Emergent gene order in a model of modular polyketide synthase system.

Ben Callahan, Mukund Thattai, Boris Shraiman

December 2, 2008

Abstract

Polyketides are widespread biologically active heteropolymers produced by ‘assembly line’-like multi-protein complexes formed by modular polyketide synthases (PKS). The nature of the polyketide product is encoded in the order of particular types of PKS proteins along the assembly line, suggesting that the diversity of polyketides derives from the combinatorial rearrangement of complexes formed by PKS modules. Remarkably, the order of PKS genes on the chromosome follows the order of PKS proteins along the assembly line: this fact is commonly referred to as ‘colinearity’. Here we propose a possible evolutionary origin for this colinearity and demonstrate the mechanism using a computational model of PKS evolution in a population. Assuming that polyketide biosynthesis is under a continuous selective pressure to evolve new products and that new polyketide pathways are formed through horizontal transfer and recombination of PKS encoding DNA, we demonstrate the existence of a broad range of parameters for which colinearity is expected to emerge. Although in our model colinearity confers no direct fitness advantage it is established and maintained through a ‘secondary’ selection mechanism, as a trait which increases the probability of formation of long functional PKS assemblies through recombination. Consequently, colinearity ‘hitch-hikes’ on the successful genotypes which periodically sweep through the continuously evolving population. In addition to the computer simulation of the simple model of PKS evolution we provide a mathematical framework describing the secondary selection mechanism, which generalizes beyond the context of PKSs.

1 Introduction

Polyketides are a class of structurally and functionally diverse heteropolymers found in bacteria, protozoa, plants and animals. In single-celled organisms these secondary metabolites mediate a variety of interactions between cells and their environment [22]; as channels of cell-to-cell communication between conspecifics [1], as anti-microbial agents against competitors, and as immuno-suppressors or virulence factors [6] between pathogens and their hosts. The enormous diversity of natural polyketides could be a result of a ‘chemical arms race’ during
inter-species and host-pathogen conflict. Alternatively, the ability to generate chemical diversity might be an end in itself, increasing the likelihood of discovery of biologically potent molecules [4].

In bacteria, the chemical diversity of polyketides is achieved through a unique combinatorial biosynthesis mechanism. A large class of polyketides are generated by ordered complexes of modular polyketide synthase (PKS) proteins via the step-by-step polymerization of acylthioester monomers such as malonyl-CoA and methylmalonyl-CoA [23]. Each step of chain extension is performed by a single PKS catalytic module, with different classes of modules adding and specifically modifying different monomer building blocks. PKS proteins consist of three obligatory catalytic domains involved in sequential elongation of the nascent chain. This ‘minimal’ PKS leaves the added subunit as an keto group. Other types of PKS involve up to 3 additional catalytic domains (keto reductase, dehydratase, enoyl reductase) which act sequentially to modify the chemical form of the added subunit. A number (up to around 15) of different PKS proteins can form a linear multiprotein complex, held together by specific interaction between the C- and N-terminal docking domains [7, 28, 3, 27] as shown in Figure 1. The order of catalytic modules in this complex thus determines the chemical structure of the polyketide chain. This structure means that combinatorial sets of four basic types of PKS assembled in a defined order are, at least in principle, capable of synthesizing any of the \(4^L\) different polyketide products (with \(L\) being the number of PKS in a linear complex. The viability of these combinatorial rearrangements depends crucially on the observed substrate tolerance of catalytic modules to accept and extend a wide range of precursors [23]. Experimental efforts to explore combinatorially generated polyketides have had success and are being actively investigated [16]. The mechanism and modularity of the polyketide biosynthetic system suggests that combinatorial exploration of biochemical space may be the integral element of its evolution, facilitated by recombination and horizontal gene transfer (HGT) [27].

Evidence from comparative genomics suggests that gene duplication, HGT and homologous recombination have played key roles in the evolution of bacterial PKS gene clusters [11, 22]. Incongruities between the phylogenetic trees of bacterial species (determined by 16S ribosomal DNA) and of iterative PKSs in their genomes (closely related to modular PKSs) strongly suggest a history of HGT in these systems [19]. Modular PKS genes on bacterial chromosomes are invariably found in giant clusters, probably due to a combination of three factors: in-situ gene-duplication [11, 22], HGT-driven clustering, as predicted by the ‘selfish operon’ hypothesis [13], and the need for transcriptional co-regulation [21]. Since PKS protein domains from different organisms retain a high degree of sequence identity, homologous recombination is expected to drive the shuffling and swapping of PKS genes between multiple DNA strands. Several examples of HGT, as well as of gene-swapping due to homologous recombination, have been inferred from sudden transitions in sequence identity along PKS gene clusters [22]. PKS gene clusters appear to be extremely dynamic, but on timescales far beyond those we can directly observe, so the processes driving their evolution...
can only be studied by indirect signatures. We focus our attention on one particularly informative signature: the ordering of genes in a PKS cluster.

