EVOLUTION AS A STATISTICAL OPTIMIZATION ALGORITHM

Gregor Kjellström
Hagvägen 29A, 141 70 Huddinge, Sweden

Received 21 November 1994, 18 October 1995

ABSTRACT: The maximum-entropy-principle according to the second law of thermodynamics may play an important role in the evolution of natural systems. It may both reduce the time needed by evolution and shorten the DNA-message needed for an increased complexity. The natural process and certain efficient statistical optimization processes may have some features in common.

* * *

Introduction

The Darwinian evolution of natural systems (in short the natural process or evolution) is a process of tremendous complexity. It has been a source of fascination and inspiration to many philosophers and researchers in the past, and will continue to be so in the future. Today, research is highly diversified and include different fields such as molecular biology (Alberts et al., 1994) and population genetics (Ridley, 1993).

A recent discipline is the use of statistical algorithms for the solution of artificial optimization problems. Examples of such algorithms are *Genetic Algorithms* (Goldberg, 1989) and an algorithm known as *Gaussian adaptation* (GA) (Kjellström, 1970; Kjellström & Taxén, 1981, 1992). Such algorithms have been used for instance to improve the quality of signal processing systems. The efficiency of the GA-process mainly relies on the maximum-entropy-principle (MEP), and it has also been described as a second order approximation of the natural process in the sense that not only the averages but also the variances of phenotypes in large populations may be adapted in such a way as to make survival more efficient. The purpose here is to see to what extent MEP may have improved the efficiency of the natural process.

One of the difficulties encountered by such processes (including the natural one) may be described as a problem of high dimensionality. How is it possible for a random process to find its way through a set of feasible DNA-messages written in an alphabet of four letters and, by now, having a length of about 10 billion symbols? A possible answer may be that the process has a history, a memory and that it learns from preceeding successes. In such case the process may hardly be seen as a pure random process, except for a short initial phase of development.

Unfortunately, messages or strings are unsuitable for a simple description of the process, so we prefer a parametric model. Such models are also frequently used by biologists (Lande, 1979). As examples of such parameters we may consider morphological polygenic characters. In addition, the main concern of natural selection is hardly DNA-messages solely, but rather phenotypes that for instance express beauty, strength and the ability to escape predators or to withstand climatic variations.

The set of acceptable points (individuals) in the parameter space is usually defined by a probability function known as the *fitness of the individual*, $s(x)$, i. e. the probability that the individual - having the phenotype vector $x^T = (x_1, x_2, \ldots, x_n)$, where x^T is the transpose of x - will become a parent of a new individual in the progeny population. Another definition is the *region of acceptability* ($\mathcal{A}$), defined by $s(x) = q =< 1$ for all x belonging to $\mathcal{A}$ and $s(x) = 0$ for all other x. But, this is not a serious limitation, because $\mathcal{A}$ may be defined over a lattice of small hypercubes, where $\mathcal{A}$ occupies a fraction of each hypercube equal to $s(x)$. Making the hypercubes sufficiently small, $\mathcal{A}$ may approximate any probability function to any degree of precision. For the sake of simplicity, we use $\mathcal{A}$ for most of the paper.

As has already been mentioned, the difficulty has to do with dimensionality and specifically with the shape and extension of $\mathcal{A}$. Unfortunately, the $\mathcal{A}$ of the natural process is unknown, but a good first metaphor may be found as follows: Most natural systems (as well as artificial) have to satisfy a large set of restrictions or requirements. In a first approximation, these requirements usually manifest themselves as curved walls in a parameter space, and since they are seldom parallel, they cut each other

* * *

Evolutionary Theory 11:105-117 (January, 1996)
The editors thank R. Lande and another referee for help in evaluating this paper.

somewhere, and there a wedge or cone will be formed. Arms races (Dawkins, 1988) between predators and prey for instance, also tend to squeeze the cone from the sides. Because of inertness, the statistical process might easily get trapped at the apex and the population might become extinct, which makes efficiency an important factor of survival.

Another good metaphor of $\mathcal{A}$ is a region surrounded by a contour line of a map of a mountain chain (Eigen, 1992). In such case the criterion for acceptable points is determined by the altitude. Thus, the process might have to climb a mountain crest or penetrate a valley in the value landscape, enforced by a raising or falling of the altitude. Figure 1 shows a possible $\mathcal{A}$ together with two circles representing the dispersion of phenotypes in two different populations.

The theorem of Gaussian adaptation

Let $v(x)$ be a Gaussian probability density function with mean m and moment matrix M, i. e.

$$v(x) = \gamma \exp\{-(m - x)^T M^{-1}(m - x)/2 \}$$

where γ is a constant such that the integral of $v(x)$ over the whole space equals 1. Since phenotypes are often Gaussian distributed in a large population, we may let circles represent circles (ellipsoids) of concentration of the Gaussian distributed phenotypes in two (many) dimensions. The *entropy* (which is a measure of disorder) of the Gaussian is defined as:

$$H = \log\{ (2\pi e)^n \det(M) \}^{1/2},$$ where n is the number of dimensions.

This means that when the volume of the ellipsoid of concentration increases, then the entropy will also increase like the entropy of an expanding gas. This corresponds to larger genetic differences between individuals in the population in analog with larger distances between molecules in the gas as well.

