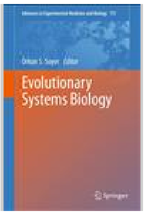
EVOLUTIONARY SYSTEMS BIOLOGY
Advances in Experimental Medicine and Biology, 2012, Volume 751, 195-224, DOI: 10.1007/978-1-4614-3567-9_10



Toward a Theory of Multilevel Evolution: Long-Term Information Integration Shapes the Mutational Landscape and Enhances Evolvability

Paulien Hogeweg

Chapter 10
Toward a Theory of Multilevel Evolution: Long-Term Information Integration Shapes the Mutational Landscape and Enhances Evolvability

Paulien Hogeweg

Abstract Most of evolutionary theory has abstracted away from how information is coded in the genome and how this information is transformed into traits on which selection takes place. While in the earliest stages of biological evolution, in the RNA world, the mapping from the genotype into function was largely predefined by the physical–chemical properties of the evolving entities (RNA replicators, e.g. from sequence to folded structure and catalytic sites), in present-day organisms, the mapping itself is the result of evolution. I will review results of several *in silico* evolutionary studies which examine the consequences of evolving the genetic coding, and the ways this information is transformed, while adapting to prevailing environments. Such multilevel evolution leads to long-term information integration. Through genome, network, and dynamical structuring, the occurrence and/or effect of random mutations becomes nonrandom, and facilitates rapid adaptation. This is what does happen in the *in silico* experiments. Is it also what did happen in biological evolution? I will discuss some data that suggest that it did. In any case, these results provide us with novel search images to tackle the wealth of biological data.

1 Introduction

Much of current research in biology is on the physical and biochemical basis of information processing in cells. This information processing leads to the transformation of the inherited genotypic information to a living organism enough adapted to its environment to survive.

P. Hogeweg (✉)
Theoretical Biology and Bioinformatics Group, Utrecht University,
Padualaan 8, 3584CH Utrecht, The Netherlands
e-mail: p.hogeweg@uu.nl

O.S. Soyer (ed.), *Evolutionary Systems Biology*, Advances in Experimental Medicine and Biology 751, DOI 10.1007/978-1-4614-3567-9_10,

Most of these processes were unknown to Darwin, when he formulated the theory of evolution by natural selection. Since Darwin's time, and the development of population genetics, the major paradigm of evolutionary biology has been to largely ignore, or at least drastically simplify, the way information is coded and transformed. Transporting the "small phenotypic variations" envisioned by Darwin, to allele frequencies and nucleotide substitutions, a direct connection between the level of mutations and the level of observation was largely maintained. Because of, or despite of, this simplification, evolutionary theory could remain the cornerstone of biological thinking through all the changes in understanding the underlying processes in biological systems.

Recent advances in high-throughput techniques are producing a wealth of data on the structure of genomes, regulatory networks, protein interaction networks, all types of posttranscriptional and posttranslation modifications, etc., which all together determine the genotype to phenotype mapping. On the basis of this wealth of data, systems biology tries to understand the working of present-day organisms, using a combination of data analysis, mathematical/computational modeling, and experiments. Combining systems biology and evolutionary theory is fruitful in at least three different ways. In the first place for analyzing the high-throughput data and understanding the functioning of current life-forms, an evolutionary perspective provides very powerful tools. For example, phylogenetic profiling of genes can be used to predict the functioning of the genes in the same process/pathway when they are (repeatedly) lost in the same lineages [33]. Also, multilevel evolutionary modeling can help to zoom in to the relevant parameter values governing regulatory interactions [62]. Secondly, the high-throughput data have shed exciting new light on what did happen in long-term evolution and what does happen in short-term evolution. For example phylogenetic reconstruction of fully sequenced genomes have highlighted the unexpected importance of gene loss in adaptive evolution (e.g., [11, 23, 28]), and short-term evolutionary experiments have shown the frequent occurrence of large-scale mutations like gross chromosomal rearrangements (GCRs) [15], and massive changes in transcription in very short-term adaptation [16]. In this chapter, we explore a third meaning of the term evolutionary systems biology, namely, how insights obtained by systems biology can enrich the theory of evolution itself. In particular, we want to investigate the effects of complex, multilevel genotype–phenotype mapping, and its evolution, on evolutionary dynamics. We seek "generic patterns," i.e., we seek a baseline for what we should expect given our current knowledge or, to use the words of Koonin [39], universal laws governing evolving systems. Koonin looks for such "universal laws" by examining the data. We look for such generic patterns by studying models with many degrees of freedom and observing, against the background of the implemented mutation selection procedure, the emerging evolutionary patterns.

We use nonsupervised modeling (or nongoal-directed modeling) [24, 26]. This concept can best be explained by analogy with nonsupervised pattern analysis (or nonsupervised learning), as opposed to supervised pattern analysis. In nonsupervised pattern analysis (e.g., cluster analysis), a description is given, and patterns that are not predefined are sought, whereas in supervised pattern analysis, a pattern (e.g., a classification) is given, and a description is sought which allows the recognition of the classes. Likewise in nonsupervised modeling, the model does not try to find an explanation for predefined phenomena, but instead structured objects, possible transformation and interactions are defined, and the emerging patterns are studied, focusing on those patterns which are not implemented or represented in the model directly. Accordingly, in nonsupervised evolutionary modeling, we are not interested in fitness attained, but in the structural side effects of attaining fitness.

The advantage of such an approach is that we can find, like in the pattern analysis counterpart, truly unexpected patterns. Moreover, apparently unrelated phenomena may appear as the side effects of the same basic processes. Another advantage is that we can retain some of the complexity which is the hallmark of biological systems, e.g., large genomes, and the complexity of the mapping of genome into the phenotype.

In formulating these models, we adhere to the well-known dictum "models should be as simple as possible, but not more so".[1] We think that abstracting from the multilevel nature of biological systems constitutes a too drastic simplification. Instead, we study the consequences of the multilevel nature in models which are as simple as possible.

An apparent disadvantage is that we can only study particular examples. That is in fact what Darwin did and what biologist still do in studying a limited number of model organisms. I will argue that by studying well-chosen examples, we can attain more generality than by molding our models into too much generality beforehand.

In line with this methodology, I will review in this chapter a number of specific models we studied recently and later point out more general patterns in the results. I will first review the by now classical results of the shape of fitness landscapes of high-dimensional genotype spaces and a complex structural mapping of genotype to fitness, as gleaned from studying RNA landscapes. Next, I will use a more flexible genotype representation, adding successive layers in the mapping from the genome to the structure and/or dynamics which determines fitness. We show that the properties of the fixed landscapes still hold but are significantly enriched in this more open-ended setting. Moreover, new patterns arise, which indicate that surprising features gleaned from phylogenetic studies may be generic patterns of multilevel evolution. Finally, adding an ecological level, I probe how new levels of selection emerge and how these levels of selection may feedback on the genome, generating a more complex genomic organization.

Together, these examples start to outline the contours of a theory of multilevel evolution and suggest that the multilevel nature of biological systems allows for long-term information integration. A striking consequence of this long-term information integration is that mutation and selection are no longer independent: the

[1] This dictum is often attributed to Einstein (e.g., [42]), although he has never said it in this form. Nevertheless, it remains a nice pointer to emphasize that on the one hand, models should not incorporate unnecessary detail, but on the other hand should not overlook (and therewith obscure) essential features of the process modeled.

types of mutations which can/will happen in evolved systems, as well as their effect, are shaped by past selection. In other words, "random mutations are not random" in evolved systems.

2 High Dimensional Genotype Space with Nonlinear, Redundant Mapping from Genotype to Phenotype

A hallmark of biological systems is the very large genotype space. An often used visualization of evolutionary processes makes use of the concept fitness landscape, first introduced by Sewell Wright [71]. However, our intuition about landscapes in general and fitness landscapes in particular is strongly biased to lower dimensional space. This bias can be highly misleading. In the beginning of the 1990s RNA sequence to secondary structure mapping became a prototype to "peer" into a realistic high-dimensional genotype–phenotype mapping [20, 31, 53]. It was chosen because it was the only realistic genotype–phenotype mapping which can be readily computed and because of the inherent interest of RNA as both information carrier and catalyst and thereby its central role in early evolution. The genotype–phenotype mapping can be brought in the landscape metaphor by defining a distance function between secondary structures. Taking one secondary structure as reference, the distance to that structure can be taken as the "height" associated with every genotype. This distance can also be interpreted as fitness to study evolutionary dynamics. An other useful representation of the RNA landscape is in terms of connected graphs of identical structures mapped on the genotype space. Both these images will be used intermingled in what follows, where we first describe features of the RNA landscape and then the consequences of these features on the evolutionary dynamics.

2.1 Shape of the RNA Landscape

By considering RNA sequences of fixed length, and allowing only base substitutions, the landscape metaphor can be applied. Fitness landscapes are often characterized in terms of "ruggedness" (e.g., Kauffman's NK landscapes [36]). Ruggedness can be quantified in different ways, but it reflects correlation between height and genotypic similarity and in low dimensional landscapes is associated with number of local peaks. Because an evolutionary process can get stuck on such a local peak, ruggedness is in general thought of as hindering evolutionary optimization. It turns out that RNA landscapes combine smoothness and ruggedness in interesting ways, as detailed in the following:

- *Redundancy*. The mapping is redundant as can be seen in that the sequences consist of four different nucleotides, whereas the secondary structure can be represented as a string with three symbols (the so-called bracket notation).

