Simulations in Evolution. III. Randomness as a Generator of Opportunities

by Bernard Testa**, Andrzej J. Bojarski†, and Stefan Mordalski†

**) Department of Pharmacy, Lausanne University Hospital, BH04-CHUV, Rue du Bugnon 41, CH-1011 Lausanne (phone: +41 21 314 4237; fax: +41 21 314 4992; e-mail: Bernard.Testa@chuv.ch)

†) Department of Medicinal Chemistry, Institute of Pharmacology of the Polish Academy of Sciences, 12 Smetna St., PL-31343 Krakow (e-mail: bojarski@if-pan.krakow.pl, stefanm@if-pan.krakow.pl)

In Neo-Darwinism, variation and natural selection are the two evolutionary mechanisms which propel biological evolution. Our previous reports presented a histogram model to simulate the evolution of populations of individuals classified into bins according to an unspecified, quantifiable phenotypic character, and whose number in each bin changed generation after generation under the influence of fitness, while the total population was maintained constant. The histogram model also allowed Shannon entropy (SE) to be monitored continuously as the information content of the total population decreased or increased. Here, a simple Perl (Practical Extraction and Reporting Language) application was developed to carry out these computations, with the critical feature of an added random factor in the percent of individuals whose offspring moved to a vicinal bin. The results of the simulations demonstrate that the random factor mimicking variation increased considerably the range of values covered by Shannon entropy, especially when the percentage of changed offspring was high. This increase in information content is interpreted as facilitated adaptability of the population.

1. Introduction.

– The genius of Darwin was to recognize that evolution is the resultant of two phenomena, variation and selection [1]. Selection (natural and artificial) implies the preferential survival and reproduction of some individuals in a population, the so-called ‘fitter’ individuals. At Darwin’s time, macroscopic variation was well-documented but not understood in biochemical terms. Today, variation within a species is understood to cover both existing phenotypic polymorphism and genetic mutation which adds to it. A third source of variations are the environment-induced epigenetic changes, i.e., hereditary changes in gene expression without changes in the DNA sequence. These changes are long-lasting and involve functionally relevant genome modifications such as DNA methylation and histone deacetylation [2][3].

We view variation and selection as the expression of two irreducible ‘forces’ which propel the evolution of inorganic and living matter in space and time [4][5]. The first ‘force’ is a drive toward variation, divergence, and dispersion. It results in expansion and the opening of new virtualities, potentialities, and opportunities (Stuart Kauffman’s ‘adjacent possible’ [6]). But left uncountered, it also leads to a state of total uniformity and loss of order. There is analogy, if not identity, with entropy understood here as a capacity to change [7][8], i.e., in a broader sense than the original thermodynamic one [9–11].

The second ‘force’ is recognition, namely the intrinsic capacity of entities to recognize each other and interact either attractively or repulsively [12–14]. There are,
for example, gravity (attractive) and magnetism (attractive or repulsive), and the
electrostatic forces which play such an essential role at the (bio)chemical levels.
Recognition is thus what ‘informs’ existing entities. In other words, recognition forces
are the expression of information; their observable consequences are constraints.

Recognition forces and their resulting constraints occur at two levels, internal and
external. At the external level, constraints arise from the interactions between a system
and its environment. Indeed, the latter will influence the relative stability of the various
states a system can exist in, for example the many conformational states of a molecule,
the many behavioral states of an organism, or the distribution of a given phenotype in a
population. And, in the context of biological evolution, recognition between an
organism and its environment defines the degree of fitness of the former, and hence its
probability of survival and reproduction, and, in the long run, determines how a
population of organisms will evolve and simultaneously modify its environment.

In previous reports \[15\][16], we used histograms to simulate the evolving
distribution of a given quantifiable phenotypic character in a model population. The
initial distribution of the character was taken to be Gaussian, and it evolved generation
after generation under the influence of the two factors driving evolution, namely
variation and selection.

The first influencing factor was variation, namely changes in the distribution of the
character which enlarge the variability from which natural selection can choose. This
was modeled by assuming that, during each generation cycle, each bin lost a fixed
percentage of its content to vicinal bins, thus mimicking the fact that offspring differ
from their parents. The second influencing factor was natural selection, which
constrains this variability by favoring the survival and fertility (i.e., fitness) of the
better adapted organisms. This was modeled by assuming a distribution of the fitness
factor different from that in the initial population (generation 0). This fitness factor was
applied as a multiplication factor during each cycle. The resulting histograms shifted
from cycle to cycle en route to the final fitness pattern. The total population was kept
constant by normalization generation after generation. The histogram model also
allowed Shannon entropy (SE) to be monitored continuously as the information
content of the total population decreased or increased.