Suppose that the part of the distribution belonging to $\mathcal{A}$ will be selected, while the rest is omitted. Evidently, there will be a displacement of the center of gravity of the distribution. We will get a selection differential, $S = m^* - m$, where m^* is the centre of gravity of the selected part of the distribution, pointing into the same direction as the arrow in figure 1. Further, if $\mathcal{A}$ is constant, if m is replaced by m^* and the process is iteratively repeated, we will sooner or later reach a position where $S = 0$, i. e. m and m^* will coincide. In the artificial GA-algorithm, m may simply be replaced by m^* in every iteration (generation). Alternatively, m may be updated after every single acceptable individual point x according to the formula:

$$m := (1 - a)m + ax,$$

where a is a scalar $<< 1$. The inverse of a may also represent the number of individuals in the population.

In previous models of evolution, the response to selection (R) usually depends on selection (S) in some way. For instance, in the truncation selection described by Hartl, 1981, R equals h^2S, where h^2 is the heritability of the trait. But because $\mathcal{A}$ is unknown, we prefer a different kind of estimate here. In our simplified model, dominance and environmental effects will also be ignored. The production of a new generation is supposed to proceed as follows:

1 A set of parents are selected from the original population.
2 Random mating among parents is used to produce offspring. Assuming no selection or mutation at this stage, we may use the Hardy-Weinberg law, which states that the frequencies of alleles will be

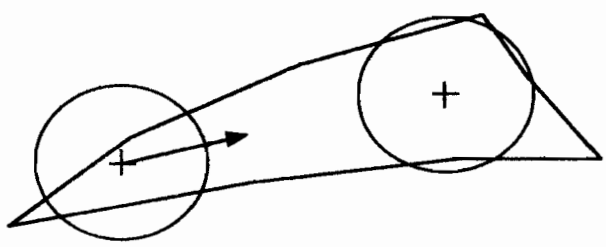

Figure 1. A region of acceptability formed like a wedge. Circles represent the dispersion of phenotypes in two different populations.

transmitted to the offspring with no changes. Thus, the centre of gravity of the offspring, m, will coincide with m^*.

Mutations like for instance inversions or transpositions will alter the sequential order of the genes and the frequencies of alleles. But assuming that the gene contribution to the phenotype will not change because of these mutations, we conclude that m will coincide with m^*.

This symmetry seems to be violated in the case of a varying $s(x)$. But, observe that the two expressions of m^* below are equivalent -

$$m^* = \frac{\int_{\mathcal{A}} x\, v(m - x)\, dx}{\int_{\mathcal{A}} v(m - x)\, dx} = \frac{\int x\, s(x)\, v(m - x)\, dx}{\int s(x)\, v(m - x)\, dx}$$

Thus, we conclude that m will coincide with m^* also in this case. If $s(x)$ is unknown, the number of offspring may serve as an approximation. Note, however, that every parent must be given a weight proportional to the number of offspring.

Using the rules of genetic variation such as crossing-over and inversion, biologists (Crow & Kimura, 1970; Lande, 1979) have shown that the natural process maximizes the *mean fitness of the population* (P):

$$P(m) = \int_{\mathcal{A}} q\, v(m - x)\, dx = \int s(x)\, v(m - x)\, dx$$

An alternative explanation is given by a theorem known as the *theorem of Gaussian adaptation* (Kjellström, 1970; Kjellström & Taxén, 1981, 1992). As was earlier mentioned, $\mathcal{A}$ is unknown, but it does not matter, as the theorem is valid for all $\mathcal{A}$ and for all Gaussian distributions.

In its simplest form, the theorem states that P is maximized by the condition $m^* = m$. The theorem also states - which is perhaps less well known - that the entropy or dispersion of the Gaussian is maximized, while P is kept constant.

This is a purely geometrical statement about regions and normal distributions and has, primarily, nothing to do with biology, but as long as Gaussian polygenic characters are concerned and the natural process strives to an equilibrium where $m^* = m$, we may say that the process - with good approximation - strives to fulfill the conditions of the theorem with respect to variations in m. Even if evolution is never in a state of perfect equilibrium, it may for long periods of time be in a state very similar to equilibrium, with small changes in m and M, and therefore we expect the theorem to be applicable also to the natural process.

The advantage of the theorem is that it provides a simple link between the maximization of mean fitness, the maximization of entropy and the efficiency of the process. From a technical point of view, the idea with MEP is that if we compare two similar Gaussian processes in the same region working at the same value of P, but having different m and entropy, then the average time or work needed to find a new individual point inside $\mathcal{A}$ is the same for both. If the entropy may only be changed by a multiplication of M by some scalar factor, then the process having the higher entropy has a higher mobility, because the average distance between independently sampled individual points is longer. Assuming random mating, the average effect of higher entropy in a population will be larger a distance between parents and longer jumps from parent to offspring in the phenotypic space. Thus, for any actual value of P, the centering of m, such that m coincides with m^*, has the general effect of increasing the mobility of the process. Because of higher mobility and because such a process covers a larger volume in parameter space, we also expect it to react quicker to unforeseen changes in $\mathcal{A}$. Intuitively a high mobility process should also be more efficient, and should in principle cause m to converge faster, even if this may not be true in every particular situation. But, in contrast to classical definitions of efficiency, that rely on the real convergence of a parameter, our new concept related to entropy is in principle independent of such a convergence, because the mobility may still be high even though m and M does not change with time. The mobility is hidden like the mobility of the molecules of a gas in a closed room. If the molecules move fast, they will leave the room quickly when the door is opened; otherwise they may leave the room slowly. However, as will be shown in the next section, efficiency also depends on P, independent of the entropy we have just discussed.

