On the evolution of phenotypic exploration distributions

Marc Toussaint

Institut für Neuroinformatik, Ruhr-Universität Bochum, ND 04, 44780 Bochum—Germany
mt@neuroinformatik.ruhr-uni-bochum.de

In *Foundations of Genetic Algorithms, Volume 7 (FOGA VII)*, 169-182, 2003.

Abstract. In nature, phenotypic variability is highly structured with respect to correlations between different phenotypic traits. In this paper we argue that this structuredness can be understood as the outcome of an adaptive process of phenotypic exploration distributions, similar to the adaptation of the search distribution in heuristic search schemes or Estimation-of-Distribution Algorithms. The key ingredient of this process is a non-trivial genotype-phenotype mapping: We rigorously define non-triviality, in which case neutral traits (as a generalization of strategy parameters) influence phenotype evolution by determining exploration distributions. Our main result is the description of the evolution of exploration distributions themselves in terms of an ordinary evolution equation. Accordingly, the "fitness" of an exploration distribution is proportional to its similarity (in the sense of the Kullback-Leibler divergence) to the fitness distribution over phenotype space. Hence, exploration distributions evolve such that dependencies and correlations between phenotypic variables in selection are naturally adopted by the way evolution explores phenotype space.

Keywords: genotype-phenotype mapping, non-trivial neutrality, evolvability, self-adaptation

1 Introduction

Evolution explores a space of creatures (phenotypes) by successive mutation, recombination, and selection. The exploration is governed by the variational topology of possible mutations on phenotype space. For many evolutionary algorithms this variational topology is prefixed ad hoc by specifying phenotypic mutation operators. In contrast, in the case of a *non-trivial* genotype-phenotype mapping (GP-map), and in particular in nature, phenotypic variational topology is *not* fixed. An impressive example has been analyzed by Schuster (1996) and Fontana & Schuster (1998): Protein folding induces very complex topologies on the space of secondary protein structures which depend on the genetic representative (i.e., primary structure) the secondary structure is encoded with. Obviously, the choice of the genetic representative within a neutral set that encodes for a phenotype decisively determines the possibilities of phenotypic mutations and innovations. Consequently is has been argued that the genetic representations in todays organisms, in particular their complexity with respect to dependencies, gene interactions, epistatis, etc., are not a mere incident but the outcome of an adaptive process that optimized these representations with respect to the phenotypic variability and "innovatability" they induce (Wagner & Altenberg 1996).

To formalize that idea of genetic representations and their influence on phenotypic variability is the basis of this paper. The major result is a description of exactly that adaptive process that explains the direction of adaptation, i.e., explains why genetic representations have evolved in nature that encode highly structured and adapted distributions of phenotypic variability: The evolution equation that governs the (self-) adaptation of exploration distributions shows that they are selected with higher probability the better they *match* the fitness distribution over phenotype space; in particular they are selected more likely if they exhibit a correlational structure similar to the correlations between phenotypic variables in selection. Hence, exploration distributions evolve such that dependencies and correlations between phenotypic variables in selection are naturally adopted by the way evolution explores phenotype space.

A new formalism for neutrality clarifies the conceptual relation to other topics in theoretical evolutionary computation. We introduce a genotype-phenotype mapping ϕ such that fitness (generally any kind of selection relevant character) does only depend on the phenotype; two genotypes are called equivalent if they have the same phenotype. Such an equivalence relation is related to a projection similar to schemata or coarse grainings (Radcliffe 1991; Stephens & Waelbroeck 1999). The question is whether this projection is compatible with evolution dynamics, i.e., whether "coarse-grained" evolution can be described without tracking microscopic traits (Vose 1999, chapter 17). We define a genotype-phenotype mapping non-trivial iff phenotype equivalence is *not* compatible, i.e., when phenotype evolution *cannot* be understood (i.e., modeled) without tracking neutral traits. We prove that this is the case whenever phenotypic mutation probabilities depend on neutral traits—vice versa, the "only" impact of neutral traits on phenotype evolution is that they determine exploration distributions. This leads to the embedding of neutral sets in the space of exploration distributions. Eventually, this formalism allows to derive the evolution equation for exploration distributions.

Section 2 introduces the formalism of phenotype equivalence and discusses in general terms under which conditions phenotype evolution depends on neutral traits. Section 3 is an extended discussion of our approach and relations to various topics in evolutionary computation, e.g., the concept of self-adaptation, strategy parameters, and evolving genetic representations. Section 4 describes the embedding of neutral traits in a larger space of exploration distributions appropriate to formulate an evolution equation for them. Section 5 then derives this equation, which we call *σ-evolution.* The equation exhibits a delay effect of evaluation of exploration traits which is well-known for conventional strategy parameters and which we generalize to n-th order delay effects in the appendix. A brief summary concludes.

2 Phenotype equivalence: When does phenotype evolution depend on neutral traits?

The following notation is adapted from Vose (1999, chapter 17). Let Ω denote the space of genotypes; which may be any space, e.g., the space of strings of arbitrary length, trees, or grammars, or even the union of these spaces (which would allow a self-adaptive "change of representation", see later sections). We interpret evolution as discrete-time stochastic dynamics of a population $p^{(t)}$ of genotypes. At every time t, $p^{(t)} \in \Lambda^\Omega$ evolves as

$$p^{(t+1)} = \mathcal{G}\, p^{(t)} \ ,$$

where $\mathcal{G} : \Lambda^\Omega \to \Lambda^\Omega$ is the evolution operator defined on the space (simplex) Λ^Ω of distributions over Ω. In the finite population case, $p^{(t)}$ is a normalized finite sum of delta-distributions and $\mathcal{G}$ is stochastic (due to sampling); here we keep to the infinite population approach, where the population is an arbitrary point in the simplex and $\mathcal{G}$ is deterministic. Later, we will assume that $\mathcal{G}$ is composed of a mutation and a selection operator, the latter of which depends on a fitness function $f : \Omega \to \mathbb{R}$.

