundancy does not influence the mutant genotype. Mechanisms of anti-redundancy purge mutant genomes from the population, leaving only wildtypes to replicate. The generational axis refers to iterations of genome replication, whereas the developmental axis refers to the non-replicative production of phenotype from genotype.

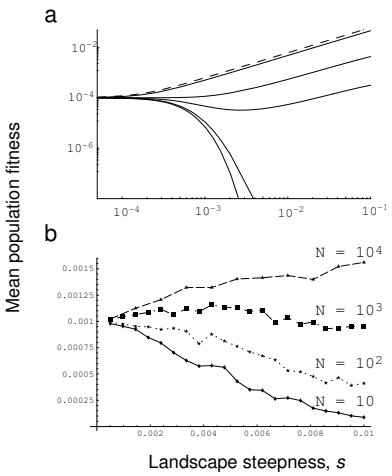

Figure 2: The theoretical relationship between redundancy and equilibrium mean fitness for populations of various sizes ($N = 10$, 100, 500, 1000, 10000, and N infinite). Small populations benefit from redundant (*i.e* flatter) landscapes, but large population prefer anti-redundant (*i.e* steeper) landscapes. The curves in the figure, given by Eqs 9, 8 and 10, correspond to genome length $L = 10^4$ and mutation rate $u = 5 \cdot 10^{-5}$. **b** The relationship between redundancy and equilibrium mean fitness as observed from individual-based computer simulations of the quasispecies equation ($L = 1000$, $u = 5 \cdot 10^{-4}$). Each individual is characterized by its hamming distance from wildtype. The mean population fitness is computed by averaging the last 20% of 10,000 generations with selection and mutation. In each discrete generation, N parents are chosen probabilistically from the previous generation according to their relative fitnesses. The offspring of a parent is mutated according to Eq. 7. The numerical studies confirm the theoretical prediction: small populations prefer shallow landscapes, while large ones prefer steep landscapes.

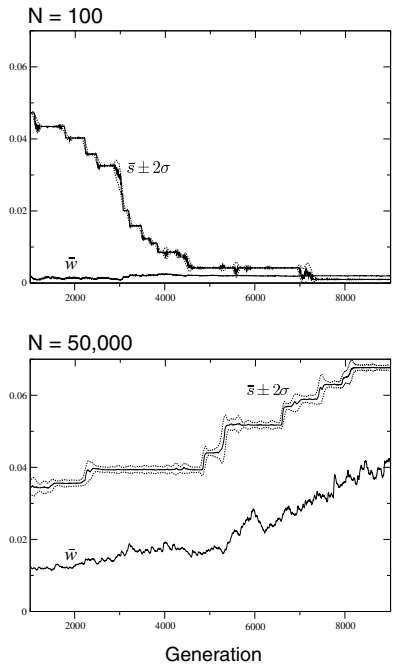

Figure 3: The evolution of redundancy in a small population (top) and of anti-redundancy in a large population (bottom). We perform quasispecies simulations in which individuals are characterized both by their hamming class k and their individual landscape steepness s. We choose genome size $L = 500$ and per-base mutation rate $p = .001$. On a slow timescale (probability .0005 per replication), an individual's landscape is heritably mutated (uniformally within $[s - .005, s + .005]$). For the small population, all individuals begin as wildtypes with a fairly steep landscape $s = 0.05$. For the large population, all individuals begin as wildtypes with $s = .025$. After an initial transient, in both cases the population's mean landscape steepness $\bar{s}$ evolves towards its preferred level. Simultaneously, the mean population fitness $\bar{f}$ increases; although it increases less dramatically for the small population. Throughout the evolutionary timecourse, the within-population variance in redundancy (σ^2) is small. In addition, over time the small population becomes de-localized ($\bar{k}$ increases), whereas the large population becomes increasingly localized (data not shown).

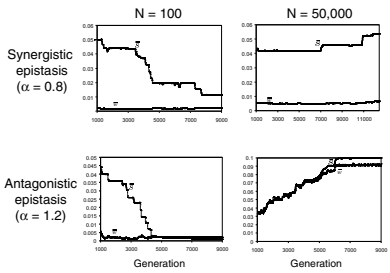

Figure 4: The relationship between redundancy and equilibrium mean fitness for epistatic landscapes. Results are derived from individual-based computer simulations as in Fig. 1b. Neither synergistic (top) nor antagonistic (bottom) epistasis alters the preference for shallow landscapes in small populations and steep landscapes in large populations. Note that the effect of population size on preferred landscape is more dramatic under antagonistic epistasis.