A proof of the theorem may be found in Kjellström, 1970, or Kjellström & Taxén, 1981. But, without digging too deep into the references, we may convince ourselves of the validity of the theorem by rather simple arguments. As, if m is slightly displaced from its optimal position (the center of the circle to the right in figure 1), then P will decrease, but may be kept constant if the radius of the circle is decreased correspondingly (i. e. if the entropy of the normal distribution is decreased). This means that both the mean fitness and the entropy are simultaneously maximized for any Gaussian distribution by satisfying the condition $m^* = m$.

Generally speaking, the entropy will increase with the length of the DNA-message. This is strong evidence, that entropy has really increased during the 4 billion years of natural evolution, because most DNA-messages today are much longer than they were for the first living creatures (Brooks & Wiley, 1986). This is in contrast to some researchers stating that living systems represent a higher degree of order. It may be that some kind of physical entropy may decrease in natural systems, while it increases in the universe as a whole, but the entropy of the DNA-message has no doubt increased for most species.

Evidently, the space angle from the apex of a conical $\mathcal{A}$ will have considerable impact on the selection differential of the process. If the space angle is too small, the selection differential might vanish in the statistical uncertainty, because the number of individuals in a population is always limited, and evolution may cease to work (see figure 2a).

The theorem of GA, however, provides an opportunity to negotiate this problem by an adaptation of the second order moments (i. e. the moment matrix M, see figure 2b). The theorem also states that if the moment matrix M of the offspring is proportional to the moment matrix M^* of the parents, then the entropy is maximized also with respect to the second order moments. M may in principle be updated after every step y - leading to a feasible point $x = m + y$ - according to:

$$M := (1 - a)M + ayy^T,$$

In order to guarantee a suitable increase in entropy, y should be Gaussian distributed with moment matrix $\mu^2 M$, where the scalar $\mu > 1$ is used to increase the entropy. But M will never be used in the calculations. Instead we use the matrix W defined by $WW^T = M$. Thus, we have $y = Wg$, where g is Gaussian distributed with the moment matrix μU (U is the unit matrix). W and W^T may be updated by the formulas:

$$W = (1 - b)W + bygT \quad \text{and} \quad W^T = (1 - b)W^T + bgy^T,$$

because multiplication gives

$$WW^T = (1 - 2b)WW^T + 2byy^T,$$

where terms including b^2 have been neglected. Thus, putting $a = b/2$, M will be indirectly adapted with good approximation. In practice it will suffice to update W only.

The adaptation of M will certainly improve efficiency, but since $\mathcal{A}$ is unknown, the magnitude of a possible improvement is always uncertain. As will be shown later, the rules of genetic variation such as crossing-over or transposition may restrict the adaptation. Another restriction is the limited number of individuals, which makes the estimation of moments uncertain. But, in a space of thousands of

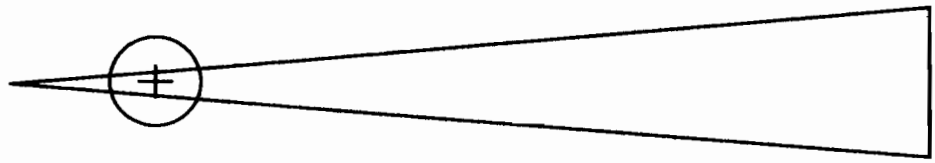

Figure 2a. Process unable to move because S is too small.

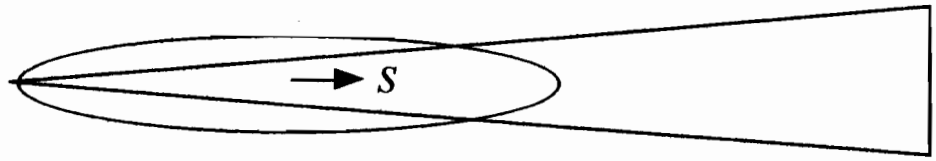

Figure 2b. Adaptation of variability may increase S by many orders of magnitude.

dimensions, the selection differential may probably, despite these restrictions, increase by many orders of magnitude.

The theorem of efficiency

Let us shortly discuss the general properties of a possible measure of efficiency, $E(P)$. Unfortunately, not even a perfect adaptation of M will necessarily imply an efficient process. For instance, if the mutation rates and variabilities are very small, and P is close to its maximum possible value q ($s(x)$ is here supposed to be constant $= q =< 1$ over $\mathcal{A}$), then even if M is adapted for maximum entropy, according to the theorem of GA, the corresponding process will hardly be able to move and will therefore be inefficient. A very high mutation rate and a large variability gives a P close to 0 and the corresponding process is inefficient because lots of time and resources are wasted on individuals who can never be selected. Between these extremes we expect to find the most economic compromise, a medium variability (see figure 3), making the process most efficient. This compromise is expressed in the *theorem of efficiency*.

We expect that a suitable measure of efficiency should not depend on linear transformations of the whole process. As an example, consider the random walk inside $\mathcal{A}$ to the left in figure 4 (a square in this case), which may be seen as a very simple asexual parametric evolution of one individual in two parameters. In addition, there is no reason to believe that the efficiency of the process will change only because we look at it from below at an angle $< 90^{o}$, which makes it look like the process to the right. In order to make the usefulness of MEP clearer, we may observe that all steps are Gaussian distributed with equal variances in the two parameters, which corresponds to a maximum entropy process. Thus, if $\mathcal{A}$ has to be searched without any presumptions about the location of an interesting sub-region, efficiency may hardly be increased by using longer steps in any particular direction. As a consequence, the process to the right is also maximally efficient. This means that the adaptation of M for maximum entropy, at the same value of P, corresponds to a linear transformation of the whole process. Specifically, if $\mathcal{A}$ is a square (or hypercube of many dimensions), then the process will, after a certain period of adaptation, attain maximum efficiency with respect to the theorem of GA, independent of linear transformations of $\mathcal{A}$. In this particular case the process is also statistically independent in different parameters.

