# An Efficient Implementation of Geometric Semantic Genetic Programming for Anticoagulation Level Prediction in Pharmacogenetics

Mauro Castelli[1,4] Davide Castaldi[2] Leonardo Vanneschi[1,2,3]

Ilaria Giordani[3] Francesco Archetti[3] Daniele Maccagnola[3]

(1) ISEGI, Universidade Nova de Lisboa, 1070-312, Lisboa, Portugal
(2) D.I.S.Co., Università di Milano-Bicocca, Milano, Italy
(3) Consorzio Milano Ricerche, 20126, Milano, Italy
(4) INESC-ID, 1000-029, Lisboa, Portugal

## Categories and Subject Descriptors

I.2.2 [**Artificial Intelligence**]: Automatic Programming

## General Terms

Algorithms, Performance

## Keywords

Genetic Programming, Geometric Semantic Operators, Oral Anticoagulation Therapy

In the last few years researchers have dedicated several efforts to the definition of Genetic Programming (GP) [**?**] systems based on the semantics of the solutions, where by semantics we generally intend the behavior of a program once it is executed on a set of inputs, or more particularly the set of its output values on input training data (this definition has been used, among many others, for instance in [**?**, **?**, **?**, **?**]). In particular, new genetic operators, called geometric semantic operators, have been proposed by Moraglio et al. [**?**]. They are defined s follows:

DEFINITION 1. **(Geometric Semantic Crossover).** *Given two parent functions $T_1, T_2 : \mathbb{R}^n \rightarrow \mathbb{R}$, the geometric semantic crossover returns the real function $T_{XO} = (T_1 \cdot T_R) + ((1 - T_R) \cdot T_2)$, where $T_R$ is a random real function whose output values range in the interval $[0, 1]$.*

DEFINITION 2. **(Geometric Semantic Mutation).** *Given a parent function $T : \mathbb{R}^n \rightarrow \mathbb{R}$, the geometric semantic mutation with mutation step $ms$ returns the real function $T_M = T + ms \cdot (T_{R1} - T_{R2})$, where $T_{R1}$ and $T_{R2}$ are random real functions.*

As Moraglio et al. demonstrate in [**?**], these operators have interesting properties: crossover produces and offpring whose semantic vector is on the line that joins the two semantic vectors of the parents; mutation induces a unimodal fitness landscape on any problem consisting in finding the match between a set of input data and a set of known outputs (like for instance regression and classification), which clearly facilitates GP evolvability.

Nevertheless, as stated by Moraglio et al. [**?**], these operators have a serious limitation: by construction, they always produce offspring that are larger than their parents. The size of the individuals in the population grows exponentially with generations and thus these operators are unusable in practice. The solution suggested in [**?**] to overcome this drawback is to integrate in the GP algorithm a "simplification" phase, aimed at transforming each individual in the population into an equivalent (i.e. with the same semantics) but possibly smaller one. Even though this is an interesting and challenging study, depending on the language used to code individuals simplification can be very difficult, and it is often a limited solution to code growth.

For this reason, in [**?**] a new implementation of these operators has been presented. Given the strict length limit of this publication, we have no space to describe it here. We just outline the fact that in this implementation only the individuals in the initial population and a set of other random ones are stored using the traditional tree representation. All the other individuals that will be generated by GP in the next generations are stored using a system of memory references. Furthermore, the semantics of each individual is also stored (which is done easily using the semantics of the ancestors and applying the definition of the operators). In this way, we are able to calculate fitness without having to evaluate the trees. This implementation has a computational complexity that is linear with the number of performed generations and with the population size. Its complexity does not depend on the size of the individuals in the population, and this makes it even more efficient than standard GP.

In this paper the objective is to evaluate the regression performance of this new GP system (Geometric Semantic GP, or GS-GP, from now on) on an application in pharmacogenetics, comparing the results with the ones obtained by standard GP. In partcular, we are interested in applying GS-GP to Oral Anticoagulation Therapy (OAT). The prediction of appropriate oral anticoagulant level of medical drugs is a particularly relevant issue given that it has to be individually determined for each patient. For this reason, studying new and powerful computational methods is of paramount importance. Evolutionary computation techniques have already been used in pharmacokinetics so far (see for instance [**?**, **?**]), but to the best of our knowledge never for this particular application.

