Measures of Diversity for Populations and Distances Between Individuals with Highly Reorganizable Genomes

Claudio Mattiussi claudio.mattiussi@epfl.ch
Markus Waibel markus.waibel@epfl.ch
Dario Floreano dario.floreano@epfl.ch
Institute of Systems Engineering
Swiss Federal Institute of Technology of Lausanne (EPFL)
1015 Lausanne, Switzerland

Abstract

In this paper we address the problem of defining a measure of diversity for a population of individuals whose genome can be subjected to major reorganizations during the evolutionary process. To this end, we introduce a measure of diversity for populations of strings of variable length defined on a finite alphabet, and from this measure we derive a semi-metric distance between pairs of strings. The definitions are based on counting the number of substrings of the strings, considered first separately and then collectively. This approach is related to the concept of linguistic complexity, whose definition we generalize from single strings to populations. Using the substring count approach we also define a new kind of Tanimoto distance between strings. We show how to extend the approach to representations that are not based on strings and, in particular, to the tree-based representations used in the field of genetic programming. We describe how suffix trees can allow these measures and distances to be implemented with a computational cost that is linear in both space and time relative to the length of the strings and the size of the population. The definitions were devised to assess the diversity of populations having genomes of variable length and variable structure during evolutionary computation runs, but applications in quantitative genomics, proteomics, and pattern recognition can be also envisaged.

Keywords

Evolutionary computation, genetic programming, variable length genomes, population diversity, substring diversity, Tanimoto distance, Jaccard similarity, linguistic complexity, nucleotide diversity.

1 Introduction

In evolutionary computation (EC) there is often the need to measure the diversity of two or more individuals of a population. This necessity can be dictated by many reasons, for example: the desire to prevent premature convergence of the population; the utility of restarting or stopping an evolutionary algorithm when the population diversity drops below a certain threshold; the requirement of evolving a population of distinct Pareto-optimal solutions in a multi-objective optimization problem; the effort of maintaining a population able to adapt rapidly to a changed environment in the case of dynamic problems; and still many others (de Jong et al., 2001; Leung et al., 1997;

Tomassini et al., 2004; Wineberg and Oppacher, 2000; Wineberg and Oppacher, 2003a; Wineberg and Oppacher, 2003b).

The diversity of individuals and populations can be measured either in the genotype or in the phenotype space. When the phenotype or the genotype are constituted by a fixed number p of real parameters, the standard tools of mathematical analysis and those of cluster analysis in the p-dimensional real space R^p can be directly applied for the definition of a measure of diversity (Theodoridis and Koutroumbas, 2003). It is often the case, however, that the structure of the phenotype does not lend itself well to such a straightforward approach; this happens, for example, when the phenotype is a structure - say, a network - with variable topology and number of parameters. In those cases, one is left with the option of either defining a specialized distance between such phenotypic structures, or focus on the genotype space, where the structure of the elements is usually much simpler, for example a sequence of characters. We will assume in the rest of the paper that the elements of the genotype space, i.e., the genomes of the individuals, are finite character strings over a finite alphabet. Note that there are contexts, where the genomes have structure more complex than a string, and where the measures defined below and based on the count of strings and substrings cannot be applied directly but must be adapted to the particular genome structure. We will describe below an example of an extension of the string-based diversity measure to the case of tree-based genomes used in the field of genetic programming (Keijzer, 1996; Langdon and Poli, 2002).

If the strings that constitute the genomes have fixed length and uniform structure, the definition of a diversity measure for two individuals is typically based on the use of the Hamming distance (although other approaches are possible, see for example (Leung et al., 1997)), that is, on the count of the number of mismatches between the pair of strings that constitute the genomes of the individuals. With the expression "genomes having uniform structure", we mean that all the individuals have a genome with the same number of subblocks, or genes, with the same phenotypic meaning for the subblocks, and arranged in the same order in the genome. A diversity measure for the whole population can be then obtained from the diversity measure for pairs of individuals by combining all the pairwise distances between individuals (Morrison and De Jong, 2002; Wineberg and Oppacher, 2003a; Wineberg and Oppacher, 2003b).

The case of strings with variable length but still with uniform structure can be treated similarly, provided the Hamming distance is generalized to permit the comparison of strings having different length. A good candidate for this generalization is the so-called *edit distance*, which is based on the use of three elementary operations - insertion, deletion, and substitution of characters - to transform a string into another. A cost is associated with each elementary operation, and the distance is defined as the minimum cost of the sequence of operations that leads from one string to the other. The correspondence established by this minimum cost sequence of operations is called a *global alignment* of the two strings, and algorithms with computational cost proportional to mn exist to perform this task, where m and n are the lengths of the two strings (Gusfield, 1997; Keller and Banzhaf, 1994; O'Reilly, 1997; Sankoff and Kruskal, 1983).

The problem of defining a measure of diversity for a population becomes more complicated if we assume that besides having variable length, the genomes of the individuals may also have different structure. For example, the genome of one individual might have a certain set of genes arranged in a certain order, and another individual might have the same set of genes but ordered differently, or it might even have a different set of genes. In these cases, the methods described above for genome comparison

cannot be applied. On the other hand, variable structure genomes are particularly interesting, especially in the light of the high reorganizability of biological genomes, which has been observed as a response of organisms to a crisis and is coming more and more into focus as one of the key players in the evolutionary potential of organisms (Shapiro, 2002). If, despite its variability, one is aware of the presence of a set of structural motifs (such as, for example, promoter sequences and specific protein coding regions in biological genomes) in the genome of every individual, the global string alignment approach based on the edit distance can still be applied by localizing and comparing one by one the corresponding motifs in the two genomes, and by assigning a cost to unmatched motifs. Alternatively, one can resort to algorithms that implement *local alignment* in place of global alignment (Gusfield, 1997; Sankoff and Kruskal, 1983). The difference is that the costs that were associated with the elementary operations in the definition of the edit distance, are now interpreted as rewards for matches, and the best matching subsequences - presumably corresponding to functionally significant motifs - are located and evaluated by the algorithm itself within the genomes of pairs of individuals. The computational cost of these approaches, however, becomes rapidly unmanageable, especially in an EC perspective where the population is a highly dynamic entity, whose diversity must be repeatedly calculated during an evolutionary run.

The approach adopted in the present paper tries to bring together the best of both worlds, by defining a measure of diversity for individuals and populations that applies to genomes with variable length and structure, but neither assumes the knowledge of the genome structure, nor incurs the computational cost entailed by the automatic identification of the actual genome motifs. To achieve this result, it looks for *potential* motifs contained in the genomes, and bases its measure of diversity on a combination of their number. This leads to a definition that applies to generic genomes, and which is capable of taking into account at least partially the structure of the genomes, while remaining computationally inexpensive. Moreover – as can be expected from a diversity measure – the definition gives a minimal value of diversity for the case of uniform populations and a maximal value for pairwise maximally distinct individuals, and it becomes a (semi-metric) distance when the population reduces to two individuals.

Despite being targeted to highly reorganizable genomes, the measures of diversity that will be defined apply also to simpler kinds of genomes, such as those having fixed length and uniform structure. Moreover, the definitions can find applications beyond EC, in any domain that requires the comparison of sequences of symbols - such as genomics, proteomics, chemical structure similarity assessment, and pattern recognition in general - and, even more generally, to domains that require the comparison of generic collections of "individuals" that can each be associated in a meaningful way with a set of features.