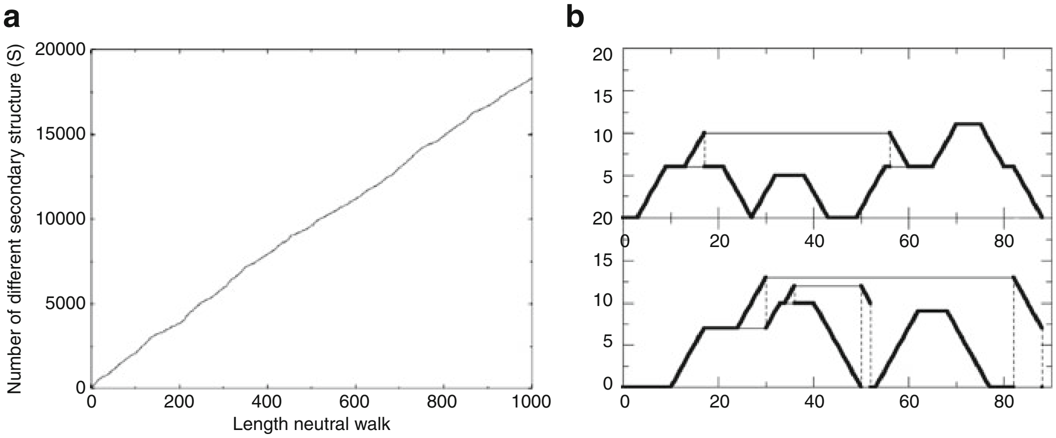


Fig. 10.1 *Innovations along the neutral path.* (**a**) the number of novel structures seen along the neutral path through mutations. No leveling off is observed (adapted from [30]). *Right*: an example of the "meeting" of two different functions on the same sequence. The two different folds have no single base pair in common, and their enzymatic function has been tested in vitro. Full functionality is reached through one point mutation [52]. *Upper panel*: ligase fold. *Lower panel* HDV fold. The foldings are represented by the mountain plot representation [25], extended to display pseudoknots. The mountain plot representation facilitates comparison of structures by preserving the primary sequence along the *x* axis from 5′ to 3′ end. Base pairs are on the same *y* value; *horizontal stretches* represent single-stranded regions. Pseudoknots are superimposed and indicated by *thin horizontal lines* and *vertical boundaries*: e.g., in the HDV fold, six bases of 3′ end fold back on the 5′ bulge of nucleotides 24–30

Moreover, there are many additional constraints, e.g., "matching" brackets. Nevertheless, a sample of a million random sequences of length 70 typically has 999,919 different structures (26 sequences do not fold) (see also [22]).

- *Mutational neighborhood: Smoothness*. Nevertheless, for length 70, ca 30% of the 1 point mutants fold into the identical structure. For longer sequences, this percentage saturates at 20%, whereas for length 30 sequences, it is about 50%. Somewhat farther away, the number of identical structures decreases somewhat less than exponential, but at distance 5, no more than 0.5% folds in the identical structure [58].
- *Mutational neighborhood: Ruggedness*. On the other hand, a single point mutation may also change the structure completely in the sense that not a single Watson–Crick base pairing is conserved. Figure 10.1b shows a beautiful experimentally verified example [52]. The sequences of two functionally different ribozymes were changed "toward each other," till finally, a sequence was obtained which can fold in both structures, which are still functional. One point mutation in each direction recovers full functionality. Note that in this case, it is not a standard secondary structure, but it contains pseudoknots, which are not considered in the computational experiments: nevertheless, the described properties of mutational neutrality and sensitivity apparently hold for these more complicated structures as well [52].
- *Neutral networks*. Identical structures with genotypic Hamming distance 1 or 2 percolate through sequence space [53], forming a so-called neutral network. The

percolation means that a sequence can change entirely while still keeping the same structure, and a certain structure is relatively close to any random initial sequence.
- *Intertwined networks*. The neutral networks of different structures are intertwined in the sense that typically somewhere on their neutral networks, any two structures "meet," i.e., are in each others, close mutational neighborhood. This is shown in Fig. 10.1a: along a path on the neutral network, new structures occur in the neighborhood in a constant rate [30].

In other words, the landscape is very rugged, as one step can bring us from maximum to minimum height. Nevertheless, they are smooth as well: no local peaks as there are always identical structures nearby (we can stay on one level).

2.2 Evolutionary Dynamics on RNA Landscapes

Evolutionary dynamics on RNA landscapes was studied by using distance to a target structure as fitness criterion. The consequences of the shape of the landscape are profound:

- *Dynamics on neutral network*. An evolving population will spend much time diffusing on a neutral network. This diffusion is similar to the neutral evolution on a flat landscape, as first described by Kimura [38], but the diffusion coefficient scales with the connectivity of the neutral net [32]. The population can travel a very long way over the neutral network in the time it will take to cross a fitness barrier. For a neutrality corresponding to a random RNA of length 70, this would amount to more than 10^9 neutral sequences explored in the time that a "ditch" of width three mutations and a depth of just 1% can be crossed (for mutation rate 10^{-6}) [63]. In other words, the problem of local peaks in lower dimensional spaces can be avoided by large detours in high-dimensional spaces. Interestingly, the random walk on the neutral network "is going somewhere," namely, to a region of the neutral network that is smoother [31], i.e., has higher connectivity than the average. To be more precise, the neutrality "seen" by the population after prolonged residence on the neutral network converges to the largest eigenvalue of the connectivity matrix [64]. For random sequences of length 70, this amounts to an increase in fraction of neutral neighbors of ca 0.3 to larger than 0.4. In other words, the robustness against mutations increases over evolutionary time. This is well known from experimental evolutionary studies in that populations which adapt to a certain environment initially have a very high mutational load (low robustness) relative to the wild type [51].
- *Neutral networks and adaptation/innovation*. Adaptation from a random sequence to an arbitrary structure shows periods of constant fitness, punctuated by adaptive steps [19, 31]. During constant fitness, the population diffuses over the neutral network. When it "meets" a structure closer to the target, it moves up to this new neutral network. In other words, the properties described above about

diffusion on the neutral network hold most of the time, i.e., the evolutionary process is dominated by neutral drift. However, this neutral drift helps adaptation because it prevents the population to get stuck on a local optimum, and the population can explore a huge amount of the genotype space. Moreover, in doing so, indeed more and more different structures are encountered (Fig. 10.1a). As Zuckerkandl [72] emphasized in his Kimura memorial lecture entitled "Neutral and nonneutral mutations: the creative mix," this result reconciles the neutral and adaptive theory of (molecular) evolution. A step from one neutral network to another can involve a complete change in the structure as we have seen above. The entanglement of the different networks ensures that the evolutionary process is capable of drastic innovations.

- *Evolution of robustness and evolvability*. The amount of neutral network explored depends not only on mutation rate but also on the connectivity of the network, as mentioned above. Since the population moves during evolution toward parts of the network that are more highly connected, the potential for exploration is increased as well. This leads to larger population variability at any point in time, as well as more movement over time. Accordingly, the chance of "meeting" a new neutral network with higher fitness increases as well. In other words, the chance of adaptation (and the potential of innovation) will increase over evolutionary time. Intuitively, it has long been assumed that mutational robustness and evolvability are incompatible with each other. These results show, on the contrary, that both features, increased mutational robustness and increased evolvability, emerge automatically from basic mutation selection processes in fitness landscapes, as exemplified by the RNA landscapes.

2.3 "Just" RNA?

The above described features of evolutionary systems were derived from studying one specific molecule, RNA. The observed features led Schuster to conclude that RNA is an "ideal evolvable molecule" [53]. Unfortunately, they have often been interpreted as features "just" for the RNA landscape—interesting as they are as such. However, such an interpretation is much too narrow: RNA was used as a paradigm system for some of the hallmarks of evolving biological systems [18, 19], exemplifying systems with large genomes, and highly nonlinear, and redundant genotype–phenotype mapping. Indeed, in subsequent work, these features of landscape structure and evolutionary dynamics have been rediscovered in lattice-based models of protein folding (e.g., [17, 48, 49]), the genotype–phenotype mapping of metabolic networks (e.g., [34]), and regulatory networks (e.g. [5, 13, 14], and see below). Indeed, the important insights on the reconciliation of neutralism and selectionism [67] as well as the compatibility of robustness and evolvability [68] and the origins of innovations [69] have recently been reemphasized by Andreas Wagner [66, 69] on the basis of systematic studies on RNA and protein folding, as well as the structure of regulatory networks and metabolic networks.

The conclusion is that these features first derived from studying a specific example (RNA) are indeed generic properties of biological evolution. Nevertheless, they were previously overlooked in "more general" models of evolution because of various simplifications (low dimensionality, linear mapping, random ruggedness, etc.). This demonstrates that an in depth study of a particular example can lead to more generalizable conclusions than models in which simplifications were made for the purpose of being general. This endorses the nonsupervised modeling approach that we advocate.

3 Evolutionary Structuring of Genomes and Regulomes and Mutational Spaces

In the studies described above, only point mutations were considered. Whole-genome sequencing studies have shown that genomes are much more flexible than previously thought. Duplication and deletion of stretches of DNA are rampant. Even in short-term evolutionary adaptation, GCR plays a major role. The static picture of a genotype space, and of adaptive walks navigating this space by point mutations, is clearly not all evolution is about. To explore the consequences of such more dynamic genomes for evolutionary dynamics, we use as basic representation a genome with genes and transcription factor binding sites (TFBS). Mutations are at the genome level and include duplications and deletion of stretches of DNA, representing genes, binding sites, or GCR, as well as point mutations changing the specificity of genes and/or binding sites. These genomes can code a regulatory network, and therewith gene expression. Because, given these mutational operators, the genotype space is not predefined, such a system is, strictly speaking, not amenable to analysis in terms of fitness landscapes or in terms of standard dynamical systems.

They are however amenable to nonsupervised modeling: given the structure of the genomes and genetic operators, we study the emerging phenomena. I will review models with an increasing number of levels above the genome. First, I will discuss a model without selection, where we study how the process of duplication and deletion by itself structures the topology of regulatory networks. Next, we add selection, and we study adaptation to changing environments in gene expression, and finally, we add a layer of metabolism, evolving regulation to maintain homeostasis in a variable environment. We will show that structuring of genomes and regulomes during evolution leads to the evolution of evolvability in ways which go beyond the increase of evolvability through increased neutrality discussed above. We will compare the results to short-term in vitro experimental evolution and to longtime phylogenetic patterns observed in fully sequenced genomes.

