Epistasis can lead to fragmented neutral spaces and contingency in evolution

Steffen Schaper¹, Iain G. Johnston²,³ and Ard A. Louis¹,*

¹Rudolf Peierls Centre for Theoretical Physics,
²Department of Physics and CABDyN Complexity Centre, and
³Oxford Centre for Integrative Systems Biology, Department of Biochemistry, University of Oxford, Oxford, UK

In evolution, the effects of a single deleterious mutation can sometimes be compensated for by a second mutation which recovers the original phenotype. Such epistatic interactions have implications for the structure of genome space—namely, that networks of genomes encoding the same phenotype may not be connected by single mutational moves. We use the folding of RNA sequences into secondary structures as a model genotype–phenotype map and explore the neutral spaces corresponding to networks of genotypes with the same phenotype. In most of these networks, we find that it is not possible to connect all genotypes to one another by single point mutations. Instead, a network for a phenotypic structure with n bonds typically fragments into at least 2^n neutral components, often of similar size. While components of the same network generate the same phenotype, they show important variations in their properties, most strikingly in their evolvability and mutational robustness. This heterogeneity implies contingency in the evolutionary process.

Keywords: epistasis; neutral evolution; neutral networks; molecular evolution; robustness; evolvability

1. INTRODUCTION
The course of evolution is shaped by the complex interaction between random mutations that change genotypes and natural selection that acts on variation between phenotypes. Progress in evolutionary theory is thus predicated on gaining further understanding of the structure of genotype–phenotype (GP) maps [1]. These mappings exhibit many non-trivial properties. For example, as emphasized by Kimura [2], many mutations are neutral—they do not appreciably change the phenotype or fitness—leading to a many-to-one redundancy in the transformation from genotypes to phenotypes that has profound consequences for evolution. In the neutral theory of evolution, genetic changes that are invisible to selection [3] can build up over time and may constitute the majority of mutations in an evolutionary lineage. Evidence for the abundance of neutral mutations can be found, for example, in homologous proteins that differ in sequence, but perform the same or very similar tasks in different organisms [2].

Epistasis describes another important property of GP maps: the phenotypic effect of a genetic change at a single locus may depend on the values of other genetic loci. That such dependencies should exist is not at all surprising. Given the many multi-scale physical processes involved in translating a genotype into a phenotype, it is rather the absence of epistasis that might be expected to be the exception to the rule.

Recent advances in high-throughput techniques and in bioinformatics have facilitated many new experimental studies of epistasis. For example, Lunzer et al. [4] studied the leuB gene that codes for β-isopropylmalate dehydrogenase in both Escherichia coli and Pseudomonas aeruginosa. These two homologous proteins differ at 168 positions, but when the mutations were implemented individually in E. coli, 63 of them were found to be individually deleterious, suggesting rampant epistasis, as their overall effect is neutral. Other recent studies have found large-scale epistasis in HIV-1 virus genes [5,6], and in mitochondrial transfer RNA from eukaryotes [7]. These three examples constitute only a very small snapshot of a much larger body of literature, which suggests that epistasis is widespread throughout the living world [8,9].

The ubiquity of epistasis also implies that neutral evolution can play a key role in facilitating the genotypic background which allows evolution to climb an adaptive peak [10]: a set of mutations can be initially neutral, but when the environment or the genotype changes, they may either be adaptive themselves, or bring a population closer to potential adaptive innovations. In other words, neutral evolution may enhance evolvability, the ability of an organism to facilitate heritable phenotypic changes [11]. For example, in a recent paper, Hayden et al. [12] showed that allowing a population of ribozymes to accumulate neutral mutations greatly increased the population’s ability to adapt to a new environment, and that this enhanced evolvability could be traced to ‘cryptic’ variation that arose neutrally.

