Multi-objective evolutionary algorithm for Evaluation of Shape and Electrostatic Similarity

S. Puertas-Martín[1,a)], J. L. Redondo[1,b)], H. Pérez-Sánchez[2,c)] and P. M. Ortigosa[1,d)]

[1]*Supercomputing: Algorithm Research Group, Department of Computer Science, ceiA3, University of Almería, Almería, Spain*
[2]*Bioinformatics and High Performance Computing Research Group, Department of Computer Science, Universidad Católica San Antonio de Murcia, Murcia, Spain*

[a)]Corresponding author: savinspm@ual.es
[b)]jlredondo@ual.es
[c)]hperez@ucam.edu
[d)]ortigosa@ual.es

Abstract. Information in chemistry field is increasing each year implying new databases and more available information. Different techniques are adopted to learn how to manage that information. Ligand-Based Virtual Screening methods help in this process. It consists on processing a compound against large databases containing up to millions of chemical compounds evaluating one or more properties. After screening, compounds with the most similar descriptors to those of the target compound are chosen for in-vitro analysis. There exist a large number of molecular descriptors to compare molecules, and in literature, they usually analyze them individually. Regarding that, in this work in progress, we propose a multi-objective algorithm where two descriptors are considered, shape and electrostatic potential similarity. Using these two objective functions, the new algorithm aims to achieve an optimal Pareto-front that allows expert eye to select the most suitable compounds according to the properties of the target compound. Different techniques have been developed to assure fast performance keeping high-quality results. The new algorithm has been compared with different algorithms from the state-of-the-art to evaluate the quality of its results using well-known databases.

Introduction

The objective of Virtual Screening (VS) techniques is to predict which subset of the compound database will be more similar to a target compound without having to perform in-vitro analysis, which is more costly and time-consuming. Depending on the information available from the compound database, two kinds of screening can be performed, Ligand-Based Virtual Screening (LBVS) and Structure-Based Virtual Screening (SBVS).

SBVS methods require detailed information on the reference compound. On the contrary, LBVS techniques only need information about the active and inactive compounds. That information is known as descriptors. Many descriptors can be used to compare compounds, as Shape similarity, Electrostatic similarity, Atomic property fields, Aromatic potential, Desolvation potential, etc.

The main concept of virtual screening is to compare properties of compounds to pre-select them. In the best scenario, a target compound is compared against thousands (even millions) of compounds in a database, evaluating one or more descriptors. Some of these descriptors are related to molecule position in 3D space. Therefore, it can be said that these algorithms have to find the optimal alignment that gets the highest value from the descriptor. Also, algorithms must make an effort in translation and rotation techniques to lose as little resources as possible in such operations.

However, despite the number of descriptors available, many methods continue to use them independently [1, 2], which has a negative impact on the selection process. Different studies have shown that considering different descriptors simultaneously improves the selection process. Hence, multi-objective solutions, as the one shown in this paper, are of great interest in this field.

Optimization problem to solve

In this work in progress, a new multi-objective algorithm has been developed to optimize the shape and electrostatic similarity simultaneously.

The mathematical functions to obtain the first one objective function uses a Gaussian representation to model atoms [3, 4, 5]. However, for the sake of completeness, this main function is written in the following equation:

$$V_{AB}^{g} = \sum_{i \in A, j \in B} w_i w_j v_{ij}^{g} \tag{1}$$

where w_i and w_j are weights associated with the atoms i and j, respectively and v_{ij}^{g} is a product of Gaussian representations overlapping of two different molecule atoms:

$$v_{ij}^{g} = \int g_i(r) g_j(r) \boldsymbol{dr} = \int p e^{-(\frac{3p\pi^{1/2}}{4\sigma_i^3})^{2/3}(\boldsymbol{r}-\boldsymbol{r_i})^2} p e^{-(\frac{3p\pi^{1/2}}{4\sigma_j^3})^{2/3}(\boldsymbol{r}-\boldsymbol{r_j})^2} \boldsymbol{dr} \tag{2}$$

where p is a parameter that controls the softness of the Gaussian spheres, i.e., the height of the original Gauss function and σ is the radius of the atom. More precisely, the radius represents the well-known van der Waals radius [6].

