Emergent colinearity in a model of modular polyketide synthases

Ben Callahan, Mukund Thattai, Boris Shraiman

September 25, 2008

Abstract

Polyketides are widespread, biologically active heteropolymers assembled by complexes of modular polyketide synthase(PKS) proteins. We implement a model of this system and explore its evolutionary behavior while subject to recombination in a changing environment. Under appropriate population parameters the combinatorial exploration of novelty allows a finite population to maintain high fitness indefinitely. Accompanying this maintenance of fitness is the phenomenon of emergent colinearity, a correlation between genetic order and the functional order of the product polyketide. Colinearity is a known characteristic of modular PKS systems. In our model it arises despite the absence of phenotypic effect. We understand this by observing that colinearity enhances the likelihood of recombination forming novel high-fitness phenotypes. We are able to quantify this effect, successive selective sweeps drive our evolving populations towards a fixed point in our colinearity parameter y. The existence and location of this fixed point is found to depend only on the density of states $\rho(y)$ and the likelihood of recombination creating high-fitness phenotypes, $q(y)$.

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 [23]: as channels of cell-to-cell communication between conspecifics [1], as anti-microbial agents against competitors, and as immuno-suppressors or virulence factors [7] between pathogens and their hosts. The enormous diversity of natural polyketide products might result from 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 [24]. 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, each containing one or more catalytic modules, are strung together into ordered multi-protein complexes through specific interactions between their N- and C-terminal 'head' and 'tail' domains [8, 29, 3, 28]. The order of catalytic modules in the multi-protein complex thus determines the order of monomers in the polyketide chain, as seen in figure 1.

The mechanism and modularity of this biosynthetic system allows a combinatorial exploration of biochemical space. Given J classes of catalytic modules, J^L different polyketides of length L are available by reordering the PKSs in a functional 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 [24]. Experimental efforts to explore combinatorially generated polyketides have had success and are actively being investigated [17]. The combinatorial access to diversity is also available to evolution via recombinant and HGT processes [28].

Evidence from comparative genomics suggests that gene duplication, horizontal gene transfer (HGT) and homologous recombination have played a key role in the evolution of bacterial PKS gene clusters [12, 23]. 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 sytems [20]. Modular PKS genes on bacterial chromosomes are invariably found in giant clusters, probably due to a combination of three factors: in-situ gene-duplication [12, 23], HGT-driven clustering, as predicted by the 'selfish operon' hypothesis [14], and the need for transcriptional co-regulation [22]. 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 [23]. 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 [11, 26]. Typically gene order is selection-neutral, so the degree of conservation can be used to estimate phylogenetic distances [27]. In the case of physically interacting proteins, however, it is seen that gene order is conserved even over relatively long timescales [11]. This effect holds in PKS ssytems, a majority of PKS proteins which physically interact are encoded contiguously [28], 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' [21].

We propose that HGT and homologous recombination are the main drivers in the combinatorial search for novel polyketide products. We specifically ask if this hypothesis is sufficient to generate the observed 'colinearity rule' of modular polyketide synthases. We implement a model of modular PKS genes and proteins, in a population of competing bacteria subject to HGT and homologous recombination, where selection rewards the generation of novel polyketide products. In our model gene order has no impact on individual fitness. Nevertheless, we find that under a wide range of parameters the system is capable of continuous innovation, and that this 'evolving' regime is characterized by emergent colinearity. If gene order is so strikingly maintained, it must confer a strong evolutionary advantage. We seek to elucidate the specific mechanism producing this advantage, and the conditions under which it arises.

2 The Model

Fisher proposed [6] that proteins which collaborate in some function, but are encoded far apart, are susceptible to disruption by recombination. Selection would therefore tend to increase the linkage between their genes. This model requires a high degree of variation at each locus, as well as frequent recombination [14]. While PKS proteins do physically interact, and their genes do show high allelic variation, it is unlikely that the recombination rate betweeen gene clusters is so high as to impose any significant cost due to the possibility of disruption. However, the evolution of modular PKSs depend on another aspect of recombination, its role in the exploration of phenotypic novelty. When recombinant offspring drive the search for novelty the impact of recombination on long-term fitness becomes significant, even when the 'recombination load' is negligible. Our model incorporates this action of recombination, which ultimately is responsible for the emergence of colinearity we observe.

We created a simplified model of the modular PKS system that retains the key features while being computationally tractable. Our PKS proteins consist of three regions, head and tail binding domains and a central catalytic module occuring in two flavors. The gene is arranged likewise, with head and tail domains on the termini of the protein and gene. The restriction in flavors of catalytic modules allows polyketides to be represented as binary strings, each bit representing the 'monomer' at that position, as is seen in figure 2b. The head and tail domains 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 and the $+1$ class is a special terminator class which does not bind.