The representations of the evolutionary process generated by this approach were
evocative but unrealistically regular. In the present study, we introduced a random
factor in the percentage of individuals whose offspring differed from their parents and
hence found themselves in a different bin. Such a situation is clearly closer to reality
that the preset, constant percent in our previous simulations. With this new simulation
tool in hand, our objective was to monitor, generation after generation, the changes in
Shannon entropy taken as an index of the information content of the population. The
results demonstrate that the random factor broadened considerably the range of values
covered by Shannon entropy, a broadening interpreted as an enhanced capacity of the
population to adapt to changing environmental conditions.

2. Layout of the Study. – The model involves a virtual population of 100,000
organisms evolving over 100 generations. An unspecified, quantifiable phenotypic
character allowed the individuals to be categorized into bins, 13 for the initial
population (generation 0). The distribution of bins was a Gaussian histogram taken to
reflect a unimodal distribution of the relative fitness corresponding to each bin. The initial population distribution is shown as grey columns (bins) in Fig. 1. At the beginning of the simulation, ambient conditions were assumed to change suddenly, for example, a fast migration of the population to a warmer climate, resulting in a corresponding change in the relative fitness of the organisms. This new fitness distribution, to which the population must adapt, is represented by the black dots in Fig. 1.

The transition from one generation to the next was computed bin after bin by a Perl (= Practical Extraction and Reporting Language) [17] application using two variables. The first variable (designated ‘$M\%$’) was the percent of individuals in each bin who differed from their parents and hence went into another bin. This variable was meant to mimic at each generation the percent of the population showing a phenotypic change. This variable $M\%$ was assigned an integer value but was under the control of a random factor (designated ‘$R$’). When the random factor ‘$R$’ was ‘off’, the preset integer value of $M\%$ remained constant throughout the simulation. But when the random factor ‘$R$’ was ‘on’, it randomly assigned an integer value between 0 and $M\%$ to each bin. In practice, the resulting effective value of $M\%$ was thus expected to average half the preset value, as indirectly observed. The following $M\%$ values were explored: 2, 5, 10, 20, 30, 40, and 50%, meaning that, when the random factor was on, the present values of $M\%$ had to be set to double the explored value (i.e., to 4, 10, 20, 40, 60, 80, and 100%, resp.).

The second variable (designated ‘$N$’) was the even number of neighbor bins into which the preset value of $M\%$ was distributed. This variable was meant to mimic the percent of phenotypically changed offspring at each generation. In practice, $M\%$ was distributed in equal parts into the $N/2$ neighboring bins on the right and the $N/2$
neighboring bins to the left. Should the number of bins on one side be insufficient, further bins on the other side received a share. The $N$ parameter was given the values 2, 4, 6, and 8.

The combinations of settings defining each simulation are presented in matrix form in the Table, together with the codes used to designate each simulation.

<table>
<thead>
<tr>
<th>Random factor</th>
<th>Settings of the M% variable</th>
<th>Effective value of M% (%)</th>
<th>Settings of the N variable and code of the individual simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>off</td>
<td>2</td>
<td>off-2%-2</td>
<td>on-2%-2</td>
</tr>
<tr>
<td>on</td>
<td>-2</td>
<td>on-4%-2</td>
<td>off-4%-2</td>
</tr>
<tr>
<td>off</td>
<td>5</td>
<td>off-5%-2</td>
<td>on-5%-2</td>
</tr>
<tr>
<td>on</td>
<td>-5</td>
<td>on-10%-2</td>
<td>off-10%-2</td>
</tr>
<tr>
<td>off</td>
<td>10</td>
<td>off-10%-2</td>
<td>on-10%-2</td>
</tr>
<tr>
<td>on</td>
<td>20</td>
<td>on-20%-2</td>
<td>off-20%-2</td>
</tr>
<tr>
<td>off</td>
<td>-20</td>
<td>on-40%-2</td>
<td>on-40%-2</td>
</tr>
<tr>
<td>on</td>
<td>30</td>
<td>off-30%-2</td>
<td>off-30%-2</td>
</tr>
<tr>
<td>on</td>
<td>-30</td>
<td>on-60%-2</td>
<td>on-60%-2</td>
</tr>
<tr>
<td>off</td>
<td>40</td>
<td>off-40%-2</td>
<td>off-40%-2</td>
</tr>
<tr>
<td>on</td>
<td>-40</td>
<td>on-80%-2</td>
<td>on-80%-2</td>
</tr>
<tr>
<td>off</td>
<td>50</td>
<td>off-50%-2</td>
<td>off-50%-2</td>
</tr>
<tr>
<td>on</td>
<td>-50</td>
<td>on-100%-2</td>
<td>on-100%-2</td>
</tr>
</tbody>
</table>