On the other hand, the electrostatic function used belongs to Zap Toolkit by OpenEye [1, 7]. It calculates the electrostatic potential of a compound by a numerical solution of the Poisson equation [8]. Electrostatic potential between two compounds is obtained by the following function:

$$E_{AB} = \int \phi^{A}(r) \phi^{B}(r) \Theta^{A}(r) \Theta^{B}(r) \boldsymbol{dr} \approx h^3 \sum_{ijk} \phi_{ijk}^{A} \phi_{ijk}^{B} \Theta_{ijk}^{A} \Theta_{ijk}^{B} \tag{3}$$

Notice that the score obtained from Equation 1 and 3 depends on the number of atoms of the two compared molecules, i.e., the higher this number, the longer the value of similarity. Actually it lies in the interval $[0, +\inf)$. To be able to measure the grade of similarity between compounds, independently of the number of atoms that compose them, the Tanimoto Similarity [9] value is computed:

$$Tc = \frac{Score_{AB}}{Score_{AA} + Score_{BB} - Score_{AB}} \tag{4}$$

where $Score$ represents the value obtained from evaluating the pair of compounds AA, AB or BB with the objective functions 1 or 3.

The Multi-Objective Evolutionary Algorithm

In previous works, we have developed mono-objective evolutionary algorithms to provide solutions to this kind of problems where several techniques have been designed to reduce execution time while keeping good quality solutions. Some of these are the specific creation of the initial solutions, individual customized search space limits for each pair of molecular compounds and simulated annealing techniques to reduce the search space of the solutions in each iteration of the algorithm. This knowledge combined with techniques of multi-objective algorithms such as the establishment of Pareto-ranks, the determination of Non-dominated solutions and the use of distance metrics to get the estimation of the density of the solution during the algorithm execution has been included in a new multi-objective algorithm allowing to evaluate several objective functions simultaneously. In Algorithm 1 the structure of the implemented algorithm is shown.

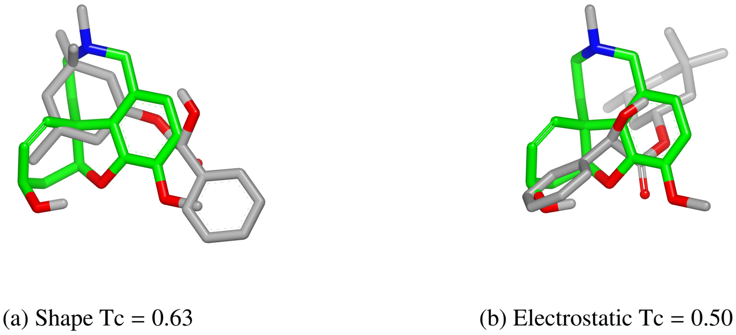
```
Data: Two molecular compounds and a configuration file
Define_specific_configuration;
Init_solution_lists;
while termination criteria are not satisfied do
    Create_new_solution(evals);
    if length(population_list) > $L_{max}$ then
        Select_solution(population_list);
    Improve_solution(population_list);
    Update_external_list;
    if length(external_list) > $L_{max}$ then
        Select_solution(external_list);
    Improve_solution(external_list);
if length(external_list) < $L_{max}$ then
    Compose_pareto;
```

Algorithm 1: Multi-objective algorithm.

Preliminary results and conclusions

The database used for computational experiments belongs to DrugBank [10]. These kinds of experiments require many computational resources so that the experimentation will be carried out on Bullion S8 equipment, specifically 8 Intel Xeon E7 8860v3 (16 cores each one), 2.3 TB of memory and 2x 300 GB of storage using SAS system.

