COST Action CA15140 on Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice: STSM Report

STSM title: Design of neutral representations for evolutionary search
Cost STSM Reference number: COST-STSM-CA15140-35053
Grand period: 2016-11-01 to 2016-11-30
Applicant: Vida Vukašinović
Home institution: Jožef Stefan Institute
Home country: Slovenia
Host: Carlos Fonseca
Host institution: University of Coimbra
Host country: Portugal

Purpose of the STSM:

The role of neutrality in evolutionary search has been widely considered in the literature, but it remains unclear to what extent it may contribute to search performance. In previous studies, a promising family of representations based on error-control codes was proposed, leading to sets of representations with various degrees of neutrality, synonymity, connectivity, and locality. The main objectives of this STSM, which is the first step in the long-term goal of the proposed collaborative research work, were to produce a mathematical formalisation capable of exposing more of the structure possessed by the aforementioned sets of neutral representations and to develop an efficient algorithm for the complete enumeration of such sets.

Description of the work carried out during the STSM:

First, the family of representations constructed using mathematical tools from error-control coding theory introduced in [1] was studied. Second, a mathematical formalization of the whole problem including the definitions of neutrality, synonymity, connectivity, and locality was provided and the implicit formalization of the above mentioned representations was replaced by an explicit mathematical formalisation. Attention was paid to uniform representations exhibiting the same properties in every neutral network. Equivalence classes of representations having some identical properties were identified and formalised mathematically, and suitable proofs were given. The insights gained support the development of efficient techniques for the enumeration of neutral representations exhibiting various degrees of these properties.

Description of the main results obtained:

In this work we study a representation r, which is a surjective mapping from the genotype space G to the phenotype space P. In particular, genotype space G is an n-dimensional vector space over $\mathbb{Z}_2$, while phenotype space P is a k-dimensional vector space over $\mathbb{Z}_2$. In both cases, the addition $\oplus$ of two vectors is the component-wise XOR operation and the scalar multiplication by the elements from $\mathbb{Z}_2$ is the AND operation. We say that a representation r is redundant if $n > k$. Further, *mutation* is a bilinear mapping $m : G \times G \to G$, where $m(g, e) = g \oplus e$ for each $g, e \in G$. The single point mutation of the i-th component g is denoted $m_i(g) = m(g, e_i)$, where all components of $e_i \in G$ are 0 except for the i-th component. A mutation m_i on genotype g is *neutral* if $r(g) = r(m_i(g))$, i.e., if the occurrence of mutation m_i on the genotype g does not change the corresponding phenotype. If for each pair $g_1, g_2 \in M \subseteq G$ there exists a sequence of genotypes $h_1 = g_1, h_2, \ldots, h_\mu = g_2$, where $h_j \in M$ for all $j = 1, \ldots, \mu$ and neutral mutations m_{i_j} for $j = 1, \ldots, \mu - 1$, such that $h_{j+1} = m_{i_j}(h_j)$, then M is called a *neutral network* [4]. We say that a neutral network is a set of genotypes connected by single point neutral mutations.

In order to try to characterise the types of redundant representations which may be beneficial for evolutionary algorithms, different properties have been introduced [3]. We took a closer look at uniformity, connectivity, synonymity, and locality.

A representation r is said to be uniformly redundant if $|r^{-1}(p_1)| = |r^{-1}(p_2)|$ for all $p_1, p_2 \in P$. In other words, if a representation is uniformly redundant, all phenotypes are represented by the same number of genotypes.

Connectivity c_r of representation r is defined as

$$c_r = \frac{1}{|P|} \sum_{p \in P} \left| r(N(r^{-1}(p))) \right|,$$

where $N(r^{-1}(p)) = \{g \in G \setminus r^{-1}(p) \,|\, \exists h \in r^{-1}(p) : d_G(g, h) = 1\}$, i.e. the set of all genotypes not in $r^{-1}(p)$ which are at distance one to at least one genotype in $r^{-1}(p)$. Connectivity measures the number of phenotypes which are accessible from a given phenotype by a single point mutation. Single point mutations m_i on $r^{-1}(p)$ have the potential to change the phenotype by changing the genotype in a single bit.

Synonymity of representation r is defined as

$$s_r = \frac{1}{|P|} \sum_{p \in P} \frac{1}{\binom{|r^{-1}(p)|}{2}} \sum_{\{g,h\} \subseteq r^{-1}(p)} d_G(g, h).$$

Representation r is said to be synonymously redundant if genotypes assigned to the same phenotype are similar to each other, i.e. if s_r is small.

Locality of representation r measures to what extent neighbouring genotypes correspond to neighbouring phenotypes. We say that two genotypes (resp., phenotypes) are adjacent if the distance between them is minimal, i.e. they differ in exactly one bit. Rothlauf [3] claimed that a representation has perfect locality if adjacent genotypes correspond to adjacent phenotypes and hence he summed over all pairs of adjacent genotypes the absolute value of the distance between the corresponding phenotypes minus the minimum distance between adjacent phenotypes, which in our case equals one. By his definition, adjacent genotypes from the same neutral network, i.e. those that are mapped to the same phenotype, contribute to higher locality values, which we consider an unwanted effect. Hence, we propose the following definition of locality, where the subtraction of the minimum distance between any two adjacent phenotypes is omitted and the sum is normalised by the number of adjacent genotypes in order to treat representations from genotype spaces of different sizes equally:

$$l_r = \frac{1}{2^{\ell-1}\ell} \sum_{\{g_1,g_2\}\subseteq G : d_G(g_1,g_2)=1} d_P(r(g_1), r(g_2)).$$

Representation r is said to be highly local if adjacent genotypes correspond to adjacent phenotypes, i.e., if l_r is small.