Gene order is generally conserved between closely related bacterial species, but this conservation is rapidly lost as the species diverge, even for genes within individual operons [10, 25]. Typically gene order is selection-neutral, so the degree of conservation can be used to estimate phylogenetic distances [26]. In the case of physically interacting proteins, however, it is seen that gene order is conserved even over relatively long timescales [10]. This effect holds in PKS systems, a majority of PKS proteins which physically interact are encoded contiguously [27], so the order of proteins in a PKS complex closely matches the order of their genes on the chromosome. When applied to PKS gene clusters, this is sometimes described as the ‘co-linearity rule’ [20].

We propose that HGT and homologous recombination are the main drivers in the combinatorial search for novel polyketide products. We suggest this hypothesis explains the observed ‘colinearity rule’ of modular PKS gene order. Below we use a theoretical model of PKS pathway evolution under continuous selection pressure to demonstrate how colinearity could arise and be maintained in a population. Due to a lack of direct data bearing on the process of PKS evolution, our model will be as simple as possible while retaining the key modular and combinatoric aspects of PKS pathways. We shall assume PKS evolution to be driven by a Red Queen-type [?] paradigm of continuous ‘arm race’ between polyketides and the environmental pressures they ameliorate. Thus, the selective advantage of a particular polyketide product decays with time, opening an opportunity for a PKS pathway producing a novel advantageous polyketide to sweep the population. In our model PKS pathways are produced by recombination-induced shuffling of PKS genes. Although in our model gene order has no impact on individual fitness, we see that it emerges and is maintained by what one may call a ‘secondary selection’ mechanism. Colinearity facilitates creation of long new PKS pathways (via recombination) and is established and maintained by ‘hitch-hiking’ on the successful pathway genotype it has helped create. Colinearity of PKS gene order is a genetic trait which in effect enhances the population’s environmental response as measured by the long-term average fitness. Below we shall define the model and use numerical simulations to determine a ‘phase diagram’ which defines the conditions under which secondary selection maintains colinearity, as a function of key parameters such as rate of HGT/recombination and the characteristic time beyond which fitness advantage conferred by a given pathway is lost. We shall then provide a probabilistic description of the dynamics which quantitatively explains the action of secondary selection on colinearity.

2 The Model

Let us consider a simplified model of the modular PKS system that retains three key features: a) modularity and combinatorial diversity; b) possibility of constructing new pathways by recombination of PKS genes; c) continuous selective
pressure favoring emergence of novel products. In our model, PKS proteins consist of three regions: head and tail docking domains and a central catalytic module occurring in two flavors (see Figure 2). The gene is arranged likewise, with head and tail domains flanking the catalytic module. The restriction in ‘flavors’ of catalytic modules allows the products to be represented as binary strings, each bit representing the ‘monomer’ at that position, as is seen in Figure 2b. This reduction in the number of flavors is not essential and chosen solely for computational convenience. The head and tail domains \( H_i, T_j \) are drawn from a set of \( N_c + 1 \) different classes with \( N_c = 15 \) in simulations shown. Binding is exclusive with the corresponding domain of the same class (i.e. protein ‘i’ binds to protein ‘j’ only if \( H_i = T_j \)) and the \( H_i = 0 \) and \( T_j = 0 \) is a special terminator class which does not bind to anything.

In the model, individuals have circular chromosomes (or "plasmids") with \( L \) PKS genes. We will consider haploid populations of \( N \) such individuals propagating in discrete, non-overlapping generations. (Simulations shown below were done with \( N = 1000 \) and \( L = 12 \).) Simulated population dynamics is based on selecting individuals for the offspring generation with the probability proportional to their fitness (BEN: IS THIS HOW YOU DO IT?? IS THERE A REASON WHY YOU DON’T USE F-W SAMPLING? WHAT’s THE STORY WITH POPULATION SIZE?? ). Recombination is implemented by introducing a probability \( r \) for each member of the offspring generation to recombine with another member so chosen. Recombination is reciprocal and homologous in the sense that exchanged segments are the same length and begin and end with the same genetic region, as seen in figure 2a. This ensures that the size and structure of chromosomes is constant under recombination. All relative rotations of recombining chromosomes are equally likely.

