Universal Information Distance for Genetic Programming

M. Gaudesi*, G. Squillero*, A. Tonda#

* Politecnico di Torino, Italy – http:/www.cad.polito.it/
#INRA UMR 782, MALICES, France – http://www.iscpif.fr/

GOAL: propose a new distance metric for Genetic Programming, based on the normalized symmetric difference between the *information* contained in the genome.

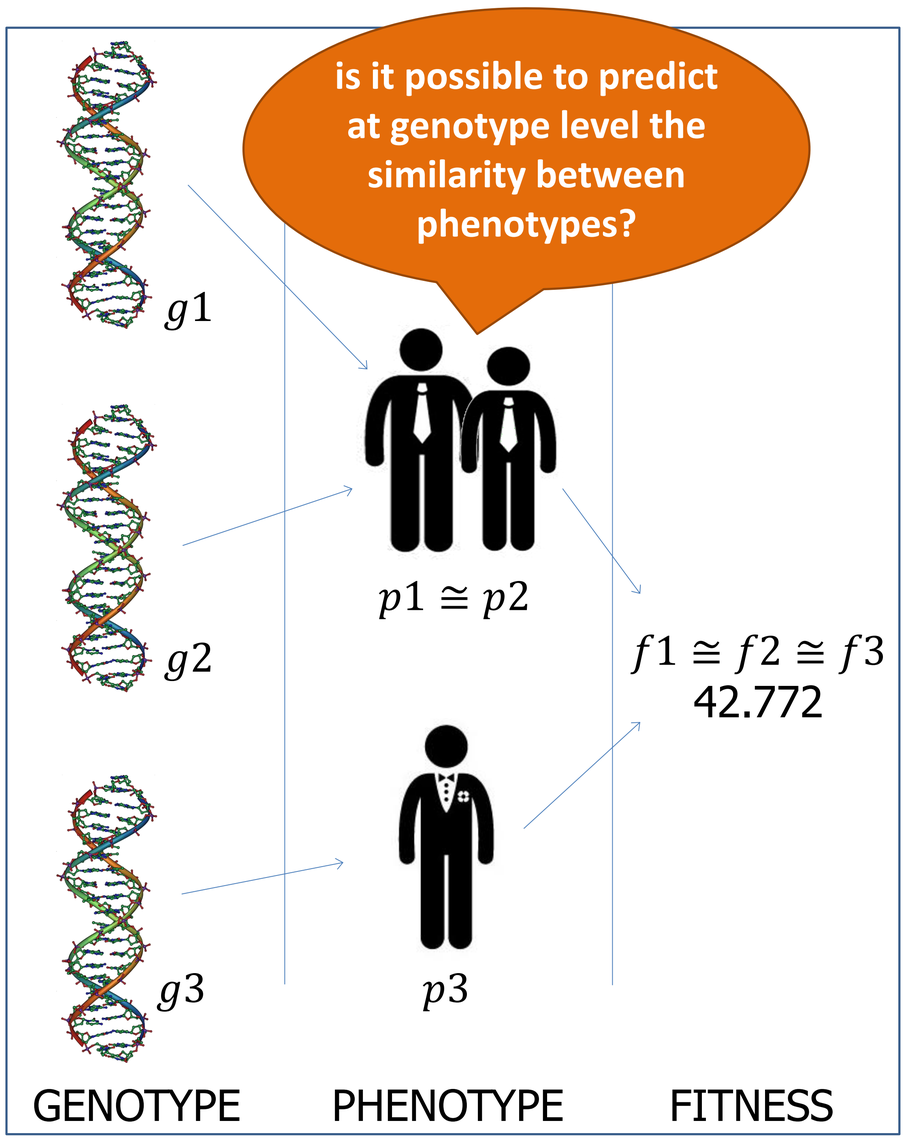
Proposed Approach:

Considering two individuals I_i and I_j, the *UID* is defined as: $UID(I_i, I_j) = \frac{|S(I_i) \,\Delta\, S(I_j)|}{|S(I_i)|+|S(I_j)|}$, where