Individuals have circular chromosomes of L PKS genes, $L = 12$ in simulations shown. The fitness of an individual is one plus a sum over the contributions of polyketides produced by complexes of these L PKS proteins,

$$f = 1 + \sum_k C(k)\Delta f(k)$$

The contribution of a polyketide, indexed by k, is a product of its concentration $C(k)$ and its fitness effect $\Delta f(k)$. The concentration of a polyketide is proportional to the probability that the complex which catalyzes its production is fully constituted. We assume same class head/tail pairs bind with equal probability to all available partners. Figure 2b shows the product polyketides and associated concentrations of an example individual. There is a special case when complexes recursively 'loop', resulting in indefinitely long PKS chains. This is considered degenerate and suppressed by assigning this case zero fitness.

The generation of novelty is rewarded by considering a constantly changing environment. This manifests itsef in the time dependence of polyketide fitness effect. When first appearing in the population polyketides have an initial fitness effect $\Delta f_0(k)$. Once present, the fitness benefit of a polyketide decays exponentially with a time constant τ, $\Delta f(k) = \Delta f_0(k) e^{-(t-t_{k0})/\tau}$. This is a result of the environment changing in time away from that in which the polyketide was initially beneficial, for this reason we call τ the environmental decorrelation time. Since all fitness contributions from polyketides present in the population decay with time there is constant pressure to find novel, and therefore higher fitness, polyketides.

Uniform fitness decay is a crude approximation, but it can be motivated by considering that environments encountered by individuals at different times will vary, and fitness is conditional on environment. For example, as bacteria gain resistance to a particular antibiotic the fitness benefit of that antibiotic decays. Ultimarely while this is simplistic it is sufficient to motivate the search for novelty, and our observation of emergent colinearity depends only on this search occurring, not on the specifics mechanism incentivizing it. Finally, we choose a simple initial fitness landscape, all polyketides of length L^* have initial fitness effect s, no other polyketides affect fitness, $\Delta f_0(k) = s\delta(|k| - L^*)$. In simulations shown $L^* = 7$ and $s = .1$ but results are qualitatively similar for different L^*, s.

We consider haploid populations of N such individuals propagating in discrete, non-overlapping generations, in simulations shown the populations have size $N = 1000$. Our offspring distribution is geometric rather than the more familiar binomial distribution of the Wright-Fisher model, the only impact of this difference is a factor of two in the branching process result for probability of escaping low number stochasticity. Recombination is implemented by introducing a probability r for each member of the child 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.

3 Results

When exploring the parameter space of the model we encountered three distinct dynamical behaviors. There is an 'evolving' behavior, in which the population is characterized by high-fitness and colinearity. There is a 'quiescent' behavior, in which populations fall into a permanent quiescent state of low fitness and no colinearity. Finally we observe a 'static' behavior characterized by high-fitness and no colinearity when time dependence is removed from the fitness. Figure 3 shows population trajectories typical of these behaviors.

The characteristic saw-tooth fitness of an evolving population is seen in figure 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 a particular L^* polyketide, for brevity we say this polyketide dominates the population. As a result, the population average fitness decays exponentially as the fitness effect of this polyketide decays. The decrease in population average fitness increases the relative fitness advantage of novel L^* polyketides, increasing the chance for a recombinant expressing one to sweep. Once enough individuals expressing novel L^* polyketides are produced to ensure escape from low number stochasticity a selective sweep will occur, 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 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 impossible to create L^* polyketides out of the short fragments remaining in the population. In figure 3b we see a population undergo this transition.

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^* polyketide, transitions to other L^* 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^* 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.

We now introduce a quantification of colinearity to facilitate an investigation of the phenomenon. We choose as our measure the mean genetic distance

between interacting head/tail pairs, in units of chromosome length,

$$\overline{d} = \frac{\sum_h \sum_t \delta_{ht} dist(h,t)}{\sum_h \sum_t \delta_{ht}}$$

In the functional complex head/tail pairs are bound, their distance from each other is zero. The greater the genetic distance between head/tail pairs the greater the deviation from the functional ordering and hence from colinearity. To put it simply, the higher $\overline{d}$ is for an individual, the less colinear is that individual's chromosome.

The phenotype of an individual consists of the PKSs encoded by its chromosome. Remember, gene order is irrelevant, all rearrangements of a given set of PKS genes are phenotypically identical and hence should be considered equiprobable. Our expectations for $\overline{d}$ are informed by considering this ensemble of genic rearrangements. In the case of an individual with a single head/tail pair, $\overline{d}$ will be uniformly distributed on $[0, 1/2]$ since the greatest distance between two points on a circle is $1/2$ of its circumference. As the number of head/tail pairs increases $\rho(\overline{d})$ will approach a Gaussian centered on $1/4$. We now introduce a transformation of $\overline{d}$, the colinearity y,

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

This value lies on the interval $[-1, 1]$. The density of states in the ensemble of genic rearrangements $\rho(y)$ is roughly Gaussian centered at $y = 0$ with the variance decreasing as the number of head/tail pairs increases, the case of an individual 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. The ensemble average of $y = 0$ indicates no correlation between the genetic and functional ordering. $y < 0$ indicates a chromosome on which genes of interacting head/tail pairs 'repel'. We can now utilize this measure, and its population average, to discuss the evolution of colinearity. Going forward, when we refer to a population exhibiting colinearity we mean that the population average colinearity (y) is greater than zero.