Phenotype equivalence. Let $\tilde\Omega$ be a phenotype space and $\phi : \Omega \to \tilde\Omega$ a surjective genotype-phenotype map (GP-map) such that fitness depends only on the phenotype, $f(g) = \tilde f(\phi(g))$, $\tilde f : \tilde\Omega \to \mathbb{R}$. Generally, ϕ may be non-injective. We define two genotypes g_1 and g_2 equivalent iff they have the same phenotype,

$$g_1 \equiv g_2 \iff \phi(g_1) = \phi(g_2) \ .$$

The set of equivalence classes $\Omega/\!\equiv$ is one-to-one with the set $\tilde\Omega = \phi(\Omega) = \{\phi(g) | g \in \Omega\}$ of phenotypes. Thus, we use the phenotype $x \in \tilde\Omega$ to indicate an equivalence class

$$[x] := \phi^{-1}(x) = \{g \in \Omega \,|\, \phi(g) = x\} \ ,$$

which is also called *neutral set* or phenotypic class. As we map genotypes on phenotypes, we can also map genotype populations on phenotype populations: The mapping Ξ projects genotype distributions to distributions on the phenotype space,

$$\Xi : \Lambda^\Omega \to \Lambda^{\tilde\Omega} : \, p \mapsto \Xi p \ , \quad \Xi p\,(x) = \sum_{g \in [x]} p(g) \ . \tag{1}$$

By this projection, the equivalence relation carries over to distributions in Λ^Ω: two distributions p_1, p_2 are called equivalent iff they induce the same distribution over the phenotype space,

$$p_1 \hat\equiv p_2 \iff \Xi p_1 = \Xi p_2 \ .$$

Again, the quotient space $\Lambda^\Omega/\hat\equiv$ of equivalence classes in Λ^Ω is one-to-one with the space $\Lambda^{\tilde\Omega}$ of distributions over the phenotype space.

Compatibility. Based on this formalism we investigate the following question: When $\mathcal{G}$ describes the evolutionary process on the genotype level and if we observe only the projected process on the phenotype level, can we model the phenotypic evolution process without reference to the genotype level? I.e., is the genotype level irrelevant for understanding the phenotypic process? Formally, this amounts to whether $\mathcal{G}$ is compatible with $\equiv$ or not: An operator $\mathcal{G} : \Lambda^\Omega \to \Lambda^\Omega$ is compatible with an equivalence relation $\equiv$ iff

$$p_1 \hat\equiv p_2 \implies \mathcal{G}(p_1) \hat\equiv \mathcal{G}(p_2) \ , \tag{2}$$

which is true iff there exists an operator $\tilde{\mathcal{G}}$ such that the diagram

$$\begin{tikzcd} \Lambda^\Omega \arrow[r, "\mathcal{G}"] \arrow[d, "\Xi"'] & \Lambda^\Omega \arrow[d, "\Xi"] \\ \Lambda^{\tilde\Omega} \arrow[r, "\tilde{\mathcal{G}}"] & \Lambda^{\tilde\Omega} \end{tikzcd} \tag{3}$$

commutes, i.e., $\Xi \circ \mathcal{G} = \tilde{\mathcal{G}} \circ \Xi$. This means that, in the case of compatibility, one can define a process $\tilde{\mathcal{G}} : \Lambda^{\tilde\Omega} \to \Lambda^{\tilde\Omega}$ solely on the phenotypic level that equals the projected original process $\mathcal{G}$: $\tilde{\mathcal{G}}$ represents the phenotypic version of evolutionary dynamics and the population Ξp of phenotypes evolves independent of the neutral traits within the genotype population p. Accordingly, we make the following definition:

Definition 2.1 (Trivial neutrality, trivial genotype-phenotype mapping). Given a genotype space Ω, an evolutionary process $\mathcal{G} : \Lambda^\Omega \to \Lambda^\Omega$, and a genotype-phenotype mapping $\phi : \Omega \to \tilde\Omega$, we define the GP-map to be *trivial* iff phenotype equivalence $\equiv$ commutes with the evolutionary process $\mathcal{G}$. In that case we also speak of *trivial neutrality.*

The meaning of this definition should become clear from the definition of compatibility: In the case of trivial neutrality, the evolution of phenotypes can be completely understood (i.e., modeled) without referring at all to genotypes, in particular, neutral traits are completely irrelevant for the evolution of phenotypes. Based on this formalism we can derive exact conditions for the case of trivial neutrality:

Theorem 2.1. *Let the evolution operator $\mathcal{G} = \mathcal{F}\,\mathcal{M}$ be composed of selection and mutation only (no crossover), and let $\mathcal{M}$ be given by the conditional probability $\mathcal{M}(g'|g)$ of mutating a genotype g into g'. Then, neutrality is trivial iff*

$$\forall x \in \tilde\Omega \ : \ g_1, g_2 \in [x] \Rightarrow \Xi\mathcal{M}(\cdot|g_1) = \Xi\mathcal{M}(\cdot|g_2) \ .$$

In other words, neutrality is non-trivial if and only if there exists at least one neutral set $[x]$ such that the projected exploration distribution $\Xi\mathcal{M}(\cdot|g) \in \Lambda^{\tilde\Omega}$ is non-constant over this neutral neutral set (i.e., differs for different $g \in [x]$).
(The footnote[1] explains the $\cdot$ notation.)