2 Population diversity

The genomes that we consider in this paper are strings of characters. Let us denote with s_{i_j} the string that constitutes the genome of the individual i_j of the population $P = \{i_1, i_2, \ldots, i_n\}$. In the following, we will identify an individual with its genome, and define diversities and distances for individuals in terms of their genome only. We are interested in counting the number of potential motifs contained in each individual genome and in the population genome. Since we do not assume any knowledge of the genome structure, the potential motifs of a string s_i constituting the genome of individual i, are all its substrings, that is, all the strings of characters that appear contiguously in s_i. Let us denote with $\mathcal{S}_i$ the set of substrings of s_i, and with $|\mathcal{S}_i|$ its

cardinality. Correspondingly, the potential motifs of a set of individuals, for example the population $P = \{i_1, i_2, \ldots, i_n\}$, are all the substrings that appear in the strings $\{s_{i_1}, s_{i_2}, \ldots, s_{i_n}\}$. We will denote this set of substrings with $\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}$ and its cardinality with $|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|$. Note that $\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}} = \bigcup_{j=1}^{n} \mathcal{S}_{i_j}$.

Example 1: Consider three individuals i_1, i_2, i_3, whose genomes are the strings $s_{i_1} = aba$, $s_{i_2} = abbc$, $s_{i_3} = babc$. We have:

- $\mathcal{S}_{i_1} = \{a, ab, aba, b, ba\}$, $|\mathcal{S}_{i_1}| = 5$
- $\mathcal{S}_{i_2} = \{a, ab, abb, abbc, b, bb, bbc, bc, c\}$, $|\mathcal{S}_{i_2}| = 9$
- $\mathcal{S}_{i_3} = \{b, ba, bab, babc, a, ab, abc, bc, c\}$, $|\mathcal{S}_{i_3}| = 9$
- $\mathcal{S}_{\{i_1,i_2,i_3\}} = \{a, ab, aba, b, ba, abb, abbc, bb, bbc, bc, c, bab, babc, abc\}$, $|\mathcal{S}_{\{i_1,i_2,i_3\}}| = 14$

We define the measure $D(P)$ of diversity of a population $P = \{i_1, i_2, \ldots, i_n\}$ as follows:

$$D(P) = D(\{i_1, i_2, \ldots, i_n\}) = n \frac{|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|}{\sum_{j=1}^{n} |\mathcal{S}_{i_j}|} \tag{1}$$

In words, the diversity of the population is defined as n times the ratio of the total number of substrings in the population genome (that is, considering only once those appearing in the genome of multiple individuals) to the cumulative number of substrings in the genome of the individuals considered separately.

Example 1 *(continued):* For the population constituted by the three individuals i_1, i_2, i_3 defined above, we have:

$$D(\{i_1, i_2, i_3\}) = 3 \frac{|\mathcal{S}_{\{i_1,i_2,i_3\}}|}{|\mathcal{S}_{i_1}| + |\mathcal{S}_{i_2}| + |\mathcal{S}_{i_3}|} = 3 \frac{14}{5 + 9 + 9} \approx 1.83$$

Let us analyze the properties of this definition by examining some particular cases:

- *Homogeneous population*. The population is constituted by n individuals having the same genome s and, consequently, the same set of substrings $\mathcal{S}$. Therefore, the set of substrings of the population coincides with the set of substrings of each individual: $\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}} = \mathcal{S}_{i_j} = \mathcal{S}$. >From the definition of D follows that

$$D(\{i_1, i_2, \ldots, i_n\}) = n \frac{|\mathcal{S}|}{\sum_{j=1}^{n} |\mathcal{S}|} = 1 \tag{2}$$

  an intuitively appealing result, since the population corresponds actually to a collection of clones of a single individual.

Note that the converse of this property is also true, that is, if the measure of diversity of a population is unitary, all the individuals have necessarily the same genome. This can be proved by the following argument. If $D = 1$ then from the definition of D follows that $n\,|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}| = n|\bigcup_{j=1}^{n} \mathcal{S}_{i_j}| = \sum_{j=1}^{n} |\mathcal{S}_{i_j}|$. Let us assume that there exists

a pair of individuals such that $S_{i_j} \neq S_{i_k}$. This means that there exist at least one substring that belongs to one of these two sets but not to the other. This substring will be counted n times in $n|\bigcup_{j=1}^{n} \mathcal{S}_{i_j}|$, but less than n times in $\sum_{j=1}^{n} |\mathcal{S}_{i_j}|$. Thus the condition $n|\bigcup_{j=1}^{n} \mathcal{S}_{i_j}| = \sum_{j=1}^{n} |\mathcal{S}_{i_j}|$ could not be realized, which contradicts our assumption. This proves that all the individuals in the population must have the same set of substrings in their genome, and, consequently, that they must have the same genome.

- *Population of pairwise maximally distinct genomes*. The population is constituted by individuals whose genomes, considered by pairs, do not have any substring in common, that is, $S_j \cap S_k = \emptyset$ for $j \neq k$. This means that each substring belonging to $\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}$ belongs only to one of the sets S_j and therefore $|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}| = \sum_{j=1}^{n} |\mathcal{S}_{i_j}|$, from which follows that $D(\{i_1, i_2, \ldots, i_n\}) = n$.

As before, the converse is also true, that is, if $D(\{i_1, i_2, \ldots, i_n\}) = n$ the individuals have, pairwise, no substrings in common. This can be proved by the following argument. If $D = n$ then from the definition of D follows that $|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}| = |\bigcup_{j=1}^{n} \mathcal{S}_{i_j}| = \sum_{j=1}^{n} |\mathcal{S}_{i_j}|$. Let us assume that there exists a pair of individuals such that $S_{i_j} \cap S_{i_k} \neq \emptyset$. This means that there exists at least one substring that belongs to both sets. This substring will be counted only once in $|\bigcup_{j=1}^{n} \mathcal{S}_{i_j}|$, but at least twice in $\sum_{j=1}^{n} |\mathcal{S}_{i_j}|$. Thus the condition $|\bigcup_{j=1}^{n} \mathcal{S}_{i_j}| = \sum_{j=1}^{n} |\mathcal{S}_{i_j}|$ could not be realized, which contradicts our assumption.

With analogous deductions, it can be proved that the values of diversity obtained in these two cases constitute actually a bound for $D(P)$, that is, that we always have $1 \leq D(P) \leq n$, where n is the size of the population. This fact, along with the interpretation of Equation 1 in terms of average number of substrings that will be presented shortly, suggests the interpretation of the value of $D(P)$ as the number of *equivalent individuals* of the population. For example, the three individuals of Example 1 above, correspond to about 1.83 equivalent individuals, a population of clones of a single individual corresponds to 1 equivalent individual., and a population of n individuals that have, pairwise, no genetic motifs in common, corresponds to n equivalent individuals.

- *Population with two kinds of genomes*. As a final example, consider a population constituted by a fraction α of its n individuals having the genome s', and the remaining fraction $(1 - \alpha)$ of individuals having a genome s'' that has no substrings in common with s'. We obtain

$$D(P) = \frac{|\mathcal{S}'| + |\mathcal{S}''|}{\alpha|\mathcal{S}'| + (1-\alpha)|\mathcal{S}''|} \tag{3}$$

  If $\alpha = 0.5$, that is, if the population is equally divided into individuals of type s' and individuals of type s'',we have $D(P) = 2$ independently from the values of $|\mathcal{S}'|$ and $|\mathcal{S}''|$. The same is approximately true when $|\mathcal{S}'| \approx |\mathcal{S}''|$ and α varies. If, on the other hand, we have $|\mathcal{S}'| \gg |\mathcal{S}''|$ or $|\mathcal{S}'| \ll |\mathcal{S}''|$, we can obtain almost any value of $D(P)$ in the range $(1, n)$.