An extended practice in LBVS when the shape and electrostatic similarity are used is to optimize one descriptor and later evaluate the other one. This approach can sometimes have errors by not considering the information from both descriptors simultaneously. An example of this problem is shown in Figure 1. Two compounds are compared, the compound reference DB00674 in green and the compound DB04838, in gray. In Figure 1-a the shape similarity has been optimized while in Figure 1-b was the electrostatic potential similarity. It can be seen that the solutions obtained are different from each other in both the position and value of the objective function.



(a) Shape Tc = 0.63

(b) Electrostatic Tc = 0.50

FIGURE 1: A comparison of the final result of shape and electrostatic similarity considering the compounds DB00674 (query compound, in green) and DB04838 (in gray).

On the other hand, Figure 2 shows a solution obtained by our multi-objective algorithm evaluating the compound DB09237 and DB01619. It shows how different solutions are distributed depending on the importance given to the functions. This result supports one of the main motivations of this work, which is none other than the importance of considering different features of the compounds simultaneously since depending on the final application, some may be more interesting than others.

In this work in progress, a bi-objective evolutionary algorithm will be presented. It is specifically designed to solve LBVS problems where rotations and translations of compounds are required including different methods related to the problem and its objective functions to optimize, shape and electrostatic similarity. A comprehensive computational study has been analyzed using well-known databases, and according to preliminary computational results, the performance of the new algorithm is promising.

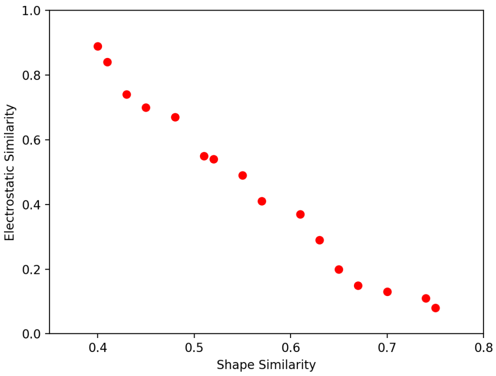


FIGURE 2: A Pareto-front solution evaluating the compounds DB09237 and DB01619.

ACKNOWLEDGMENTS

This work was funded by grants from the Spanish Ministry of Economy and Competitiveness (TIN2015-66680-C2-1-R, CTQ2017-87974-R), Junta de Andalucía (P12-TIC301), Fundación Séneca–Agencia de Ciencia y Tecnología de la Región de Murcia under Projects 19419/PI/14 and 18946/JLI/13.
Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA−CIEMAT), funded by the European Regional Development Fund (ERDF). CETA−CIEMAT belongs to CIEMAT and the Government of Spain. Juana López Redondo is a fellow of the Spanish 'Ramón y Cajal' contract program, co-financed by the European Social Fund. Savíns Puertas Martín is a fellow of the Spanish 'Formación de Profesorado Universitario' program, financed by the Spanish Ministry of Education, Culture and Sport.

REFERENCES

[1] J. Boström, J. A. Grant, O. Fjellström, A. Thelin, and D. Gustafsson, Journal of Medicinal Chemistry **56**, 3273–3280 (2013).

[2] C. A. Nicolaou and N. Brown, Drug Discovery Today: Technologies **10**, 427–435 (2013).

[3] J. A. Grant and B. T. Pickup, The Journal of Physical Chemistry **99**, 3503–3510 (1995).

[4] J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP, Journal of Computational Chemistry **17**, 1653–1666 (1996).

[5] X. Yan, J. Li, Z. Liu, M. Zheng, H. Ge, and J. Xu, Journal of Chemical Information and Modeling **53**, 1967–1978 (2013).

[6] A. Bondi, The Journal of Physical Chemistry **68**, 441–451 (1964).

[7] O. S. Software, (2018).

[8] C. Böttcher, Elsevier **1** (1973).

[9] P. Jaccard, Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 241–272 (1901).

[10] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, Nucleic Acids Research **34**, D668–D672 (2006).

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, IEEE Transactions on Evolutionary Computation **6**, 182–197 (2002).

[12] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization," in *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems* (International Center for Numerical Methods in Engineering, 2001), pp. 95–100.