When the small square inside $\mathcal{A}$ has been located, we may say that a specified amount of information has been obtained.This information should not be confused with the *information* as defined by the well known *information theory*, introduced by Shannon (see references in Brooks & Wiley, 1986), but if the information is measured as the logarithm of some volume, then the concepts are closely related. This follows from the fact that entropy may be seen as the logarithm of a volume and that average information is equivalent to entropy. The points traversed by the process may also represent a specified amount of information, which may be used for the estimation of m, M or, for instance, the speed of the process. Therefore, in order for E to be a proper measure of efficiency, we require that if some process has achieved a specified amount of information, at the expense of time or work, then efficiency should be proportional to the inverse of the time or work.

Let us summarise the requirements:

1 The process is statistically independent in the n different parameters and equally efficient in all parameters.
2 All individuals have equal cost in time or work.
3 E is a continuous function of P, $0 < P < q =< 1$, and is positive in the P-interval, except at the end points, where E becomes zero. The derivative $dE(P)/dP$ at the point $P = q$ is < 0.
4 If some specified amount of information can be obtained in a certain amount of time or work, then E is proportional to the inverse of time or work.
5 E is independent of linear transformations of the whole process.

The theorem states that all measures of efficiency, that satisfy the conditions above, are asymptotically proportional to $-P \log(P/q)$, when the number of dimensions increases, and are maximized by $P = q \exp(-1)$.

A proof has been given by Kjellström, 1991, but we will repeat some of the arguments here. First assume that $q = 1$. Since all parameters are equally efficient, we may say that E_n is the efficiency of a

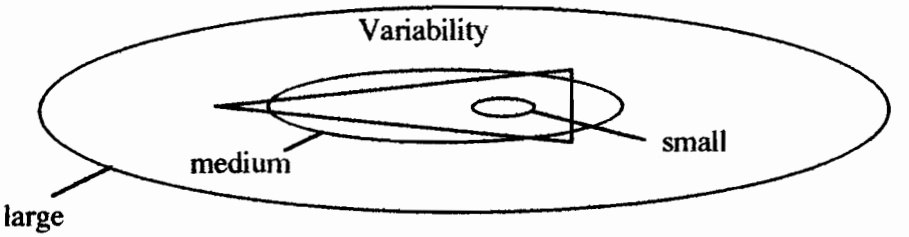

Figure 3. The most economic medium variability.

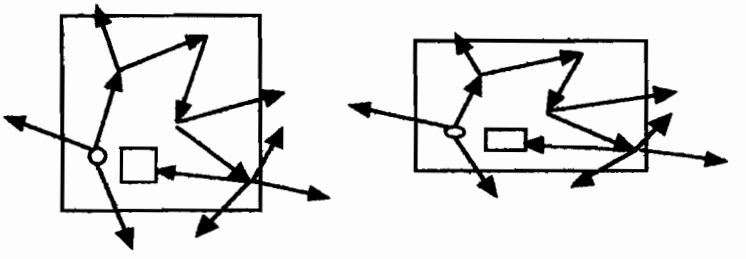

Figure 4. Process. Transformed process.

single parameter in a n-dimensional system. Further, let

$E_1 = -P \sum \gamma_i [\log(P)]^i$, where summation is extended over $i \rightarrow 1, 2, \ldots \infty$,

be the efficiency in a one-dimensional system. Provided that γ_1 is > 0 we have $[dE_1(P)/dP]_1 < 0$. Therefore E_1 may approximate any measure of efficiency, to any degree of precision, without violating the conditions, except possibly the condition no. 4. But this condition is fulfilled by a certain recursion formula as follows: Suppose that a new parameter x_{n+1} is added without any changes to the process characteristics in the existing n parameters. In principle the efficiency is the same as before, except that P will decrease by factor p, where p is the probability of success in a single parameter, i. e. $p^n = P$. Thus, the work or time needed to obtain the same information will be increased by factor $1/p$. This leads to the recursion:

$E_{n+1}(pP) = p\, E_n(P)$ or $E_n(P) = P^{1/n}\, E_{n-1}(P^{1-1/n})$.

Starting with $E_1(P)$, the efficiency may now be calculated for any n. As has been shown,

$nE_n(P) \rightarrow -\gamma_1 P \log(P)$ as n increases.

When q becomes < 1 over the same $\mathcal{A}$, the effect will be a shift of P-scale to P/q and the work or time is increased by factor $1/q$ to arrive at the same result. Thus $nE_n(P)$ becomes

$-\gamma_1 P \log(P/q)$. See also figure 5.

The optimal value of $P = q \exp(-1)$ fairly well corresponds to a difference of entropy between offspring and parents by one unit in every generation, provided that the entropy is measured in natural logarithms. This may be achieved by suitable heritable mutation rates. But of course, the natural process is free to look for other mutation rates, if they are more suitable from the efficiency point of view. Besides, efficiency will hardly be needed in a static environment. Note also that even though the optimal value of P depends on q, the optimal mutation rate does not.

Because of the independence of linear transformations, the theorem is also valid for a large class of processes that are not strictly statistically independent, but that may be linearly transformed into such a process. In a more complicated region, the conditions may still be fulfilled with good approximation, and mobility may still be high.

Examples

The applications of the theorem, however, might not always be easy. E. g. does the speed of the

random walk in the hypercube already discussed, satisfy the conditions? Evidently, the fifth condition is not satisfied. On the other hand, only the first four conditions are used during the proof, so if these are satisfied, the speed may still have the same asymptotic behaviour.