Proof. Since selection depends only on phenotypes it is obvious that the selection operator $\mathcal{F}$ commutes with phenotype projection. The composition of two compatible operators is also compatible (Vose 1999, Theorem 17.4). Hence, we need to focus only on the mutation operator $\mathcal{M}$:

Let us consider the mutational process given by

$$p^{(t+1)}(g') = \sum_g \mathcal{M}(g'|g)\, p^{(t)}(g)$$

Following definition (2) of compatibility we investigate what happens under projection Ξ:

$$\Xi p^{(t+1)}(x') = \sum_{g'\in[x']} \sum_g \mathcal{M}(g'|g)\, p^{(t)}(g) = \sum_g \sum_{g'\in[x']} \mathcal{M}(g'|g)\, p^{(t)}(g) = \sum_g \Xi\mathcal{M}(x'|g)\, p^{(t)}(g) \ .$$

We distinguish two cases:

First case: For all neutral sets $[x]$, let $\Xi\mathcal{M}(\cdot|g)$ be constant over the neutral set, i.e., independent of $g \in [x]$, and we can write $\Xi\mathcal{M}(\cdot|g) = \Xi\mathcal{M}(\cdot|x)$. It follows:

$$\Xi p^{(t+1)}(x') = \sum_x \sum_{g\in[x]} \Xi\mathcal{M}(x'|g)\, p^{(t)}(g) = \sum_x \Xi\mathcal{M}(x'|x) \sum_{g\in[x]} p^{(t)}(g) = \sum_x \Xi\mathcal{M}(x'|x)\, \Xi p^{(t)}(x) \ .$$

Hence, $\mathcal{M}$ is compatible with phenotype equivalence; the diagram (3) commutes with the "coarse-grained" mutation operator $\widetilde{\mathcal{M}}$ given by $\Xi\mathcal{M}(x'|x)$.

Second case: Let there exist at least one neutral set $[x]$ with different genotypes $g_1, g_2 \in [x]$ such that the corresponding projected exploration distributions are not equal, $\Xi\mathcal{M}(\cdot|g_1) \neq \Xi\mathcal{M}(\cdot|g_1)$. Further, consider the two single genotype populations $p_1^{(t)}(g) = \delta_{g,g_1}$ and $p_2^{(t)}(g) = \delta_{g,g_2}$, which are phenotypically equivalent, $\Xi p_1^{(t)} = \Xi p_2^{(t)}$. Their projected offspring populations though are different: $\Xi p_1^{(t+1)} = \Xi\mathcal{M}(\cdot|g_1) \neq \Xi p_2^{(t+1)} = \Xi\mathcal{M}(\cdot|g_2)$. Hence, in the second case $\mathcal{M}$ is not compatible with phenotype equivalence. □

[1] We use the $\cdot$ as a "wild card" for a function argument. E.g., given a function $f : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto f(x, y)$, if we want to fix y and consider the function on that hyperplane, we write $f(\cdot, y) : \mathbb{R} \to \mathbb{R} : x \mapsto f(x, y)$. Accordingly, for a conditional probability we write $\mathcal{M} : \Omega \to \Lambda^\Omega : g \mapsto \mathcal{M}(\cdot|g) \in \Lambda^\Omega$.

3 Behind this formalism

The variability of variational topology.

We introduced $\mathcal{M}(g'|g)$ as the conditional probability distribution describing mutations from g to g'. Actually, this mutation probability is the *only* "structure" given on Ω; there exists no other relevant a priori topological or metric structure. I call it the *variational structure* on Ω and often it is associated with a *variational topology* on Ω, where g' may be defined a neighbor of g if the transition probability $\mathcal{M}(g'|g)$ is greater than some lower bound (Stadler, Stadler, Wagner, & Fontana 2001). Fundamental aspects of evolutionary transitions are often discussed in terms of the variational topology (Schuster 1996; Fontana & Schuster 1998; Reidys & Stadler 2002), and we will also use this intuitive language.

The role of the projection Ξ becomes more evident: while $\mathcal{M}(\cdot|g)$ describes the variational structure on Ω, its projection $\Xi\mathcal{M}(\cdot|g)$ describes the variational structure on $\tilde{\Omega}$. The projection Ξ may be interpreted as a *lift* of variational structure from the genotype space onto the phenotype space—similar to the lift of a metric structure from a base space onto a Riemannian manifold. There is, however, a crucial difference: the GP-map ϕ need not be one-to-one. If ϕ is non-injective, there exist different genotypes g_i that map to the same phenotype; and thus there exist different neighborhoods U_{g_i} that map to potentially *different* neighborhoods of the *same* phenotype. Hence, the variational topology on phenotype space is generally not fixed but variable and depends on the genotypic representation g_i that induces the topology!

On evolving genetic representations.