We define the fitness of an individual is terms of a sum over the contributions of all possible products each corresponding to a possible chain assembly of the \( L \) model PKS proteins. Specifically: \( f = 1 + \sum_k C_k \Delta f_k \) where \( k \) labels different products, each contributing fitness \( \Delta f_k \) weighted by the probability \( C_k \) that the complex which catalyzes its production is fully constituted. \( C_k \) is determined on the assumption that same class head/tail pairs bind with equal probability to all available partners. Figure 2b shows the product polyketides and associated concentrations of a sample individual. There is a special case when complexes recursively ‘loop’, resulting in indefinitely long PKS chains: we chose to suppress such situations by assigning them zero fitness. The choice of \( \Delta f_k \) should reflect the reasonable expectation that strongly beneficial products must have sufficient complexity, e.g. must be sufficiently long. In principle there would be a certain probability of finding a product of length \( L' \) with fitness above any given value. Here, we will consider only a very simple fitness landscape, all polyketides of length \( L^* \) have initial fitness effect \( s \) and no other polyketides affect fitness, \( \Delta f_0(k) = s\delta(|k| - L^*) \). In simulations shown below \( L^* = 7 \) and \( s = 0.1 \) but results are qualitatively similar for different \( L^* \), \( s \). We shall return to the subject of the choice of the fitness function in the Discussion section.

To facilitate generation of novel pathways as expected in a constantly changing environment, we make fitness assigned to a given product decay with time.
Thus, when it first appears in the population "polyketide" $k$ has initial fitness contribution $\Delta f_k(0)$ which subsequently decays exponentially with a time constant $\tau$ so that $\Delta f_k(t) = \Delta f_k(0)e^{-t/\tau}$. This is a result of the environment changing in time and we call $\tau$ the "environmental decorrelation time". Choosing this form of selection pressure is motivated by thinking of an antibiotic polyketide: as bacteria in the environment gain resistance to a particular antibiotic the fitness benefit of that antibiotic decreases. Since all fitness contributions from polyketide products present in the population decay with time there is constant pressure to find novel, and therefore higher fitness products.

In order to define gene order as a quantitative trait we introduce a quantitative measure of colinearity. It is natural to define colinearity based on the mean distance between DNA positions of interacting head/tail pairs on the chromosome. For computational convenience and to avoid introduction of unnecessary parameters, here we shall measure this distance in terms of the number of PKS units separating an interacting head/tail pair (i.e. $H_i = T_j$), which we denote by $\text{dist}(H_i, T_j)$. Thus a head/tail pair that are adjacent on the chromosome are at distance zero. Explicitly, the average head/tail pair distance, normalized to the total number of PKS units is:

$$\overline{d} = \frac{\sum_{i,j} \delta_{H_i T_j} \text{dist}(H_i, T_j)}{L \sum_{i,j} \delta_{H_i T_j}}$$

If chromosomal order of PKS modules follows their order in the pathway, their distance from each other is zero and $\overline{d} = 0$ which corresponds to perfect colinearity. More generally, the higher $\overline{d}$ is for an individual, the less colinear is that individual’s chromosome.

**BEN:** I DON’T QUITE UNDERSTAND WHAT YOU’RE DOING WITH ACTUALLY POSITIONS ALONG THE CHROMOSOME, YOU HAVE NOT DESCRIBED THE PLACEMENT OF THE PKS ON THE CHROMOSOME, OR THEIR LENGTH, ETC. I ASSUME YOU’RE DOING BUSINESS IN TERMS OF DOMAIN UNITS WITH TOTAL LENGTH "L"?