The speed may be defined as the average length of steps divided by the average time needed for a new parent, $1/P$. The first two conditions do not cause any problems, but in the third condition, it may be difficult to check if $dE(P)/dP$ is < 0 at the point $P = q$. But if the variances of the Gaussian distributed steps are equal to zero, then we have $P = q$ and the speed is equal to zero. As soon as the variances become > 0, then we will have a positive speed and a $P < q$. Thus $dE(P)/dP$ should be < 0. Condition no. 4 is more difficult to check, but guided by the proof, we observe that the component of speed in a single parameter satisfies the condition. Since the total speed is proportional to the component, we conclude that the speed satisfies the condition.Thus, the asymptotic behaviour of the speed will be in accordance with the theorem.

For another example, consider the case where the process behaves as before, except in the parameter x_1, in which the process is pushed in such a way that $q = 0.5$. (See figure 6). Note that the theorem is applicable here even though the process is not exactly in a state of equilibrium. The P-value for maximum speed equals $1/(2e)$. The same result has been obtained by Rechenberg, 1973, but such a fast push will hardly be recommended.

In the next example we consider a case in which $s(x) = \exp(-x^T x)$, $m = 0$ and M is a diagonal matrix with equal elements. The process, which is in a state of perfect equilibrium, strives to increase the entropy in every generation, while the selection strives to decrease it by the same amount. If the restriction due to $s(x)$ is suddenly removed in some parameter, then the entropy will increase. A measure defined as the increase in entropy divided by the time or work, $1/P$, certainly satisfies the conditions. It has earlier been shown by another method (Kjellström, 1991), that this measure is proportional to $-P \log(P/q)$ for all n in this case.

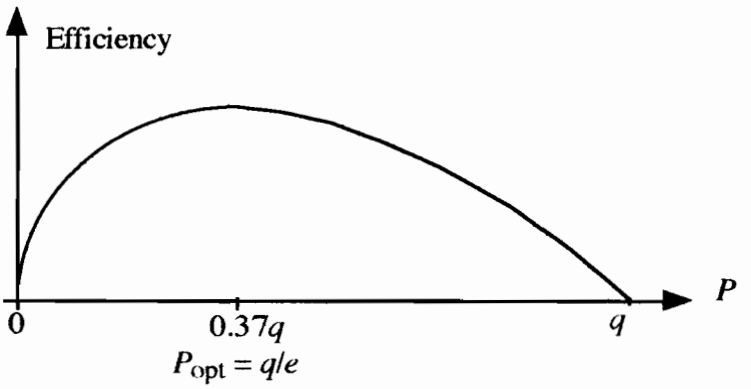

Figure 5. Relation between efficiency and mean fitness.

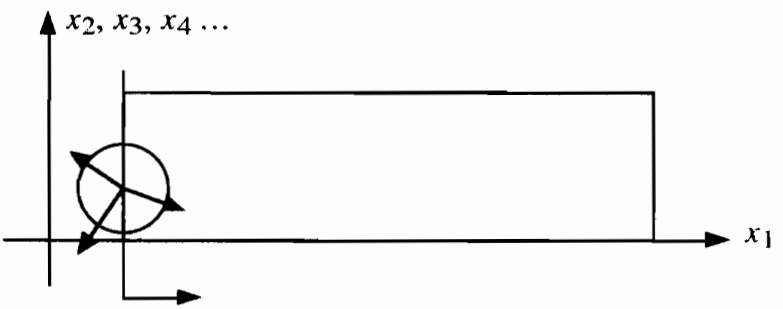

Figure 6. A random walk pushed in the direction x_1.

The model

That efficiency is of considerable importance in a changing environment has been demonstrated on test examples of optimization (Kjellström & Taxén, 1992). But the question still remains: To what extent will the natural process be able to improve efficiency? As long as the ontogenetic program is not known in every detail, it will certainly be difficult to give a comprehensive answer. But we can investigate a model that might give some insight into the problem.

The basis of the model is the assumption that the real ontogenetic program is a modified replay of evolution. The phenotype may still be the sum of a large number of small pleiotropic Gaussian distributed contributions. The ontogenetic program (like evolution itself) is seen as a random walk inside some region of acceptability; the end point representing the phenotype of the adult individual. While the evolution over 4 billion years has to make a number of time-consuming tests for each feasible step, the ontogenetic program may carry out the feasible steps already tested in a couple of years.

Assuming that different genes express themselves as different steps, this model reflects the development of many different organs in parallel, because one single step may simultaneously affect many different parameters. But, will the simultaneous usage of different genes be considered in the model? To some extent, Yes! If some sequence of genes, all representing singular steps, are already at hand, parallel usage of genes may be introduced, for instance, by the addition of a new gene. This means that the step, expressed by the new gene, may be affected by some of the old genes and that its contribution may be included in the new step. For the sake of simplicity, because we do not know the direction or length of any step, we assume that the resulting step is Gaussian distributed as before.

A difficult problem, however, is that the physical gene order along a chromosome and the temporal order of gene expression in development is not known. But, we will investigate two different cases here. Firstly, we assume that consecutive evolutionary steps are distributed at random along the same chromosome, and secondly that they are represented by consecutive genes on the chromosome. For the sake of simplicity we also assume that the steps already taken will never be modified. Mutations are effectuated by means of crossing-over, inversion, transposition and addition (or withdrawal) of genes only.

The row below is an example of a random sequence of ten steps along a chromosome, i. e. the first step is expressed by gene no. 2, the second step by gene no. 8 etc.