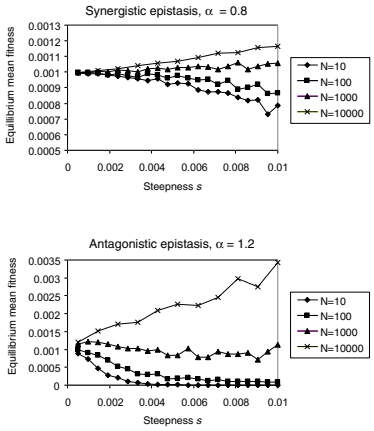

Figure 5: The evolution of redundancy in a small population (left) and of anti-redundancy in a large population (right) under synergistic (top) and antagonistic (bottom) epistasis. As in Fig 2, we perform quasispecies simulations in which individuals are characterized both by their hamming class and their individual landscape steepness s. As in Fig. 2, large populations evolve steep landscapes and small populations evolve flatter landscapes. The effect of population size is more dramatic under antagonistic epistasis.

References

[1] Abeliovich, H.,Klionsky, D. J. Autophagy in yeast: mechanistic insights and physiological function. Microbiol Mol Biol Rev **65**(3), 463-79, table of contents. (2001)

[2] Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139**(2), 1067-76. (1995)

[3] Ancel, L. W.,Fontana, W. Plasticity, evolvability, and modularity in RNA. J Exp Zool **288**(3), 242-83. (2000)

[4] Bergstrom, C. T.,Pritchard, J. Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. Genetics **149**(4), 2135-46. (1998)

[5] Bourguet, D. The evolution of dominance. Heredity **83**(Pt 1), 1-4. (1999)

[6] Brookfield, J. F. Genetic redundancy. Adv Genet **36**, 137-55 (1997)

[7] Burger, R. The Mathematical Theory of Selection, Recombination and Mutation, Wiley, New York (2000)

[8] Buss, L. W. The evolution of individuality, Princeton University Press, Princeton N. J (1987)

[9] Cali, B. M., Kuchma, S. L., Latham, J.,Anderson, P. smg-7 is required for mRNA surveillance in Caenorhabditis elegans. Genetics **151**(2), 605-16. (1999)

[10] Chomyn, A., Martinuzzi, A., Yoneda, M., Daga, A., Hurko, O., Johns, D., Lai, S. T., Nonaka, I., Angelini, C.,Attardi, G. MELAS mutation in mtDNA binding site for transcription termination factor causes defects in protein synthesis and in respiration but no change in levels of upstream and downstream mature transcripts. Proc Natl Acad Sci U S A **89**(10), 4221-5. (1992)

[11] Chosmky, N. (1981) Principles and parameters in syntactic theory. Explanations in Linguistics. Edited by Hornstein, N., and Lightfoot, D., Longman, London

[12] Cui, J., Shen, F., Jiang, F., Wang, Y., Bian, J.,Shen, Z. [Loss of heterozygosity and microsatellite instability in the region including BRCA1 of breast cancer in Chinese]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi **15**(6), 348-50. (1998)

[13] Eggertsson, G.,Soll, D. Transfer ribonucleic acid-mediated suppression of termination codons in Escherichia coli. Microbiol Rev **52**(3), 354-74. (1988)

[14] Eigen, M. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**(10), 465-523. (1971)

[15] Eigen, M., Gardiner, W., Schuster, P.,Winkler-Oswatitsch, R. The origin of genetic information. Sci Am **244**(4), 88-92, 96, et passim. (1981)

[16] Elena, S. F.,Lenski, R. E. Epistasis between new mutations and genetic background and a test of genetic canalization. Evolution Int J Org Evolution **55**(9), 1746-52. (2001)

[17] Felsenstein, J.,Yokoyama, S. The evolutionary advantage of recombination. II. Individual selection for recombination. Genetics **83**(4), 845-59. (1976)

[18] Fewell, S. W., Travers, K. J., Weissman, J. S.,Brodsky, J. L. The action of molecular chaperones in the early secretory pathway. Annu Rev Genet **35**, 149-91 (2001)

[19] Fisher, R. A. The possible modification of the response of the wildtype to recurrent mutations. Am. Nat **62**, 115-126 (1928)