Without selection pressure favoring particular gene order one expects recombination to randomize gene order: i.e. make all possible rearrangements of a given set of PKS genes equiprobable. In this case distance between any head/tail pair will be equidistributed (among integer values) between 0 and $L/2$ and the distribution of $\overline{d}$ in the ensemble of possible "chromosomes" will (given sufficiently large $L$) approach a Gaussian centered on $1/4$. It is useful to define the measure of colinearity, $y$, of the genomic PKS arrangement as the deviation of $\overline{d}$ from the mean expected in a random ensemble:

$$y = \frac{1/4 - \overline{d}}{1/4}$$

**CHECK** This value lies on the interval $[-1, 1]$. The density of states in the ensemble of randomly ordered genes $\rho(y)$ is approximately Gaussian centered at $y = 0$ with the variance decreasing as the number of head/tail pairs increases. The case of an individual $\overline{d} = 7$ encoding one $L^* = 7$
polyketide is pictured in figure 6. When $y > 0$ the interacting head/tail genes are closer than expected if arranged randomly. Hence, when we refer to a population exhibiting colinearity we mean that the population average colinearity ($y$) is greater than zero.

3 Results

Model dynamics exhibits three fundamental regimes of behavior. One is a continuously evolving, Red Queen-type state characterized by high average fitness and high colinearity. The other is a “quiescent” state, where population can not keep up with the continuous selection pressure and falls into a state with low fitness and no colinearity. In addition, if the continuous selection pressure to evolve is turned off by removing the time dependence of fitness, we observe a “static” high fitness state with no colinearity. Figure 3 shows population trajectories typical of these behaviors. The evolving Red Queen (RQ) state exhibits a characteristic saw-tooth behavior of the population-averaged fitness (Fig 3a). This fitness trajectory is produced by the successive selective sweeps of individuals encoding novel $L^*$ polyketides. After a sweep the population is dominated by individuals expressing this high fitness $L^*$ product. Subsequently the population average fitness decays exponentially following the decay of fitness benefit due to this polyketide. The decrease in population average fitness increases the relative fitness advantage of individuals expressing novel recombinant $L^*$ pathways, increasing the chance for their progeny to sweep the population. If a new beneficial genotype avoids extinction, selective sweep occurs, and the cycle continues.

The exponential decay of fitness after a sweep leads to arbitrarily small selective pressure if enough time passes without a novel polyketide being found and swept. When selection becomes negligible the population dynamics changes, drift becomes dominant. In a selection-free, drift-dominated population the encoded PKS pathways, and hence product polyketides, are broken into smaller and smaller fragments by recombination. The end result of this is the quiescent (Q) state, a population state in which individuals have few if any interacting head/tail pairs and produce only very short polyketides. In our finite population this state is essentially permanent, it becomes probabilistically difficult to create $L^*$-long PKS chains out of the short fragments remaining in the population. In figure 3b we see a population undergo this transition. To the extent that the quiescent state is an “absorbing” state in a finite population there is always a chance that the RQ state will fail to generate a sweep before recombination and drift destroys ordered PKS chains within individuals driving the causing the population to be trapped in the Q state, however, as we shall see below provided that the rate of recombination falls in a certain range, this probability decreases exponentially with $N$.

The case of an unchanging environment, $\tau \to \infty$, serves an important role for comparison. As we might guess, the dynamics in this case are relatively static. Drift is the dominant mode of change in population composition, but
the constant selection prevents any quiescent-like state developing. Drift occurs among individuals expressing the current \(L^\ast\) polyketide, transitions to other \(L^\ast\) polyketides are very rare without the incentivization present in a changing environment. The only relevant fitness effect is the recombination load. Most recombination events break up the genes responsible for producing the \(L^\ast\) polyketide resulting in a low fitness recombinant. Accounting for this we should have a population average fitness of approximately \(\langle f \rangle = (1 + s)(1 - r)\) consistent with what we observe in figure 3c.

Colinearity emerges spontaneously and is maintained in the evolving, RQ, phase. Figure 3 shows, in addition to fitness, the population average colinearity as function of time. The evolving population in Fig. 3a maintains colinearity through repeated selective sweeps. Excursions to non-colinear genetic realizations do occur, but are temporary and long-term time averages of colinearity in the RQ phase are significantly greater than the \(y = 0\) expected in an equiprobable ensemble.

In contrast to the RQ state, quiescent state populations do not generate novel genotypes fast enough to keep up with the continuous degradation and slip to a much lower level of average fitness. Quiescent state also can’t maintain colinearity. In figure 3b we see pre-existing colinearity decaying away after the population transitions into the quiescent state. Once gone it does not return and long-time average of the colinearity in the Q state is zero. Maintenance of colinearity requires the population to be repeatedly swept by high fitness genotypes, which on average are more colinear than random. In a quiescent state, the population is unable to assemble novel high fitness pathways on the time scale of \(\tau\), repeated sweeps do not occur and any colinearity that may be initially present eventually disappears because of genome reshuffling by recombination. A similar situation arises in the static state where selective pressure maintains one particular high fitness \(L^\ast\) pathway, but, because its fitness does not decay with time there is no pressure for generation of new products. In that case, reshuffling of PKS gene order has no effect and colinearity decays to zero.