The GP-map is commonly been thought of "the choice of representation". When applying evolutionary algorithms for problem solving, this choice of representation is usually crucial for the algorithm's efficiency. Even if one has two isomorphic representations, their effects on evolution may be very different since the algorithm's designer will usually define the mutation operators according to a seemingly natural topology on that representation (like the hypercube topology on bit strings). A common but misleading conclusion is that adaptability of exploration strategies requires an adaptability of the GP-map (Altenberg 1995; Wagner & Altenberg 1996). However, as our formalism clarifies, adaptability of exploration can be achieved by a *fixed*, but *non-trivial* genotype-phenotype mapping. In this case, a variation of exploration strategy does not occur by varying the GP-map but by neutral variations in the genotype space. For example, the genotype space may be considered very large, comprising *all* thinkable genetic representations of phenotypes (strings *and* trees *and* grammars, etc.). In that case, different neutral traits literally correspond to different genetic representations—such that a neutral variation allows for a change of representation although the genotype-phenotype map *as a whole* remains fixed. Of course, this presumes that there exist possible *neutral* mutations between these representations—in the case of the artificial evolution of strings, trees, and grammars, this can be realized and is subject to current research (Toussaint 2001). (See also the *symbiotic composition* approach (Watson & Pollack 2002).) In the case of natural evolution, the neutral transition from RNA to DNA genomes in early evolution is a corresponding paradigm. In contrast, a self-adaptation of the whole genotype-phenotype mapping hardly makes sense because it is generally inconsistent to speak of a mapping from genotype to phenotype being parameterized by the genotype.

Strategy parameters as a special case.

In a less formal account (Toussaint & Igel 2002) we already discussed the implications of an exploration distribution $\mathcal{M}(\cdot|g)$ that depends on neutral traits (i.e., genotypic representations

$g \in [x]$ in a given neutral set): It allows the exploration strategy (e.g., the variational topology) to *self-adapt*. In traditional approaches, in particular evolution strategies, so-called *strategy parameters* play the part of neutral traits—strategy parameters are a direct parameterization of mutation operators which are themselves part of the genotype (Angeline 1995; Smith & Fogarty 1997). In these approaches, the genotype space is a Cartesian product $\Omega = \tilde{\Omega} \times Z$ of the phenotype space and the space Z of neutral strategy parameters. In some sense, our formalism generalizes the concept of strategy parameters to the case where the genotype space can not be decomposed into a product of phenotype and strategy parameter spaces but consists of arbitrarily interweaved neutral sets.

Examples.

Assume that we encode (phenotype) strings of length n by (genotype) strings of length $n+1$, such that the GP-map simply ignores the genotype's last symbol. Obviously, this encoding is non-injective. But what is more important is that neutrality is trivial in that case, i.e., $\mathcal{G}$ commutes with Ξ. Hence, the encoding really is redundant—the additional bit has no effect whatsoever on phenotype evolution. Many investigations that aim to argue against neutral encodings actually only investigate such trivial neutrality; they forget that neutrality (in their terms "redundancy") can have an absolute crucial impact on phenotypic variability in the case of a non-trivial GP-map.

In contrast, non-trivial neutrality is implicit in many models of evolutionary computation. Evolution strategies that make use of strategy parameters are the most basic paradigm, which we have already mentioned. But non-trivial neutrality occurs also in many other models. An excellent example are grammar-like genotype-phenotype mappings. Here, the same final phenotype can be represented by different sets of "developmental" rules. Depending on this representation, exploration in the space of phenotypes is very different. The example given in the footnote[2] nicely demonstrates the variability of variational topology on the phenotype space. In (Toussaint 2001), neutral variations between such equivalent grammar-type representations have been introduced and proved efficient.

Evolvability.

All of this is strongly related to the discussion of *evolvability* in the biology literature. Many discussions are based on some relation between neutrality and evolvability (Schuster 1996; Kimura 1983) but to my knowledge there hardly exists a generic theoretical formalism to investigate such issues. (Analytical approaches to describe canalization, i.e. the evolution of mutational robustness of certain characters, based on neutral variations have been proposed by Rice (1998) and Wagner, Booth, & Bagheri-Chaichian (1997).) Following Wagner & Altenberg (1996), evolvability denotes the capability to explore further and further *good* individuals during evolution, which seems possible only when the exploration strategy itself is adapted in favor of evolvability during evolution. In our formalism this raises the question of how neutral traits, and thereby exploration distributions, do actually evolve. We will propose an answer to this question in section 5, which requires the formalism introduced in the next section.

[2] Let the string *abab* be represented (1.) by the grammar {start-symbol→x, x→ab} or (2.) by the grammar {start-symbol→*abab*}. If mutations are only rhs symbol flips, then the phenotypic variational neighbors of *abab* are either (1.) {$*$, $*b*b$, $a*a*$} or (2.) {$*bab$, $a*ab$, $ab*b$, $aba*$}, where $*$ means a symbol flip. These are quite different topologies! See (Toussaint 2001, chapter 5) for details.

4 Embedding neutral sets in the variety of exploration distributions

Our goal is to describe the evolution of neutral traits and thereby the evolution of exploration distributions $\mathcal{M}(\cdot|g)$. In simple cases where the genotype space decomposes, $\Omega = X \times Z$, there seems no conceptual difficulty to do this: the evolutionary process as described on the genotype space may be projected on the hyperplane Z. This is the case for conventional strategy parameters and enabled Beyer (2001) to even derive analytical solutions for this process. However, since in general neutral traits live in arbitrarily interweaved neutral sets it seems tedious to find a description of their evolution. We are missing some embedding space to formulate equations—this actually reflects that we are missing a uniform way of interpreting and modeling neutral traits.