In the context of evolution, it is helpful to quantify epistasis in terms of the fitness that selection can act on¹ [9]. Epistasis manifests in many different ways. In this paper, we concentrate on just two of these. Consider, for example, a simple two-allele two-locus system with alleles a or A at locus one, and b or B at locus two. If the transition from ab to AB increases fitness then sign epistasis [13] describes the situation where either aB or Ab has a lower fitness than ab, whereas reciprocal sign epistasis [14] occurs if both
intermediate genetic states have a lower fitness than $ab$. Sign epistasis constrains the potential pathways that evolution can take towards high fitness phenotypes [15], whereas reciprocal sign epistasis is a necessary, but not sufficient, condition for peaks in fitness landscapes [16]. Even in this simple biallelic two-locus system, one can imagine other epistatic scenarios [13,14], and the potential for complexity increases greatly as more genetic loci are considered.

The considerations above frame the main question to be addressed in this paper: if epistasis constrains the pathways of adaptive mutations, can it also constrain the potential for neutral mutations to facilitate adaptation?

Although epistasis can have many different consequences for neutral evolution, in this paper, we will in particular focus on the role of neutral reciprocal sign epistasis: consider our biallelic system—if both $ab$ and $AB$ have the same fitness, but $Ab$ and $aB$ are unviable, then the only way to get directly from $ab$ to $AB$ is through double-mutations. In this context, it is helpful to define neutral networks (NNs) [17]: sets of genotypes that share the same phenotype. If we are in the regime of strong selection and weak mutation, the main case we consider in this paper, then double-mutations will be very rare. One consequence of this neutral reciprocal sign epistasis will be that an NN that contains $ab$ and $AB$ may be fragmented into separate neutral components (NCs). If the NN is fragmented into several NCs, this raises further questions like: Are these NCs homogeneous or heterogeneous? Does the potential for innovation depend on which NC a population finds itself in?

There are many potential causes of neutral reciprocal sign epistasis. For example, any mechanism that resembles a lock and a key may need two compensatory mutations, one for the lock, and the other for the key, in order to restore function. In his classic paper on compensatory mutations, Kimura [18] considered the case of two interacting amino acid sites for which a mutation in either amino acid is deleterious, but where a double-mutation can restore the function. Although these two sites are physically close in the folded state, they may be far away along the protein backbone, and it is hard to be sure that a correlated set of mutations at other positions may not allow the two sites to change by single mutational states. Thus, just as is the case for fitness peaks [16], neutral reciprocal sign epistasis is a necessary, but not sufficient condition, for disconnected NNs.

In this context, it is also important to remember that the GP map is typically characterized by very high dimensions, a property whose consequences have been of recent theoretical interest [19–21]. Briefly put, isolated fitness peaks are less likely to occur in high-dimensional landscapes; instead, long neutral ridges feature much more prominently. NNs can be identified with these ridges, and by traversing these networks, populations can explore large proportions of genotype space without having to cross fitness valleys. Similar arguments suggest that even when neutral reciprocal sign epistasis breaks a pathway between two genetic configurations, there may nevertheless be other pathways that connect up the NN. Thus, an investigation of NCs necessitates either a fairly complete description of the GP map, or alternatively, a good enough understanding of local topology to ensure that an NN is disconnected.

For these reasons, we concentrate in this paper on a computationally tractable and biologically motivated GP mapping. RNA strands can fold into well-defined three-dimensional structures driven by the specific bonding between AU, GU and GC base pairs, as well as stacking interactions between adjacent bases. The RNA secondary structure describes the bonding pattern of a folded RNA strand of length $L$. There exist efficient and reliable algorithms that predict secondary structure from primary sequence by minimizing the free energy. For the work presented here, we use the RNAFold program, v. 1.8.4 [22]. This system describes a map from a genotype of length $L$ to a phenotype that is characterized by the secondary structure. It has been extensively studied, generating many important insights into evolutionary theory [17,23–26]. The RNA map has the advantage that for modest values of $L$, one can perform an exhaustive enumeration, and from this completely characterize the connectivity of the NNs [27].