The two key parameters which control the behavior of the system are the recombination rate \(r\) and the environmental decorrelation time \(\tau\). In figure 4 we display time-averaged fitness and colinearity of populations evolved under a range of \(r, \tau\). There is a clear separation into two regimes: a high fitness, high colinearity regime corresponding to the evolving RQ behavior, and a low-fitness non-colinear regime corresponding to the quiescent state. A basic understanding of the RQ regime’s boundaries can be gained by considering the dual effects of recombination in this system. First there is the simple recombination load, relevant at higher recombination. When an individual encoding a novel polyketide is formed, its spread is hindered by recombination, which generally destroys the phenotype. This introduces a requirement that \(r < s\) for a novel individual to have a non-zero chance of sweeping, where \(s\) is the largest available selective advantage. If novel recombinants cannot sweep, evolving behavior cannot be maintained. This immediately restricts the evolving regime and is seen as the boundary near \(r = s\).

On the other hand, recombination in our system is the sole source of pheno-
typic novelty. The persistence of evolving behaviors requires that novel individuals encoding $L^*$ products be created in sufficient number to escape stochastic extinction and sweep before drift forces the population into the quiescent state. Thus the population must produce and sweep at least one novel $L^*$ recombinant genotype in a time proportional to $\tau$. The number of such genotypes created each generation is proportional to $Nr\tau$, and the probability of one sweeping once created is approximately $s$. Hence, the persistence of evolving behavior requires that $Nr\tau > C$ for some constant $C$, which describes the second boundary in figure 4 at $\log Nr + \log s\tau = C$. This argument also explains the dependence of the RQ regime boundaries on the population size $N$ and the strength of selection $s$. The mechanism supporting the establishment and maintenance of colinearity in the RQ state is a little more subtle and will be discussed below.

4 Secondary selection and colinearity.

The emergence of colinearity is an interesting phenomenon because in our model colinearity has no direct fitness benefit and therefore is not spread simply by selection. Furthermore colinearity is degraded by recombination and therefore must be constantly reinforced. Colinearity spreads by hitch-hiking on novel long pathways. Colinear genomes, or portions of genomes, make better building blocks for constructing novel long polyketides via recombination, as shown schematically in Figure 5. As a result, recombinant genomes encoding long pathways exhibit higher than average colinearity. Selective sweeps through which these genomes take over the population therefore increase the average level of colinearity and counteract recombinant reshuffling which attenuates it.

Let us quantify the dynamics of colinearity in a population as it goes from one selective sweep to another. After a selective sweep the population is approximately clonal. We shall assume that the sweep time scale of $\log(N)/s$ is small compared to the other time scales in the problem, assign time zero to the sweep event and consider subsequent dynamics of the population as the fitness benefit of the pathway that just swept the population decays with time. Let the population distribution of colinearity be $\xi(y, t)$ with the initial condition $\xi(y, 0) = \delta(y - y_0)$ where $y_0$ is the colinearity of the genotype that swept at time $t = 0$. As recombination reshuffles genomes [BEN: SOMEWHERE WE MUST EXPLAIN WHY THIS HAPPENS EVEN WITH NOMINALLY "CLONAL" POPULATIONS!!], variation is created in $y$ which we describe with a diffusion equation. Accounting for the density of states in $y$, $\rho(y)$, we obtain the following expression,

$$\partial_t \xi(y, t) = D(r) \partial_y \left[ \partial_y \xi(y, t) - \xi(y, t) \frac{d}{dy} \ln \rho(y) \right]$$

which describes relaxation of the $\xi(y, t)$ distribution to the random ensemble $\rho(y)$ shown in FIGURE ??? BEN???. This equation describes the change in distribution of $y$ after a sweep.

Next we need to parameterize the influence of colinearity on the rate of generation of novel long pathways. Let function $q(y)$ be the probability that a
recombinant with colinearity $y$ encodes a novel $L^*$ polyketide. We expect this function to be monotonic increasing in $y$. We determine it exactly: HOW IS THIS DONE? DOESN’T IT DEPEND ON THE WHICH POPULATION YOU SAMPLE?? in silico and the result, along with $\rho(y)$, is displayed in figure 6.