Colinearity arises spontaneously and is maintained in our simulations. Returning to figure 3 we see, in addition to fitness, the population average colinearity in time. In figure 3a, the evolving population, we see initial colinearity maintained through several selective sweeps. This effect is characteristic of evolving populations and persists as long as evolving behavior continues. Excursions to non-colinear genetic realizations do occur, but are temporary. Long-term time averages of the colinearity of evolving populations are significantly greater than the $y = 0$ expected in an equiprobable ensemble. The evolving behavior both maintains colinearity and generates it when not initially present.

Quiescent populations do not create or 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, long-time averages of the colinearity in these populations are zero. This is not unexpected, these states

have no selective pressure to differentiate between different genic arrangements so the population average goes to the ensemble average. We see something similar in the static state. Selective pressure maintains the class of L^* polyketides, but there is no incentive to explore this space. Without this search for novelty, colinearity is not created.

A systematic exploration of the model parameters allows us to understand the conditions under which these different behaviors obtain. The two key parameters are the recombination rate r and the environmental decorrelation time τ. In figure 4 we display time-averaged fitness and colinearity of populations evolved under a range of r, τ. There is a clear separation into two regimes: a high fitness, colinear regime corresponding to the region of parameter space in which our finite populations maintain evolving behavior, and a low-fitness non-colinear regime corresponding to the region of parameter space in which populations fall into the quiescent state. We are now in position to address quantitatively the location of these regime boundaries and to understand why colinearity emerges in the evolving regime.

A basic understanding of the evolving 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 encodinging a novel polyketide is formed its spread is hindered by recombination, which generally destroys the phenotype. If we consider these recombination events as 'deaths' for the lineage, this introduces a requirement that $r < s$ for a novel individual to have a non-zero chance of sweeping, since s is the largest available selective advantage. If novel recombinants cannot sweep, evolving behavior cannot be maintained. This restricts the evolving regime and is seen as the boundary near $r = s$.

Recombination is not only an obstacle, in our system it is the sole source of phenotypic novelty. The persistence of evolving behaviors requires that novel individuals encoding L^* polyketides be created in sufficient number to ensure one can escape low-number stochasticity and sweep before drift forces the population into the quiescent state. We understand this heuristically as a requirement for the population to produce and sweep at least one novel L^* recombinant in a time proportional to τ. The number of such individuals created each generation is proportional to $Nr\tau$, and the probability of one sweeping once created is approximately s. So, to maintain evolving behavior it is required that $Nrs\tau > C$ for some constant C. This describes the second boundary in figure 4 at $logr + log\tau = C$.

Ths observation of colinearity is non-trivial because, by our model construction, it has no impact on fitness. Colinearity is a purely genotypic characteristic. There is a key insight that allows us to understand its emergence even without phenotypic effect, colinear individuals are disproportionately represented in the class of novel L^* recombinants which form the basis for each selective sweep. Colinear genomes, and colinear portions of genomes, make better building blocks for constructing novel long polyketides via recombination, as represented schematically in figure 5. These building blocks bring their colinearity with them to their offspring. Since selective sweeps occur within this class of

novel individuals, the successive sweeps of the evolving regime serve to select for colinearity. This effect balances against the reduced number of such ordered states, nevertheless there is still a real and measurable impact on genetic structure, as we saw in figure 4.

4 Analysis

The dynamics of our model are understood by quantifying the process of a population, recently swept by an individual encoding an L^* polyketide, generating and then sweeping an individual encoding a novel L^* polyketide. The impact of this process on the expected y of the population will inform our expectations of y at long times. The time dependence of the process will describe the viability boundaries and inform our expectations of the fitness at long times.

After a selective sweep the population is approximately clonal. We consider sweeps to be instantanteous, a reasonable approximation given that the sweep time scale of $log(N)/s$ is small compared to the other time scales in the problem. We assign time zero to the sweep event and consider the fitness effect of the swept polyketide decaying from that time. We write the population distribution in colinearity and time as $\xi(y,t)$, thus our initial condition is $\xi(y,0) = \delta(y - y_0)$. We define $p_{sweep}(y_0,t)$ to be the probability that the next sweep occurs at time t after the previous individual with colinearity y_0 swept.

As recombination acts on our population, 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,

$$\partial_t \xi(y,t) = D(r)\partial_y \left[\partial_y - \frac{\rho^{'}(y)}{\rho(y)} \right] \xi(y,t)$$

This equation describes the change in distribution of y after a sweep. It is not yet sufficient to determine which individual, and which y, is the seed of the next sweep. We must also account for our earlier observation that an individual's colinearity influences its likelihood to be a novel high-fitness recombinant. We seek a function accounting for this effect we will call $q(y)$, the likelihood a recombinant with colinearity y encodes a novel L^* polyketide.

We can compute, using the implementation of our model, the well-defined function $q(y_m, y_f; y)$, which accounts for the colinearities of the recombining 'mother' and 'father'. We argue that the projection of this function onto the child colinearity alone, $q(y) \equiv q(y,y;y)$, captures the relevant effects. This is justified by understanding that even if the parental colinearities are different, the portions of their chromosome contributed to the recombinant offspring had colinearity y, and these portions are what are relevant for the potential of the child. We expect this function to be monotonic increasing in y. We determine it exactly in silico and the result, along with $\rho(y)$, is displayed in figure 6.