The paper is organized as follows: in §2, we establish that the NNs of most RNA secondary structure phenotypes are fragmented into disconnected NCs. We identify an important source of this fragmentation to be a particular kind of neutral reciprocal sign epistasis that arises from the biophysics of the GP map: converting a pyrimidine–purine base pair (e.g. GC) into a purine–pyrimidine pair (e.g. CG) in an RNA stem motif cannot proceed by single mutations without passing through an intermediate of a different structure. By exhaustive enumeration of length $L = 15$ RNA sequences, we can study detailed properties of the NCs. We establish that many NNs can be split into multiple components with no particular NC being dominant. We also show that the fragmentation of these NNs will be sustained under crossover moves, implying that our results may be relevant for populations in both asexual and sexual regimes.

We next examine some consequences of this fragmentation of NNs in §3. We show that the size of a given NC component correlates with a measure of its robustness to genetic mutations. Since a typical NN is fragmented into multiple NCs of different size, this implies that the robustness of a given population will depend on which NC it is on, and not only on its phenotype. Similarly, we find that the number of phenotypes accessible within one point mutation of the NCs, a measure of their evolvability [28], varies significantly between different NCs in a given NN. This heterogeneity leads us to conclude that the evolutionary fate of a population is contingent on the NC it occupies in genotype space. Finally, in §4, we discuss our main results, and look beyond the RNA secondary structure GP map to consider which conclusions may hold for a wider class of systems in including gene-regulatory networks, proteins and the genetic code.

2. RNA NEUTRAL NETWORKS ARE FRAGMENTED

The structure of NNs in RNA has been extensively studied previously [17,23,27,29]. Here, we briefly repeat some key results of this earlier work that are relevant for our investigations (see also electronic supplementary material, table S1):

— The number of structures is much smaller than the number of sequences $4^L$. The number of structures increases with $L$, but at a much slower rate than the number of sequences.
— The distribution of NN sizes is highly skewed, i.e. a minority of the phenotypes occupy a majority of the genotypes. For a given $L$, NNs with more than average size are called large, and the corresponding secondary structures are said to be frequent. While the absolute number of frequent structures increases with $L$, the fraction of sequences folding into frequent structures goes up, while the fraction of NN’s that are large goes down.

— The fraction of sequences that fold into the trivial structure (that is the structure that has no bonds) decreases with $L$.

The connectivity of NNs can be studied under the simplifying assumption that a network is made up of randomly chosen points on a genotypic hypercube. Analyses using graph theory then suggest that larger NNs are likely to be fully connected, while small ones are likely to be fragmented [21,30].

However, for RNA secondary structure, it is important to also take the biophysics of bonding into account. In principle, each bond can be formed by one of six different nucleotide pairs: GC, CG, AU, UA, GU and UG. Point mutations can potentially connect GC ↔ GU ↔ AU and CG ↔ UG ↔ UA, but these two subspaces cannot be connected together by point mutations without breaking a bond. This type of neutral reciprocal sign epistasis suggests that for a structure with $n$ bonds, we can expect the order of $2^n$ disjoint sets of compatible sequences. This argument is independent of sequence length, and given that longer sequences may generate more structures with more bonds, we expect the average number of NCs per NN to grow with $L$ (see electronic supplementary material, table S2).

We therefore predict that virtually all NNs for RNA secondary structure should be fragmented. By contrast, if double-mutations (base-pair swaps) are allowed, then the results from random graph theory give a good estimate of the connectivity of an NN [30]. But, in nature, base-pair swaps are expected to be very rare [7]. Although the fact that RNA secondary structure NNs are not fully connected has been widely acknowledged in the literature [27,29,31], the potential consequences of this fragmentation have not yet been fully explored.

In order to determine the connectivity of the NNs, we start from a random sequence in the NN and follow all neutral mutations that can be accessed [27]. Sequence space grows exponentially with length $L$, so that this exhaustive approach is feasible only for relatively short sequences; we will mainly present results for sequence length up to $L = 15$, but will also consider other lengths where appropriate. As has been done in many other studies [24,29], we ignore the trivial structure with no bonds. Our $L = 15$ system has 431 distinct secondary structures (at a folding temperature of $37^\circ C$) of which 86 or about 20 per cent are large. The large structures cover 93 per cent of the folding sequences. By exhaustively searching through all of the sequence space, we are able to identify all 12526 components, so that there are on average about 29 components per neutral network. Figure 1 shows how these are distributed among the different NNs. The largest number of components is 216 for a relatively infrequent structure ranked 206th, and only a few small structures have a single NC (the largest has rank 333). We summarize the data in electronic supplementary material, tables S1, S2 and figures S2, S3. The NCs can be individually ordered, and are even more skewed than the NNs. Overall, 1120 NCs (less than 10%) are larger than average, but together they cover 95 per cent of non-trivial genotype space.