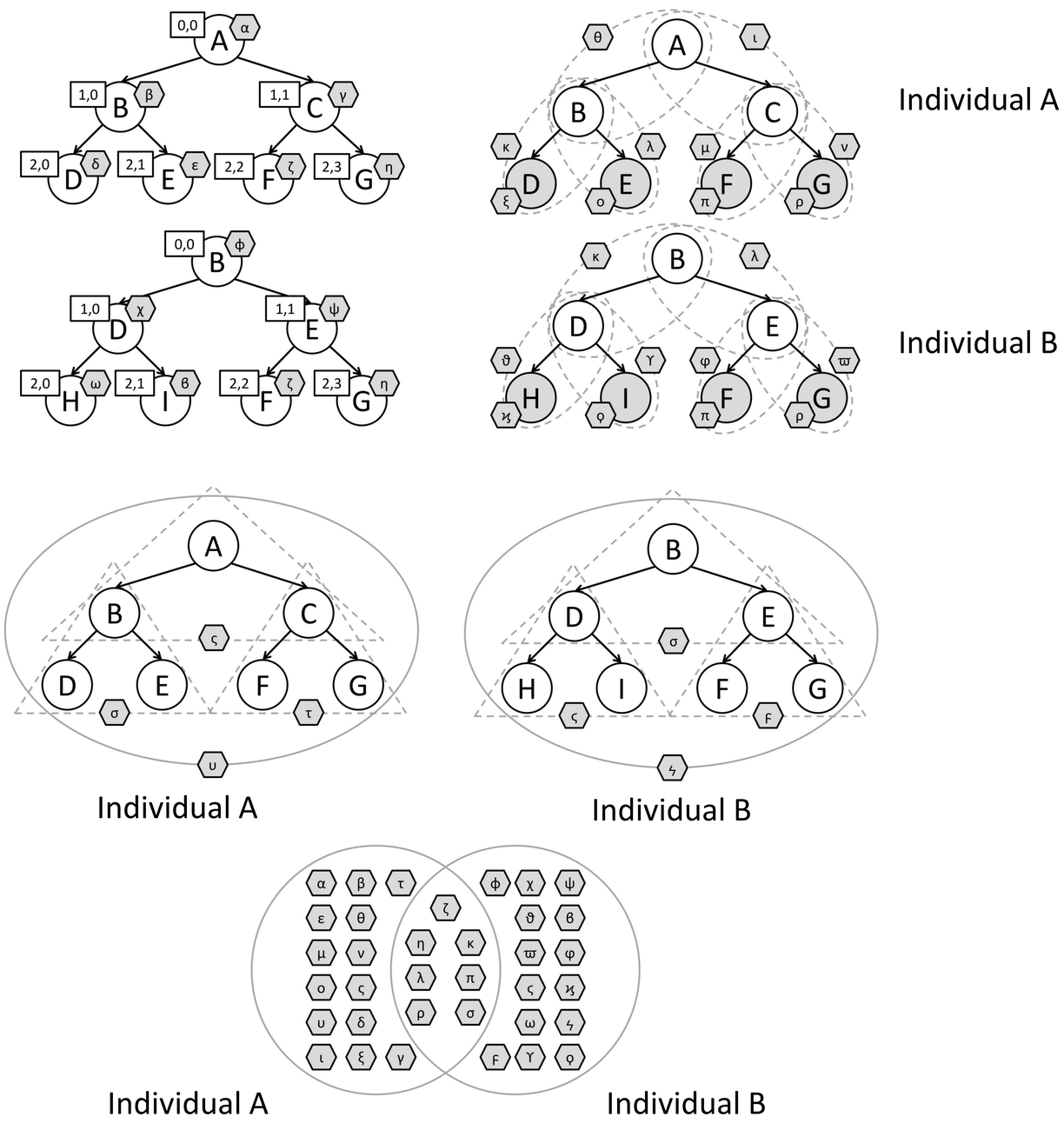
- $S(I)$ represents a set of symbols associated to the individual I
- Δ is the symmetric difference defined by the formula: $A \,\Delta\, B = (A \cup B) - (A \cap B)$
- $|S|$ denotes the cardinality of set S

Four kinds of symbols are used to characterize a GP individual (see figure «Symbols Adopted»):

Node and position	Symbols encode the content of a node and its absolute (*x*, *y*) position inside the GP tree, where *x* is the level of the node, and *y* its order inside the level;
Node and child	Symbols encode node *A* and one of its children *B*, whitout considering its relative position;
Node and children	Symbols encode the content of node *A* and all its children, also taking into account their position with respect to the parent node;
Node and grandchildren	Symbols encode the content of node *A*, all its children, and all its children's children, taking into account their position with respect to the parent node.