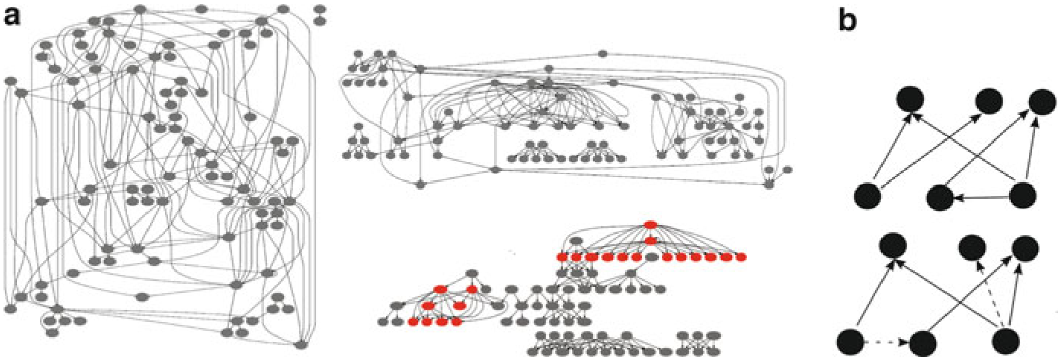


Fig. 10.2 *Network structuring through random mutations*. (**a**) The transformation of a small toy network by random duplication and deletion of genes and TFBS is shown: a random network is transformed in a hierarchical structured network. The *red nodes* are part of the neutrally generated feed forward motifs. (**b**) Connectivity preserving transformation: only by changing two links simultaneously preservation of the connectivity profile can be guaranteed (adapted from [6])

3.1 From Random Mutations to Nonrandom Networks

Classical evolutionary theory assumes that random mutations lead to random phenotypes unless guided by positive selection or constrained by negative selection. This is indeed true to a large extent when we consider point mutations only. Given that other genomic changes (mutational operators) play at least as large a role as point mutations, a better visualization of the mutational part of the mutation selection process is to see it as a stochastic dynamical system governed by the mutational operators as the transition rules. The attractors of these dynamical systems may have a very distinct, and counter intuitive, structure. The consequences of random duplications and deletions of genes and of TFBS were studied by [6, 65], by simply implementing them together with point mutations which change the specificity of the TFBS or the transcription factors.

There have been quite a few network models which showed that certain type of network transformations leads to networks with certain features in common with biological networks (e.g., [2, 37, 47]). The above described model differs from most of these in that a clear separation of genotype and phenotype is maintained, where mutations take place at the genome level and not directly at the network level. This has important consequences, for example, a change in gene specification impacts on many network connections. Although one can implement this at the network level, such a rule should seem to be ad hoc, but it is a default choice given the underlying genome structure.

Figure 10.2a shows the transformation of a toy random network when subjected to these mutations. The resulting network is clearly much more hierarchically organized than the initial network. Thus, we should conclude that random mutation leads to a hierarchically structured network.

Moreover, when duplication/deletion rates of binding sites are larger than those of genes, and we initiate the process with a random network which corresponds in terms of genome size, number of transcription factors, and average connectivity to the yeast transcription regulation network, we see the following results:

- Like the coexpression network of yeast, the coexpression network resulting from this mutational process has a small world, scale-free architecture [65].
- Like the in-degree of the yeast transcription network, the in-degree of the networks generated by the mutational process follows a power law with exponent 2 [6].
- Like in the yeast network, there appears to be an overrepresentation of feed-forward motifs (FFL) in the network [6]. Moreover, the higher-order organization of these feed-forward loops is of the type called "multi-output" by [35], like it is in yeast. In the toy model of Fig. 10.2a, the nodes belonging to such feed-forward loops are colored red. They appear in the mutational process when a hub gene is duplicated, and a connection between the two duplicates is established. In the yeast network, we see this architecture, for example, in the cell cycle genes.

Because of these multiple similarities of the model with the yeast regulatory network, it is tempting to conclude that these features are the result of neutral processes in yeast as well. However, the dichotomy between neutral and adaptive processes is too naive. In the remainder of this chapter we demonstrate a tight mutual dependence on mutation and selection: what is neutral can a side effect of selection, and vice versa. The conclusion that random networks are quite special does however hold.

The important observation here is that comparing network structures with "random" networks is often very misleading. In testing the overrepresentation of the FFL in the empirical networks, they were randomized keeping the degree profile constant, i.e., the number of edges of the nodes was held constant [43]. Figure 10.2b depicts a transformation step which preserves this profile. It is clear that such a double step is unlikely both by mutation and by selection. Moreover, there is no reason to suppose that the degree profile is selected for!

3.2 Evolution of Evolvability: Mutational Priming

Evolutionary experiments show that adaptation to new environments often is a surprisingly fast process. High-throughput experiments on yeast adaptation to a new environment have shown that over a short time span, adaptation occurs and involves massive changes at both at the level of the gene expression [16] and at the genome level [15]. Expression of about 10% of the genes changes, and duplication and/or deletions of large stretches of the genome (GCR) are observed repeatedly, although also single gene duplications can lead to the massive and "appropriate" gene expression change. Similar changes in gene expression occur in independent evolutionary experiments, and several GCRs re-occur in several experiments. In this section, we explore whether these features, unexpected as they were when first observed, are in fact generic properties of *evolved* evolutionary systems.

Crombach and Hogeweg considered two questions separately: (1) can genomes organize themselves so that few mutations can cause fast adaptation [8], and (2) can regulomes organize themselves such that mutations can cause fast adaptation [9].

For both questions, we extend the basic model of genome evolution introduced in the previous section with selection to a randomly fluctuating environment. The selection criterion is simply the matching of available gene products to the prevailing environment. No sensors of the environment are implemented such that adaptation can occur by evolution only.

3.2.1 Evolution of Genome Organization

In this model, we focus on genome organization—and exclude regulatory interactions. Adaptation to the environment requires that the copy number of the genes matches the environment. Part of the genes are housekeeping genes that are always needed in the same amounts, whereas the one or two sets of other genes should be present in one or two copies dependent on the environment. Indeed, gene duplications/deletion often act in early phases of adaptation through dosage effects [21]. We use a diploid genome, and the set of mutational operators used above is extended by mutations related to retrotransposon dynamics. Transposons are duplicated including their long terminal repeats (LTR) and inserted at a random position in the genome. Deletion of retroposons is always by single-stranded annealing, which leaves a single LTR in the genome. LTRs can be deleted as well. GCR happens through double-stranded breaks at LTRs, which are repaired by randomly reattaching chromosome segments (for further details, see [8]).

Figure 10.3 demonstrates the dramatic increase in evolvability during evolution. While early on the population cannot adapt to the prevailing environment before the next environmental switch, late in evolution adaptation is quite fast, and the population is well adapted most of the time. Thus, while early in evolution the population cannot adapt through evolution, it can evolve evolutionary adaptation. The fast evolution is due to the clustering of the housekeeping genes and of the variable genes and flanking these groups by LTRs such that GCR occurs more often in between these clusters. In other words, the random mutations are not random anymore, but favor the duplication or deletion of coherent sets of either housekeeping or variable genes and not a mixture. Such GCRs are either adaptive or very maladaptive, and selection is therefore efficient. Interestingly, this mechanism resembles the one observed in the evolutionary experiment mentioned above [15] where many of the observed GCR in the adapted populations were associated with LTRs.

However, an important difference between the model and the experiments is with respect to the relation between gene expression and gene duplication or deletion. In the simple model, these are assumed to be identical. In the experiments, however,

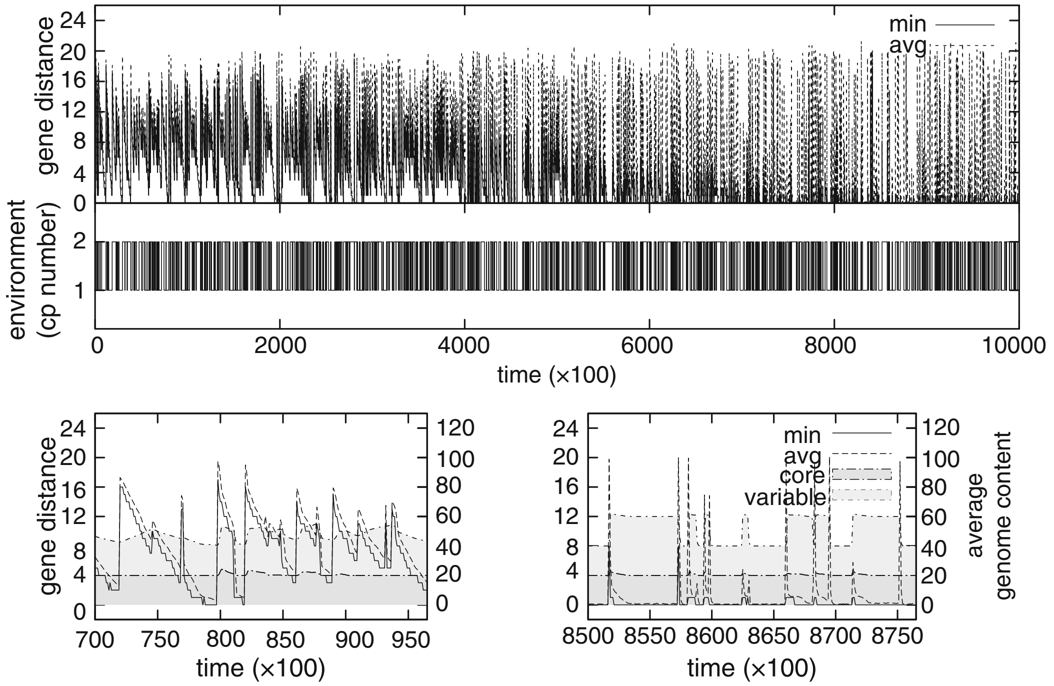


Fig. 10.3 *Evolution of evolvability through genome organization*. *Upper panels*: fitness over time (expressed as distance to target), while below, the switching of the environment is shown (Poisson distribution with $p = 10^{-3}$). Below blowups are shown of the adaptive process early in evolution (*left*) and late in evolution (*right*). Early in evolution, the population is maladapted almost all of the time, whereas late in evolution, it is well adapted most of the time (figure courtesy of A. Crombach)

this is not the case. Although gene expression of duplicated genes is more often enhanced than repressed (and the reverse is true for deleted genes), some duplicated genes are underexpressed and some deleted genes are overexpressed (see Fig. 10.4). This is evidently because of transcription regulation. The power of evolution of transcription regulation to make the effect of random mutation biased toward an adaptive direction is discussed in the next section.

