# A Hybrid GA-MLR Algorithm for Automated Descriptor Selection in QSAR Modeling

Chandan Srivastava †, Alberto Fernández †, Robert Rallo ‡, Josep Cester † and Francesc Giralt †

† *BIOCenit Research Lab, Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain*

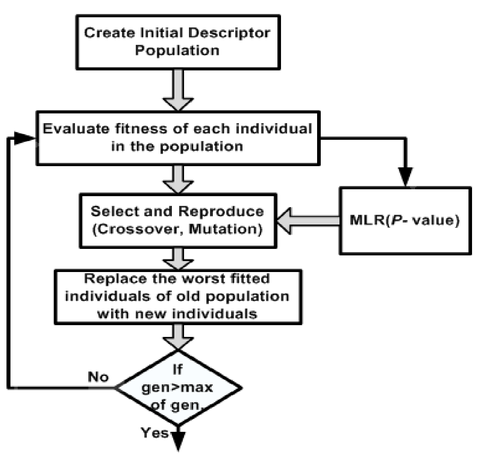
‡ *BIOCenit Research Lab, Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain*

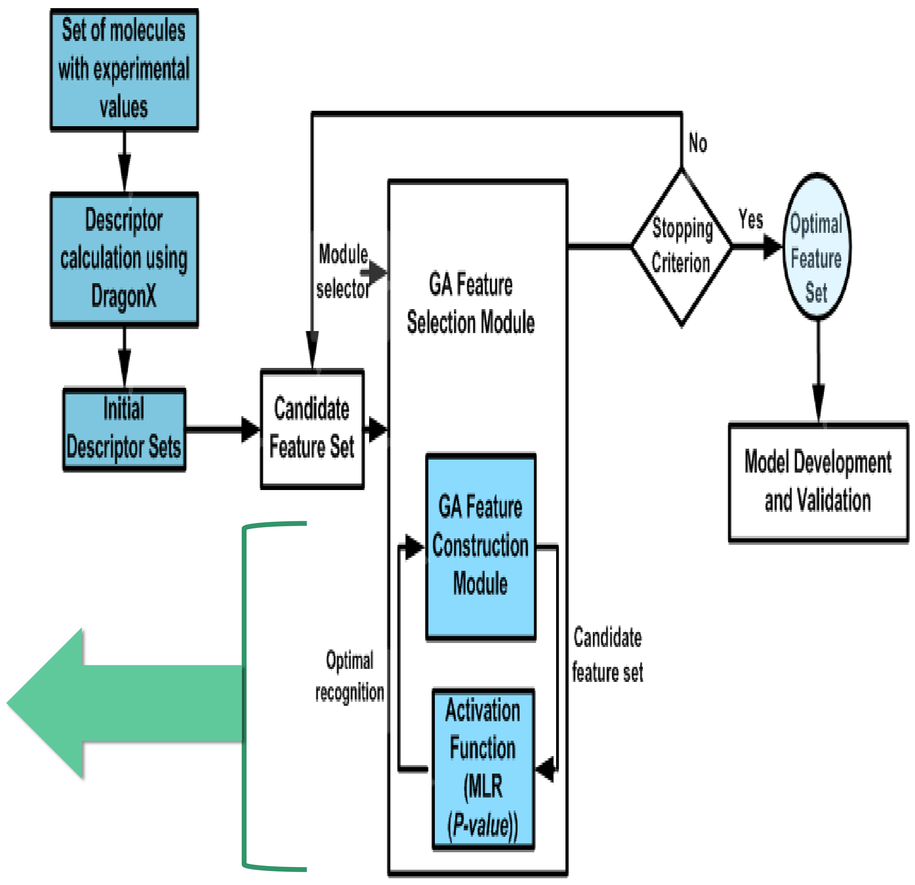
## Motivation and Objectives

- The development of Quantitative Structure-Activity Relationships (QSAR) involves the analysis of high dimensional feature spaces
- The selection of truly relevant feature subsets is fundamental for the establishment of accurate and reliable QSARs suitable for regulatory purposes
- The goal of the current work is to develop and validate a new wrapper approach for feature selection based on the integration of genetic algorithms (GA) with multilinear regression (MLR) analysis

## Architecture of the GA-MLR Feature Selector

- The genetic algorithm operates on a set of molecular descriptors (population).
- New candidate subsets (individuals) are generated from the initial population using crossover and mutation operators.
- The optimal subset is identified by a fitness function based on the *p-value* of a MLR model.
- Stopping criteria are either the number of generations or a fitness function threshold.