This last example reveals the limitations of the measure of diversity defined above, and provides further insight into its operation. We can rewrite Equation 1 as follows

$$D(P) = \frac{|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|}{\frac{1}{n}\sum_{j=1}^{n} |\mathcal{S}_{i_j}|} = \frac{|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|}{\overline{|\mathcal{S}_i|}} \tag{4}$$

This shows that the measure of diversity compares the total number of different substrings in the population genome to the average number of substrings $\overline{|\mathcal{S}_i|}$. Therefore, if one or a few individuals possess a number of substrings that greatly exceeds the average number of them in the population, the formula overestimates the diversity of the population since it implicitly distributes evenly the substrings among the individuals.

2.1 Linguistic complexity

The diversity measure defined above is loosely related to the concept of *linguistic complexity* for a string defined on a given alphabet A. The linguistic complexity of a string s is defined as the ratio of the number of substrings of s, to the maximum number of substrings that can be obtained from a string of the same length on the same alphabet (Trifonov, 1990; Troyanskaya et al., 2002). In the spirit of our definition of population diversity given by Equation 1, we can generalize the concept of linguistic complexity from single strings to populations, as follows

$$LC(P) = LC(\{i_1, i_2, \ldots, i_n\}) = \frac{|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|}{\max\limits_{P'_A \sim P} |\mathcal{S}_{\{i'_1,i'_2,\ldots,i'_n\}}|} \tag{5}$$

where $\max_{P'_A \sim P} |\mathcal{S}_{\{i'_1,i'_2,\ldots,i'_n\}}|$ is the maximum number of substrings that can be obtained with a population P'_A built on the same alphabet A of P, and having the same number of individuals and with the same length.

The value of the linguistic complexity $LC(P)$ complements the information constituted by the value of the diversity $D(P)$. $LC(P)$ gives a measure of how well the population realizes the potential of motif existence constituted by the kind and number of its individuals. In other words, it gives an idea of how effectively the population is exploring the genome space, relative to what can be done with the same number of individuals, with the same genome lengths, and on the same alphabet. Thus, it is a relative measure of diversity, whereas $D(P)$ – which estimates the number of different individuals that the population contains – looks more like an absolute one. For example, a population P of a thousand random binary strings of length two will contain almost certainly all the six possible substrings of length one and two, and will therefore result in a value of $LC(P) = 1$ that testifies the fulfillment of the population potential. On the other hand, the six possible substrings, given the expected value of two and a half substrings for binary string of length two, will result in $D(P) \approx 2.4$, which suggests that only a handful of different individuals exist in the population but does not inform us about how many could be housed by a population having the same structure.

3 Distance between individuals

The diversity of a population is closely connected to the concept of distance between individuals. For example, a measure of diversity for a population can be obtained summing all pairwise distances between its individuals. Moreover, the distance between individuals can be used to define the distance between populations (Wineberg and Oppacher, 2003a). Hence, it is worth considering the possibility of using the substring count approach to define a distance between individuals belonging to populations with genomes of variable length. We will start by trying to derive a distance d between individuals applying our formula for population diversity (Equation 1) to pairs of individuals $\{i_1, i_2\}$. From the inequalities given above, we know that $D(\{i_1, i_2\})$ satisfies the inequality $1 \leq D(\{i_1, i_2\}) \leq 2$, with the lower bound achieved only for identical

individuals. Hence, for the expression

$$d(i_1, i_2) = D(\{i_1, i_2\}) - 1 = 2\,\frac{|\mathcal{S}_{\{i_1,i_2\}}|}{|\mathcal{S}_{i_1}| + |\mathcal{S}_{i_2}|} - 1 \tag{6}$$

we have $d(i_1, i_2) \geq 0$, with $d(i_1, i_2) = 0$ if and only if $i_1 = i_2$. Moreover, d is obviously symmetric in its arguments, that is $d(i_1, i_2) = d(i_2, i_1)$ for any pair of individuals. The triangle inequality $d(i_1, i_2) + d(i_2, i_3) \geq d(i_1, i_3)$, however, which would qualify d as a metric and its value as a distance between individuals and between strings, is *not* satisfied. For example, for the individuals i_1, i_2, i_3, with genomes $s_{i_1} = baaaa$, $s_{i_2} = baaaab$, and $s_{i_3} = aaaab$, we have $d(i_1, i_2) = d(i_2, i_3) \approx 0.217$, and $d(i_1, i_3) \approx 0.444$, so that $d(i_1, i_2) + d(i_2, i_3) < d(i_1, i_3)$. This makes of d a *semi-metric distance* in the space of strings. Note that in the following we will not mention explicitly the qualifier "semi-metric" for d. The distance thus defined satisfies the inequality $0 \leq d(i_1, i_2) \leq 1$.

Example 2: Consider the three individuals i_1, i_2, i_3, with genomes $s_{i_1} = abab$, $s_{i_2} = abcb$, and $s_{i_3} = cbab$. We have

- $\mathcal{S}_{i_1} = \{a, ab, aba, abab, b, ba, bab\}$, $|\mathcal{S}_{i_1}| = 7$
- $\mathcal{S}_{i_2} = \{a, ab, abc, abcb, b, bc, bcb, c, cb\}$, $|\mathcal{S}_{i_2}| = 9$
- $\mathcal{S}_{i_3} = \{c, cb, cba, cbab, b, ba, bab, a, ab\}$, $|\mathcal{S}_{i_3}| = 9$
- $\mathcal{S}_{\{i_1,i_2\}} = \{a, ab, aba, abab, b, ba, bab, abc, abcb, bc, bcb, c, cb\}$, $|\mathcal{S}_{\{i_1,i_2\}}| = 13$
- $\mathcal{S}_{\{i_1,i_3\}} = \{a, ab, aba, abab, b, ba, bab, c, cb, cba, cbab\}$, $|\mathcal{S}_{\{i_1,i_3\}}| = 11$

  from which we obtain

$$d(i_1, i_2) = 2\,\frac{|\mathcal{S}_{\{i_1,i_2\}}|}{|\mathcal{S}_{i_1}| + |\mathcal{S}_{i_2}|} - 1 = 2\,\frac{13}{7+9} - 1 = 0.625$$

$$d(i_1, i_3) = 2\,\frac{|\mathcal{S}_{\{i_1,i_3\}}|}{|\mathcal{S}_{i_1}| + |\mathcal{S}_{i_3}|} - 1 = 2\,\frac{11}{7+9} - 1 = 0.375$$

This example shows a characteristic of d that appears at first disturbing. We can consider both the genome of i_2 and that of i_3 as obtained from that of i_1 with a single character substitution, and yet the distance of i_1 from i_2 is greater than that of i_1 from i_3. The reason is that the substitution that leads from the genome of i_1 to that of i_2 is located towards the center of the genome, whereas that leading from i_1 to i_3 is located at one extreme of it. This permits the first substitution to create a bigger set of motifs $\mathcal{S}_{\{i_1,i_2\}}$ relatively to $\mathcal{S}_{\{i_1,i_3\}}$, and this is reflected in the difference of distances. As we observed in the introduction, to define our diversity measure we do not assume any knowledge of the structure of the genomes. Hence, the number of potential motifs produced by a substitution can depend on its position in the genome. If we knew where the genes or the motifs boundaries are located, we could exclude from the count of substrings those crossing those boundaries, and obtain a new definition of diversity and distance still based on the count of substrings but now taking into account our knowledge of the genome structure. In that case, the phenomenon illustrated by Example 2 would not appear. The definition of d, on the other hand, does not suffer the problem of "extraordinary genomes" mentioned in the previous section. If two individuals i_1 and i_2 are such that $|\mathcal{S}_{i_1}| \gg |\mathcal{S}_{i_2}|$, we obtain simply $d(i_1, i_2) \approx 1$, as expected.

3.1 Population diversity as sum of pairwise substring distances

We can use d to define a new measure of population diversity D' based on the traditional pairwise comparison of individuals (Wineberg and Oppacher, 2003a), as follows:

$$D'(P) = D'(\{i_1, i_2, \ldots, i_n\}) = \frac{2}{(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d(i_j, i_k) \quad , \quad n > 1 \tag{7}$$

Note that this definition takes advantage of the fact that $d(i, i) = 0$ and of the symmetry of d, while the multiplying factor keeps the values of D' in the range $[0, n]$.