We now propose such an embedding. To simplify the notation let us "not distinguish" between two genotypes $g_1, g_2 \in [x]$ which induce the same exploration distribution $\mathcal{M}(\cdot|g_1) = \mathcal{M}(\cdot|g_2) \in \Lambda^\Omega$. Formally, this means that we define another equivalence relation. Not distinguishing equivalent g's means considering only the evolution equation on the respective quotient space. It is clear the $\mathcal{M}$ commutes with this equivalence and of course also selection does. Since these circumstances are rather obvious we skip introducing formal symbols and, from now on, just assume that all g's in $[x]$ induce different distributions $\mathcal{M}(\cdot|g)$.

Thus, there exists an bijection between $[x]$ and the set

$$\overline{[x]} = \{\mathcal{M}(\cdot|g) \,|\, g \in [x]\} \subset \Lambda^\Omega$$

of exploration distributions. It is this bijection that we want to emphasize because it defines an embedding of neutral sets in the space of exploration distributions. Specifically, it defines an embedding of the non-decomposable genotype space Ω in a *product* space of "Phenotype × Distribution":

Definition 4.1 (σ-embedding). Given a genotype space Ω, a genotype-phenotype mapping $\phi : \Omega \to \tilde{\Omega}$, and a mutation operator $\mathcal{M} : \Omega \to \Lambda^\Omega$, we define the σ-embedding as:

$$\Omega \to \tilde{\Omega} \times \Lambda^\Omega \;:\; g \mapsto (x, \sigma) = \big(\phi(g), \mathcal{M}(\cdot|g)\big)\;. \tag{4}$$

The embedding space is $\bar{\Omega} = \tilde{\Omega} \times \Lambda^\Omega$. Note that this mapping is injective (but of course not surjective) and thus there exists a one-to-one relation between the genotype space Ω and the subset $\{(x,\sigma) \,|\, x \in \tilde{\Omega},\, \sigma \in \overline{[x]}\} \subset \bar{\Omega}$. The injectiveness allows to directly associate a genotype distribution $p \in \Lambda^\Omega$ with a distribution over $\bar{\Omega}$ by

$$p(x,\sigma) = \begin{cases} 0 & \text{if } \sigma \notin \overline{[x]} \\ p(g) & \text{if } \sigma \in \overline{[x]},\ x = \phi(g), \text{ and } \sigma = \mathcal{M}(\cdot|g)\;. \end{cases} \tag{5}$$

The product structure of the embedding space is the key to formulate the evolution equation of exploration distributions in the next section. The embedding also offers a new formal way of modeling neutral traits as specifying an exploration distribution $\sigma \in \overline{[x]}$; the neutral set $[x]$ being nothing but an isomorphic copy of the space $\overline{[x]} \subset \Lambda^\Omega$ of exploration distributions.

5 σ-evolution: The evolution of exploration

How in principle can the exploration distribution be adapted in order to allow for evolvability, i.e., successful exploration? There seems to be a hidden consensus in the class of so-called Estimation-of-Distribution Algorithms (EDAs) (Pelikan, Goldberg, & Cantú-Paz 2000; Mühlenbein, Mahnig, & Rodriguez 1999; Baluja & Davies 1997). All of these approaches use probabilistic models (Bayesian networks, dependency trees, or factorized distributions instead of population and mutation operators) to directly describe the exploration distribution. Their strategy to adapt this distribution is straightforward: in the forthcoming exploration, make solutions more probable that have been selected currently. This means that the exploration distribution has been adapted to become "more similar" to the distribution of selected solutions. Formally, such similarity can be captured by a distance measure in the space of distributions, the *relative entropy* or *Kullback-Leibler divergence.* In fact, Pelikan, Goldberg, & Cantú-Paz (2000) directly work with a variant of the Kullback-Leibler divergence specifically defined on their Bayesian models. These approaches are by no means self-adaptive, adaptation occurs by an explicit ("external") algorithm, not implicitly by evolutionary selection. Could the mere interaction of mutation and selection induce a similar kind of adaptation of exploration? We will argue so in the following.

Let us write the evolution equation for mutation and selection as

$$p^{(t+1)}(g') = \sum_{g\in\Omega} \frac{f^{(t)}(g')}{\bar f^{(t)}}\, \mathcal{M}(g'|g)\, p^{(t)}(g)\ . \tag{6}$$

We embed the equation in $\bar\Omega$ and, according to equations (4) and (5), identify the exploration distribution $\mathcal{M}(x',\sigma'|g)$ with $\sigma(x',\sigma')$,

$$p^{(t+1)}(x',\sigma') = \sum_{x\in\tilde\Omega} \sum_{\sigma\in\Lambda^\Omega} \frac{\tilde f^{(t)}(x')}{\bar f^{(t)}}\, \sigma(x',\sigma')\, p^{(t)}(x,\sigma)\ .$$

This allows to run the summation over all possible distributions $\sigma\in\Lambda^\Omega$; note that $\sigma\notin\overline{[x]} \Rightarrow p^{(t)}(x,\sigma)=0$. We now benefit from $\bar\Omega$ being a product space: The summations commute and executing summation over x gives

$$p^{(t+1)}(x',\sigma') = \sum_{\sigma} \frac{\tilde f^{(t)}(x')}{\bar f^{(t)}}\, \sigma(x',\sigma')\, p^{(t)}(\sigma)\ .$$

Here, $p^{(t)}(\sigma)$ is the marginal distribution of $p^{(t)}(x,\sigma)$, well defined because $\bar\Omega = \tilde\Omega\times\Lambda^\Omega$ is a product space. Summing over x' and decomposing the mutation probability $\sigma(x',\sigma') = \sigma(x'|\sigma')\,\sigma(\sigma')$ we finally get

$$p^{(t+1)}(\sigma') = \sum_{\sigma} \frac{\sum_{x'} \tilde f^{(t)}(x')\,\sigma(x'|\sigma')}{\bar f^{(t)}}\, \sigma(\sigma')\, p^{(t)}(\sigma)\ .$$