We are now in a position to write the expected change in y after a sweep,

$$E(y-y_0) = \int dt\, p_{sweep}(y_0,t) \frac{\int dy\, (y-y_0)\xi(y,t)q(y)}{\int dy\, \xi(y,t)q(y)}$$
$$= \int dt\, p_{sweep}(y_0,t) E^*(y-y_0,t)$$

$E^*(y-y_0,t)$ is the expected change in y conditioned on the next sweep occuring at time t. We utilize the fact that the diffusion in y is slow, in the sense that it remains local and dominated by the initial conditions on the sweep to sweep time scale. This occurs, and is confirmed in silico, because most recombinants do not encode an L^* polyketide, hence are quickly eliminated by selection. The diffusion prior to a selective sweep is among the limited class of individuals encoding one particular L^* polyketide. This allows $E^*(y-y_0,t)$ to accept a linear response approximation,

$$E^*(y-y_0,t) \approx \partial_t E^*(y-y_0,t)|_{t=0} \cdot t$$

We solve for $\partial_t E^*(y-y_0,t)|_{t=0}$ by use of the diffusion equation for $\xi(y,t)$ and the initial conditions. When we insert that into the equation for $E(y-y_0)$ we get,

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

Now we have gotten somewhere, the term in brackets is solely responsible for the sign of the expected change in colinearity, the rest is necessarily positive. This term, shown in figure **??** is sufficient to understand the long-term behavior of y after a succession of sweeps. It defines a high y fixed point that exists for our model, towards which repeated selective sweeps force the population. We see 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 evolutionary transitions which drive it towards the fixed point. We calculate the location of this fixed point, using $q(y)$ and $\rho(y)$ determined in silico, to be $y_\infty \approx 0.35$. This is in good agreement with the observed colinearity of populations in the evolving regime, as can be seen in figure 4.

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 process of sweeping a novel individual is two-fold, first the creation of a novel high-fitness individual and then that individual's escape from low-number stochasticity. In each generation there will be on average rN recombinants, each with approximately $q(y_0)$ chance of being a novel high-fitness individual. If such an individual is created, its low number behavior can be modeled by a branching process to determine its probability to escape low number stochasticity, p_{esc}. The initial fitness is $1+s$, but p_{esc} depends on the population average fitness $\langle fit \rangle$. These populations remain largely clonal, the population average fitness is essentially equal to the fitness of the recently swept individual, $\langle fit \rangle = 1 + s \cdot exp(-t/\tau)$. So, with negative values of the following equation implicitly set to zero,

$$p_{esc}(t) = \frac{(1+s-\langle fit \rangle)(1-r)}{1+(1+s-\langle fit \rangle)(1-r)} \approx s(1-e^{-t/\tau}) - r + O(rs)$$

We can now write $p_{sweep}(y_0, t)$, in the relevant limit where less than one novel individual is expected to be created and swept each generation,

$$p_{sweep}(y_0, t) = \left[e^{-\int_0^t dt\, Nrq(y_0) p_{esc}(t)} \right] Nrq(y_0) p_{esc}(t)$$

We see in $p_{esc}(t)$ that a waiting time is enforced before the next sweep can occur, the fitness must decay enough that novel individuals have a selective advantage exceeding r. When we also consider that a sweep, assuming it occurs, must occur before fitness decays so far to be negligible, we see that τ basically sets the sweep to sweep time scale in our evolving populations.

The boundaries of the evolving regime also come out of $p_{sweep}(t)$. 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 $\int_0^{\tau\, log(Ns)} dt\, p_{sweep}(t) > T$ where our threshhold 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, $Nrs\tau q(y_0) > C$.

5 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 [25, 10, 9] 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 [14, 15]. 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 applications 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 [18, 19]. PKS enzymes are very large, even on the order of the ribosome [16], 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. Our approach provides a clear criteria for whether significant enrichment of a characteristic is expected, the existence of a high fixed point. 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 ρ 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 [21, 16] and is a quantifiable phenomenon [28]. 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 [13, 30]. 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.

References

[1] Michael B Austin, Tamao Saito, Marianne E Bowman, Stephen Haydock, Atsushi Kato, Bradley S Moore, Robert R Kay, and Joseph P Noel. Biosynthesis of dictyostelium discoideum differentiation-inducing factor by a hybrid type i fatty acid-type iii polyketide synthase. *Nat Chem Biol*, 2(9):494–502, 2006.

[2] David Botstein. A theory of modular evolution for bacteriophages. *Annals of the New York Academy of Sciences*, 354:484–491, November 1980.

[3] R. William Broadhurst, Daniel Nietlispach, Michael P. Wheatcroft, Peter F. Leadlay, and Kira J. Weissman. The structure of docking domains in modular polyketide synthases. *Chemistry & Biology*, 10(8):723–731, August 2003.