[20] Fontana, W.,Schuster, P. Continuity in evolution: on the nature of transitions. Science **280**(5368), 1451-5. (1998)

[21] Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D.,Schuster, P. RNA folding and combinatory landscapes. Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics **47**(3), 2083-2099. (1993)

[22] Haigh, J. The accumulation of deleterious genes in a population–Muller's Ratchet. Theor Popul Biol **14**(2), 251-67. (1978)

[23] Hartman, J. L. t., Garvik, B.,Hartwell, L. Principles for the buffering of genetic variation. Science **291**(5506), 1001-4. (2001)

[24] Hartwell, L. H.,Kastan, M. B. Cell cycle control and cancer. Science **266**(5192), 1821-8. (1994)

[25] Higgs, P. G. Genet. Res. Cam **63**, 63-78 (1994)

[26] Hoffmann, F. M. Drosophila abl and genetic redundancy in signal transduction. Trends Genet **7**(11-12), 351-5. (1991)

[27] Jablonka, E.,Lamb, M. J. Epigenetic Inheritance and Evolution. The Lamarckian Dimension, Oxford University Press, Oxford (1995)

[28] Kacser, H.,Burns, J. A. The molecular basis of dominance. Genetics **97**(3-4), 639-66. (1981)

[29] Karlin, S.,Mrazek, J. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol **182**(18), 5238-50. (2000)

[30] Karlin, S.,Mrazek, J. What drives codon choices in human genes? J Mol Biol **262**(4), 459-72. (1996)

[31] Kimura, M. DNA and the neutral theory. Philos Trans R Soc Lond B Biol Sci **312**(1154), 343-54. (1986)

[32] Kimura, M. The neutral theory of molecular evolution. Sci Am **241**(5), 98-100, 102, 108 passim. (1979)

[33] Krakauer, D. C. Stability and evolution of overlapping genes. Evolution **54**, 731-739 (2000)

[34] Krakauer, D. C.,Jansen, V. Red queen dynamics of protein translation. J. theor. Biol (in press 2002)

[35] Krakauer, D. C.,Mira, A. Mitochondria and germ-cell death. Nature **400**(6740), 125-6. (1999)

[36] Krakauer, D. C.,Nowak, M. A. Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol **10**(5), 555-9. (1999)

[37] Krakauer, D. C.,Sasaki, A. Noisy clues to the origin of life. (2002)

[38] Kurland, C. G. Codon bias and gene expression. FEBS Lett **285**(2), 165-9. (1991)

[39] Leigh, E. G. (1999) in Levels of Selection in Evolution (Keller, L., ed), pp. 15-31, PUP, Princeton

[40] Levine, A. J. p53, the cellular gatekeeper for growth and division. Cell **88**(3), 323-31. (1997)

[41] Lewin, B. Genes V, OUP, Oxford (1994)

[42] Lewontin, R. C. The problem of genetic diversity. Harvey Lect **70**(Series), 1-20. (1974)

[43] Li, X.,Noll, M. Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions. Nature **367**(6458), 83-7. (1994)

[44] Lynch, M., Burger, R., Butcher, D.,Gabriel, W. Journal of Heredity **84**, 339-344 (1993)

[45] Maconochie, M., Nonchev, S., Morrison, A.,Krumlauf, R. Paralogous Hox genes: function and regulation. Annu Rev Genet **30**, 529-56 (1996)

[46] Maynard Smith, J.,Szathmary, E. The major transitions in evolution, W. H. Freeman, San Francisco (1995)

[47] Muller, H. J. Our loads of mutations. Am. J. Hum. Genet **2**, 111-176 (1950)

[48] Normanly, J.,Bartel, B. Redundancy as a way of life - IAA metabolism. Curr Opin Plant Biol **2**(3), 207-13. (1999)

[49] Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F. P.,Olsson, O. Overlapping genes. Annu Rev Genet **17**, 499-525 (1983)

[50] Nowak, M. A., Boerlijst, M. C., Cooke, J.,Smith, J. M. Evolution of genetic redundancy. Nature **388**(6638), 167-71. (1997)

[51] Pulak, R.,Anderson, P. mRNA surveillance by the Caenorhabditis elegans smg genes. Genes Dev **7**(10), 1885-97. (1993)

[52] Rutherford, S. L.,Lindquist, S. Hsp90 as a capacitor for morphological evolution. Nature **396**(6709), 336-42. (1998)