3.2.2 Evolution of Regulome Organization

Here we focus on gene expression. To this end, the dynamics of transcription regulation was added to the model framework described above. Accordingly, the edges of the transcription regulation networks have a weight (encoded in the binding sites), and the genes have an activation threshold, all of which are subject to evolution. The expression pattern of the genes (on–off) in the attractor of this network should match the environment. The required state in the two different environments differs in the expression of nine genes. For further details, see [9].

Like in the previous example, the adaptation rate to the alternative environment is dramatically increased over evolutionary time (from more than 1,000 time step to almost immediate adaption). Figure 10.5 shows that this increase in adaptation

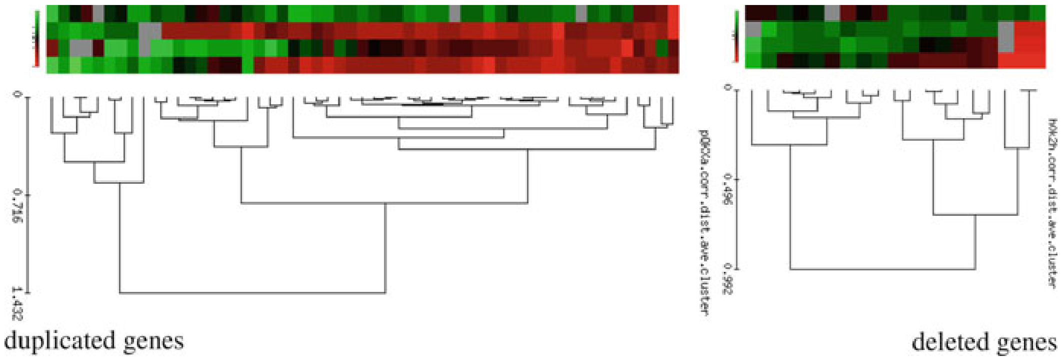


Fig. 10.4 *Expression of duplicated and deleted genes in experimental evolution*. We extracted the genes which were duplicated (275 genes) and those which were deleted (77 genes) in experiment 1 in [15] and selected from these genes those which were significantly differentially expressed relative to the ancestor, as observed in the corresponding experiment 1 reported in [16]. The two dendrograms show the 76 duplicated and the 19 deleted differential expressed genes, respectively. They are clustered according to their expression relative to the ancestor in the three replicate evolutionary experiments, and the ancestor, reported in [16]. The *upper part* shows the expression levels (*red* overexpressed, *green* underexpressed), the *lower part* the clustering. We see that the expression patterns are similar in the replicate experiments. Moreover, we see that, although duplicated genes are more often overexpressed, and deleted genes underexpressed, there are clear exceptions, consistent over replicate experiments. Note that in the replicate experiments, other genomic changes took place. The *upper row* of the heat plot is the expression of the ancestor

rate is accomplished through the effect of almost all types of mutations, as follows. Both early in evolution and late in evolution, most mutations are neutral. Early in evolution, the nonneutral mutations are evenly distributed between positive and negative effects. In contrast, late in evolution, there is a clear overrepresentation of mutations with a positive effect. Moreover, these mutations often have a large positive effect: a relatively large proportion even shifts the gene expression from one target to the other target (changing the expression of all nine differential expressed genes). In particular, duplication and deletions of a single gene can cause such a full switch quite often. These mutations change the attractor landscape in such a way that the attractor with optimal gene expression in the one environment becomes a point in the domain of attraction of the attractor corresponding to optimal gene expression in the other environment. Thus, the adaptation to switching environment can be accomplished immediately by repeatedly duplicating and deleting of the same gene. We called such a gene an “evolutionary sensor.”

We conclude that the evolution of transcription network organization results in nonrandom effect of random mutations. In the light of these results, the similar effects of different mutations (whether large-scale mutations or not) observed in the yeast experiments become less surprising, as does the weak correspondence between gene duplication and overexpression (respectively, gene deletion and under-expression) reported above.

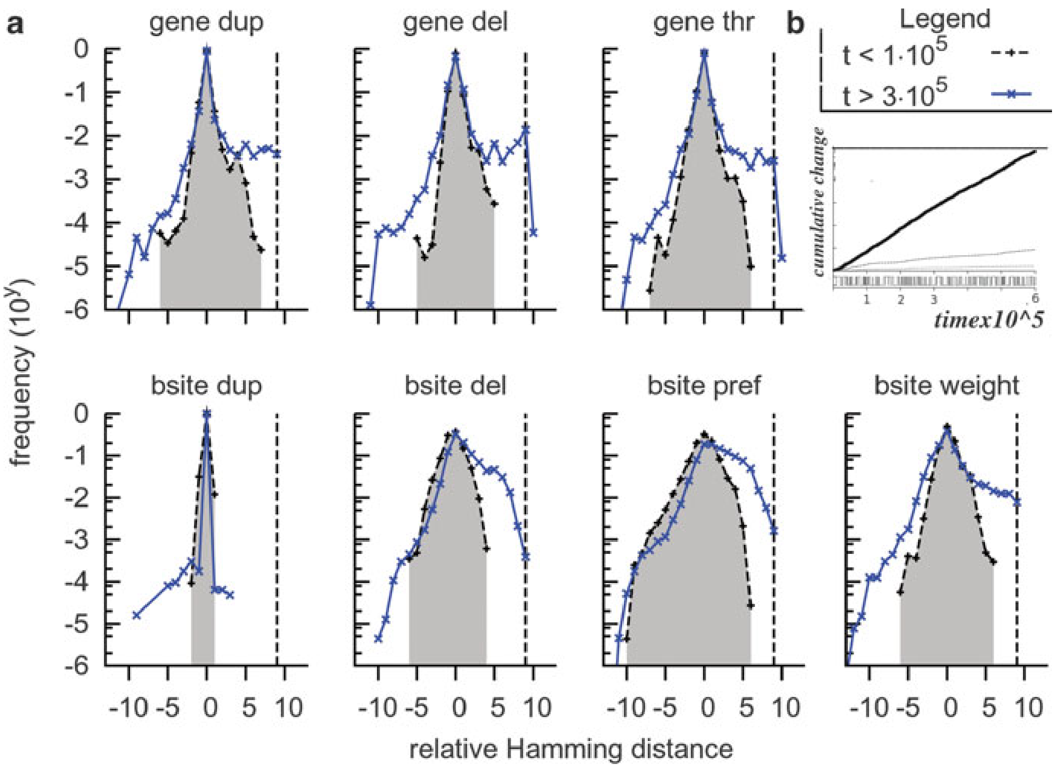


Fig. 10.5 *Mutational priming*. The effect of the various mutational operators is studied along the line of descent. (**a**) Histograms of the effect of mutations around the ancestor lineage. The *gray shaded area* early in evolution ($t < 10^{-5}$), the *blue line* late in evolution ($3.10^{-5} < t < 6.10^{-5}$). X axis: positive (negative) approach toward opposite target, Y axis frequency. Most mutations are and remain neutral; however, late in evolution, there is a clear bias to beneficial mutations (large steps in the right direction). (**b**) Cumulative change over time: despite strong adaptation, neutral mutations (*thick line*) strongly dominate the amount of change (adapted from [9])

3.3 Evolution of Evolvability: Beyond Increased Variability

When we compare these results with those obtained in fixed landscapes, we see that all the results obtained there hold, but are also extended. Like in the fixed landscapes, there are neutral networks: neutral mutations in fact dominate (Fig. 10.5b). Moreover, drastic changes in the phenotype require only single mutations, and evolvability increases during evolution. However, the increase of evolvability is essentially different from the increase in population variability due to an increase of neutrality. In the examples discussed here, evolution actually increases (the effect of) mutational changes "in the right direction." This happens either at the level of mutations themselves or through regulatory effects. In the first example, there were more GCRs which increase/decrease the number of variable genes. In the second example, the effect of almost all the implemented mutations is biased toward the opposite target. Thus, through genome and regulome organization "random mutations are not random," but biased toward beneficial mutations. These results appear to reflect the observations in short-term evolution in yeast mentioned above, where both mechanisms appear to be present. We should note however that yeast, unlike these models, can adapt to prevailing conditions by sensing the environment and thereby trigger changes of attractors of the gene regulatory networks. In the yeast experiments, changes in expression patterns were measured after regulatory adaptation. It was noted however [16] that the evolutionary adaptation partially reflects regulatory adaptation: e.g., genes in respiratory pathways are overexpressed, and genes in fermentation pathway are underexpressed relative to the ancestor (which is allowed to regulatory adapt to the poor environment) in the strains evolutionary adapted to the poor nutrient conditions. It seems likely that evolution of such direct regulation helps the evolution of evolutionary adaptation: they work via the same regulatory network. The reported computational experiments show however that this help is not needed to shape regulatory networks so that only one or a few mutations are needed for appropriate attractor switching and, moreover, so that many different mutations can accomplish this switch.

An obvious and important objection could be that evolution is only toward targets "which have been seen before" and therefore is not "real" evolution. This is true, but one should realize that to a large extent, the experimental evolution of yeast reflects this situation: it is likely that yeast has had to adapt to low nutrient concentrations in its evolutionary history! Even if not exactly toward the experimental conditions, a similar evolutionary response should at least increases fitness.

Nevertheless, the objection is relevant, and we will discuss evolution of evolvability to novel circumstances in the next section, where we add new layers to the transition between the genome and the phenotype.

3.4 Genome Size Dynamics and Evolvability of Virtual Cells

In the previous examples, we equated a gene-expression state with a fitness in a certain environment. In the next example [10], we add more flexibility and more layers as we define an evolving entity which actually has to cope with a changing environment and can "choose" how to do it. Thus, we add an important new level and therewith degrees of freedom of the evolutionary process. We evolve (virtual) cells instead of just networks. These virtual cells should evolve regulatory adaptation to maintain a stable internal state despite wide fluctuations in the external environment.

The virtual cells [44] have anabolic and catabolic enzymes, transporters, and consume one resource, which fluctuates widely (three orders of magnitude) in the environment, and passively diffuses through the cell membrane. The cell copes with this environment when it can keep the concentration of the resource (A) and of an energy carrier (X) at a predefined value, i.e., if it can maintain homeostasis. Catabolic enzymes convert resource into X, and X is used by anabolic enzymes to convert resource to building blocks and by the transporters to transport resource into the cell. The proteins are encoded in the genome and associated with TFBS. Transcription factors regulate gene expression depending on their binding to ligands

A and X. Mutations include duplication and deletion of stretches of the genome, as well as changes in the binding constants etc. The genome is translated in a set of ODE; the intracellular concentration of resource and energy carrier in the fixed point of the intracellular dynamics determines fitness (homeostasis) (for further details, see [10, 44]).