The definition of representation mapping in this work is primarily motivated by the theory of error-control codes. The proposed representations were first introduced by Fonseca and Correia [1] who explained the motivation details and their connection to error-control codes. Such representations are especially interesting because they are able to exhibit various degrees of neutrality, connectivity, and locality. In this work we propose their direct mathematical formulation without implicit introduction through error-control encodings [2].

Definition 1. Let i be an inclusion of phenotype space P in genotype space G and Z be a transversal of all cosets of $i(P)$ in G. A representation r is **compatible** with i and Z if the following conditions hold:

1. Z forms a neutral network in G,
2. $r(z_0) = 0_P$,
3. for each $g \in i(P)$ it holds that $r(z_0 \oplus g) = r(z_i \oplus g)$ for every $i = 0, \ldots, \tau - 1$.

Among all representations which are compatible with i and Z, we pay special attention to those whose restriction $r|_{i(P)} = i^{-1} \circ m(\cdot, z_0)$.

Definition 2. A representation r is said to be **fully compatible** with inclusion i and transversal Z if r is compatible with inclusion i and transversal Z and $r|_{i(P)} = i^{-1} \circ m(\cdot, z_0)$.

The first important step in this work was to show that if a representation r is fully compatible with inclusion i and transversal Z, it suffices to compute connectivity, synonymity, and locality locally on one phenotype $p \in P$ in order to compute the connectivity, synonymity, and locality values of representation r.

Theorem 1. *Let i be a linear inclusion of P into G. Let r be a representation which is fully compatible with inclusion i and transversal Z. Then,*

$$s_r(p_1) = s_r(p_2), \quad c_r(p_1) = c_r(p_2), \quad l_r(p_1) = l_r(p_2)$$

for every $p_1, p_2 \in P$.

Second, for a given inclusion i we provide sufficient conditions for the representations to possess the same properties, i.e. connectivity, synonymity, and locality. This greatly reduces the number of representations that need to be considered and the most important result of this part is summarised within the following theorem.

Theorem 2. *Let i be a linear inclusion of P into G. Let r, r' be representations which are fully compatible with inclusion i and transversals Z, $Z \oplus c$, respectively. Then,*

$$s_{r'}(p) = s_r(p), \quad c_{r'}(p) = c_r(p), \quad l_{r'}(p) = l_r(p)$$

for every $p \in P$.

Third, we identified equivalence classes of representations exhibiting the same connectivity and synonymity, while it is expected that locality may vary from representation to representation. The main result of this part is given by the following theorem.

Theorem 3. *Let i be a linear inclusion of P into G. Let π be a permutation of n elements and G_π its permutation matrix such that $G_\pi\, i(P) = i(P)$. Let r, r' be representations which are fully compatible with inclusion i and transversals Z, $G_\pi\, Z$, respectively. Then,*

$$s_{r'}(p) = s_r(p), \quad c_{r'}(p) = c_r(p)$$

for every $p \in P$.

Based on the insights from this work, an algorithm for efficient enumeration of representations for a given inclusion i was conceptually developed and most parts of the algorithm are already implemented. In the next 2–3 months we plan to merge separate parts of the algorithm, and perform an enumeration of the representations associated with the Hamming(15,11) error-control code. The database containing the representations thus found will provide the basis for future collaboration.

Future collaboration with the host institution (if applicable):

In the near future our collaboration will be continued via skype meetings and on the sidelines of the COST Action's MC and other meetings. As a possible future direction for the continuation of the work done so far, we recognize the exploration of the obtained representation database and the identification of promising representations by studying parallels between RNA sequence to secondary structure mappings found in theoretical biology and genotype-phenotype mappings used in evolutionary computations. In this regard we will try to establish collaboration with researchers from theoretical biology in Europe. We also find other representation properties relevant, such as accessibility and its study under the proposed neutral representations is a possible direction for future collaboration.

Foreseen publications/articles resulting from the STSM (if applicable):

First results obtained during this STSM were already presented by Carlos Fonseca on the 6th December in the international CoLaB Workshop (Mathematics of Complex Systems: from precision medicine to smart cities) in Coimbra. More information is available on https://www.mat.uc.pt/colab2016. Further, we plan to submit the results obtained during this STSM in a suitable international journal within next 3 months.

References

[1] Fonseca, C. M.; Correia, M. B. 2005. Developing redundant binary representations for genetic search, In Proc. CEC05, 372-379.

[2] Lin, S.; Costello, D. J. 1983. Error control coding: fundamentals and applications, Prentice Hall.

[3] Rothlauf, F., Goldberg D. E. 2003. Redundant representations in evolutionary computation, Evolutionary Computation, vol. 11, no 4, 381-415.

[4] Schuster, P.; Fontana, W.; Stadler, P. F.; Hofacker I. L. 1994. From sequences to shapes and back: A case study in RNA secondary struc-

tures, Proceedings of the Royal Society B, Biological Sciences, vol. 255, 279-284.