We summarize this in

Theorem 5.1 (σ-evolution). *Given the evolutionary process (6) on the genotype space Ω, the evolution of exploration distributions is described by the projection of the process on Λ^Ω given by the σ-evolution equation*

$$p^{(t+1)}(\sigma') = \sum_{\sigma} \frac{\big\langle \tilde f^{(t)}\,,\,\sigma(\cdot|\sigma')\big\rangle}{\bar f^{(t)}}\, \sigma(\sigma')\, p^{(t)}(\sigma)\ , \tag{7}$$

where $\langle f\,,\,g\rangle := \sum_{x'\in\tilde\Omega} f(x')\,g(x')$ denotes the scalar product in the function space L^2.

σ-evolution describes the transition of a parent population $p^{(t)}(\sigma)$ of exploration distributions to the offspring population $p^{(t+1)}(\sigma')$ of exploration distributions. Therein, the term $\sigma(\sigma')$ corresponds to the mutation operator on Λ^Ω (recall that $\sigma(x',\sigma')$ corresponds to a mutation distribution $\mathcal{M}(x', \sigma'|g)$), and the equation matches the standard evolution equation in the form "$p^{(t+1)} = \mathcal{F}\,\mathcal{M}\,p^{(t)}$". In the following we discuss three aspects of the most interesting part of this equation, the fitness term $\big\langle \tilde f^{(t)} \,,\, \sigma(\cdot|\sigma') \big\rangle$.

The scalar product as quality measure. A term $\big\langle \tilde f^{(t)} \,,\, \sigma \big\rangle$ is a measure for σ that we call *σ-quality.* In the first place, it is the scalar product of $\tilde f^{(t)}$ with σ in the space of functions over $\tilde\Omega$. The scalar product is a measure of the similarity and thus, *σ-quality measures the similarity between the exploration distribution and fitness.* σ-quality is very similar to the concept of effective fitness (Nordin & Banzhaf 1995; Stephens & Vargas 2000).

A delay effect. However, equation (7) exhibits that the term $\big\langle \tilde f^{(t)} \,,\, \sigma(\cdot|\sigma') \big\rangle / \bar{\tilde f}^{(t)}$ is actually the fitness term for the *offspring* σ'. Thus, the fitness one has to associate with an offspring is the σ-quality *of its parent* under the condition that σ has in fact generated the offspring σ' ($\sigma(\cdot|\sigma')$ is the parent's phenotypic exploration distribution). Roughly speaking, the offspring is selected according to the quality of its parent. This circumstance can be coined a *first order delay* of evaluation. The quality of an individual's exploration distribution is not rewarded immediately by higher selection probability of this individual itself. It is rewarded when its offspring are selected with higher probability. This delay effect is well-known in the context of evolution strategies, see the footnote.[3] With our formalism it is straightforward to generalize it to arbitrary *degrees of equivalence* leading to n-th order delay, see the appendix.

The information theoretic interpretation. Since σ is actually a probability distribution we can also give an information theoretic interpretation. Let us introduce the *exponential fitness distribution* as

$$F^{(t)} = \frac{\exp \tilde f^{(t)}}{C^{(t)}} \;, \quad C^{(t)} = \sum_{x'} \exp \tilde f^{(t)}(x') \;.$$

One would also call $F^{(t)}$ the soft-max or Boltzmann distribution of $\tilde f^{(t)}$. σ-quality can now be rewritten as

$$\big\langle \tilde f^{(t)} \,,\, \sigma \big\rangle + \ln C^{(t)} = \sum_{x'} \sigma(x') \, \ln F^{(t)}(x') = -D\big(\sigma \,\big\|\, F^{(t)}\big) - H(\sigma) \;,$$

where D denotes the *relative entropy* or *Kullback-Leibler divergence*[4] between two distributions, and H the entropy,

$$D\big(p \,\big\|\, q\big) = \sum_x p(x) \, \ln \frac{p(x)}{q(x)} \;, \quad H(p) = -\sum_x p(x) \, \ln p(x) \;.$$

[3] For evolution strategies it is a common approach to *first* mutate the strategy parameters z before mutating the objective variables x according to the new strategy parameters (Schwefel 1995; Bäck 1998). In our formalism this means that in equation (7) the evaluated distribution $\sigma(\cdot|\sigma') = \Xi\mathcal{M}(\cdot|x, z')$ is "similar" (with same strategy parameter z') to the offspring's distribution $\Xi\sigma'(\cdot) = \Xi\mathcal{M}(\cdot|x', z')$. However, the evaluated and the offspring's exploration distributions still differ significantly because they depend on the objective variables x and x', respectively. It is thus questionable to state that this method cancels the delay effect: the real σ-quality of strategy parameters z' becomes evident only in the next generation in combination with the offspring's objective parameters x'.

[4] Relative entropy $D\big(p \,\|\, q\big)$ is a measure for the loss of information (or gain of entropy) when a "true" distribution p is represented (approximated) by a model distributions q. For example, when $p(x, y)$ is approximated by $p(x)\,p(y)$ one loses information on the mutual dependence between x and y. Accordingly, the relative entropy $D\big(p(x, y) \,\|\, p(x)\,p(y)\big)$ is equal to the mutual information between x and y. Generally, when *knowing* the real distribution p, one needs on average $H(p)$ bits to describe a random sample. If, however, one knows only an approximate model q, then one will need on average $H(p) + D\big(p \,\|\, q\big)$ bits to describe a random sample. The loss of knowledge about the true distribution induces an increase of entropy and thereby an increase of description length for random samples.