3.2 Tanimoto distance

We can define another distance between individuals based on the counting of substrings, using the Tanimoto measure of similarity between two generic sets X and Y (known also as *Jaccard similarity* (Levandowsky and Winter, 1971)), which is defined as (Theodoridis and Koutroumbas, 2003)

$$\sigma_t(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{8}$$

The similarity of two individuals can therefore be defined as

$$\sigma_t(i_1, i_2) = \frac{|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2}|}{|\mathcal{S}_{i_1} \cup \mathcal{S}_{i_2}|} = \frac{|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2}|}{|\mathcal{S}_{\{i_1, i_2\}}|} \tag{9}$$

from which we can derive the Tanimoto substring distance between two strings

$$d_t(i_1, i_2) = 1 - \sigma_t(i_1, i_2) = 1 - \frac{|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2}|}{|\mathcal{S}_{\{i_1, i_2\}}|} \tag{10}$$

This distance can be substituted to d in Equation 7 to obtain a Tanimoto diversity measure D'_t for a population of strings

$$D'_t(P) = D'_t(\{i_1, i_2, \ldots, i_n\}) = \frac{2}{(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_t(i_j, i_k) \quad , \quad n > 1 \tag{11}$$

Example 2 *(continued):* For the population constituted by the three individuals i_1, i_2, i_3, with genomes $s_{i_1} = abab$, $s_{i_2} = abcb$, and $s_{i_3} = cbab$, introduced above, we have

- $\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2} = \{a, ab, b\}$, $|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2}| = 3$
- $\mathcal{S}_{i_1} \cap \mathcal{S}_{i_3} = \{a, ab, b, ba, bab\}$, $|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_3}| = 5$

  from which we obtain

$$d_t(i_1, i_2) = 1 - \frac{|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2}|}{|\mathcal{S}_{\{i_1, i_2\}}|} = 1 - \frac{3}{13} \approx 0.77$$

$$d_t(i_1, i_3) = 1 - \frac{|\mathcal{S}_{i_1} \cap \mathcal{S}_{i_3}|}{|\mathcal{S}_{\{i_1, i_3\}}|} = 1 - \frac{5}{11} \approx 0.55$$

3.3 Generalized distance and diversity

To define the Tanimoto substring distance, we specialized the general definition given by Equation 8 to the case of strings. We can go in the opposite direction with our distance d between strings defined by Equation 6, and see it as particular case of the distance between two generic sets X and Y, defined by

$$d(X,Y) = 2\,\frac{|X \cup Y|}{|X| + |Y|} - 1 = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y| + |X \cap Y|} \tag{12}$$

This reveals the similarity of d with the Tanimoto distance d_t, which corresponds to

$$d_t(X,Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} \tag{13}$$

Contrary to the case of d, however, d_t satisfies the triangle inequality and is therefore a metric distance (Levandowsky and Winter, 1971; Lipkus, 1999).

In an analogous way, we can interpret Equation 1 as a particular case of the following measure of diversity for a collection (or multiset (Monro, 1987)) $\{X_1, X_2, \ldots, X_n\}$ of finite and not all empty sets X_j

$$D(\{X_1, X_2, \ldots, X_n\}) = n\,\frac{|\bigcup_{j=1}^{n} X_j|}{\sum_{j=1}^{n} |X_j|} \tag{14}$$

This means that D can be used to measure the diversity of a generic population of "individuals" i_j, provided there is a way to associate with each of them a set X_j which is representative of its relevant substructures and features for the application at hand. For example, remaining in the realm of strings, we could deem more meaningful for the assessment of the diversity of a population of them, the use of the set of subsequences (i.e., of characters that do not necessarily appear contiguously) instead of the set of substrings of the individual strings; in the field of image classification, one could associate with each image a set of its subimages, and so on.

3.4 Tree-based representations and genetic programming

As an example of an application of the generalized distances and diversity measures defined in the previous subsection to genetic representations not based on strings we can consider the field of genetic programming (GP). In GP the genome of individuals usually has the structure of a tree (Langdon and Poli, 2002). Following an approach proposed in (Keijzer, 1996) we can associate with each individual i_j the set X_j of all the subtrees of the tree that constitutes its genome. With this choice, Equation 14 gives a subtree-based measure of diversity for GP populations. Similarly, interpreting X and Y as the set of subtrees of the trees that constitute the genomes of two individuals, Equation 13 becomes a Tanimoto distance and Equation 12 becomes a subtree distance between individuals.

4 Implementation issues

The definitions of diversity and distance given above are practically useful for EC runtime calculations only if there exist efficient ways to compute the number of substrings of a string and of a collection of strings.

The number of substrings of a string can be calculated efficiently building the so-called *suffix tree* of the string. This is a data structure that compactly represents the

substring structure of a string and which is based on a less compact structure called *trie.* The trie associated with a string is a rooted directed tree where the edges are labeled by letters, any path down the tree spells a substring of the string, such that all paths from root to leafs are suffixes of the string and all suffixes of the string are labels of paths from the root (Crochemore and Rytter, 2002) (Figure 1). Note that paths corresponding to suffixes do not end necessarily in a leaf. Nodes of the trie where paths corresponding to suffixes end are called *essential nodes.* The property of a trie that interests us here is the fact that the number of its edges corresponds to the number of substrings of the string (Troyanskaya et al., 2002).

The trie associated with a string can be compacted by suppressing non-branching non-essential nodes and associating with the edges of the tree thus obtained the substrings obtained from the chain of original edges (Figure 1). The resulting structure is called the *suffix tree* of the string (Crochemore et al., 2001; Crochemore and Rytter, 2002; Gusfield, 1997). Since the labels associated with the edges of the suffix tree correspond to substrings of the original string, they can be substituted with pointers to the substring within the string. This allows a further compaction of the suffix tree relative to the corresponding trie (Figure 1).

Several algorithms exist that build the suffix tree of a string with a computational cost that grows linearly with the length of the string, both in terms of computation time and memory occupation. Two popular algorithms are Ukkonen's (Ukkonen, 1995) and McCreight's (McCreight, 1976): a detailed description of these algorithms including the pseudocode can be found in (Crochemore et al., 2001; Crochemore and Rytter, 2002; Gusfield, 1997). These algorithms build the suffix tree by adding successively the characters that correspond to the edges of the trie to which the suffix tree corresponds. Hence, to obtain the number of substrings of a string we must simply count the number of characters added during the construction of its suffix tree. This permits an efficient computation of the number $|\mathcal{S}_{i_j}|$of substrings of the genome of an individual of a population.

The algorithms that build the suffix tree of a single string can be modified to allow the construction of a structure representing the suffixes of a collection of strings which is called the *generalized suffix tree* (Gusfield, 1997, p. 116). The basic idea is to append to each string in the collection an end string marker not belonging to the alphabet from which the strings are formed. To build the generalized suffix tree, one starts by constructing the suffix tree of the first end-marked string using one of the algorithms listed above. Then, the second end-marked string in the collection is matched against the existing suffix tree starting from the root, until a mismatch occurs. At this point the construction of the suffix tree resumes with the first non-matched character of the string. Proceeding in this way for all the strings in the collection results in the construction of the generalized suffix tree. Counting the number of characters added during the construction one obtains the number of substrings of the strings in the collection and, in particular, the number of substrings $|\mathcal{S}_{\{i_1,i_2,\ldots,i_n\}}|$ of the genomes of a population.