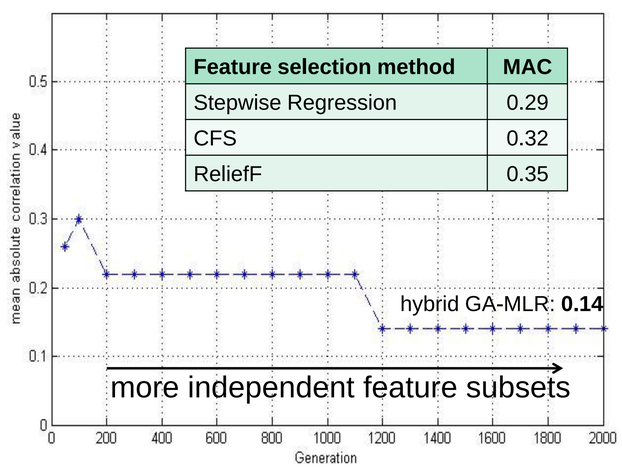




### Performance Assessment:

**Mean Absolute Correlation (MAC) for each combination of descriptors**

$$C = 2 * \frac{\sum abs(D_{i,j})}{[N * (N-1)]}$$

where *D* are the up-diagonal elements of the correlation matrix and N is total number of descriptors selected.




| Feature selection method | MAC |
|---|---|
| Stepwise Regression | 0.29 |
| CFS | 0.32 |
| ReliefF | 0.35 |

Evolution of the MAC for increasing number of generations

## Case Study: Modeling Biodegradation in Water

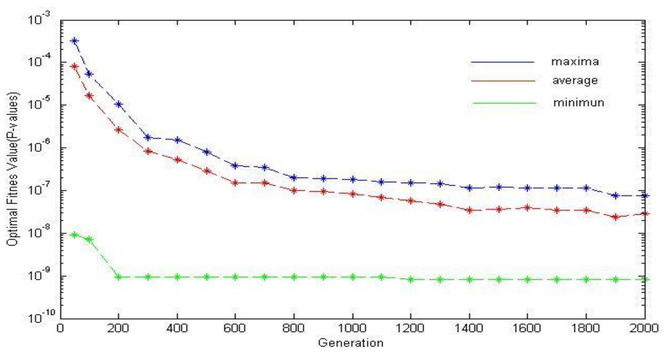
### 1. Data Sources:

**endpoint:** MITI-I biodegradation
**chemical Space:** 1063 compounds
(71% persistent, 29% biodegradable)
**type of descriptors:** Constitutional
**total number of descriptors** (calculated by DragonX): 40