Hence, σ-quality is proportional to the negative of the divergence between the exploration distribution σ and the exponential fitness distribution $F^{(t)}$ minus the entropy of exploration. We summarize this:

Corollary 5.2. *The evolution of exploration distributions (σ-evolution) naturally has a selection pressure towards*

- *minimizing the KL-divergence between exploration and the exponential fitness distribution, and*
- *minimizing the entropy of exploration.*

This also means that, assuming fixed entropy of σ, the exponential fitness distribution is a fix point of σ-evolution that corresponds to the quasispecies (Eigen, McCaskill, & Schuster 1989). We aimed at this interpretation because we find it establishes a bridge to numerous approaches and discussions already present in the literature. First of all we understand the relation between self-adaptive σ-evolution and the deterministic adaptation schemes of the exploration distribution in Estimation-of-Distribution Algorithms (Pelikan, Goldberg, & Cantú-Paz 2000; Baluja & Davies 1997; Mühlenbein, Mahnig, & Rodriguez 1999). Actually, σ-evolution realizes a similar kind of adaptation—minimization of the Kullback-Leibler divergence between exploration and fitness—but in a self-adaptive way.

Further, the question of how variational properties evolve has also been raised in numerous variations (pleiotropy, canalization, epistatis, etc.) in the biology literature. These discussions aim at understanding how evolution can handle to introduce correlations or mutational robustness or functional modularity in phenotypic exploration. Our answer is that variational properties evolve as to approximate the selection distribution. If, for example, certain phenotypic traits are correlated in the selection distribution F, then the Kullback-Leibler divergence decreases if these correlations are also present in the exploration distribution σ.

6 Summary

Why have we applied Vose's formalism of compatibility? The literature is full of discussions on the purpose or relevance of neutrality or redundancy in genetic representations. There is a lack of a formal basis on which to ground these discussions. The approach to formalize neutrality by means of equivalence classes is not new—but the strict application of simple and fundamental arguments concerning the compatibility of $\mathcal{G}$ with phenotype equivalence allows to derive under which conditions neutrality influences phenotype evolution: non-dependence of (projected) mutation on neutral traits is equivalent to compatibility of the evolution equation with phenotype equivalence, which we termed trivial neutrality.

Why have we introduced the σ-embedding of neutral traits? For a description of the evolution of neutral traits it is essential to have an embedding space in which to formulate the evolution equation. In simple cases where the genotype space decomposes (strategy parameters) an embedding space of neutral sets is obvious. To generalize to arbitrary neutral sets and arbitrary genotype-phenotype mappings we introduced the σ-embedding, i.e., we embedded neutral sets in the space of probability distributions over the genotype space (*exploration distributions*).

Finally, what have we learned about σ-evolution? We derived an evolution equation for exploration distributions. The selection term can be interpreted as a fitness of exploration distributions that measures the (scalar product) similarity to the fitness function. In terms of information theory, σ-evolution minimizes the Kullback-Leibler divergence between the exploration distribution and the exponential (Boltzmann) fitness distribution, and minimizes the entropy of exploration. We argued that this result provides a bridge between heuristic search approaches to adaptation, conventional self-adaptation, and theoretical biology concerning the origin of epistatis.

Appendix — n-th order equivalence and n-th order delay

How could we consider also higher order delays? Eventually, how could evolution arrange to *increase the probability for children to generate with high probability children that generate with high probability ...et cetera... children with high fitness.* We propose the following. What we considered up to now was the case when two genotypes g_1, g_2 are equivalent, $g_1 \equiv g_2$, because they share the same phenotype. As a result we found a first order delay of selection of neutral characters. Now consider the case when two genotypes are *2nd order equivalent* because their phenotypes are the same, $g_1 \equiv g_2$, *and* their *phenotypic* explorations are the same, $\Xi\sigma_1 = \Xi\sigma_2$, i.e. $\sigma_1 \hat{\equiv} \sigma_2$. Then, these two exploration distributions have, in fact, the same σ-quality and there is no immediate difference in selection between them. However, when writing equation (7) for *two* time steps we find

$$p^{(t+2)}(\sigma'') = \sum_{\sigma',\sigma} \frac{\left\langle \tilde{f}^{(t+1)}\,,\, \sigma'(\cdot|\sigma'')\right\rangle}{\bar{f}^{(t+1)}} \, \frac{\left\langle \tilde{f}^{(t)}\,,\, \sigma(\cdot|\sigma')\right\rangle}{\bar{f}^{(t)}} \, \sigma'(\sigma'')\, \sigma(\sigma')\, p^{(t)}(\sigma) \;.$$

Which means that there exists a longer term difference in selection between σ_1 and σ_2 iff the expected exploration distributions σ' *of their offspring* are not equivalent in the sense

$$\sum_{\sigma'} \sigma_1(\sigma')\,(\Xi\sigma') \neq \sum_{\sigma'} \sigma_2(\sigma')\,(\Xi\sigma') \;.$$