5 Experimental results and comparisons

We now compare the measures of diversity introduced above with those currently used in EC and in other fields that make use of string comparison. In what follows we will denote with n the size of the population, with l the length of the genomes of the individuals, and with A the alphabet over which the genomes are defined.

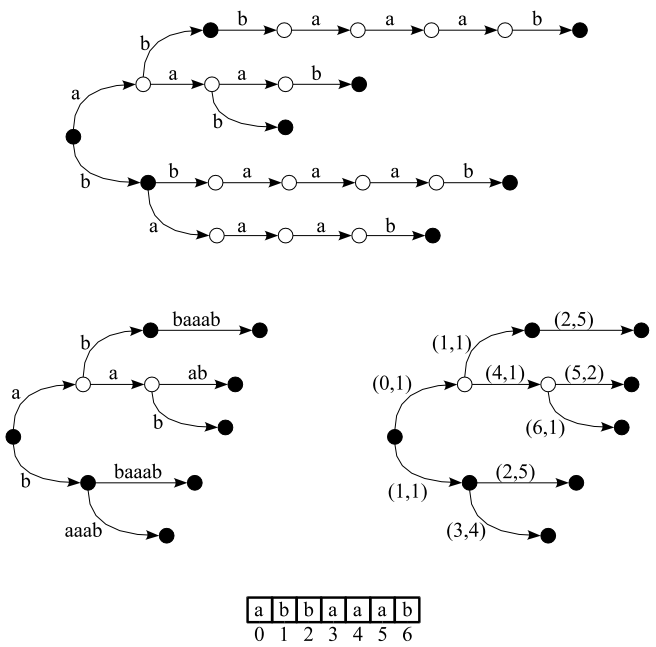

Figure 1: The trie and suffix tree of the string *abbaaab*, whose characters are indexed from 0 to 6 (*bottom*). The trie (*top*) has a single letter associated with each edge and the number of its edges corresponds to the number of substring of the string. Nodes represented in black are *essential nodes* and correspond to suffixes of the string, which can be obtained by traversing the tree from the root (which corresponds to the empty suffix) to the node. The trie can be compacted by suppressing non-branching non-essential nodes and associating with the new edges the substrings obtained from the chain of original edges. This gives the suffix tree of the string (*center, left*). An alternative, more compact representation of the suffix tree can be obtained by substituting the substrings associated with each edge with a pair of integers (p, l) that gives the position of the start of the substring in the original string, and the length of the substring (*center, right*).

5.1 Computational cost

We consider the following kinds of diversity measures for populations of n individuals

- *Leung-Gao-Xu diversity*. Leung, Gao, and Xu (Leung et al., 1997) introduced a measure of diversity $D_\lambda(P)$ for populations whose individual genomes are binary strings s_j of fixed length l. $D_\lambda(P)$ is defined as the number of components of the string of integers $\sum_{j=1}^{n} s_j$ (with the sum performed componentwise in $\mathbb{N}$) whose values are not equal to 0 and n. The time computational complexity of the direct implementation of this definition is $O(l \cdot n)$.

- *Moment of inertia diversity, pairwise Hamming diversity*, and *entropic diversity*: We group under a unique heading three diversity measures that are slight variations on the same theme (Morrison and De Jong, 2002; Wineberg and Oppacher, 2003b). In the experiments that will follow, we will only report the results of the moment of inertia measure, since the other two require almost the same computation time

and have a value that differs from the moment of inertia only by a scaling factor and, possibly, by terms of second order and higher in the character frequencies (Wineberg and Oppacher, 2003b).

– The *moment of inertia diversity* for populations of binary strings of fixed length l is defined by

$$D_m(P) = \sum_{i=1}^{l} \sum_{j=1}^{n} (s_{ij} - c_i)^2 \tag{15}$$

where s_{ij} is the character in position i of the j-th string, c_i is the i-th coordinate of the centroid

$$c_i = \frac{1}{n} \sum_{j=1}^{n} s_{ij} \tag{16}$$

and the operations are performed in $\mathbb{R}$. The time complexity of the direct implementation of Equation 15 is $O(l \cdot n)$ (Morrison and De Jong, 2002).

- The *pairwise Hamming diversity* is defined as

$$D_h(P) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_h(i_j, i_k) \tag{17}$$

where $d_h(i_j, i_k)$ is the Hamming distance between the individual genomes. $D_h(P)$ is defined for populations with genomes of fixed length l over an arbitrary finite alphabet. The time complexity of the naive implementation of Equation 17 is $O(l \cdot n^2)$ but there exist implementations of this measure that reduce the time complexity to $O(l \cdot n)$ (Morrison and De Jong, 2002; Wineberg and Oppacher, 2003b). For the case of binary genomes it can be shown (Morrison and De Jong, 2002) that $D_h(P) = n\, D_m(P)$. Hence, $D_h(P)$ differs from $D_m(P)$ only by a scaling factor and can be implemented with the same complexity. For arbitrary genomes, $D_h(P)$ can be calculated with $O(l \cdot n)$ complexity using the following expression (Wineberg and Oppacher, 2003b)

$$D_h(P) = \frac{n^2}{2l} \sum_{k=1}^{l} \sum_{\alpha \in A} f_k(\alpha)(1 - f_k(\alpha)) \tag{18}$$

where $f_k(\alpha)$ is the frequency of the character α at the position k in the population genomes.

- The *entropic diversity* is defined as

$$D_e(P) = -\frac{1}{l} \sum_{k=1}^{l} \sum_{\alpha \in A} f_k(\alpha) \log f_k(\alpha) \tag{19}$$

where $f_k(\alpha)$ is the frequency of the character α at the position k in the population genomes. The first term of the Taylor expansion of $D_e(P)$ is proportional to the pairwise hamming distance $D_h(P)$, which is therefore a good approximation. $D_e(P)$ can be implemented with time complexity $O(l \cdot n)$ (Wineberg and Oppacher, 2003b).

- *Substring diversity*. It is defined by Equation 1. Using suffix trees it can be implemented with time complexity $O(l \cdot n)$.

- *Pairwise substring diversity* and *pairwise Tanimoto diversity*. They are defined by Equation 7 and Equation 11, respectively. Using suffix trees to compute the number of substrings, and then applying the definitions directly, the time complexity of the implementation is $O(l \cdot n^2)$.

To give an idea of the actual computation time of a typical implementation on a present day personal computer, we have plotted in Figure 2 the computation time for these measures of diversity as a function of genome length and population size for a randomly generated population of fixed genome length over a binary alphabet. The substring diversity measure is implemented with the McCreight suffix tree algorithm (McCreight, 1976). As anticipated, the results for the pairwise Hamming diversity and for the entropic diversity are not shown, since they are almost indistinguishable from the values obtained for the moment of inertia diversity.

The curves of Figure 2 confirm the predicted time complexities and show that the substring diversity measure $D(P)$, although more computationally expensive than existing diversity measures for fixed length genomes, is sufficiently inexpensive to be usable for runtime diversity measurements on present day computers. Note that the pairwise substring and pairwise Tanimoto curves are almost coincident, and that their $O(l \cdot n^2)$ complexity makes the direct implementation of these measures rapidly impractical for runtime diversity assessment when the population size grows.

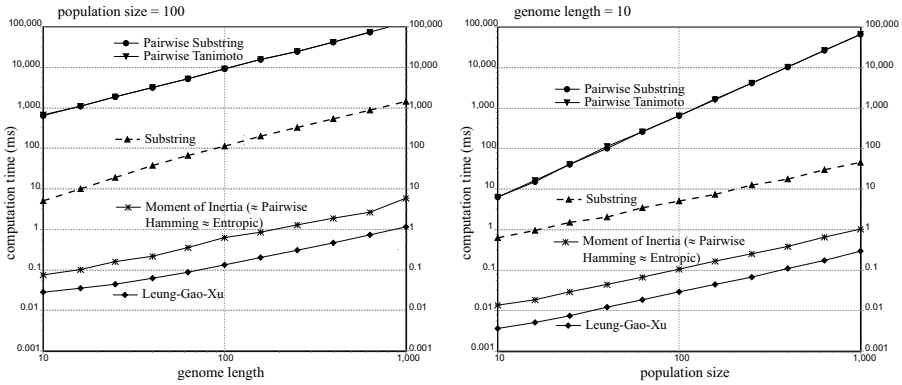

Figure 2: Computation time vs. genome length and population size of different measures of diversity for the implementations described in the text and run on a PC with Pentium III microprocessor clocked at 850MHz.