The process of sweep generation is two-fold: a novel high-fitness genotype must arise and then avoid stochastic extinction before deterministic dynamics of selection takes over in spreading it through the population. In each generation there will be on average $rN$ recombinants, each with approximately $q(y)$ chance of having high-fitness. Probability of escaping extinction is described by a stochastic branching process yielding \[ p_{esc}(t) \approx (s - \bar{s}(t) - r)\theta[s - \bar{s}(t) - r] = (s(1 - e^{-t/\tau}) - r)\theta(s(1 - e^{-t/\tau}) - r) \]

which becomes nonzero when the fitness advantage relative to the mean fitness of the population $s - \bar{s}(t)$ exceeds the rate of recombination. (This expression is valid for $s, \bar{s}, r << 1$ and $\theta(x) = x$ for $x > 0$ and is zero otherwise.) The time dependence of mean fitness $\bar{s}(t) = se^{-t/\tau}$ corresponds to the assumed exponential decay of fitness benefit of a new product after it has swept the population.

We can now write down the probability distribution of colinearity associated with the next sweep that follows a sweep with $y_0$:

\[
\chi(y|y_0) = \int_0^\infty dt N r p_{sw}(t|y_0) q(y) \xi(y,t|y_0) \exp\left[-\int_0^t dt N r Q(t|y_0)\right]
\]

which allows us to express the expected change in $y$ from one sweep to the next:

\[
E(y - y_0) = \int dy (y - y_0) \chi(y|y_0) \quad \text{so that}
\]

\[
E(y - y_0) = \int dt p_{sw}(t|y_0) \frac{\int dy (y - y_0) \xi(y,t|y_0) q(y)}{\int dy \xi(y,t|y_0) q(y)}
\]

\[
= \int dt p_{sw}(t|y_0) E^*(y - y_0, t)
\]
where \( E^*(y - y_0, t) \) is the expected change in \( y \) conditioned on the next sweep occuring at time \( t \).

Let us assume that \( y \) changes slowly on the time scale of an interval between successive sweeps. This assumption holds in the RQ state (away from the boundaries) and is confirmed by simulations. This allows a linear approximation for \( E^*(y - y_0, t) \):

\[
E^*(y - y_0, t) \approx t \frac{\partial}{\partial t} E^*(y - y_0, t) |_{t=0}
\]

To determine \( \frac{\partial}{\partial t} E^*(y - y_0, t) |_{t=0} \) one uses the diffusion equation for \( \xi(y,t|y_0) \) and its initial condition, which yields:

\[
E(y - y_0) = D(r) \left[ \frac{2 q'(y_0)}{q(y_0)} + \frac{\rho'(y_0)}{\rho(y_0)} \right] \int dt \, t \, p_{sw}(t|y_0)
\]

The term in brackets is solely responsible for the sign of the expected change in colinearity - the rest being necessarily positive - and hence is sufficient to understand the long-term behavior of \( y \) after a succession of sweeps. The latter is governed by a stable fixed point, defined by the vanishing right hand side of Eq. ADD NUMBERS. As shown in figure 7 there are two fixed points. The high colinearity fixed point is the stable one which means that repeated selective sweeps force the population towards that value of \( y \). Using \( q(y) \) and \( \rho(y) \) determined in silico we determine the location of this fixed point to be \( y_\infty \approx 0.38 \). This is in good agreement with the observed colinearity of populations in the evolving regime, as can be seen in figure 4. We note that the existence of this fixed point depends only on the form of \( q(y) \) and \( \rho(y) \), other model parameters only determine whether or not the population goes through the repeated sweeps which drive it towards the fixed point.

5 Crossover between the evolving and quiescent regimes. NEEDS WORK

??? The time dependence of the selective sweeps in our model is a much more particular problem, sensitive to many of the model parameters. In particular, it depends on the full definition of \( q(y) \), whereas overall multipliers independent of \( y \) factor out of the determination of the fixed point. In practice, determining \( q(y) \), rather than just its functional dependence, requires one to estimate a prevalence of beneficial phenotypes, something very difficult to do in real organisms. Our model also contains an important peculiarity impacting long-time behavior, the fitness effects of beneficial phenotypes decay all the way to zero, which allows for the drift-induced quiescent state. These specifics of our model indicate that the distribution of transition times, and the viability criteria determined from it, are not generalizable. However, for completeness, we outline how to calculate the values relevant to our simulation results.