This difference in selection after two generations may be coined a second order delay effect. Generally, let us define two genotypes, (x_1, σ_1) and (x_2, σ_2), n-th order equivalent iff their phenotype and all their expected phenotypic exploration distributions after $0..n$–2 mutations are the same,

$$\begin{aligned}(x_1,\sigma_1) \equiv^n (x_2,\sigma_2) \qquad \Longleftrightarrow \qquad & x_1 = x_2 \;, \quad n \geq 2 \Rightarrow \Xi\sigma_1 = \Xi\sigma_2 \;, \\ & \forall_{1\leq k\leq n-2}: \;\; \textstyle\sum_{\sigma^1,..,\sigma^k} \sigma_1(\sigma^1)\, \sigma^1(\sigma^2) \,..\, \sigma^{k-1}(\sigma^k)\,(\Xi\sigma^k) \\ & \qquad\qquad = \textstyle\sum_{\sigma^1,..,\sigma^k} \sigma_2(\sigma^1)\, \sigma^1(\sigma^2) \,..\, \sigma^{k-1}(\sigma^k)\,(\Xi\sigma^k)\end{aligned}$$

It follows that, if two genotypes are n-th order equivalent, then there exists no difference between the phenotypic dynamics of their evolution for the next $n-1$ generations. Thereafter though, with a delay of n generations, their evolution differs because the expected exploration distributions of the n-th order offspring of these two genotypes are not phenotypically equivalent.

Acknowledgments

The work was supported by the German Research Foundation DFG under grant *SoleSys-II* SCHO 336/5-2.

References

Altenberg, L. (1995). Genome growth and the evolution of the genotype-phenotype map. In W. Banzhaf & F. H. Eeckman (Eds.), *Evolution and Biocomputation: Computational Models of Evolution*, pp. 205–259. Springer, Berlin.

Angeline, P. (1995). Adaptive and self-adaptive evolutionary computations. In M. Palaniswami, Y. Attikiouzel, R. Marks, D. B. Fogel, & T. Fukuda (Eds.), *Computational Intelligence: A Dynamic Systems Perspective*, pp. 152–163. IEEE Press.

Bäck, T. (1998). On the behavior of evolutionary algorithms in dynamic environments. In D. Fogel, H.-P. Schwefel, T. Bäck, & X. Yao (Eds.), *Proc. of Fifth IEEE Int. Conf. on Evolutionary Computation (ICEC 1998)*, pp. 446–451. IEEE Press.

Baluja, S. & S. Davies (1997). Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. In *Proc. of Fourteenth Int. Conf. on Machine Learning (ICML 1997)*, pp. 30–38.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies.* Springer, Berlin.

Eigen, M., J. McCaskill, & P. Schuster (1989). The molecular quasispecies. *Advances in chemical physics* **75**, 149–263.

Fontana, W. & P. Schuster (1998). Continuity in evolution: On the nature of transitions. *Science* **280**, 1431–1433.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press.

Mühlenbein, H., T. Mahnig, & A. O. Rodriguez (1999). Schemata, distributions and graphical models in evolutionary optimization. *J. of Heuristics* **5**, 215–247.

Nordin, P. & W. Banzhaf (1995, 15-19). Complexity compression and evolution. In L. Eshelman (Ed.), *Genetic Algorithms: Proc. of Sixth International Conf. (ICGA 1995)*, pp. 310–317. Morgan Kaufmann, Pittsburgh.

Pelikan, M., D. E. Goldberg, & E. Cantú-Paz (2000). Linkage problem, distribution estimation, and Bayesian networks. *Evolutionary Computation* **9**, 311–340.

Radcliffe, N. J. (1991). Equivalence class analysis of genetic algorithms. *Complex Systems* **5**, 183–205.

Reidys, C. M. & P. F. Stadler (2002). Combinatorial landscapes. *SIAM Review* **44**, 3–54.

Rice, S. H. (1998). The evolution of canalization and the breaking of von Bear's laws: Modeling the evolution of development with epistatis. *Evolution* **52**, 647–656.

Schuster, P. (1996). Landscapes and molecular evolution. *Physica D* **107**, 331–363.

Schwefel, H.-P. (1995). *Evolution and Optimum Seeking.* John Wiley & Sons, New York.

Smith, J. & T. Fogarty (1997). Operator and parameter adaption in genetic algorithms. *Soft Computing* **1**, 81–87.

Stadler, B. M., P. F. Stadler, G. P. Wagner, & W. Fontana (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology* **213**, 241–274.

Stephens, C. & J. M. Vargas (2000). Effective fitness as an alternative paradigm for evolutionary computation I: General formalism. *Genetic Programming and Evolvable Machines* **1**, 363–378.

Stephens, C. & H. Waelbroeck (1999). Codon bias and mutability in HIV sequences. *Molecular Evolution* **48**, 390–397.

Toussaint, M. (2001). Self-adaptive exploration in evolutionary search. Technical Report IRINI-2001-05, Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany.

Toussaint, M. & C. Igel (2002). Neutrality: A necessity for self-adaptation. In *Proc. of IEEE Congress on Evolutionary Computation (CEC 2002)*, pp. 1354–1359.

Vose, M. D. (1999). *The Simple Genetic Algorithm.* MIT Press, Cambridge.

Wagner, G. P. & L. Altenberg (1996). Complex adaptations and the evolution of evolvability. *Evolution* **50**, 967–976.

Wagner, G. P., G. Booth, & H. Bagheri-Chaichian (1997). A population genetic theory of canalization. *Evolution* **51**, 329–347.

Watson, R. & J. Pollack (2002). A computational model of symbiotic composition in evolutionary transitions. *Biosystems, Special Issue on Evolvability.*