5.2 Fixed genome length

To give an example of computation of the various kinds of diversities in an actual EC setting, we performed a series of runs of a genetic algorithm on a function optimization problem. The genomes considered in this section have fixed length. Hence, both the conventional non-string-based and the string-based diversity measures can be used. The goal is to show that in this simple setting – the only one where the comparison can be performed – the string-based measures give results comparable to those of the conventional measures.

The function to be optimized was the two-dimensional sine envelope sine wave function (Leung et al., 1997) defined by

$$f(x_1, x_2) = 0.5 - \frac{\sin\sqrt{x_1^2 + x_2^2} - 0.5}{(1 - 0.001(x_1^2 + x_2^2))^2} \tag{20}$$

The optimization was performed in the domain $[-100, 100] \times [-100, 100]$, where the unique global maximum is $f(0, 0) = 1$. Each parameter x_i was binary encoded by a string of length 22, so that $l = 44$. The population size was $n = 50$. The algorithm used tournament selection with tournament size $t = 2$, with mutation probability $p_m = 0.01$ and crossover probability $p_c = 0.85$. Figure 3 reports the results average over ten runs, starting with randomly generated populations.

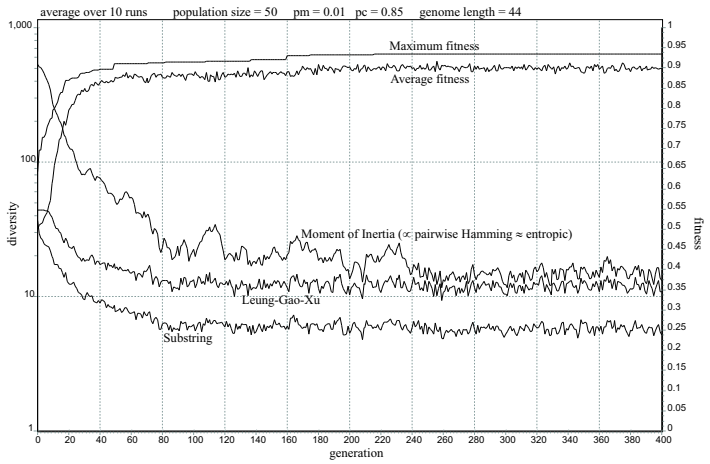

Figure 3: The various population diversity measures computed on the population evolved with a genetic algorithm applied to a function optimization experiment.

Figure 3 does not show the pairwise Hamming diversity, since it differs from the moment of inertia diversity only by a scale factor, nor the entropic diversity, since, after suitable scaling, it is almost indistinguishable from the moment of inertia diversity. The values of the pairwise substring and pairwise Tanimoto diversities are close to those of the Leung-Gao-Xu diversity. To avoid an excessive overlapping of curves, all the substring-based diversities are represented in Figure 4 for the same set of runs of Figure 3. In this case, the substring diversities for the genes that represent each parameter were also separately computed as *substring columnwise diversity*. This was done to show that the substring approach can be applied to evaluate the diversity of each genome substructure when these are known. Note that the substring diversity calculated on the whole genome is not the sum of the two substring columnwise diversities, which is instead the case for most kinds of diversity measures currently used, since they obtain the global diversity summing the contribution of each locus in the genome string.

Figures 3 and 4 show that in this conventional setting constituted by populations with fixed genome length, the measures of diversity based on the substring approach behave very much like the familiar diversity measures based on the Hamming distance, or based on the count of the converged bits like the Leung-Gao-Xu measure. In

other words, when applied to fixed genome length populations the substring-based measures conform to the behavior expected from a conventional measure of diversity.

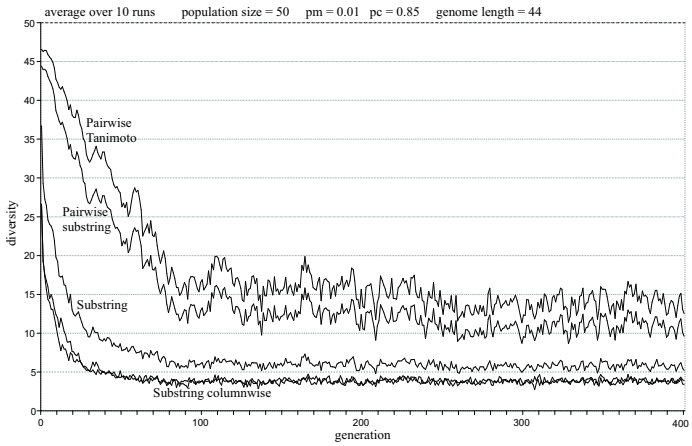

Figure 4: The substring-based population diversity measures computed for the evolutionary runs of Figure 3. The substring columnwise curves refer to the genomic diversity computed for each parameter separately.

5.3 Variable genome length

So far we have considered only examples of population diversity calculations made on fixed genome length populations. This was necessary to allow the comparison with conventional diversity measures, which apply only to the fixed genome length case. However, the measures of diversity introduced here find their justification mainly in the case of populations with variable genome structure, where the approaches based on the Hamming distance cannot be used. To illustrate the applicability of the substring approach to the variable genome length - which is just the first step towards a truly variable structure, to which the substring approach still applies - we modified the encoding of the function optimization experiment described above to allow variable length encoding of each parameter. More precisely, the parameters x_i were encoded by a binary string whose length was allowed to be different for the two parameters of an individual, and to vary from individual to individual. We started the evolutions with randomly generated populations where each parameter was represented by a gene with length randomly chosen between 2 and 20 bits. To allow the variation of the genome length of individuals during the evolution, the mutation operator was extended to include character insertion and deletion with probabilities $p_i = p_d = 0.001$, in addition to substitution with probability $p_m = 0.01$. Individuals that, following a mutation, were found to have an empty gene were simply removed from the population.

Figure 5 shows the result of a single evolutionary run of function optimization with variable length genome. Note that the substring-based measures of population diversity introduced above apply unaltered to the case of variable genome length, whereas the traditional population diversity measures cannot be applied to this case.

5.4 Highly reorganizable genome

In the previous examples all individuals had a genome with the same structure, that is, the number, meaning, and order of the genes was predefined and at most the length of

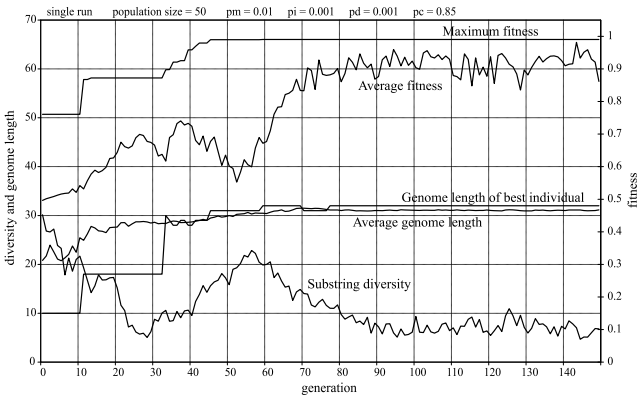

Figure 5: The substring diversity measure computed on the population evolved with a genetic algorithm applied to a function optimization experiment with variable length genome. From generations 30 to 60 there is probably an episode of migration of the population from two distinct regions of the genetic space; this is accompanied by a decrease in average fitness and is reflected in an increase of diversity of the population. Note that in this experiment the length of the genome of distinct individuals of the population can be different and, therefore, the Leung-Gao-Xu, pairwise Hamming, moment of inertia, and entropic diversity measures cannot be used.

each gene could vary. We consider now an evolutionary system where the genome of individuals can be restructured much more freely and the complexity of the structure decoded from the genome can change accordingly. The interest of exploring this kind of genomes stems from their potentially greater evolutionary openness compared to genomes with fixed structure.