??? The boundaries of the evolving regime also come out of \( p_{sw}(t) \). We immediately recognize the \( r < s \) boundary by inspecting \( p_{esc} \), it is zero at
all times when \( r \) exceeds \( s \). The second boundary on the evolving regime requires us to understand the transition into the quiescent state. This occurs once drift is free to act after selective pressure is removed. If a time greater than approximately \( \tau \log(Ns) \) passes without sweeping a novel individual the exponential decay of the fitness leads to precisely this situation. This imposes a condition that \( \int_0^{\tau \log(Ns)} dt P_{\text{sweep}}(t) > T \) where our threshold \( T \) is defined by how many expected transitions before decay into the quiescent state we require for ‘viability’. This condition simplifies when appropriate limits are taken to the inequality found heuristically above, with dependence on \( q(y_0) \) now written explicitly, \( Nrs\tau q(y_0) > C \).

6 Discussion

Adaptive evolution is a story of phenotypic exploration and selection that, in finite populations, results in a series of selective sweeps. These sweeps germinate from a select subgroup of the evolving population, those individuals which express a novel, beneficial phenotype. Thus, if there is a genetic characteristic which increases the chance of an individual being in this group, evolution will increase the prevalence of that characteristic.

The obvious case is that of a gene encoding an advantageous function. Incorporation of such a gene makes the phenotype more beneficial, therefore it will be well represented in any group of beneficial phenotypes, novel and otherwise. In our model we see an alternative to this obvious case that also results in enhanced representation in this group. The phenomenon of colinearity allows recombination to better reuse preexisting function. It increases access to an area of phenotypic space likely to be beneficial. As a result, colinearity is amplified in the one subpopulation that matters, the founders of the next selective sweep. The same process of natural selection which spreads higher fitness phenotypes also spreads colinearity, even in the absence of any phenotypic effect.

The analytic framework we employ here is generalizable to other systems and genetic characteristics. One such characteristic that suggests itself is modularity, where significant work has been done exploring its interaction with the evolutionary process. Tailed phages were an early system in which modular evolution was suggested [24, 9, 8] and “easy and continual access to... variety” as the reason for the existence of the modular architecture [2]. The Selfish Operon Model explains operon creation and maintenance as resulting from the benefit to the operon, not the organism, of access to horizontal gene transfer. This favours clustered operons because mechanisms of gene transfer are limited by the size of DNA fragments they can mobilize [13, 14]. These ideas are very similar to our understanding of emergent colinearity, which could even be viewed as high-level modularity. A genetic characteristic, with no direct fitness effect, spreads by facilitating exploration of fertile phenotypic space.

One important caveat to our analysis is that we have operated under the assumption that forward evolution is an influential process on the genome under consideration. When considering other systems the validity of this assumption
must be evaluated, however in the particular case of modular PKSs we feel it is justified. The density of interesting, accessible phenotypes is exceptional, as evidenced by the great interest in performing combinatorial variation of PKS pathways in the lab [17, 18]. PKS enzymes are very large, even on the order of the ribosome [15], while catalyzing the production of comparatively few products. Such an investment of resources would be appropriate if the return is not only a single product but also a heightened responsiveness to environmental changes and evolutionary opportunities. Additionally, PKS modules have a high level of homology with one another which increases the ability of homology preferring HGT processes to perform the sort of pathway hybridization postulated.

If influential forward evolution can be assumed in a given system, we feel that an approach similar to the one outlined here has some unique advantages. The population based perspective taken reminds us that such ‘evolvability’ enhancing characteristics are of benefit to the group. The notion of a ‘selfish’ genetic element is misleading insofar as it implies the individual or the population are disadvantaged, or even unadvantaged. This provides a clear criteria for whether significant enrichment of a characteristic is expected, the existence of a high fixed point in the recurrence equation. This expectation can be evaluated separately from the details of the time scale of the process, provided one can assume a large number of the selective sweeps have taken place, and it depends only on the functional forms of the $\rho$ and $q$ functions. Further details can be extracted if desired, such as the variation expected over time or over a group of lineages.