[4] RD Firn and CG Jones. Natural products a simple model to explain chemical diversity. *Natural Product Reports*, 20:382–391, 2003.

[5] Michael A. Fischbach, Christopher T. Walsh, and Jon Clardy. The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences*, 105:4601–4608, March 2008.

[6] RA Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, 1930.

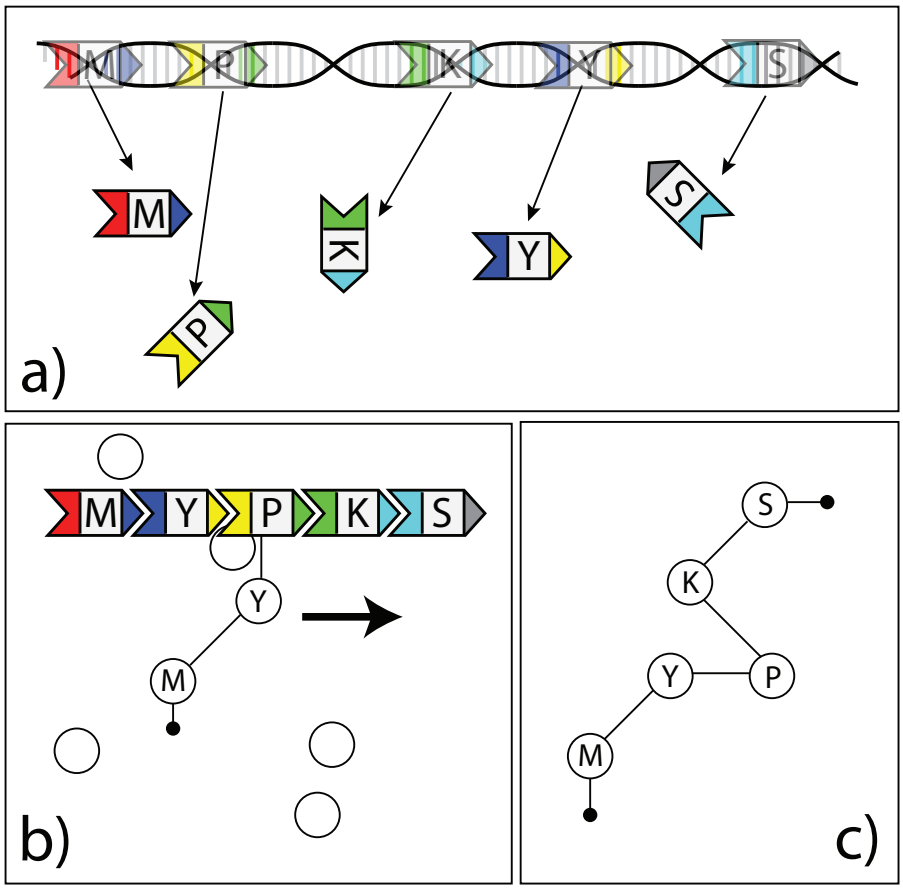

Figure 1: The passage of DNA to product polyketide is represented schematically. Panel A shows the translation of the PKS genes into PKS proteins represented by arrows. The head and tail domains are colored, binding is exclusive between corresponding domains of the same color. The flavor of chain extension 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.