The system we consider was developed to evolve networks of devices connected by links endowed with a scalar connection strength. Examples of this kind of network are neural networks (NN), analog electronic circuits, genetic regulatory networks, and many other technological and biological networks. The interested reader can find in (Mattiussi and Floreano, 2004) a more extensive discussion and description of the system with examples of evolution of electronic circuits. In the case of neural networks that we will consider here, the devices are artificial neurons and the strength of the connections corresponds to the weights of the links connecting the output of a neuron to the input of another neuron (Haykin, 1999). The basic idea of the evolutionary system is to work with a string genome and associate a substring extracted from it with each input and output terminal of the neuron. This is done by defining a set of predefined tokens that identify in the genome the devices and delimit the strings that will be extracted from the genome and associated with the terminals by the genome decoding process. The weight of the connections is determined by a preassigned function f that associates numerical values to pairs of strings (Figure 6).

With this encoding strategy the genome can be subjected to many genetic operators such as insertion, deletion, and substitution of characters; duplication, deletion and transposition of fragments; insertion of device descriptors; recombination of genomes. The genetic operators can thus create new devices and, implicitly, new connections in the network. The genome thus modified remains decodable because it is the presence of a group of tokens that determines the existence of a decodable device, irrespective of the position of the group in the genome. On the other hand the genetic operators can

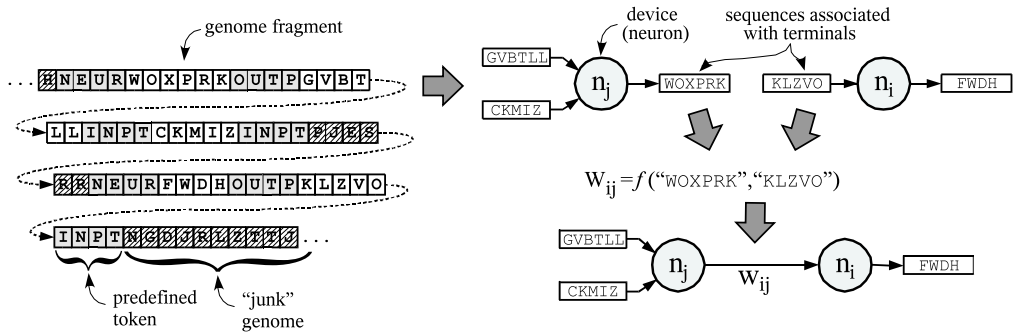

Figure 6: A fragment of genome (*left*), the devices decoded from it (*right top*), and an example of weighted connection established between the devices (*right, bottom)*. A set of predefined tokens ("NEUR", "OUTP", "INPT", ...) identifies the regions coding for devices and delimits the strings that will be associated with the terminals by the decoding process. The weight of the connections between the output and the inputs of all neurons is obtained by applying a function f that maps pairs of strings associated with the terminals, to numeric values.

invalidate the tokens corresponding to existing devices, preventing their decoding. The corresponding fragment of genome then becomes simply a fragment of "junk" genome which is available as raw material for the evolutionary process, or can be eliminated by a mutation that deletes a genome fragment. We have thus a genome where a given structure can appear in any position in the genome and can be present in several instances as more or less diverging copies. We are therefore in the presence of the kind of highly reorganizable genome whose relevance for artificial evolution was discussed in the Introduction and for which the substring-based diversity and distance measures come to full fruition.

Figure 7 reports the result of a single evolutionary run aimed at the synthesis of a neural network solving the XOR problem (Haykin, 1999, p. 175). The setup of the system provides two predefined input neurons and an output neuron. The neurons decoded from the genome are thus inserted as hidden neurons (Haykin, 1999). An exact solution is characterized by a fitness value of zero, and is first found at about generation 120. It is known (Haykin, 1999) that the simplest network that can solve the problem has a single (hidden) neuron decoded from the genome. The curve of the average number of decoded neurons testifies that the evolutionary process produces networks that are more complex than this while proceeding towards the exact solution. The same curve shows that these networks are subsequently simplified by the evolutionary process until, at about generation 300, the population is composed almost exclusively of individuals whose genome encodes a single neuron. This is due to the presence of a "dead zone" around the required output values, which eliminates any selective pressure among networks that can produce outputs within the dead zone and, in particular, does not reward networks that "associate" many neurons just to better approximate the exact output values. The curves reporting the average and maximum genome length during the evolution testify to the presence of several episodes of duplication and deletion of genome fragments. The duplication episodes in particular appear instrumental to the first attainment of an exact solution. The curve of the population substring diversity shows that the diversity remains low before an exact solution is found, that is, while the selective pressure is high. Successively, the population di-

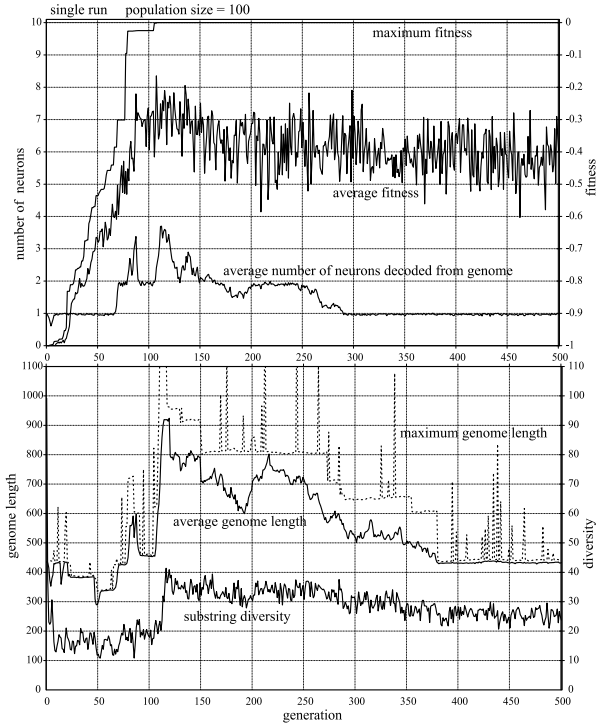

Figure 7: An evolutionary run aimed at the synthesis of a neural network for the XOR problem. The meaning of the curves is detailed in the text.

versity assumes consistently higher values relative to the pre-solution phase, irrespective of the value of genome length. This observation supports the hypothesis of the existence of several alternative solutions. Moreover, it indicates that the substring diversity measure is able to capture the essential dynamics of the population diversity. Note that all the non-substring-based diversity measures considered above are defined only for fixed genome lengths and cannot be applied to this kind of variable-length and variable-structure genome.

5.5 Nucleotide diversity and substring diversity

In molecular and population genetics the measurement of the polymorphism of a population is based on the *nucleotide diversity* measure Π, defined on a DNA sequence by

$$\Pi = \sum_j \sum_k x_j x_k \pi_{jk} \tag{21}$$

where the sum is performed on all types of different sequences, x_j and x_k are the frequencies of the j-th and k-th type of sequences, and π_{jk} is the proportion of different nucleotides between the two types of sequences (Graur and Li, 2000). It follows from this definition that Π corresponds to

$$\Pi = \frac{1}{l\, n^2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_h(i_j, i_k) \tag{22}$$

that is, it is a normalized pairwise Hamming diversity measure (Wineberg and Oppacher, 2003b) which gives the average number of nucleotide differences per site (Graur and Li, 2000).