Previous work (e.g., [45]) has shown that, counter intuitively, sparse fitness evaluation facilitates the evolution of regulation. In other words, when only a very small subset of possible environments is encountered per generation, better adaptive regulation evolves to all possible environments. Thus, regulation evolves by long-term information integration better than by direct evaluation against all relevant information. Accordingly, in our model, a cell encounters only 1–3 environments in its lifetime—and its fitness is determined by how well it maintains homeostasis in the encountered environments. However, to assess how well a cell performs, we evaluate it on a set of standard environments, spanning the entire range of variation.

In line with our nonsupervised modeling strategy, we evolve these virtual cells and observe what happens during evolution. Some striking features of the evolutionary dynamics are summarized below:

- *Early large expansion of the genome size*. A "typical" pattern of genome size dynamics is shown in Fig. 10.6a: early in evolution, there is a large expansion of the size of the genome. This pattern is more extreme in the subset of runs that do attain high fitness late in evolution (as the one shown indeed does) but is seen in almost all evolutionary runs. This is shown in Table 10.1. Those runs (ca 50%) which do attain high fitness late in evolutionary time have a significantly larger genome expansion early on than those which do not attain high fitness. The large size expansion is significantly correlated with a slight bias for beneficial duplications. However, this bias is responsible for only a small part of the expansion: most of the size increase is due to near neutral (or even harmful) mutations. Accordingly, and interestingly, the fitness during the early stages of evolution is not different between those runs which do have the large expansion or those who have less expansion or between those which reach high fitness later on and those which do not attain high fitness. We conclude that early genome expansion facilitates evolution to high fitness much later in evolutionary time.
- *Gene-loss during adaptation*. After the initial expansion, genome size reduction occurs while fitness is still increasing. While this is happening, duplications are still more likely to be beneficial than deletions; nevertheless, genome size decreases. An important driving force in gene loss is the deleterious effect of mutations of nearly neutral genes.
- *Shape of fitness landscape*. Figure 10.6b shows that the degree of neutrality is maintained, notwithstanding large fitness increase (as opposed to, e.g., [1, 63]). Even more unexpected is the increase in the frequency of lethal mutations, while the number of slightly deleterious mutations decrease. Nevertheless, this makes "sense" in that strong selection is maintained.
- *Increased evolvability to novel situations*. Once high fitness is reached in the prevailing circumstances with large fluctuation in resource availability,

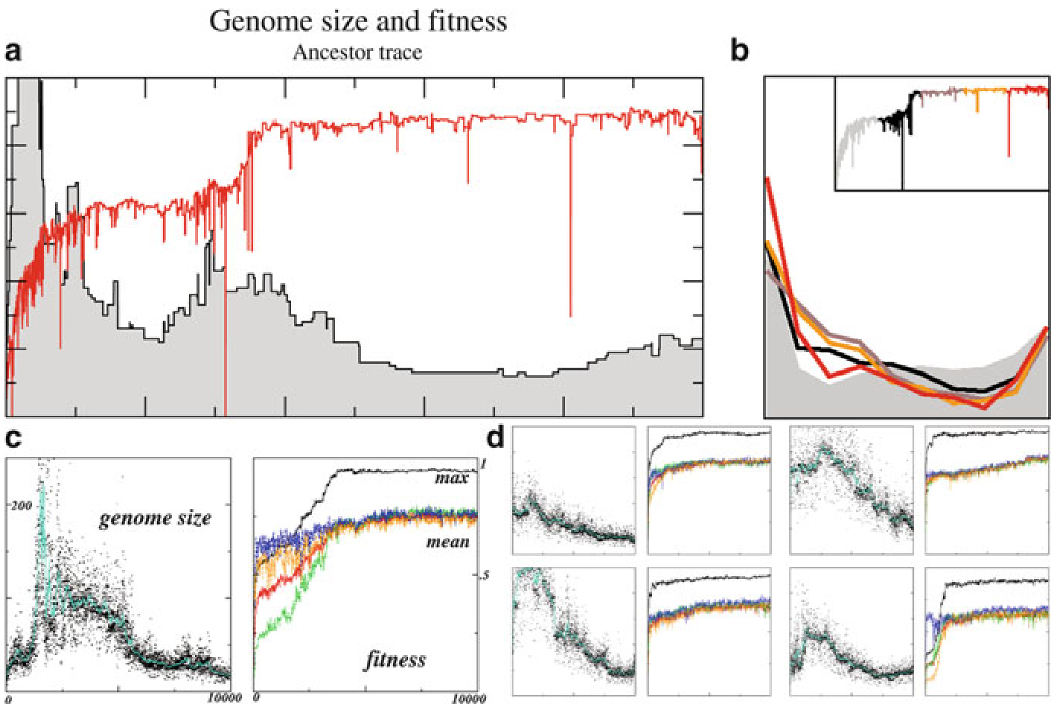


Fig. 10.6 *Genome dynamics, evolved mutational landscape, and evolvability*. (**a**) Typical evolutionary dynamics over time. *Red line* fitness along line of decent, as measured in three standard environments; *gray filled area*: genome size. We see early expansion of genome size, followed by streamlining. (**b**) Changes in mutational landscape over time. Fitness decrease by mutations in ancestral genomes, averaged over 5 time periods of 2,000 time steps (color, see *inset*). *X*-axis percentage of fitness remaining after a mutation, ranging from 0% (lethal) to 100% (neutral). *Y* axis frequency. The mutational landscape becomes more U-shaped. The frequency of neutral mutations remains constant despite fitness increase, slightly deleterious mutations decrease, and lethal mutations increase. This assures effective selection. (**c**, **d**) Fast adaptation to novel environments. (**c**) original run: high fitness is reached at $t = 3{,}800$. (**d**) four examples of an environmental switch at $t = 3{,}800$ of original run: almost immediate regain of fitness. *Left panels* genome size, *right panels* fitness: *black line* maximum fitness in population, *colored lines* average fitness in population at several resource concentrations (figure courtesy of T. Cuypers)

adaptation to entirely new circumstances is extremely fast. The new circumstances were simulated by altering the nonevolvable parameters of the model, e.g., set point of the homeostasis, diffusion of resource through the membrane, conversion ratios, and degradations rates. After these drastic changes, fitness falls to very low values, but recovery to high fitness values takes less than 100 generations, Fig. 10.6c, d shows four typical examples where different combinations of these changes were applied.

The pattern of expansion and streamlining is typical (or generic) in the following sense: (1) it occurs in our default parameter setting in those runs which attain high fitness (Table 10.1). (2) In other (mutational) parameter regimes, less often high fitness evolves, and the pattern is seen less. A high fitness filter to recognize generic patterns in evolution is appropriate as we are prone to encounter only those

Table 10.1 *Local landscapes and future fitness*. The fitness of duplications and deletions relative to the ancestral genomes. We extracted the genomes of the ancestral lineage (i.e. the lineage which gave rise to all genomes in later populations) of 74 evolutionary simulations. We subjected the ancestors to 50 duplication and 50 deletion mutations and determined the fitness as fraction of the ancestor's fitness. We compare the fitness effects in those runs which in the end reached high fitness, with those which did not evolve high fitness over the first 200 time steps. A $+$ indicates significant more in the fit runs, $-$ significant less in the fit runs, and $=$ no difference (parenthesis indicates almost significant). We observe that evolutionary trajectories which reach high fitness after 10,000 time steps have significantly more positive-effect duplication mutations in the first 100 and 200 steps than those which do not reach high fitness. They also have larger genomes, but remarkably, they do *not* have higher fitness yet in this period

Duplications			Deletions		
t = 1–100	t =101–200	ΔF	t = 1–100	t = 101–200	ΔF
$+$	$(+)$	>1.05	$=$	$=$	>1.05
$(+)$	$+$	0.95–1.05	$=$	$+$	0.95–1.05
$-$	$-$	<0.95	$=$	$-$	<0.95
Genome size			Fitness		
t =1–100	t = 101–200		t = 1–100	t = 101–200	
$+$	$+$		$=$	$=$	

organisms which indeed obtain high fitness. We have observed similar genome expansion and streamlining needed for efficient adaptation in a very different model in which LISP programs are evolved to approximate an algebraic function [12]. Although further research is needed, we expect that this is truly a generic evolutionary pattern, given enough degrees of freedom and the need for subtle regulation.

Interestingly, this pattern of genome expansion and streamlining nicely reflects one of the big surprises that emerged from the phylogenetic analysis of fully sequenced genomes: unexpected large genomes in early ancestors and a major role of gene loss in later evolving, often more complex, species. The pattern is beautifully mapped in the reconstruction of Archean genome dynamics [11]. It occurs at all different timescales. For example, a striking case is the large number of HOX genes in amphioxus, and their loss in vertebrates [28]. The pattern also occurs within one genus: gene loss dominates gene gain in all terminal branches of the Drosophila radiation [23].

It turns out that also the evolved U-shaped fitness landscape, surprising as it was to us, actually is reflected in the fitness landscape found in yeast relatively to naturally occurring mutations [70]. In the case of yeast, the pattern is even sharper than the one evolved in our virtual cells over relatively short times: only close to neutral and close to lethal mutations were observed in the experiments.

This virtual cell example again highlights the importance of long-term effects in evolution, the shaping of the mutational landscape, as well as the evolution of evolvability. The latter being to entirely new circumstances in this case.

4 Evolution Toward Multilevel Evolution

In the previous section, we studied the impact of multilevel evolution by implementing successively more complex dynamics between the level of genetic encoding and the level on which selection takes place. In other words, we bypassed the question how/why such an complex mapping did evolve. Earlier work has shown that spatial patterns which automatically emerge through local interactions, constitute a new level of selection [4, 27, 50], and indeed, the emerging waves can themselves be considered as "Darwinian entities" [57] evolving by, e.g., maximizing birthrate. At the level of replicators, this may lead to very counterintuitive evolutionary results, e.g., positive selection for early death (without any trade-offs implemented) [3]. In this section, we examine how such an emerging higher level selection feeds back on the structure of the genomes and the mapping of genomes to function.