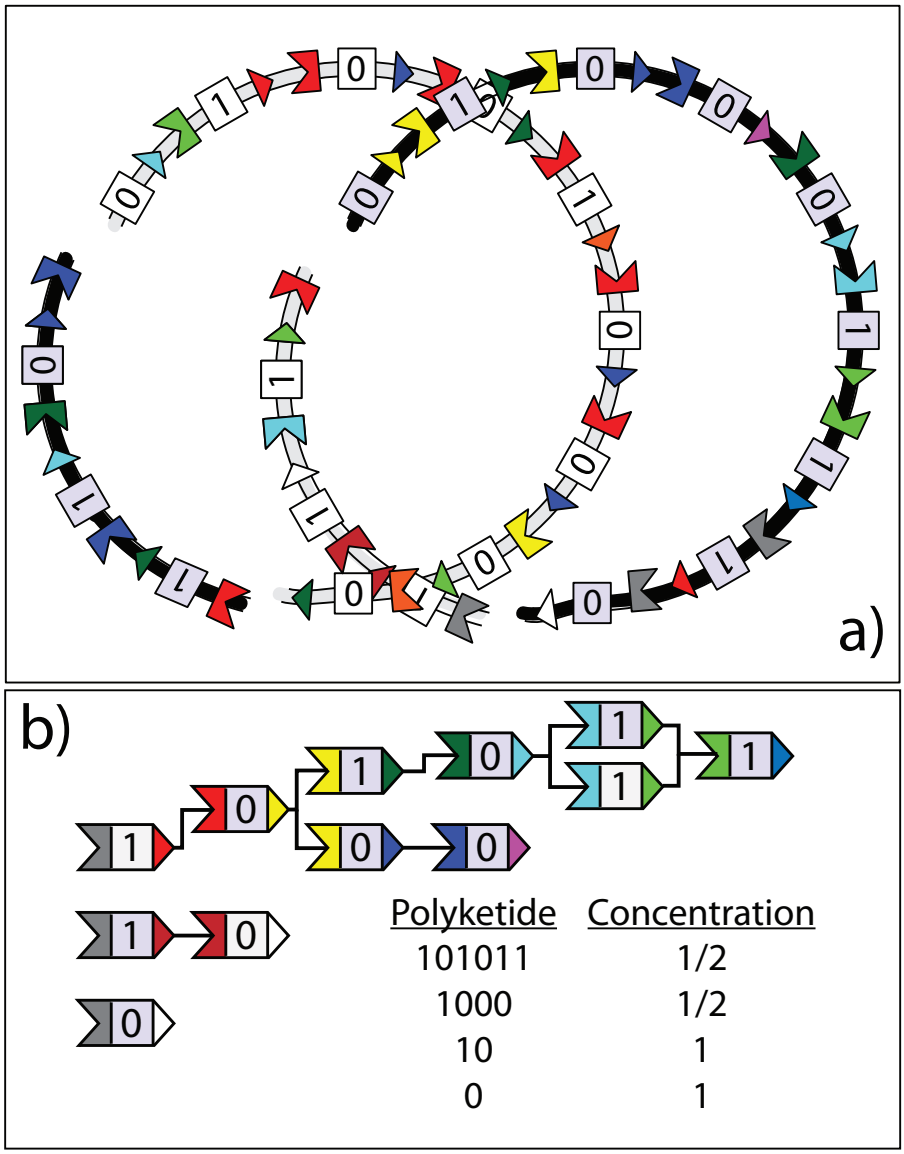

Figure 2: In Panel A we see two model individuals, one grey and one black, undergoing recombination. The genes for PKS proteins are represented by the same arrows used for the PKS proteins themselves. The circular chromosomes exchange homologous sections of DNA to form recombinant children. The modules 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.

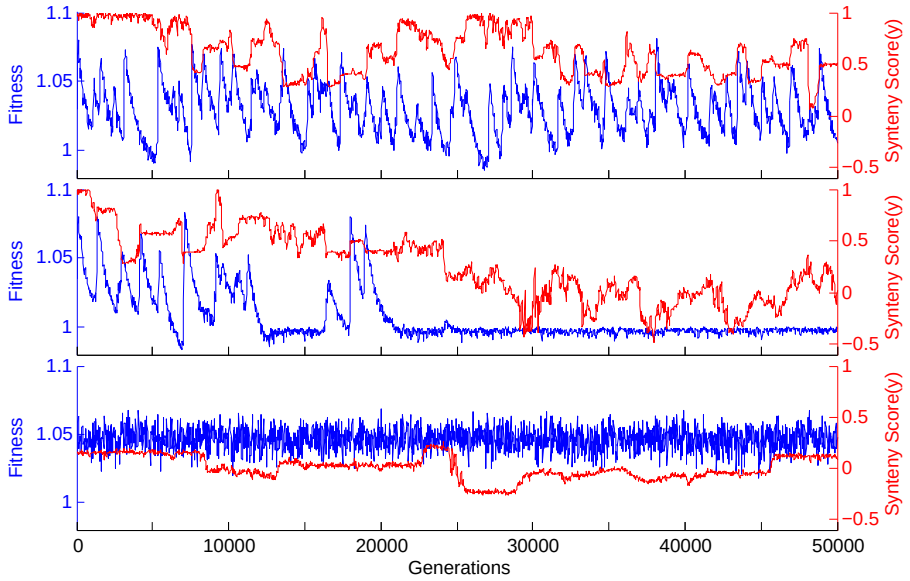

Figure 3: These evolutionary trajectories of populations represent the three dynamical behaviors available to our system. The population average fitness is in blue and population average colinearity in red. In panel A we see an evolving population, novel polyketides are created and swept, maintaining 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 τ 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 remove, $\tau \to \infty$, 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.

[7] Rajesh S. Gokhale, Priti Saxena, Tarun Chopra, and Debasisa Mohanty. Versatile polyketide enzymatic machinery for the biosynthesis of complex mycobacterial lipids. *Natural Product Reports*, 24(2):267–277, 2007.

[8] Rajesh S. Gokhale, Stuart Y. Tsuji, David E. Cane, and Chaitan Khosla. Dissecting and exploiting intermodular communication in polyketide synthases. *Science*, 284(5413):482–485, April 1999.

[9] Roger W. Hendrix. Bacteriophages: Evolution of the majority. *Theoretical Population Biology*, 61:471–480, June 2002.

[10] Roger W. Hendrix, Jeffrey G. Lawrence, Graham F. Hatfull, and Sherwood Casjens. The origins and ongoing evolution of viruses. *Trends in Microbiology*, 8:504–508, November 2000.

[11] M A Huynen and P Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11):5849–56, May 1998. PMID: 9600883.

[12] Holger Jenke-Kodama, Thomas Brner, and Elke Dittmann. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium streptomyces avermitilis. *PLoS Computational Biology*, 2(10):e132, October 2006.