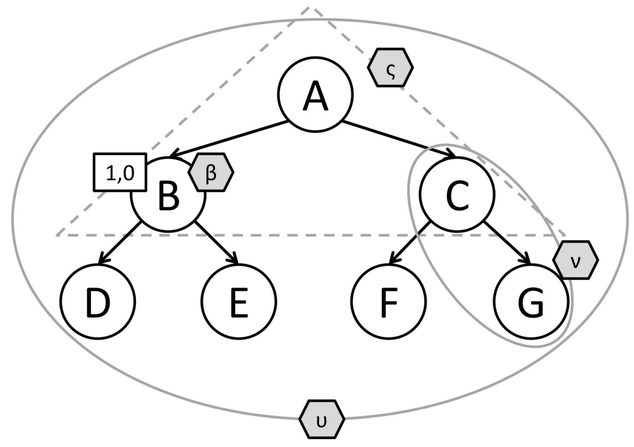


UID computed between GP individuals A and B

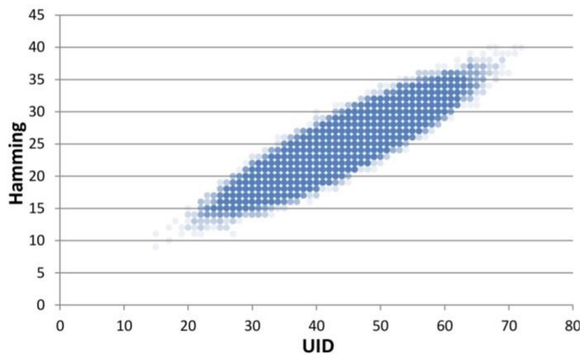


Symbols are computed and sotred in the symbol set of each individual, then a symmetric difference is performed between the two sets.
The cardinality of the resulting set is the UID.
In this case *UID(A, B) = 28*, since there are 28 symbols that are not shared between the two individuals.
Shared symbols reflect the similarities between the two individuals, namely the structure *B D E* and the terminal symbols *F* and *G*. Normalization is omitted.

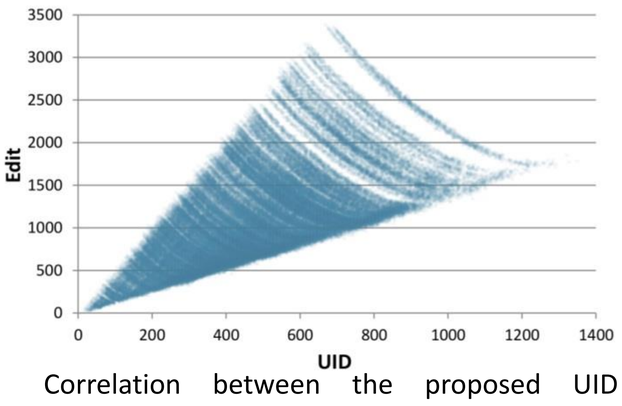
Symbols Adopted



- **Node and position**: for node *B*, in level 1 and place 0, obtaining symbol β
- **Node and child**: for nodes *CG*, resulting in symbol ν
- **Node and children**: for the structure *ABC*, that produces symbol ξ
- **Node and grandchildren**: for the whole *ABCDEFG*, resulting in symbol υ



Correlation between the proposed UID distance and Hamming distance in the standard OneMax problem (50 bits) – Sample of 500 random individuals



Correlation between the proposed UID distance and the Levenshtein distance in the Assembly OneMax problem (32 bitS) – Sample of 500 random individuals

Experimental results

radius	$k = 1,000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$
0.10	43.72	58.73	84.28	86.33	25.61
0.15	77.02	63.41	56.77	62.99	40.79
0.20	75.78	39.02	11.30	11.80	10.24

In order to validate the proposed approach, the minimal GP engine *TinyGP* [Poli, Langdon] was modified to include the UID. The table reports the percentage of fitness improvements using fitness sharing; each experiments evolved 1000 individuals for 100 generations

To speed up calculations, m_i was simply set to the number of individuals within the given radius divided by a constant *k*

Contacts:
giovanni.squillero@polito.it – alberto.tonda@grignon.inra.fr