Figure 8 shows two examples of the calculation of Π for a population of four DNA sequences. Note that the value of Π in the two cases is the same, since the number of nucleotide differences is the same, even if the differences are spread among more individuals in the second case. Conversely, the value of D is different in the two cases, since D is sensitive to the number of DNA sequences affected by the differences. Hence, the information provided by the substring diversity measure D can complement that conveyed by Π in assessing population polymorphism.

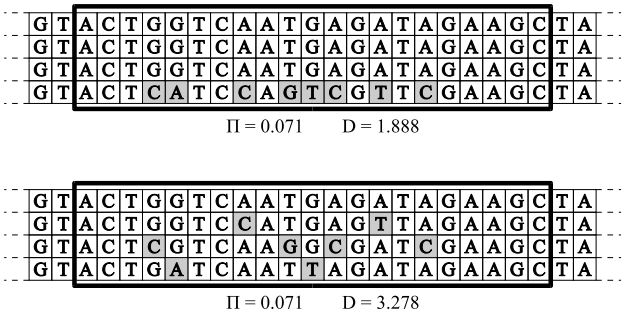

Figure 8: The nucleotide diversity measure Π is used in molecular genetics to assess the polymorphism of a population. It measures the average number of nucleotide differences per site in DNA sequences of the population. As such, Π is not affected by the distribution of the differences among the individuals. On the contrary, the substring diversity measure D is sensitive to this difference, and can therefore complement Π in assessing the polymorphism of the population.

6 Conclusion

There is a growing interest in the field of evolutionary computation in the definition of genomes that can be subject to major reorganizations – such as insertions, deletions duplications and transpositions – and still remain decodable. The reason for this interest is the hope of thereby fostering the exploration of the search space, the emergence of modularity, the reuse of evolved substructures, and, eventually, the open-endedness of the evolutionary process. At the same time, the effort to improve the performance of evolutionary algorithms results in a tendency towards the integration in the algorithms of a certain degree of control of the population diversity. It follows from this that there is a need for measures of diversity that apply to populations with genomes of variable length, and, more generally, to populations with highly reorganizable genomes.

To fulfill this need, we have defined measures of diversity and distances between individuals, that apply to populations and individuals whose genome is constituted by finite strings of variable length on finite alphabets. The definitions are based on the counting of substrings of the population genomes, considered first separately and then collectively. The motivation behind the substring counting approach is the possibility of estimating, in this way, the potential genomic motifs contained in the genomes. For example, there is the possibility of recognizing the similarity of individuals whose genomes are constituted by the same motifs but differently arranged. The measures thus obtained do not require any detailed knowledge of the structure of the genome

and, therefore, apply to generic genomes. However, information about the genome structure can be taken into account when available by applying the counting procedures to the substructures present in the genomes. The measures introduced are based on properties of the genomes that can be computed in linear space and time, thus making them suitable for runtime application during an evolutionary process. Moreover, these measures and their generalizations can be used for the assessment of the diversity of other kinds of populations, such as tree-based genetic programming populations, biological sequences, and generic collections of sets.

Supporting information

An example of implementation of the measures of diversity described above, inclusive of source code, can be downloaded from the Autonomous Systems Laboratory website, at the address `http://asl.epfl.ch/resources.php`

Acknowledgments

Many thanks to Antoine Beyeler and Jean-Christophe Zufferey for reading and commenting on the manuscript. This work was supported by the Swiss National Science Foundation, grant no. 620-58049.

References

Crochemore, M., Hancart, C., and Lecroq, T. (2001). *Algorithmique du texte*. Vuibert, Paris.

Crochemore, M. and Rytter, W. (2002). *Jewels of Stringology*. World Scientific, Singapore.

de Jong, E., Watson, R., and Pollack, J. (2001). Reducing bloat and promoting diversity using multi-objective methods. In et al., L. S., editor, *GECCO 2001*, pages 11–18, San Francisco, CA. Morgan Kaufmann.

Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 2nd edition.

Gusfield, G. (1997). *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Sunderland, MA.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NY, 2nd edition.

Keijzer, M. (1996). Efficiently representing populations in genetic programming. In Angeline, P. et al., editor, *Advances in Genetic Programming*, volume 2, pages 259–278, CAmbridge, MA. MIT Press.

Keller, R. and Banzhaf, W. (1994). Explicit maintenance of genetic diversity on genospaces. Unpublished manuscript. Available online at Citeseer.

Langdon, W. and Poli, R. (2002). *Foundations of Genetic Programming*. Springer, Berlin.

Leung, Y., Gao, Y., and Xu, Z. (1997). Degree of population diversity - a perspective on premature convergence in genetic algorithms and its markov chain analysis. *IEEE Trans. Neural Networks*, 8(5):1165–1176.

Levandowsky, M. and Winter, D. (1971). Distance between sets. *Nature*, 234(5):34–35.

Lipkus, A. (1999). A proof of the triangle inequality for the tanimoto distance. *J. of Mathematical Chemistry*, 26:263–265.

Mattiussi, C. and Floreano, D. (2004). Evolution of analog networks using local string alignment on highly reorganizable genomes. In et al., R. Z., editor, *Proceedings of the 2004 NASA/DoD Conference on Evolvable Hardware, 24-26 June 2004, Seattle*, pages 30–37, Los Alamitos, CA. IEEE Computer Society.

McCreight, E. (1976). A space-economical suffix tree construction algorithm. *J. ACM*, 23(1):262–272.

Monro, G. (1987). The concept of multiset. *Zeitschr. f. math. Logik ung Grundlagen d. Math.*, 33:171–178.

Morrison, W. and De Jong, K. (2002). Measurement of population diversity. In et al., P. C., editor, *EA 2001*, volume 2310 of *LNCS*, pages 31–41, Berlin. Springer.

O'Reilly, U.-M. (1997). Using a distance metric on genetic programs to understand genetic operators. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 5, pages 4092– 4097.

Sankoff, D. and Kruskal, J. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.

Shapiro, J. (2002). A 21st century view of evolution. *J. Biol. Phys.*, 28:1–20.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*. Academic Press, Sand Diego, CA, 2nd edition.

Tomassini, M., Vanneschi, L., Fernández, F., and Galeano, G. (2004). A study of diversity in multipopulation genetic programming. In et al., P. L., editor, *EA 2003, Artificial Evolution: 6th International Conference, Marseilles, France, October 27-30, 2003*, volume 2936 of *LNCS*, pages 243–255, Berlin. Springer.

Trifonov, E. (1990). Making sense of the human genome. In Sarma, R. and Sarma, M., editors, *Structure and Methods*, volume 1, pages 69–77. Adenine Press, New York.

Troyanskaya, O., Arbell, O.and Koren, Y., Landau, G., and Bolshoy, A. (2002). Sequence complexity of prokariotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–688.

Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica*, 14:249–260.

Wineberg, M. and Oppacher, F. (2000). Enhancing the ga's ability to cope with dynamic environments. In et al., D. W., editor, *GECCO 2000*, pages 3–10, San Francisco, CA. Morgan Kaufmann.

Wineberg, M. and Oppacher, F. (2003a). Distance between populations. In et al., E. C.-P., editor, *GECCO 2003*, volume 2724 of *LNCS*, pages 1481–1492, Berlin. Springer.

Wineberg, M. and Oppacher, F. (2003b). The underlying similarity of diversity measures used in evolutionary computation. In et al., E. C.-P., editor, *GECCO 2003*, volume 2724 of *LNCS*, pages 1493–1504, Berlin. Springer.