4.1 Mutation Rates, Mutational Landscapes and the Structure of Evolved Sequences, Populations and Ecosystems

To this end, we return to the RNA world, where the genotype is the RNA sequence and the function is determined by its secondary structure [56]. We again allow only for point mutations. However, we now add the potential for interaction between molecules. A particular secondary structure defines replicase function. If the single-stranded $5'$ end of a replicase sequence binds to the single-stranded $3'$ end of an other RNA sequence by complementary base pairing, the latter is replicated (with mutations). Note that in this model, only the structure of genomes (RNA) and reactions are defined. An interaction network between replicators may (of may not) emerge through evolved sequence complementarity. The RNA sequences are embedded in space. In one Monte Carlo simulation step, a sequence has a probability to diffuse, to decay, or to interact with other sequences. Complex formation between two adjacent sequences takes place by complementary base pairing between $5'$ and $3'$ dangling ends of the molecules; the complex can fall apart, and the complex of a replicase and another sequence (template) can lead to replication of the template, when empty space is available in the neighborhood. The replication produces the complementary strand of the template. The embedding in space allows for spatial pattern formation, depending on the interaction topology which may evolve. Without spatial pattern formation, the system would go extinct by exploitation by so-called parasites, i.e., sequences that bind more strongly than the replicases to the $5'$ end of replicases, but are not replicases as they do not fold in the predefined replicase structure. This model truly represents the nonsupervised modeling approach, maximizing evolutionary degrees of freedom and minimizing a priori specification. For further details, see [56].

Here I highlight the results which show how the coding of the sequences, and the structure of the population and ecosystem, evolves under different evolutionary regimes.

- *The shape of the quasi-species at high mutation rates*. At very high mutation rates ($\mu = 0.015$ per base), only one quasispecies survives. In order to obtain a viable system with very high mutation rates, the coding of the replicase has to be evolved by slowly increasing mutation rates: random initial replicases (i.e., a sequence which folds in the catalytic structure and whose plus and its minus string can be replicated) are over the error threshold and die out. However, through evolution, sequences that tolerate high mutations rates emerge. The survival strategy of the evolved quasispecies is NOT to maximize neutrality (and thereby increase the phenotypic error threshold) as would be the case in noninteracting RNA (and other) landscapes [58, 64], as discussed above. On the contrary, only 8% of the distance 1 mutations of the master sequence is a viable replicase, and apart from one possible neutral mutation, they are all less fit. Accordingly, the variability in the quasispecies is very low (see "C catalyst" in Fig. 10.7). This strategy evolves because it protects the quasispecies against mutations in two ways: none of its nearby mutants is a "parasite," i.e., a noncatalytic sequence of which both strands can be replicated, and most are "junk" molecules. These junk molecules are not "viable" as they cannot be replicated both as + and as − strand. However, they dilute the population and prevent parasites, which could emerge as rare (distant) mutants, to receive enough catalysis to survive. Thus, although opposite to what we saw before, also here coding structure evolves mutational robustness according to the prevailing circumstances.
- *Niche creation and alternative coding structures at lower mutation rates*. On lowering the mutation rate, speciation into several lineages occurs. The lineages are named according to the most prevalent bases in the $5'$ or $3'$ dangling ends, as detailed in Fig. 10.7. First, a strong parasite lineage evolves ("G parasite"). The parasites are not part of the catalyst quasispecies but form a separate lineage and optimize their primary and secondary structure to maximize the amount of catalysis they get in both strands. It locally outcompetes the C catalyst, and a characteristic wave structure emerges. At still lower mutation rates, a niche is created for a second catalytic species (A catalyst). The second catalyst "chooses" a very different coding strategy: it does maximize neutrality. It can afford to do so because of the lower mutation rates. The high neutrality increases population variability. This is an alternative strategy against parasitism [27] but likewise can harm self-replication. At still lower mutation rates, the latter catalyst, having decreased population variability, is parasitized. The resulting four-species ecosystem is depicted in Fig. 10.7. The spatial structuring stabilizes this strongly parasitized system and creates the niches which allow for (or demand) alternative coding strategies.

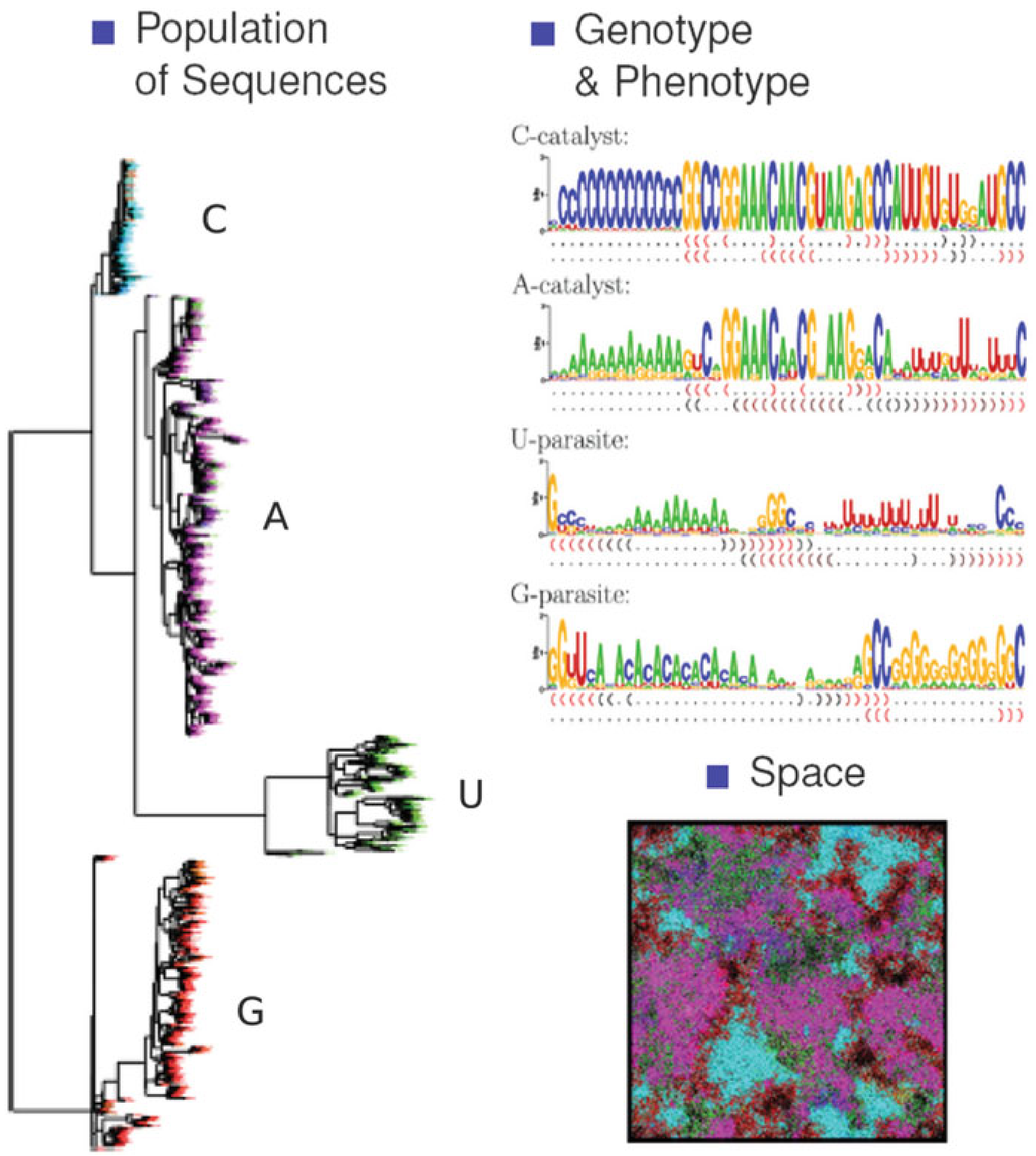


Fig. 10.7 *Evolved structure of populations, individuals, and ecosystems*. The population structure is shown in the *left panel*: four lineages (species) have evolved and stably coexist. The lineages are called C catalyst (*cyan*), A catalyst (*magenta*), U parasite (*green*), and G parasite (*red*), respectively on the basis of the prevalence of the bases at the 5′ end for catalysts and the 3′ end for parasites. The genotype and phenotype of evolved lineages is shown in the *upper right* picture as a sequence logo (using standard coloring for the bases) of the genotype and the bracket notation for the phenotype, where highly conserved base pairings are colored *red*. The spatial structure of the ecosystem is shown in the *lower right* picture. The coloring corresponds to coloring in the phylogenetic tree. The G parasite outcompetes the C catalyst, creating a niche for the A catalyst and its parasite (U parasite). Note the difference in within lineage variably (adapted from [56])

- *Mutation rates and ecosystem stabilization*. At even lower mutation rates, no stable eco-evolutionary system is maintained. Instead, a red queen dynamics is seen in which evolved parasites outcompete the resident catalyst, but an escape

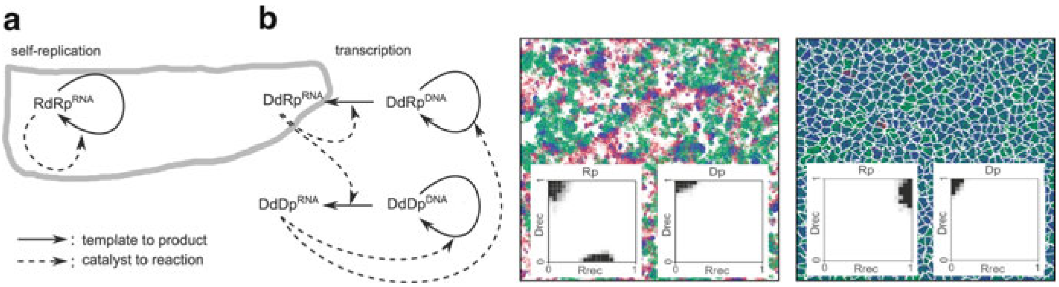


Fig. 10.8 *Evolution of DNA in the RNA world*. *Left*: schematic view of the model. (**a**) Self-replication: RNA-dependent RNA polymerase (in RNA form) (RdRP$^{\mathrm{RNA}}$) replicates itself. This represents the RNA world, and the model is initialized as such. (**b**) transcription: the RNA form of DNA-dependent RNA polymerase (DdRP$^{\mathrm{RNA}}$) transcribes both itself and DNA-dependent DNA polymerase (DdRDP$^{\mathrm{RNA}}$) from the corresponding DNA, whereas DNA-dependent DNA polymerase (DdDP$^{\mathrm{RNA}}$) replicates DNA of both polymerases. Note that other interaction schemes may evolve, e.g., reverse transcription. *Right*: evolutionary outcome of the surface system (*left*) and the protocell system (*right*). Snapshot of the space with *blue* RNA polymerase (Rp) molecules, *green* DNA polymerase (Dp), *red* parasites. DNA and RNA forms are not distinguished. *Inlay*: 2D histograms of the recognition of DNA and RNA by Rp and Dp. Both systems evolve to a combination of the self-replication and the transcription system, whereas reversed transcription is avoided since Dp recognizes only DNA and not RNA. In the surface system, Rp speciates in an RNA recognizing and a DNA recognizing lineage, whereas in the protocell system, a polyfunctional Rp evolves (this is indicated by the *gray line* in the scheme) (Adapted from [59])

mutant of the catalyst, which is less severely parasitized, takes over subsequently and so on. The stabilization of ecosystem interaction by maintaining high population diversity (by high mutation rates and or high neutrality) is an interesting feature, also seen in more simple models [61], emphasizing the interlocking timescales of ecological and evolutionary processes.