Because available resources were assumed to be constant, the total population was kept constant generation after generation; this was achieved by a simple normalization procedure at each generation. The resulting total in each bin was rounded off to the nearest unit.

The value of Shannon entropy (SE) of the population was calculated at each generation. To recall, SE is a measure of the relative information content of a data set [18]. Thus, if for datasets 1 and 2 one finds that $\text{SE}(2) < \text{SE}(1)$, this means that data set 2 contains more information than data set 1. Shannon’s theory is originally one of digital communication [19], but it is finding applications in chemistry and biology [20][21]. SE was calculated from these histograms using Eqsns. 1 and 2:

$$\text{SE} = - \sum_i p_i \log_2 p_i \quad (1)$$

where $p_i$ is the probability in each bin, obtained from the count in each bin ($c_i$):

$$p_i = \frac{c_i}{\sum_i c_i} \quad (2)$$

3. Results of the Simulations with Inactivated Random Factor. – While the effect of the random factor on the information content of the population was the main objective of the work, it was necessary to begin with an appreciation of the influence of the $N$
parameter and of the unaffected $M\%$ parameter on the dynamics and final outcome of the evolutionary process being modeled. Fig. 2 shows the variation of the population over 100 generations under five settings of the two parameters selected to span their explored range, namely off-2\%-2, off-2\%-8, off-20\%-4, off-50\%-2, and off-50\%-8 (Fig. 2, a–e, resp.).

As seen in Fig. 2, a and b, a small value of $M\%$ allows the population to adapt tightly to the new environmental conditions implied by the fitness histogram in Fig. 1. This is confirmed by the low (and stable) SE values reached after 20 or 30 generations (Fig. 3). In contrast, a large value of $M\%$ (Fig. 2, d and e, and Fig. 3) results in a distinctly smaller drop in SE, which we interpret to imply a looser adaptation to the new environmental conditions. Conditions of intermediate values of $M\%$ (e.g., Fig. 2, c)
yielded an intermediate drop in SE. Of particular note is the fact that the SE values remained constant, once the population had reached its maximal possible adaptation under the preset final conditions (starting population, \(M\%), N, and fitness profile).

A summary of the combined influence of the \(M\%) and \(N\) parameters on the stable SE values reached with the random factor off is presented in Fig. 4. In contrast to the influence of the \(M\%) parameter, that of the \(N\) parameters does not appear to be interpretable.

4. Results of the Simulations with Activated Random Factor. – All 28 simulations carried out with inactivated random factor (off) were repeated with activated random factor (on), the settings of the \(M\%) parameter now being doubled as explained above. Given the fluctuating values of SE generation after generation, each simulation was run
seven times, the individual runs being labeled from \textit{a} to \textit{g}. Fig. 5 show the variation of the population over 100 generations under a sampling of the two parameters, namely \textit{on-4\%-2}, \textit{on-4\%-8}, \textit{on-40\%-4}, \textit{on-100\%-2}, and \textit{on-100\%-8} (Fig. 5,\textit{a}–\textit{e}, resp.).

As seen in Fig. 5,\textit{a} and \textit{b}, the random factor has little influence on the population’s evolution, when the value of $M\%$ is small. At higher values of $M\%$ (Fig. 5,\textit{c}, and mainly \textit{d} and \textit{e}), a different picture emerges in that the distribution of individuals in the various bins fluctuates markedly. This is confirmed by the fluctuations seen in SE values (Fig. 6), whose amplitudes appear approximately proportional to the average value of all SE points in the right half of the plots.
To make sense of these plots, a mean SE value and standard deviation (SD) was calculated for the 7 runs, taking only generations 51 to 100 into account to ensure that the population had reached its maximal possible adaptation. In other words, a total of 350 datapoints were used to calculate a mean SE value (black dots and continuous black lines) and its SD (grey dots and continuous grey lines). The lowest and highest single values in each set of 350 points are also plotted as grey dots and broken grey lines. These various values are plotted as a function of the preset value of the $M\%$ parameter, yielding one plot for each value of the $N$ parameter (Fig. 7).