[53] Scheiner, S. M. Genetics and the evolution of phenotypic plasticity. Ann. Rev. Eco. Syst. **24**, 35-68 (1993)

[54] Schulte-Hermann, R., Bursch, W., Grasl-Kraupp, B., Marian, B., Torok, L., Kahl-Rainer, P.,Ellinger, A. Concepts of cell death and application to carcinogenesis. Toxicol Pathol **25**(1), 89-93. (1997)

[55] Sobel, D. Longitude : The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time, Penguin Books, New York (1995)

[56] Stearns, S. C.,Kawecki, T. J. Fitness sensitivity and the canalization of life history traits. Evolution **48**, 438-1450 (1994)

[57] Tautz, D. A genetic uncertainty problem. Trends Genet **16**(11), 475-7. (2000)

[58] Tautz, D. Redundancies, development and the flow of information. Bioessays **14**(4), 263-6. (1992)

[59] Urquidi, V., Tarin, D.,Goodison, S. Role of telomerase in cell senescence and oncogenesis. Annu Rev Med **51**, 65-79 (2000)

[60] van Nimwegen, E., Crutchfield, J. P.,Huynen, M. Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A **96**(17), 9716-20. (1999)

[61] Voelker, R. A., Langley, C. H., Leigh Brown, A. J., Ohnishi, S., Dickson, B., Montgomery, E.,Smith, S. C. Enzyme null alleles in natural populations of Drosophila melanogaster: Frequencies in North carolina populations. Procl. Natl. Acad. Sci. USA **77**, 1091-1095 (1980)

[62] Waddington, C. H. Canalization of development and the inheritance of acquired characters. Nature **150**, 563-565 (1942)

[63] Wagner, A. Redundant gene functions and natural selection. J. Evol. Biol **12**, 1-16 (1999)

[64] Wagner, A. Robustness against mutations in genetic networks of yeast. Nat Genet **24**(4),

[65] Wagner, G.,Alternberg, L. Complex adaptations and the evolution of evolvability. Evolution **50**, 967-976 (1996)

[66] Wagner, G. P., Booth, G.,Bagheri-Chaichiam, H. A population genetic theory of canalization. Evolution **51**, 329-347 (1997)

[67] Wallace, B. Fifty Years of Genetic Load, Cornell U. Press, Ithaca (1991)

[68] Whitlock, M. C., Phillips, P. C., Moore, G.,Tonsor, S. J. Multiple fitness peaks and epistasis. Ann. Rev. Ecol. Syst **26**, 601-629 (1995)

[69] Williamson, A. R., Premkumar, E.,Shoyab, M. Germ line basis for antibody diversity. Fed Proc **34**(1), 28-32. (1975)

[70] Wolfe, M. F.,Goldber, R. (eds) (2000) Rube Goldberg: Inventions, Simon and shuster

[71] Wolffe, A. P.,Matzke, M. A. Epigenetics: regulation through repression. Science **286**(5439), 481-6. (1999)

[72] Woodcock, G.,Higgs, P. G. Population evolution on a multiplicative single-peak fitness landscape. J Theor Biol **179**(1), 61-73. (1996)

[73] Wright, S. Evolution in Mendelian populations. Genetics **16**, 97-159 (1931)

[74] Garcia-Dorado, A., Caballero, A. On the average coefficient of dominance of deleterious spontaneous mutations. Genetics, **155**(4) 1991-2001. (2000)

[75] Deng, H.W., Lynch, M. In breeding depression and inferred deleterious-mutation parameters in Daphnia. Genetics, **147**(1) 147-155. (1998)

[76] Newbold, R.F. The significance of telomerase activation and cellular immortalization in human cancer. Mutagenesis **17**(6):539-50. (2002)

[77] van Nimwegen, E., Crutchfield, J.P., Huynen, M. Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A. **96**(17):9716-20. (1999)

[78] Wilke, C.O. Adaptive evolution on neutral networks. Bulletin of Mathematical Biology. **63**, 715-730. (2001)

[79] Palmer, R. Optimization on Rugged Landscapes. In Molecular Evolution on Rugged Landscapes: Proteins, RNA and the Immune System. Edited by A. S. Pereleson and S.A. Kauffman. Santa Fe Insititute Studies in the Sciences of Complexity. Addison Wesley. CA.

[80] Gillespie, J.H. The Causes of Molecular Evolution. OUP. Oxford. 1991.

[81] Kauffman, S.A. The Origins of Order. OUP. Oxford. (1993)