The case we study is the one of *coumarin*-derived OAT.


*GECCO'13 Companion,* July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

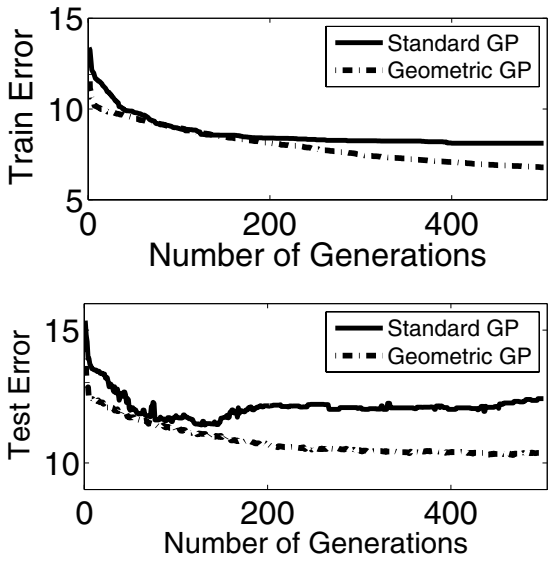


**Figure 1: Upper plot: median of training fitness for the considered techniques at each generation calculated over 30 independent runs. Lower plot: median of test fitness for the same 30 runs.**

Coumarin is commonly used for life-long therapy in the prevention of systemic embolism in patients with atrial fibrillation, valvular heart disease, and in the primary and secondary prevention of venous and pulmonary thromboembolism. It is also used for the prevention of thromboembolic events in patients with acute myocardial infarction and with angina pectoris, in patients with heart valves, and after some types of orthopedic surgery. In all subjects, a combination of personal, genetic and non-genetic factors are responsible for about 20-fold variation in the coumarin's dose required to achieve the therapeutic range of drug action. In this work, we have collected data from clinical computerized databases based on 950 genotyped over 65 years old patients in OAT. A data preprocessing approach returned 748 *cleaned* patients (i.e. with complete data, not missing any information) useful for analysis. The features of this dataset can be summarized in four main entities: personal characteristics, anamnestic features, genetic data and therapy's characteristics.

The experimental results we have obtained are reported in Figure 1, where we show the median of training and test error for STD-GP and GS-GP against generations for 30 independent runs. These figures clearly show that GS-GP outperforms STD-GP on both training and test sets. The differences between the results obtained by GS-GP and the ones obtained by standard GP are statistically significant (the statistical study is omitted to save space).

The good results that GS-GP has obtained on training data were expected: the geometric semantic operators induce an unimodal fitness landscape, which facilitates evolvability. On the other hand, these excellent results on the training set do not affect the generalization ability of the model on unseen data. This had already been observed in [**?**]. An interpretation of it can be that the geometric properties of those operators hold *independently of the data* on which individuals are evaluated, and thus they hold also on test data. In other words, geometric semantic crossover produces an offspring that stands between the parents also in the semantic space induced by test data. As a direct implication, following exactly the same argument as Moraglio et al. [**?**], each offspring is, in the worst case, not worse than the worst of its parents on the test set. Analogously, as it happens for training data, geometric semantic mutation produces an offspring that is a "weak" perturbation of its parent also in the semantic space induced by test data (and the maximum possible perturbation is, again, expressed by the $ms$ step). The immediate consequence for the behaviour of GS-GP on test data is that, while geometric semantic operators do not guarantee an improvement in test fitness each time they are applied, they at least guarantee that the possible worsening of the test fitness is bounded (by the test fitness of the worst parent for crossover, and by $ms$ for mutation). In other words, *geometric semantic operators help controlling overfitting*. We remark that, without the novel implementation that allowed us to use geometric semantic GP on these complex real-life problems, this interesting property would probably remained unnoticed.

**Acknowledgments.** This work was supported by national funds through FCT under contract Pest-OE/EEI/LA0021/2011 and by projects EnviGP (PTDC/EIA-CCO/103363/2008) and MassGP (PTDC/EEI-CTP/2975/2012), Portugal.