This example highlights the mutual dependence on multiple levels of organization. Not only do the lower levels determine the higher levels, but the higher levels feed also back on the structure of the lower levels. This mutual dependence is relative not only to the structure of the different levels of organization (genotype, phenotype, and ecosystem) but also on the shape of the mutational landscape around the selected master sequences. Currently, we are investigating how these mutual dependencies can lead to evolution of more complex, i.e., larger, genomes despite high mutation rates in this “structure-based rather than interaction-based” model.

Here we will next study the evolution of more complex genomes by mutual interactions across multiple levels and multiple timescales in a more conventional structured model of the RNA world, which targets one of the major transition in evolution, the take-over of DNA as information carrier.

4.2 The Evolution of DNA in the RNA World

One of the major transitions in evolution was the evolution of DNA in the RNA world. Whereas in the RNA world RNA acts both as catalyst and information carrier, at a certain point, a noncatalytic counterpart of RNA (DNA) evolved which carries the inherited information but is catalytically inactive.[2] The replication cycle becomes longer, involving both replication and transcription. Such a longer cycle should be slower and therefore should be disadvantageous. So, why did a transcription like system evolve? There may be chemical reasons, but here we study whether such evolution can be explained on the basis of eco-evolutionary dynamics alone. One hypothesis which has been put forward is that DNA is a more stable molecule, and the longevity might be advantageous. Here we show that this longevity is not needed to explain its evolution: the division of labor between information storage and using the information for catalysis, by itself, can explain its evolution.

We model a system of RNA and DNA polymerases, which can each exist in DNA or in RNA form [59]; see Fig. 10.8a, b. Each of them can recognize DNA and RNA. The strength of recognition is an evolvable parameter. Recognition of the template leads to complex formation and subsequently to the copying of the template into RNA or DNA dependent on the type of polymerase.[3] Notice that this setup allows for both transcription and for reverse transcription to evolve, as well as any combination of these.

We study this system in two modes, both of which include a level of selection above that of the polymerases. In the surface system, the molecules are embedded in space, and the spatial patterns which emerge constitute this higher level of selection as in the previous example. In the protocell system, the molecules are enclosed in compartments, which, dependent on the number of molecules inside, grow/shrink and can divide and die. Like in the previous case without the higher level of selection the system would quickly die because "giving catalysis" is a strong "altruistic" trait as it takes time and replicates the competing molecule instead of being replicated itself.

We first evolve the RNA world system, including a parasitic RNA which replicates 10% faster than the polymerase. This parasite goes quickly extinct in the protocell system, and it survives in the surface system, forming the characteristic wave patterns of such systems. We then introduce rare mutations of RNA polymerase to DNA polymerase. The DNA polymerase mutant can invade in both systems. After a long and interesting transient, the evolutionary dynamics stabilizes to a state shown in Fig. 10.8. In both systems, transcription-like interaction as well as

[2] DNA can, in fact, be a catalyst as well, but in the model, we define it as noncatalytic as it is in present-day systems.

[3] In this model, we do not distinguish $+$ strands and $-$ strands.

RNA replication occurs. In the surface system, the RNA polymerases have speciated into two types, one recognizing RNA and one recognizing DNA, while in the protocell system (where only a limited number of molecules occur in a protocell), a polyfunctional RNA polymerase evolves which recognizes both RNA and DNA with high affinity. In other words, a transcription-like system evolves coexisting with RNA replication. Ancestor tracing shows, however, that long-term inheritance is mainly through DNA and that this hybrid system does incorporate division of labor between information carrier and catalysts.

This division of labor evolves because it accomplishes what we called "evolutionary stabilization." This concept is most clear-cut seen in the protocell system, but also operates in the surface system. When a protocell happens to lose DNA by stochastic fluctuations, the RNA only cells replicate indeed faster because of the shorter replication cycle and expand in the population. However, at high enough mutation rates, before they take over the entire population, the recognition affinity evolves to lower values, which reduces replication rate, and the DNA containing cells take over the population again. This evolutionary deterioration of the RNA replicator system is because of the dual role of RNA as template and catalyst: by reducing recognition strength, RNA spends less time being a catalyst and more time being replicated. Although the higher level of selection prevents this selection pressure to lead to extinction, the altruistic catalytic behavior is minimized. Because DNA does not act as a catalyst, this selection pressure does not play a significant role in the transcription-like system, and catalysis is maintained at high values. This is the case as long as inheritance via DNA dominates. Accordingly, reverse transcription should be avoided, as it indeed is in the evolved systems: DNA polymerase only recognizes DNA (Fig. 10.8c).

This example shows clearly the mutual feedback between multiple levels of selection. The levels of selection "above" the replicators (waves or compartment) enable the evolution of a multilevel genotype–phenotype mapping, here the evolution of the division of labor between information storage and information usage.

Moreover, a very profound conclusion is that the major evolutionary transition from the RNA world to a DNA and RNA world could have occurred because of the evolutionary properties of this more complex replication system, rather than because of direct functional properties.

5 Discussion and Conclusions

We have studied evolution in a number of example systems which aim to be simple enough for thorough analysis, but at the same time maximize the flexibility of the evolutionary process. In these models, the structure of the genomes, as well as the transformation of the genomes to the properties on which selection operates, can evolve. This happens alongside the direct evolutionary adaption to the changing environments. A recurrent theme in all examples discussed above is that, given this flexibility, long-term information integration occurs which shapes both short-term and long-term evolutionary dynamics. We have seen this phenomenon in different guises in the various examples.

In the network evolution models in a fluctuating environment, we have seen an effect we call mutational priming: mutations and their effect are biased toward those with a large (beneficial) effect. It has often been argued that large changes are likely to be deleterious, as the mutant is prone to fall in the abyss. This is indeed also true in the RNA landscapes, where the deleterious effects of mutations are approximately additive [58]. However, as we have seen in models which are more flexible in shaping the mutations which do occur and/or the effect of these mutations through genotype–phenotype mapping, the effect of such large effect mutations may be biased to beneficial mutations. Nonbeneficial mutations become indeed more strongly deleterious and can therefore be easily weeded out by the selection process. In these examples, we only considered clonal reproduction. An often made argument for the necessity of allowing only for small changes comes from sexual reproduction, as mutants which are very different would be less likely to mate or to be able to produce viable offspring. A model using a similar genetic encoding as the one discussed here, but with obligate sexual reproduction [60], shows, however, similar long-term effects that shape the structure of the genomes, in this case such that recombination between differently adapted individuals produces offspring which is still well adapted to some environment.

The virtual cell model highlights long-term evolutionary effects by showing that a chance slight, but significant, bias to positive effects of gene duplications leads to huge increases in genome size. Many neutral and (slightly) deleterious genes hitchhike along with this genome increase, and accordingly, the genome increase does not lead to higher fitness at the time. However, the increase in genome size correlates with high fitness late in evolution, apparently due to the larger degrees of freedom in larger genomes. The increase of evolvability in these evolved large genomes was shown to make adaptation to novel conditions, never seen before by the evolving population, extremely fast. This beneficial effect of large genomes runs counter to common wisdom which assumes that larger search spaces make adaptation harder. This intuition was already countered in the case of RNA landscapes because of percolating, and intertwining, neutral networks of various functional structures. The virtual cell example suggests that at least looking back from those entities which did obtain high fitness, large genomes with a structure amenable to easy evolution have been part of their evolutionary history. On the other hand, later in evolution, large increases in genome size can be a side effect of the evolution of high neutrality, not only in the sense of evolving neutral genes but also in evolving a decrease in the deleterious effect of mutations of these neutral genes. This happens most effectively for regimes with effective selection (e.g., large populations). We see here an interesting duality with respect to the relation between population size and genome increase due to nearly neutral mutations. On the one hand, small population size decreases the selection, thus effectively rendering more mutations neutral, whereas large population size selects more effectively (faster) for a larger degree of neutrality, both primarily and secondarily (compare [40, 41]).

In the RNA replicase model, we saw that at high mutation rates, replicases evolve for which many mutations lead to loss of function, i.e., to low neutrality. This is in apparent contradiction between the classical results in RNA landscapes in which evolution of high neutrality occurs at high enough mutation rates/population sizes [64], a result which is mimicked in protein folding, regulatory and metabolic models. However, both observations fit perfectly in the more general point highlighted in the series of experiments reported here: the mutations which do occur and their effects evolve dependent on the evolutionary regime to which they are subjected. As we have seen, the low neutrality leads in this case to robustness by preventing parasite invasion, by minimizing parasite creation by mutations, as well as by decreasing the catalysis parasites can receive because of the many noncatalytic molecules which do arise through mutation from the replicase. We conclude that, like in the network models, the spectrum of mutational effects is optimized relative to the prevailing environmental challenges: here the evolved quasispecies protects itself against parasites. Moreover, unlike in the network models, the environmental challenges are not externally imposed, but arise from the evolving replicators themselves.