**Optimized subsets of molecular descriptors**

| Selected Descriptors | Descriptor information | GA-MLR Model | Stepwise Regression | CFS | Relief F |
|---|---|---|---|---|---|
| MW | molecular weight | X | - | - | - |
| AMW | average molecular weight | - | - | - | - |
| Sp | sum of atomic polarizabilities | - | - | - | - |
| ARR | aromatic ratio | - | X | X | - |
| nCIC | number of rings | - | X | - | - |
| nCIR | number of circuits | X | - | - | - |
| RBF | rotatable bond fraction | - | X | X | X |
| nSK | number of non-H atoms | - | X | - | - |
| nH | number of Hydrogen atoms | - | - | - | X |
| nS | number of Sulfur atoms | - | - | X | - |
| nBT | number of bonds | - | - | - | X |
| nO | number of Oxygen atoms | X | - | X | X |
| nCl | number of Chlorine atoms | - | - | X | - |
| nX | number of halogen atoms | X | - | - | - |
| nBnZ | number of Benzene-like rings | X | - | X | - |
| Mv | mean atomic van der Waals volume | - | - | X | - |
| Mp | mean atomic polarizability | - | X | X | - |
| nAB | number of aromatic bonds | - | - | X | - |
| nF | number of Fluorine atoms | - | X | X | - |
| nR06 | number of 6-membered rings | - | - | X | - |
| nP | number of Phosphorous atoms | - | - | - | X |
| nDB | number of double bonds | - | - | - | X |
| nN | number of Nitrogen atoms | - | X | - | X |
| Ms | mean electrotopological state | - | X | - | X |
| Ss | sum of Kier-Hall electrotopological states | - | - | - | X |
| RBN | number of roatable bonds | - | X | - | X |
| Me | mean atomic Sanderson electronegativity | - | X | - | X |
| nSK | number of non-H bonds | - | - | - | X |
| nR03 | number of 3-membered rigs | - | - | - | X |
| Sv | sum of atomic van der Waals volumes | - | X | - | - |
| Se | sum of atomic Sanderson electronegativity | - | X | - | - |

### 2. Identification of the most suitable subset of descriptors using the GA-MLR approach:




Evolution of the fitness value for increasing number of generations

| subset | # generations | Subset of descriptors# | Dimension reduction (%) | p-value | MAC |
|---|---|---|---|---|---|
| 1 | 50 | MW, AMW, Sp, ARR, nCIR, RBF, nBnZ | 82.5 | 9.18E-9 | 0.26 |
| 2 | 100 | MW, nSK, nCIC, nH, nS | 87.5 | 7.19E-9 | 0.23 |
| 3 | 101 - 1100 | MW, nSK, nBT, ARR, nO, nCl | 85.0 | 9.19E-10 | 0.22 |
| **4** | **1101 - 2000** | **MW, nCIR, nO, nX, nBnZ** | **87.5** | **8.01E-10** | **0.14** |

# using a cutoff for *p-values* of 1.0E-08

- The optimal descriptor subset (#4) combines significant dimension reduction (87.5%) with the lowest *p-value* (8.01E-10) and MAC (0.14)

### 3. Assessment of feature subset quality by QSAR model development:

| Target class | Subset of features | Accuracy | Sensitivity | Specificity | False negative rate (FNR) |
|---|---|---|---|---|---|
| biodegradable | **GA-MLR** | **89.30** | **67.81** | **97.61** | **0.13** |
| | Stepwise-regression | 81.66 | 66.23 | 91.09 | 0.17 |
| | CFS | 73.18 | 52.22 | 87.20 | 0.28 |
| | ReliefF | 51.75 | 30.58 | 72.46 | 0.48 |
| persistent | **GA-MLR** | **94.01** | **98.61** | **86.95** | **0.06** |
| | Stepwise-regression | 92.45 | 98.65 | 83.78 | 0.08 |
| | CFS | 80.12 | 90.90 | 61.17 | 0.20 |
| | ReliefF | 62.05 | 38.88 | 80.86 | 0.34 |

## Conclusions

- A hybrid GA-MLR feature selection technique has been successfully developed and validated via QSAR development
- The proposed GA-MLR method selects reduced dimension subsets of independent features that can be subsequently used for the development of QSARs suitable for bioassay waiving within the context of REACH.

## References:

[1] Advances in the Replacement and Enhanced Replacement Method in QSAR and QSPR Theories, Mercader, G. A.; Duchowicz, R. P.; Fernández, M. F.; Castro, E. A., *J. Chem. Inf. Model*, 2010.

[2] Unsupervised feature selection using incremental least squares, Liu, R.; Rallo, R.; Cohen, Y., *International Journal of Information Technology and Decision Making,* 2011.

[3] Statistical Confidence for Variable Selection in QSAR Models via Monte Carlo Cross-Validation, Konovalov, A. D.; Sim, N.; Deconinck, E.; Heyden, V. Y.; Coomans, D, *J. Chem, Inf. Model*, 2008.

[4] E-DRAGON. *Dragon 5.4*; http://www.vcclab.org/lab/edragon/ (accessed, June 1, 2007).