[13] Robert J. Juhala, Michael E. Ford, Robert L. Duda, Anthony Youlton, Graham F. Hatfull, and Roger W. Hendrix. Genomic sequences of bacteriophages hk97 and hk022: pervasive genetic mosaicism in the lambdoid bacteriophages. *Journal of Molecular Biology*, 299:27–51, May 2000.

[14] J. G. Lawrence and J. R. Roth. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–1860, August 1996.

[15] Jeffrey Lawrence. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current Opinion in Genetics & Development*, 9:642–648, December 1999.

[16] J.B. McAlpine, B.O. Bachmann, M. Piraee, S. Tremblay, A.-M. Alarco, E. Zazopoulos, and C.M. Farnet. Microbial genomics as a guide to drug discovery and structural elucidation: Eco-02301, a novel antifungal agent, as an example. *Journal of Natural Products*, 68:493–496, April 2005.

[17] Hugo G Menzella and Christopher D Reeves. Combinatorial biosynthesis for drug development. *Current Opinion in Microbiology*, 10(3):238–245, June 2007.

[18] Hugo G Menzella, Ralph Reid, John R Carney, Sunil S Chandran, Sarah J Reisinger, Kedar G Patel, David A Hopwood, and Daniel V Santi. Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nat Biotech*, 23:1171–1176, 2005.

[19] Hugo G Menzella, Sarah J Reisinger, Mark Welch, James T Kealey, Jonathan Kennedy, Ralph Reid, Chau Q Tran, and Daniel V Santi. Redesign, synthesis and functional expression of the 6-deoxyerythronolide b polyketide synthase gene cluster. *Journal of Industrial Microbiology & Biotechnology*, 33:22–8, 2006. PMID: 16187094.

[20] Mikko Metsa-Ketela, Laura Halo, Eveliina Munukka, Juha Hakala, Pekka Mantsala, and Kristiina Ylihonko. Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16s ribosomal dna genes from various streptomyces species. *Appl. Environ. Microbiol.*, 68(9):4472–4479, September 2002.

[21] Yohsuke Minowa, Michihiro Araki, and Minoru Kanehisa. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *Journal of Molecular Biology*, 368(5):1500–1517, May 2007.

[22] Morgan N. Price, Katherine H. Huang, Adam P. Arkin, and Eric J. Alm. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, 15:809–819, June 2005.

[23] Christian P Ridley, Ho Young Lee, and Chaitan Khosla. Evolution of polyketide synthases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4595–600, March 2008. PMID: 18250311.

[24] James Staunton and Kira J. Weissman. Polyketide biosynthesis: a millennium review. *Natural Product Reports*, 18:380–416, 2001.

[25] M M Susskind and D. Botstein. Molecular genetics of bacteriophage p22. *Microbiological Reviews*, 42, June 1978. PMC281435.

[26] Mikita Suyama and Peer Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in Genetics*, 17(1):10–13, 2001.

[27] J Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6):RESEARCH0020, 2001. PMID: 11423009.

[28] Mukund Thattai, Yoram Burak, and Boris I Shraiman. The origins of specificity in polyketide synthase protein interactions. *PLoS Computational Biology*, 3:e186, 2007.

[29] S.Y. Tsuji, D.E. Cane, and C. Khosla. Selective protein-protein interactions direct channeling of intermediates between polyketide synthase modules. *Biochemistry*, 40(8):2326–2331, February 2001.

[30] E. Yagil, S. Dolev, J. Oberto, N. Kislev, N. Ramaiah, and R A Weisberg. Determinants of site-specific recombination in the lambdoid coliphage hk022. an evolutionary change in specificity. *Journal of Molecular Biology*, 207:695–717, June 1989. PMID: 2547971.

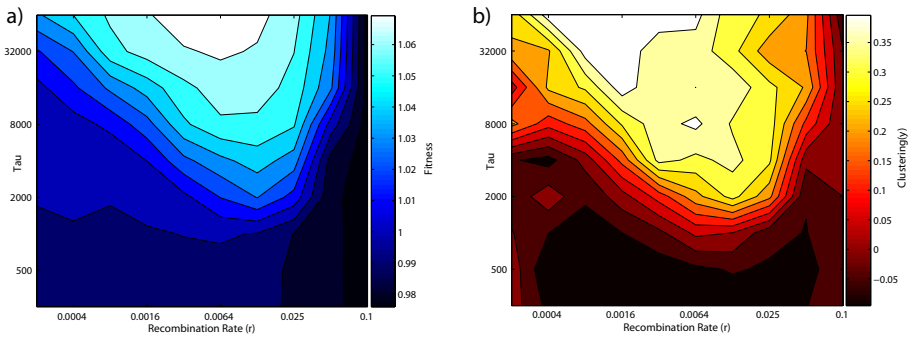

Figure 4: Long time averages of population fitness and colinearity (y) for varied recombination rate(r) and decorrelation time(τ). Averages are taken over 100 replicates of our simulation, each running for one million generations. In panel A we see the time averaged fitness. The region of high fitness 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.