Finally, in the example of the evolution of DNA in the RNA world, one of the major transitions in evolution, we have seen that the more complex transcription-like system evolves not because of its superior functional properties but because of its evolutionary properties. When the information flow is from DNA to RNA, and not (or rarely) in the reverse direction (compare Crick's "central dogma" [7]), the evolutionary pressure to minimize catalytic strength is alleviated. While the higher level of selection on waves or compartments is necessary to prevent extinction of the simple replicator system, strength of catalyzes is nevertheless minimized. The more complex and hence slower transcription-like system prevents this evolutionary deterioration and is therefore maintained.

All these cases highlight long-term information integration during evolution. Long-term processes are often banned from evolutionary inferences. For example, Maynard Smith and Szathmary explicitly state in their introduction [55] "the transitions must be explained in terms of their immediate selective advantages. . . ." Indeed without such a constraint, explanations may be generated too easily. However, evolution itself is not bound by this constraint, nor are constructive models of evolution. We have seen that long-term information integration does occur as a result of basic mutation and selection processes in the simple models studied. This is because not only adaptation to external environments occurs during evolution, but also the coding of information in the genome, as well as the transformation of the genome into selectable traits, is shaped by evolution.

One of the consequences of the shaping of the mutational landscape through evolution is that adaptive and neutral evolution is even more interwoven than was inferred from the "neutrality aids adaptation" observed first in RNA landscapes. Indeed, the types of mutations which do occur are the product of adaptation. Thus, the dynamics of neutral evolution against the background of evolved genomes is, in fact, the product of long-term evolution in which adaptive and neutral processes are intertwined. In other words, due to mutation and selection, mutation as well as selection evolves.

These features are a consequence of random mutation and selection in multilevel systems. We therefore should expect them to shape biological evolution. The data discussed on experimental evolution of yeast as well as broad evolutionary patterns gleaned from phylogenetic studies of fully sequenced genomes indeed suggest that they did shape biological evolution (see also, e.g., [29, 46, 54]). Nevertheless, a major challenge is now to find more signatures of (the consequences of) long-term information integration in the data. We strongly expect to find such signatures. If they are not found, the challenge would be to unravel the mechanisms which prevented their occurrence.

Acknowledgements I thank my (former) students, in particular Nobuto Takeuchi, Anton Crombach, Otto Corderro, and Thomas Cuypers. I reviewed their work in this chapter, and I thoroughly enjoyed working with them! I also thank my longtime collaborator Ben Hesper for his strong conceptual support.

References

1. Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. Proc Natl Acad Sci 97(9):4463
2. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509
3. Boerlijst M, Hogeweg P (1992) Self-structuring and selection: Spiral waves as a substrate for prebiotic evolution. In: In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II pp. 255–276
4. Boerlijst MC, Hogeweg P (1991) Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites. Phys D Nonlin Phenom 48(1):17–28
5. Ciliberti S, Martin OC, Wagner A (2007) Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci 104(34):13591
6. Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23(10):1931
7. Crick F (1971) Central dogma of molecular biology. Tsitologiia 13(7):906
8. Crombach A, Hogeweg P (2007) Chromosome rearrangements and the evolution of genome structuring and adaptability. Mol Biol Evol 24(5):1130
9. Crombach A, Hogeweg P (2008) Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4(7):e1000112
10. Cuypers TD, Hogeweg P (2012) Virtual genomes in flux: An interplay of neutrality and adaptability explains genome expansion and streamlining. Genome Biol Evol 4(3):212–229
11. David LA, Alm EJ (2011) Rapid evolutionary innovation during an archaean genetic expansion. Nature 480(7376):241–244
12. de Boer F, Hogeweg P (2010) Eco-evolutionary dynamics, coding structure and the information threshold. BMC Evol Biol 10(1):361
13. Draghi J, Wagner GP (2009) The evolutionary dynamics of evolvability in a gene network model. J Evol Biol 22(3):599–611
14. Draghi JA, Parsons TL, Wagner GP, Plotkin JB (2010) Mutational robustness can facilitate adaptation. Nature 463(7279):353–355
15. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natl Acad Sci 99(25):16144

16. Ferea TL, Botstein D, Brown PO, Rosenzweig RF (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci 96(17):9721
17. Ferrada E, Wagner A (2008) Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc Roy Soc B Biol Sci 275(1643):1595
18. Fontana W (2002) Modelling evo-devo with RNA. BioEssays 24(12):1164–1177
19. Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. Science 280(5368):1451
20. Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P (1993) RNA folding and combinatory landscapes. Phys Rev E 47(3):2083–2099
21. Francino MP (2005) An adaptive radiation model for the origin of new gene functions. Nat Genet 37(6):573
22. Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P (1996) Analysis of rna sequence structure maps by exhaustive enumeration I. Neutral networks. Monatsh Chem Chem Mon 127(4):355–374
23. Hahn MW, Han MV, Han SG (2007) Gene family evolution across 12 drosophila genomes. PLoS Genet 3(11):e197
24. Hogeweg P (2011) The roots of bioinformatics in theoretical biology. PLoS Comput Biol 7(3):e1002021
25. Hogeweg P, Hesper B (1984) Energy directed folding of rna sequences. Nucleic Acids Res 12(1 Pt 1):67
26. Hogeweg P, Hesper B (1989) An adaptive, selfmodifying, non goal directed modelling methodology. In: Elzas MS, Oren TI, Zeigler BP (eds) Knowledge systems paradigms. Elsevier Science, North Holland, pp 77–92
27. Hogeweg P, Takeuchi N (2003) Multilevel selection in models of prebiotic evolution: compartments and spatial self-organization. Orig Life Evol Biosph 33(4):375–403
28. Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez È, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ et al (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 18(7):1100
29. Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5(4):299–310
30. Huynen MA (1996) Exploring phenotype space through neutral evolution. J Mol Evol 43(3):165–169
31. Huynen MA, Hogeweg P (1994) Pattern generation in molecular evolution: Exploitation of the variation in RNA landscapes. J Mol Evol 39(1):71–79
32. Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness: The role of neutrality in adaptation. Proc Natl Acad Sci USA 93(1):397
33. Huynen MA, Snel B, Bork P, Gibson TJ (2001) The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. Hum Mol Genet 10(21):2463
34. Kacser H, Beeby R (1984) Evolution of catalytic proteins. J Mol Evol 20(1):38–51
35. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Topological generalizations of network motifs. Phys Rev E 70(3):031909
36. Kauffman S, Levin S (1987) Toward a general theory of adaptive walks on rugged landscapes*. J Theor Biol 128(1):11–45
37. Kim WK, Marcotte EM (2008) Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput Biol 4(11):e1000232
38. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
39. Koonin EV (2011) Are there laws of genome evolution? PLoS Comput Biol 7(8):e1002173
40. Lynch M (2007) The origins of genome architecture. Sinauer Associates, Sunderland
41. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302(5649):1401
42. May RM (2004) Uses and abuses of mathematics in biology. Science 303(5659):790

43. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824
44. Neyfakh AA, Baranova NN, Mizrokhi LJ (2006) A system for studying evolution of life-like virtual organisms. Biol Direct 1(1):23
45. Pagie L, Hogeweg P (1997) Evolutionary consequences of coevolving targets. Evol Comput 5(4):401–418
46. Pál C, Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nat Genet 33(3):392–395
47. Pastor-Satorras R, Smith E, Solé RV (2003) Evolving protein interaction networks through gene duplication. J Theor Biol 222(2):199–210
48. Renner A, Bornberg-Bauer E (1997) Exploring the fitness landscapes of lattice proteins. Pac Symp Biocomput 361–372
49. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10(12):866–876
50. Savill NJ, Rohandi P, Hogeweg P (1997) Self-reinforcing spatial patterns enslave evolution in a host-parasitoid system. J Theor Biol 188:11–20
51. Scharloo W (1991) Canalization: genetic and developmental aspects. Annu Rev Ecol Systemat 22:65–93
52. Schultes EA, Bartel DP (2000) One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. Science 289(5478):448
53. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci 255(1344):279–284
54. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E (2005) Protein structure and evolutionary history determine sequence space topology. Genome Res 15(3):385
55. Smith JM, Szathmáry E (1997) The major transitions in evolution. Oxford University Press, Oxford
56. Takeuchi N, Hogeweg P (2008) Evolution of complexity in RNA-like replicator systems. Biol Direct 3(11). doi:10.1186/1745-6150-3-11
57. Takeuchi N, Hogeweg P (2009) Multilevel selection in models of prebiotic evolution II: a direct comparison of compartmentalization and spatial self-organization. PLoS Comput Biol 5(10):e1000542
58. Takeuchi N, Poorthuis P, Hogeweg P (2005) Phenotypic error threshold; additivity and epistasis in rna evolution. BMC Evol Biol 5(1):9
59. Takeuchi N, Hogeweg P, Koonin EV (2011) On the origin of dna genomes: evolution of the division of labor between template and catalyst in model replicator systems. PLoS Comput Biol 7(3):e1002024
60. ten Tusscher K, Hogeweg P (2009) The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. BMC Evol Biol 9(1):159
61. Van Der Laan JD, Hogeweg P (1995) Predator-prey coevolution: Interactions among different time scales. Proc Roy Soc Lond B 259:35–42
62. Van Hoek MJA, Hogeweg P (2006) In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. Biophys J 91(8):2833–2843
63. Van Nimwegen E, Crutchfield JP (2000) Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? Bull Math Biol 62(5):799–848
64. Van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96(17):9716
65. Van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5(3):280–284
66. Wagner A (2005) Robustness and evolvability in living systems. Princeton University Press, Princeton
67. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9(12):965–974

68. Wagner A (2008) Robustness and evolvability: a paradox resolved. Proc Roy Soc B Biol Sci 275(1630):91
69. Wagner A (2011) The origins of evolutionary innovations: a theory of transformative change in living systems. Oxford University Press, Oxford
70. Wloch DM, Szafraniec K, Borts RH, Korona R (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast saccharomyces cerevisiae. Genetics 159(2):441
71. Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc 6th Int Cong Genet 1:356–366
72. Zuckerkandl E (1997) Neutral and nonneutral mutations: the creative mix; evolution of complexity in gene interaction systems. J Mol Evol 44:2–8