By analogy to NNs, we call an NC large if its size is more than the average in its NN. Most NNs contain several large NCs (see electronic supplementary material, figure S8 and table S3).
Robustness increases with component size

The robustness to genetic change has been widely studied in the context of NNs. In particular, van Nimwegen et al. [31] have shown how the robustness of an evolving population depends on the structure of the underlying neutral space. While the dynamic properties of a population also depend on its size and mutation rate, we consider here only the effect of the structure of the NC. To this end, we define the mutational robustness of a genotype as the fraction of mutations that leave the phenotype unchanged. In analogy to the study of Wagner [28], we calculate the robustness of an NC by averaging the genotypic robustness of all genotypes in the NC. This measure gives the expected average robustness of a monomorphic population evolving on the NC [31].

In agreement with earlier results based on sampling techniques [28], we observe a clear positive correlation between mutational robustness of an NC and its size ($r = 0.47$), as illustrated in figure 3. Hence, the larger the NC, the more likely individuals are to pass their phenotype on to their offspring after a random mutation. Given the large heterogeneity of NC sizes comprising a given NN, these results suggest that robustness estimates based on the NN as a whole will not be representative of the robustness experienced by a population confined on a given NC. For example, if a population is restricted to a small component of a very large NN, the effective mutational robustness will be (much) lower than that estimated for the NN as a whole.

(b) Evolvability varies between components of the same phenotype

The evolvability of a population is related to its ability to produce heritable phenotypic change [11]. One might naively think that the more robust a phenotype is to mutations, the harder it is for mutations to generate novelty. However, this argument ignores the ability of neutral exploration to pave the way for future adaptive innovations [10]. Wagner has proposed a proxy measure of evolvability that counts the number of phenotypes $\mathcal{E}_P$ that can be reached by a single mutation from a given NN [28]. He showed that this measure also correlates positively with the size of an NN, and argued that phenotypes with larger NNs may be simultaneously more robust and more evolvable. But if the NNs are fragmented into separate NCs, then it is in fact the NC robustness and evolvability that matter to a population, and not the properties of the whole NN, which it cannot access by point mutations. Nevertheless, we find that the average robustness and evolvability of individual NCs are positively correlated (figure 3 and electronic supplementary material, figures S10 and S11) just as was found for NNs. We can thus qualify Wagner's result [28]: robustness and evolvability are not so much correlated at the level of phenotypes (NNs), but rather the correlation holds at the level of an NC (our results yield $r = 0.81$), which can vary strongly within one NN.
Figure 4. Phenotypic neighbourhoods are heterogeneous among NCs in the same NN. (a) Here, the joint and common evolvability are shown considering all NCs of each NN. Square markers indicate fully connected NNs, for which joint and common evolvability trivially coincide. (b) Here, only the large NCs in each NN were used for the calculations. Now square markers indicate NNs with only one large NC, for which again the equality is trivial. In both panels, the black-dashed line indicates the equality of joint and common evolvability. Grey diamonds, rare NN; black diamonds, frequent NN.