7 1 6 5 9 4 3 8 2 10

Assuming, for example, that the crossing-over operator will first read 5 genes from the maternal string followed by 5 genes from the paternal string. This is a strong linkage of genes, but if we look at the time sequence of steps, the crossing-over is purely random:

Steps from maternal string 1 5 6 7 9
Steps from paternal string 2 3 4 8 10

For a rough estimate of the effects of such a process, we may assume that $\mathcal{A}$ is a rectangle such that the range of parameter x_1 is three times larger than the range of the x_2. Assuming further, that steps can contribute to the parameter values by +1 (+) or by -1 (-) only and that parents are, after some time, almost uniformly distributed over the rectangle. We therefore expect that there should be three times as many differences between maternal and paternal strings in x_1 as in x_2, on average. Because crossing-over at random will give no variability from homozygous loci, we may focus attention on the heterozygous loci only. (The order of "+" and "-" steps is irrelevant here, because we are only interested in the summation of steps). A possible average distribution of "+" and "-" steps might be:

parameter	x_1	x_2
maternal string	+ + + + + + + + - -	+ + + + + + - - - -
paternal string	+ + - - - - - - - -	+ + + + - - - - - -

The expected variability in offspring from such parents, assuming that phenotypes are expressed by single strings, may therefore be seen from the Pascal triangle (figure 7) lines 6 and 2, where every number tells us how many sequences of equally probable "+" and "-" give a certain contribution to a parameter value. The variance is equal to 6 in x_1 and equal to 2 in x_2. But the ratio between standard deviations is equal to the square root of 3 and not equal to 3 as it should be in accordance with the theorem of GA.Thus, we have got a certain adaptation, which is better than nothing, but it is not as

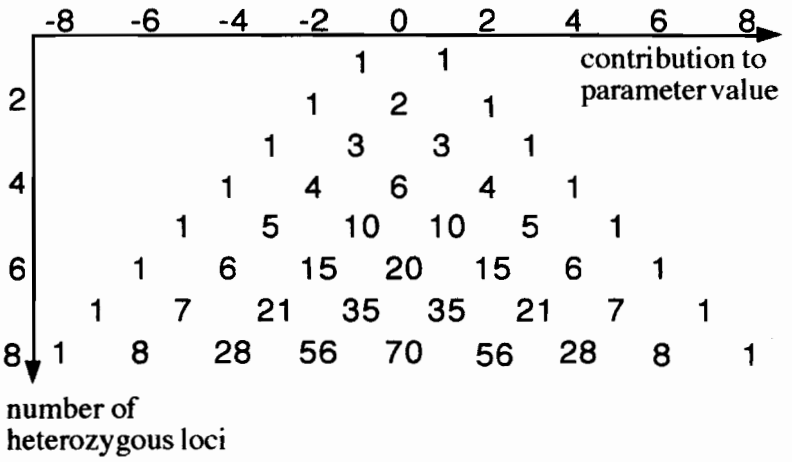

Figure 7. The Pascal triangle

good as it should be.

Suppose now, that the inversion- or transposition operators are capable of altering, not only the gene order, but also the order in which steps are expressed, then our sequence may look as follows:

1 5 4 3 2 7 6 9 8 10

i. e. the first step is expressed by gene number 1, the second step by gene number 5 etcetera. Then, if crossing-over still occurs after the fifth gene, the first five steps come from the maternal string, while the rest come from the paternal, even though the steps are not expressed in strict ascending order.

This motivates the investigation of the second case, where we assume that consecutive steps are expressed by consecutive genes along the chromosome. In such a case, the crossing-over operator will read not only thousands of consecutive genes but also thousands of consecutive steps from maternal or paternal strings. This means that thousands of steps separate the starting- and final point of every partial sequence and that these points may be seen as independently sampled in $\mathcal{A}$. As a result, the effect of crossing-over will be a summation of vectors that follows the main directions of $\mathcal{A}$, avoiding less suitable directions. The distribution of offspring from two parents will be approximately Gaussian and adapted to the shape of $\mathcal{A}$. Since the distribution of characters in the population is mainly a superposition of individual distributions, we may expect the total variability of the offspring to be approximately Gaussian and proportional to the total variability of parents in the population. Such a process has a better opportunity to maximize the entropy also with respect to the second order moments. There is also the possibility that higher order moments will be adapted, but this will not be dealt with here.

A credible assumption is that if a certain number of genes are already at hand, then evolution may proceed by mutations in the existing genes that modify the evolutionary steps of the past, and a more efficient evolution will certainly achieve better results in shorter time. But, such a method has limitations, at least if the initial conditions and $\mathcal{A}$ are not allowed to change very much, and if the single steps have a limited length.

A more reliable method, to enable the random walk to withdraw from the starting point, will be to increase the number of genes to allow new steps to be incorporated. But this means that more genetic code is needed. Unfortunately, our theory does not say anything about the efficient addition of genes, but on the whole, efficiency may reduce both the time needed by evolution and the length of the DNA-message.

Simulations

Because of the complexity, simulations prior to theoretical investigations will be preferred. A large proper simulation of a high dimensional process would require a lot of computer power and memory. It would also be necessary to measure the variability of the parents and to compare it with the variability of the offspring. Because of these difficulties we restrict ourselves to a partial simulation, assuming that the process at some time has reached a state where some 2-dimensional problem has to

be solved and that $\mathcal{A}$ is a simple rectangle defined by $|x_1| < 3$ and $|x_2| < 1$. We would like to demonstrate that the algorithm works properly, provided that there is no variability in the initial conditions. Then we expect the variabilities in the coordinate directions of the offspring from two parents to be proportional to the sides of the rectangle.