Two great limitations exist in the application of our model to bioinformatic data: the concurrent action of other evolutionary modes and sparsity of the sequencing of all PKS pathways available to HGT processes. In particular, the assumption that there is no fitness benefit to colocalizing interacting proteins is questionable. Still, it is clear that the ‘colinearity rule’ effect exists [20, 15] and is a quantifiable phenomenon [27]. If our proposed mechanism has substantially contributed to this effect we have some further expectations about modular PKS systems. Genetic mosaicity in these complexes will be common and widespread. Not only will mosaic complexes exist, but mosaic PKS genes will exist. The recombination joints will often correspond to protein domain boundaries, as has been observed in phages [12, 29]. Colinearity will be present, but not be perfect. We would like to further constrain the relative contribution of different evolutionary modes, such as mutation and duplication which have also been observed in PKS evolution [5]. We think that phylogenetic analysis at the gene and even protein domain level in these systems would be interesting and informative as to their evolutionary history, allowing better use of the predictions from models such as ours.
Figure 1: The passage of DNA to product polyketide is represented schematically. Panel A shows the translation of the PKS genes into PKS proteins represented by arrows. The head and tail domains are colored, binding is exclusive between corresponding domains of the same color. The flavor of chain extension performed by the central catalytic unit is represented by a letter. These proteins bind together in the cytoplasm to form the complexes which catalyze polyketide production. In Panel B the functional complex has assembled and polyketide production begins. Individual PKS proteins perform one cycle of chain extension and then pass the result to the next PKS in line. The result, seen in Panel C, is that the product polyketide chain is analogous to the chain of PKS proteins forming the complex.

Figure 2: In Panel A we see two model individuals, one grey and one black, undergoing recombination. The genes for PKS proteins are represented by the same arrows used for the PKS proteins themselves. The circular chromosomes exchange homologous sections of DNA to form recombinant children. The modules of one of the recombinant children are laid out schematically in Panel B to illustrate the determination of product polyketides and associated concentrations. The fitness is a sum of the fitness effects of those four polyketides weighted by the concentration.

Figure 3: These evolutionary trajectories of populations represent the three dynamical behaviors available to our system. The population average fitness is in blue and population average colinearity in red. In panel A we see an evolving population, novel polyketides are created and swept, maintaining its high fitness. The population was initialized with perfect colinearity and maintains high colinearity through the 50,000 generations. In panel B the environmental decorrelation time $\tau$ has been lowered to 500 generations from the 1000 generations in Panel A. The faster fitness decay eventually results in the population failing to find a novel polyketide quickly enough to avoid the effects of drift, causing a transition to the quiescent state. Once this occurs the initial colinearity decays away and then oscillates around the ensemble average of $y = 0$. In panel C environmental change has been removed, $\tau \to \infty$, and we see the static behavior. The fitness is roughly constant, its average value can be understood by considering recombination load with $r = .05$. The colinearity does not significantly change.
Figure 4: Long time averages of population fitness and colinearity ($y$) for varied recombination rate ($r$) and decorrelation time ($\tau$). Averages are taken over 100 replicates of our simulation, each running for one million generations. In panel A we see the time averaged fitness. The region of high fitness is roughly bounded by $r = s$, the requirement that recombination be low enough that novel polyketides can fix, and $\log r + \log \tau = C$, the requirement that sufficient phenotypic novelty is generated before drift predominates. Panel B is the time averaged colinearity $y$ for the same parameter range. The region of high colinearity corresponds to the region of high fitness, this is the region of parameter space in which the population maintains evolving behavior.

Figure 5: We expect colinearity, or the correlation of genetic order to phenotypic order, to increase the likelihood of recombination forming novel polyketides. Our reasoning is demonstrated here by example with recombinant pieces outlined and the resulting products shown. Recombination between two syntenic parents produces a long and potentially high fitness product. When non-syntenic parents recombine many head/tail bonds are cut and the recombinant child contains only fragmentary PKS complexes.

Figure 6: The probability of two parents creating a novel $L^*$ recombinant in our model is heavily dependent on the colinearity $y$. We determine this dependence in silico for our model, it is displayed in red. The blue histogram represents the density of states, $\rho(y)$. Together these two functions determine the expected long-time colinearity, $y_\infty$.

Figure 7: The expected change in $y$ after a selective sweep depends in magnitude on many of the model parameters. However, the sign of the expected change depends only on the function $2q(y)/q(y) + \rho(y)/\rho(y)$ which is plotted above, given a polynomial fit to $q(y)$. We see the two fixed points, one unstable near $y = 0$ and the stable high $y$ fixed point at $y_\infty \approx 0.38$. 

14
References