What clearly emerges from these Figures are the broad values of the standard deviations, which allow a quantitative assessment of the fluctuations seen in Fig. 6. In other words, the random factor does not influence the broad evolutionary trend a population is submitted to, but it produces a population that differs widely yet within
limits from one generation to the other. What this implies in evolutionary terms will be discussed below.

5. An Evolutionary Interpretation. – In a previous study [16], we used the simplified histogram model to simulate the influence of natural selection in a stable population challenged by the emergence of one or two fitter mutations. In these scenarios, the appearance of a mutation was part of the settings and not a monitored outcome. In other words, the focus of this paper was exclusively on the influence of natural selection independently of variation understood as the result of the many processes generating new genotypes and phenotypes. Here, we have explored the influence of random variations on the evolution of a virtual population challenged by a sudden change in ambient conditions.

As shown in Fig. 4, the N parameter did not seem to play an interpretable role in the evolution of the population. In contrast, inspection of Figs. 2–4 clearly reveals that small effective values of M% (2–5%) led to very low SE values, whereas large effective values of M% (30–50%) led to lesser drops in SE, in other words, to broader

Fig. 7. Variations and fluctuations in the information content (SE) of a population of 100,000 organisms as a function of the M% parameter, the other settings being random factor on, and N of 2 (a), 4 (b), 6 (c), or 8 (d). For the seven runs carried out under these conditions, the SE was calculated as the average of the SE of generations 51 to 100 (i.e., a total of 350 datapoints). This mean value is shown as ‘mean of of 350 datapoints’ (black dots and line). The ‘means of 350 points ± SD’ are shown as the gray dots and gray continuous lines. The two gray dotted lines connect the lowest and highest individual SE values, respectively, among the 350 datapoints.
histograms. The very low values in SE reflect a tight adaptation of the population to the new ambient conditions, whereas broader histograms reflected a more diverse population. Both outcomes can be understood as being advantageous to the survival of the population, but under different conditions. Indeed, a tight adaptation to ambient conditions can be most favorable as long as the latter do not change, while a more diverse population retains flexibility and faster adaptability should such conditions continue to change.

This point is of particular significance when examining and interpreting the role of the random factor in our simulations. This factor did not influence the time (number of generations) it took for the population to adapt. In contrast, it strongly impacted on the diversity of the population, whose information content fluctuated widely from one generation to the next, the average value of SE nevertheless remaining notably close to the stable SE value reached, when the random factor was off. Snapshots of these wide fluctuations caused by the random factor are seen in Fig. 5 and are presented as raw data in Fig. 6. In Fig. 7, however, they are quantified as mean ± SD, revealing their true magnitude.

When searching for a realistic interpretation of these results, should we conclude that the population is at a large disadvantage because randomness constantly perturbs any progress it makes toward an optimal adaptation? This could indeed be a valid interpretation, but only under permanently stable ambient conditions. This is a situation real populations do not encounter, since ambient and other living conditions are not known to remain stable for unlimited periods. In other words, a population whose adaptativity fluctuates from generation to generation may be at some disadvantage in the short term but will be favored when this population must face new challenges of various origins. This research is in line with the previous papers, showing variation-driven increases of SE at the equilibrium state. But as seen here, randomizing variation (within low levels of change) is a compromise allowing to keep the information content favorably high on the one hand, and preserving a broader phenotypic range on the other.

As mentioned in the Introduction, three biomolecular mechanisms are now known to underlie or cause variation, namely existing phenotypic polymorphism, genetic mutations, and environment-induced epigenetic changes. Our $M\%$ parameter cannot be understood as coding for genetic mutations, or at best only when $M\%$ takes very small values. In contrast, it can readily be understood as phenotypic polymorphism, especially when amplified by environment-determined epigenetic changes which influence gene expressions [22]. This is all the more convincing given recent findings that such changes ‘can be influenced by environmental factors to affect phenotype for multiple generations’, raising ‘the possibility that epigenetic states provide a substrate for natural selection’ and for ‘the rapid adaptation of species to changes in environment’ [23].

REFERENCES


Received June 15, 2012