Our intention is to use a population of pairs of sequences each having a length equal to 100 pleiotropic genes or steps, that are Gaussian distributed with mean = 0 and moment matrix equal to the unit matrix. For the sake of simplicity, phenotypes are expressed by a single sequence. In order to speed up the calculations and to reduce the need for memory it will be assumed that crossing-over always takes place after a natural multiple of 10 steps. This makes it possible to split the simulation in two hierarchical phases as follows: At the first level, a large number of random walks each of length = 10 steps inside $\mathcal{A}$ are simulated. The difference vector between the starting point and end point of each 10-step random walk is saved in a memory and finally the moment matrix of these vectors is estimated. This moment matrix may also be seen as a large pool of 10-step *super-genes* or *high-level* steps, available by crossing-over in a large population.

In the next phase of the simulation, 20 pairs of sequences are generated using the moment matrix of the high-level steps to form a small population. Each sequence is a 10-step random walk of high-level steps inside $\mathcal{A}$. Sequences for the new generation are generated by crossing-over on the pairs. Crossing-over takes place at random between high-level steps. Thus, at least 10 ordinary steps are always linked together from the maternal or paternal sequence. The only mutation used in this simulation is permutation of two arbitrary adjacent high-level steps at a rate of one permutation in each new sequence. In this way a new population of 20 pairs of sequences is generated.

Figure 8 shows the parameters of two parent strings (heavy dots) together with its offspring (small dots) after some generations. As can be seen, the dispersion of offspring is fairly symmetric around the parents and the adaptation to $\mathcal{A}$ is fairly good. Anyway, the proportion of dispersions in x_1 and x_2 is better than the square root of 3. Note also, that this is possible even though the differences in parameter values between parent strings are about the same. If we look at the the small dots as possible gametes of an individual, we also observe that an individual i may carry a set of moments of his own, m_i and M_i. The simulation shows that the model process has better opportunities to satisfy the theorem of GA at least when there is no dispersion in the initial conditions.

But, when the process enters a new phase of evolution, because the mountain crest curves off in some other direction, the dispersion already adopted by history may seriously disturb the future of the process. One possible solution to this problem is to put the mutation rates of historical genes to zero, and to add new mutable genes for the actual phase. I would not say that this is a simple task, but the natural process is in principle able to do this, because mutation rates are heritable. Another possibility may be to let the process pass through a bottleneck. In addition, individuals having a more suitable m_i and M_i are better off under the new conditions, because they might get more grandchildren, great grandchildren etc. Note also, that a high $s(x)$ is no guarantee for a large number of great grandchildren, which has a higher long term priority.

Figure 8. Offspring (small dots) from two parents (heavy dots).

The global problem

In this section we will discuss the global optimization problem, i. e. the searching for the highest peak out of a large set of peaks of a varying $s(x)$. We would like to show that the maximization of $s(x)$ is hardly recommendable and that evolution may proceed in a direction of decreasing $s(x)$ as well.

The number of peaks of $s(x)$ is unknown, but we may speculate that this number might be of the same order of magnitude as the estimated number of atoms in the universe (about 10^{80}). Such a high number of peaks may easily be generated by a simple summation of functions. Suppose that we have the sum of 100 functions, i. e.

$$f_1(x_1) + f_2(x_2) + \ldots + f_{100}(x_{100}),$$

each having 10 peaks in some interval, then the sum will have at least 10^{100} peaks. In this case the highest peak is easily found by a search in one parameter at a time, but if the system has been exposed to rotations, the search will be extremely difficult. Since we have roughly some million species and since every species may be on its way along a mountain crest or in the vicinity of some peaks of its own, we may expect that only a very tiny fraction of potential peaks have been found by evolution.

But let us use the Fourier analysis to see that evolution may be a very powerful tool in the search for the highest peak.This analysis is applicable to any number of dimensions, but here we assume that x is a single parameter. In the case of a varying $s(x)$, mean fitness may be defined as:

$$P(m,t) = \int s(x)\, \gamma \exp[\, -(m - x)^2/(4t)\,]\, dx$$

where t represents the halved variance of the parameter. As has been shown (Kostrowicki & Piela, 1991; Kjellström & Taxén, 1992) the mean fitness satisfies the diffusion equation $\Delta P(m,t) = \delta P/\delta t$, with $s(x)$ as a boundary condition at $t = 0$. Now, if $s(x) = \sin(wx)$ or $\cos(wx)$, then P becomes equal to $\exp(-w^2t)\sin(wm)$ or $\exp(-w^2t)\cos(wm)$, where w is the frequency of the sine- and cosine waves, which are also eigenfunctions of the process. Thus, if $s(x)$ is represented by its Fourier expansion, we may say that the process acts as a Gaussian low-pass filter, attenuating the high-frequency components of $s(x)$. When t (or the entropy) increases in the population, then the attenuation of the high-frequency components will also increase.

Let us investigate some small examples to see the global properties of the process. For instance, if $s(x)$ is the sum of a high- and a low-frequency component, as in figure 9a, then the high-frequency component may almost vanish and the process will climb the smooth grey ridge. Thus, the advantage of maximizing P instead of s is that the possibility of coming close to the highest peak considerably increases.

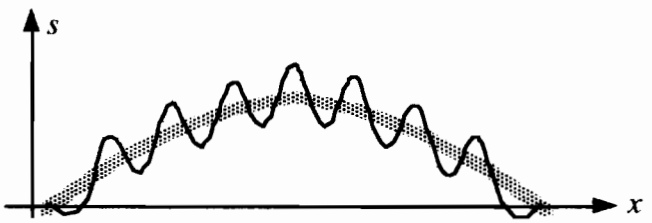

Figure 9a. The sum of a high- and low-frequency component.