Thus, different populations with the same phenotype may exhibit significantly different evolvabilities. These differences can be further quantified with the following definitions:

\[ E_{c}^{(j)} = \bigcup_{c \in P} E_{c} \]  \hspace{1cm} (3.1)

and

\[ E_{c}^{(c)} = \bigcap_{c \in P} E_{c} \]  \hspace{1cm} (3.2)

where \( E_{c} \) is the set of phenotypes that can be reached by a single mutation from NC \( c \). Thus, the joint evolvability \( E_{c}^{(j)} \) counts the number of structures that can be reached from at least one NC in the NN, while the common evolvability \( E_{c}^{(c)} \) counts the structures available from all NCs. The comparison of these two properties reveals significant heterogeneity in the phenotypic neighbourhoods of NCs in the same NN (figure 4a). In fact, the joint and common evolvability are only identical for those small NNs that are fully connected. For most NCs, a population will only be able to access a restricted subset of the entire NN’s neighbouring phenotypes: averaged over all NNs, \( F = E_{c}^{(c)}/E_{c}^{(j)} = 0.14 \). There is no significant correlation of \( F \) and NN size: \( r = 0.04, p = 0.41 \).

We can further explore this heterogeneity by restricting our analysis to the large NCs only. We expect the differences to diminish because the joint evolvability has a lower bound given by the most evolvable NC while the common evolvability cannot be larger than for the least evolvable NC, which is typically very small. As figure 4b shows, the ratio of common to joint evolvability decreases when only large NCs are taken into account: \( F_{\text{large}} = 0.37 \) with a weak correlation with NN size: \( r = 0.17, p = 4.5 \times 10^{-4} \). We note that some phenotypes are only accessible from small NCs so that the joint evolvability decreases slightly (on average by about 10%). In electronic supplementary material, figure S12, we restrict the phenotypes further to just those that are large—the same general results hold. Finally, we can ask what fraction of the joint evolvability is accessible on average from a single NC. If we consider only the large NCs of the frequent phenotypes, this fraction is on average 76 per cent (in agreement with [29]), while averaging over all NCs in all NNs brings this down to 42 per cent (see also electronic supplementary material, figure S13). Instead of requiring large NCs to be greater than the average NC in their NN, we also employed an entropy-based criterion and obtained qualitatively similar results (electronic supplementary material, §S3.3 and figure S14).

It is important to consider whether this discrepancy is an artefact of the relative short sequences that we study. Answering this question by exhaustive enumeration is unfeasible. Instead, we employed a sampling technique (electronic supplementary material, §S4) for sequences of 20 nucleotides. We find that the heterogeneity between NCs becomes even more pronounced as the sequence length increases (electronic supplementary material, figure S15).

Taken together, we have arrived at a key result: the potential for future innovation does not only depend on the current phenotype, but also on which NC a population occupies. The fact that different NCs provide access to different new phenotypes suggests a new mechanism for contingency in evolution. A dynamic setting in which this may be particularly important is a polymorphic population with genotypes from two (or more) NCs. If environmental changes are sufficiently rapid (that is faster than genetic drift), this could drive parts of the population to different phenotypes, potentially aiding diversification at the phenotype level [21].

4. DISCUSSION

We have shown how neutral reciprocal sign epistasis in RNA leads to fragmentation of NNs into multiple components. For many of the NNs, no one component dominates. Moreover, the components are heterogeneous, so that different populations with the same phenotype, but different NCs, may show large variations in robustness and evolvability.

These inferences were possible because of the tractability of the GP map between an RNA sequence and its secondary structure. An obvious question is whether our results extend to other maps. Boldhaus & Klemm [32] studied a coarse-grained Boolean threshold dynamics model [33] for the regulatory network of the
yeast cell cycle and identified nearly half a billion functional NCs, ranging in size between $6.1 \times 10^{24}$ and $4.4 \times 10^{26}$ genotypes. Interestingly, the wild-type network is part of one of the smaller NCs. It contains networks which are quite sparse and noise-resilient, indicating that there are secondary aspects in the performance of the network which can be selected for. This example also shows heterogeneity in the properties of NCs. One caveat is that the point mutations were in an abstract space with discretized interactions. It is not yet clear how a more realistic model of mutations would affect the NCs.

Recent experimental reconstructions of fitness landscapes may also open up avenues to study NCs. For example, in an important paper, Weinreich et al. [15] characterized all 32 combinations of five mutations that together increase resistance to a particular antibiotic by a factor of about $10^7$. By measuring the resistance of each possible combination, they produced a phenotype landscape; in examining their data, we found that this landscape also contains several NCs (see electronic supplementary material, figure S16).