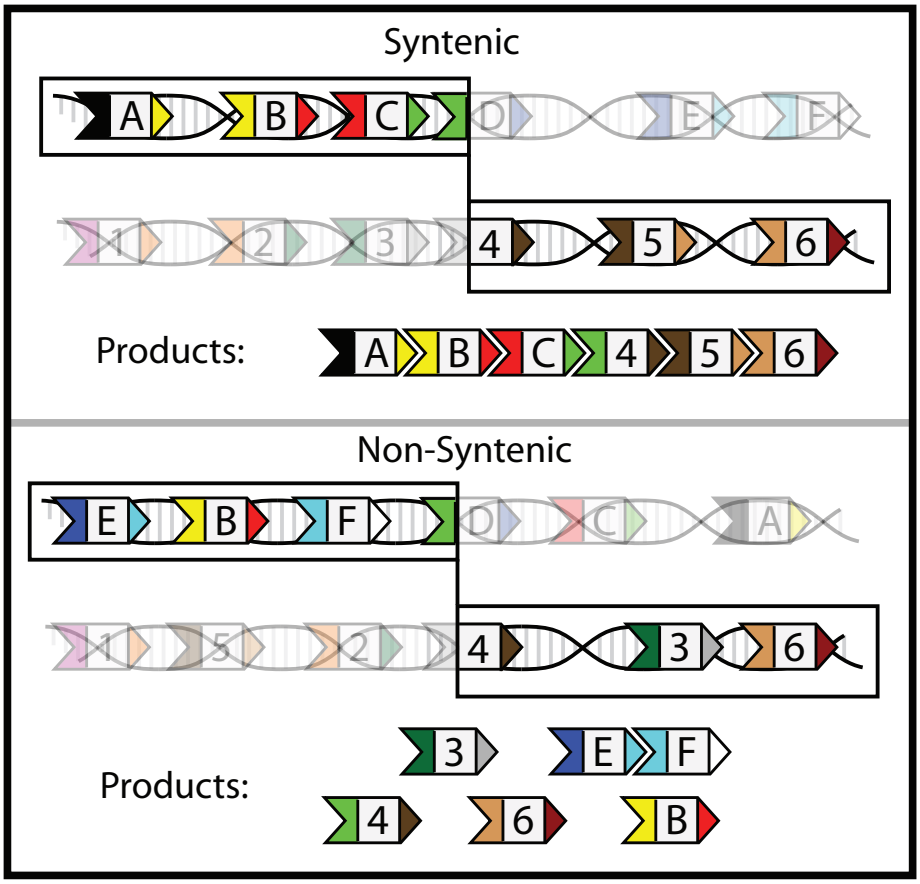

Figure 5: We expect colinearity, or the correlation of genetic order to phenotypic order, to increase the likelihood of recombination forming novel polyketides. Our reasoning is demonstrated here by example with recombinant pieces outlined and the resulting products shown. Recombination between two syntenic 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.

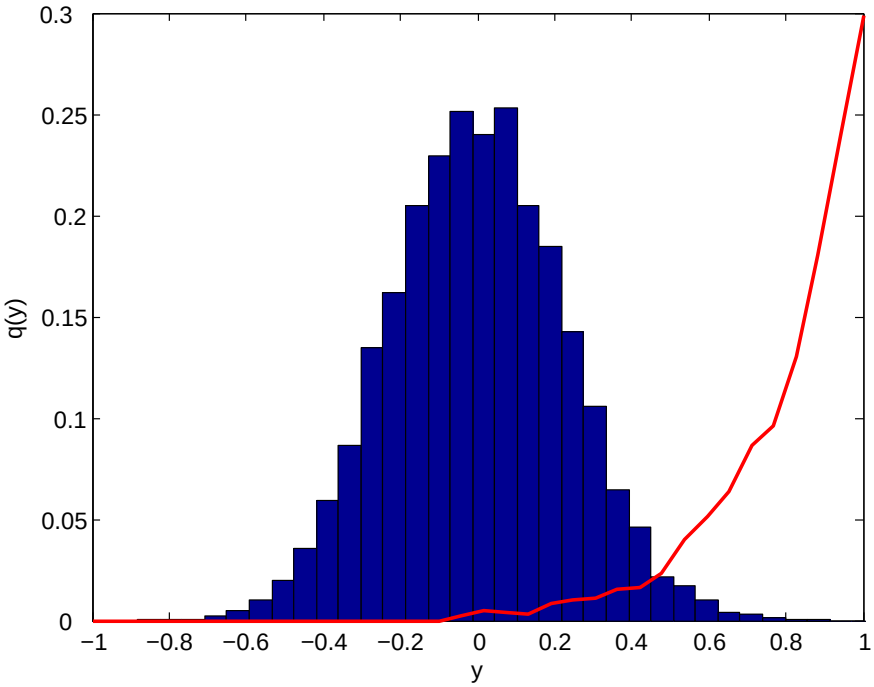

Figure 6: We find that the probability of two parents creating a novel L^* recombinant in our model is heavily dependent on the colinearity y. We determine this dependence in silico for our model, it is displayed in red. The blue histogram represents the density of states, $\rho(y)$. Together these two functions determine the expected long-time colinearity, y_∞, which we find to be approximately 0.35.