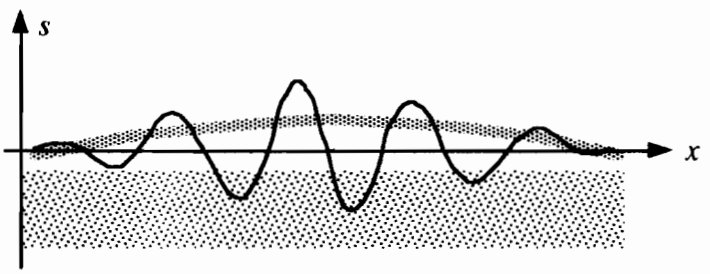

Figure 9b. An amplitude modulated carrier wave.

In the next example we consider a case where $s(x)$ behaves like a high-frequency carrier wave, having a slow variation in amplitude as in figure 9b (amplitude modulated carrier wave, AM). Since this function contains only high frequency components in a narrow band, there will be nothing left for the climber after filtration, because a Gaussian filter will not introduce any new frequency components outside the band. But if we use a rectifier or diode in combination with a filter, as in an oldfashioned AM radio reciever, the low frequency components will appear again, and the climber has a greater opportunity to find higher peaks. A similar method seems to be used by evolution in the sense that arms races between different species, forces the genetic landscape to sink into the sea (the zero level). Thus $s(x)$ becomes truncated from below, which has the similar effect as a diode. Needless to say, these principles considerably improve the possibility of finding higher peaks.

But is there any particular advantage with the Gaussian filter prior to ohter low-pass filters? Even if the question is hypothetical, we may make some observations. E. g. if $s(x)$ is constant over $\mathcal{A}$, in which case the maximization of $s(x)$ ceases to work, other filters may introduce overshoots or false maxima or minima. But the Gaussian filter will not, and the maximization of P may continue as before. This also shows, that a maximization of P, may even force the process in a direction of decreasing $s(x)$.

Discussion

Thus far the environmental effects have been ignored, so let us see how they may affect our discussions? If we restrict ourselves to external environmental changes of $\mathcal{A}$, we conclude that evolution may go in some other direction, but m and M may still be adapted to the new $\mathcal{A}$ in the same way as before. When individuals themselves affect $\mathcal{A}$, we will have complex feedback between a cultural and a genetic level, but the process will continue to maximize entropy at the genetic level, now with respect to the restrictions given by the individuals. Thus, we might face a situation in which individuals try to make $\mathcal{A}$ smaller while MEP strives to enlargen it. If the individuals were too successful in their effort, it might even lead to their extinction. We therefore conclude that MEP is the strongest force for most living species. Nevertheless, the omission of environmental effects gives little loss of generality to our discussion of the MEP, at least on the genetic level.

In reality we may hardly expect an exact representation of consecutive evolutionary steps by consecutive genes on the same chromosome. But if the number of consecutive steps linked together by the crossing-over operator increases, for instance by inversion, transposition and natural selection, then the statistical independence of the starting- and end points of maternal or paternal partial vector sums increases. This may improve the adaptation of variability to the extension of $\mathcal{A}$ and the penetrative ability of the process, which means that cones or valleys of a considerably higher dimensionality may be penetrated. Even if the rate of evolution also depends on many other factors, such as the addition of new genes, an efficient use of the MEP will certainly shorten the time needed by evolution and the length of the DNA-message.

References

Alberts, B. et al. The molecular biology of the cell. Garland Publishing, Inc., New York & London, 1994.

Brooks, D. R. & Wiley, E. O. Evolution as Entropy, Towards a Unified Theory of Biology, The University of Chicago Press, Chicago and London, 1986.

Crow, J. F. and Kimura, M. An Introduction to Population Genetics Theory, Harper and Row, 1970.

Dawkins, R. The Blind Watchmaker, Penguin Books Ltd, 1988.

Eigen, M. Steps towards life. Oxford University Press, 1992.

Goldberg, David E. Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, 1989.

Hartl, D. L. A Primer of Population Genetics. Sinauer, Sunderland, Massachusets, 1981.

Kjellström, G. Optimization of electrical Networks with respect to Tolerance Costs, Ericsson Technics, no. 3, pp. 157-175, 1970.

Kjellström, G. and Taxén, L. Stochastic Optimization in System Design, IEEE Trans. on Circ. and Syst., vol. CAS-28, no. 7, July 1981.

Kjellström, G. On the Efficiency of Gaussian Adaptation, Journal of Optimization Theory and Applications, vol. 71, no. 3, Dec. 1991.

Kjellström, G. and Taxén, L. Gaussian Adaptation, an evolution-based efficient global optimizer. In Computational and Applied Mathematics, Brezinski, I. C. and Kulish, U. Eds, Elsevier Science Publishers B. V., pp 267-276, 1992.

Kostrowicki, J. & Piela L. Diffusion Equation Method of Global Minimization: Performance for Standard Test Functions. Journal of Optimization Theory and Applications, vol. 69, no 2, may 1991.

Lande, R. Quantitative Genetic Analysis of multivariate Evolution Applied to Brain-Body Size Allometry. Evolution, 33(1), pp. 402-416, 1979.

Rechenberg, I. Evolutionsstrategie. Fromann - Holzboog, 1973.

Ridley, M. Evolution. Blackwell, Oxford. 1993.