There are two important caveats to this finding: first, the resistance scale used in the experiment is relatively coarse. Thus, the neutrality in this landscape may even be broken by relatively small populations. In general, we stress that neutrality is always an effective statement, depending on the population size [3]. Secondly, we cannot exclude the existence of neutral connections that were outside the scope of the experiment. Excluding such paths by exhaustively cataloguing all possible mutations would be prohibitive. Progress can be made by studying the biophysics of a GP map, and looking for examples of lock and key type systems. In proteins, binding sites may be potential candidates [18]. However, as reviewed by Poelwijk et al. [14], even lock and key systems can sometimes evolve in subtle ways through single mutations.

Another system to consider is the genetic code. It is interesting to note that with the exception of serine, all sets of codons coding for a particular amino acid can be reached by single synonymous point mutations. However, serine has two NCs, one made up of AGU and AGC, and the other of UCU, UCC, UCA and UCG. Given that serine often plays a key role in active sites in proteins, it may be that it cannot easily be neutrally replaced by another amino acid, so that these two NCs may indeed be separate in nature. It is noteworthy that this high NN connectivity is extremely unlikely to arise in a random genetic code with the same degeneracy as the universal code (electronic supplementary material, figures S17). As robustness correlates with NC (and not NN) size, this striking degree of NN connectivity may be a by-product of selection for other properties such as robustness of the genetic code to point mutations or translation errors [26].

We have focussed on the approximation of strong selection and weak mutations where double-mutations are excluded. However, compensatory mutations can occur if the fitness penalties are weak, or if mutation rates are high. Measuring fitness is notoriously difficult, but a recent study of compensatory mutations in mitochondrial transfer RNA estimates that transitions from GC to AU may occur through low-fitness GU and AC intermediates [7]. By contrast, switches like AU ↔ UA, GC ↔ CG and AU ↔ CG, each of which requires a transversion, were found to be very rare. These results suggest that in nature, we should expect fragmented neutral spaces in RNA to be common.

Nevertheless, a sufficiently large population and/or high mutation rate can lead to a regime in which NCs are effectively connected, so that evolutionary dynamics may be less sensitive to the effects of NN fragmentation. In the opposite limit of small populations and/or mutation rates, the average spread of a population in genotype space can be much smaller than the size of many NCs, implying that the local NC structure becomes more important. For a fixed mutation rate, there will thus be a crossover in the effect of NCs on evolutionary dynamics with increasing population size. More generally, the dependence of evolvability and neutral space exploration on dynamic parameters is an important issue which we plan to address in a future publication.

Our analysis has only considered the local phenotypic neighbourhood of individual NCs. Over longer evolutionary timescales, populations evolve from one phenotype (and hence NC) to another and traverse the phenotypic landscape. In order to understand the importance of landscape structure on such long timescales, it is necessary to study not only accessible phenotypes, but also the connectivity among NCs, which will be the focus of future studies.

In conclusion then, we have focussed on one striking effect of epistasis on neutral evolution, namely the fragmentation of neutral spaces. The heterogeneity of the resultant NCs is important both conceptually and in practice: properties such as the robustness and evolvability of an evolving population may not only depend on its phenotype, but also on which NC of that phenotype the population occupies. This sensitivity may lead to contingency in evolution: the evolutionary trajectory of a population depends not only on the occurrence of random mutations, but also on the possible innovations that are available to the NCs it happens upon.

ENDNOTES

1 In this paper, we will ignore the potentially very complex mapping from phenotypes to fitness, and simply assume that fitness can be ascribed to and differs between phenotypes.

2 There may be other causes of fragmentation, and bonds vary in energy, so not all possible combinations will lead to the same secondary structure.

3 In systems where folded structures have an adaptive advantage, it is likely that the completely unfolded strand has very low fitness, and so can be ignored. There is also a practical reason for this choice. The trivial structure is much more frequent for small $L$ than for large $L$, and so it could affect the applicability of our results for much longer structures.

REFERENCES


Epistasis, neutral spaces and contingency  S. Schaper et al.


