

COMBINED INTERACTIVE AND AUTOMATED ADAPTIVE SEARCH FOR MOLECULAR DESIGN

Michael T.M Emmerich*, Johannes W. Kruisselbrink*, Eelke van der Horst[†], Ad P. IJzerman[†], Andreas Bender[†], Thomas Bäck*

*Natural Computing Group, LIACS, University of Leiden, 2333-CA Leiden, NL

[†]Leiden/Amsterdam Center for Drug Research, University of Leiden, 2333-CC Leiden, NL

Corresponding author: Michael T.M Emmerich (michael.emmerich@pharma-it.net)

ABSTRACT

This paper discusses an algorithm and software for the adaptive search in molecular designs which can be used as a tool in the process of drug discovery. The search strategy is based on evolutionary algorithms. The search in the complex space of molecular structures can be guided by interactive selection as well as by user defined target criteria. Computations and predictions of properties of the candidate molecules are used to compute the fitness values. The vector of fitness values serves as basis for the multi-objective comparison in the selection routine of the automated evolution loop. The new approach has been tested on a set of test problems - the inverse design of well-known drugs. First results are very promising and indicate the high potential of the approach.

1. INTRODUCTION

The field of *de novo* molecular design using adaptive search techniques has recently become a very active area of research. Adaptive search techniques are a very promising approach in this domain [1,2].

The Molecule Evaluator is a software tool for the interactive, *de novo* design of drug molecules. It supports the expert in pharmacology to explore the huge space of molecular structures. This exploration is done in a semi-automated way. This paper outlines this approach. The Molecule Evaluator™ is a very powerful design tool which can be used as an aid in the highly interactive development process.

The core technology of Molecule Evaluator™ is adaptive interactive evolution. Starting from an initial set of molecules the expert selects molecules which seem to exhibit the best fit for the desired purpose. Subsequently the Molecule Evaluator™ generates variations of this set using a set of problem specific mutation and recombination operators. The variations form a new set for the next interactive selection step. This process of interactive evolution is repeated until molecules are found that satisfy the needs of the experts.

Besides a 2D visualisation of the molecules, also a number of chemical and biological properties are presented such as molecular mass, number of rotatable bonds, hydrophilicity, and bio-activity. The chemist typically would like to keep these values within

certain bounds or close to a target value, while satisfying constraints on the structure that are usually difficult to express in terms of numbers. To satisfy this need, additional filters are implemented by which only molecules that comply to the filter bounds are generated. Figure 1 shows the selection window for an example.

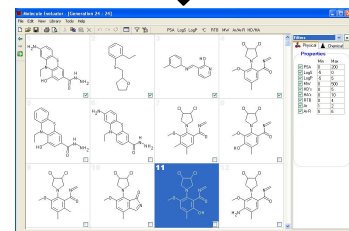
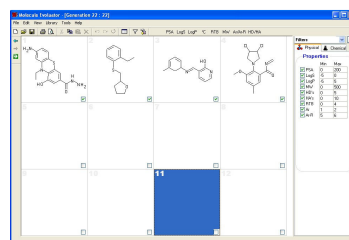


Figure 1: Interactive evolution: expert selects parents and new offspring is generated automatically.

Two major drawbacks of interactive evolution are that experts are biased and the number of molecules that can be evaluated is limited. In order to overcome these problems, we are researching and developing automated evolutionary algorithms that can be implemented in the Molecule Evaluator™. The intent is to offer a fully automated evolutionary algorithm as an extra option besides the interactive evolutionary algorithm. By combining the advantages of both, we create a very powerful design tool for molecular design.

We specifically focus on creating a method that generates diverse sets of molecules. By generating sets of diverse molecular structures we can better cope with the occasional generation of non-synthesizable or unsuitable molecular structures. By providing a broad set of candidates, the chemist can make the final decision about which structures are indeed useful or promising.

2. GENERAL FRAMEWORK

Besides the general framework, details of the approach are also of interest. The Molecule Evaluator™ can be viewed as an application of Evolutionary Algorithms in a non-standard combinatorial search space. Therefore a problem specific representation and variation operators for this space were designed. Essentially the search space can be viewed as a finite discrete space of graph like structures (the size of molecules is restricted to about 500 mw) with a complex non-regular neighborhood structure induced by the mutation operator.

The molecule representation which is used is the so-called TreeSmiles representation. The (Lisp-like) TreeSmiles notation is derived from the more strict SMILES notation commonly used by chemists. The major advantage of the TreeSmiles notation compared to other representations is that it is specifically designed to allow for an easy implementation of the genetic operators.

The Molecule Evaluator™ implements one recombination operator and eleven mutation operators.

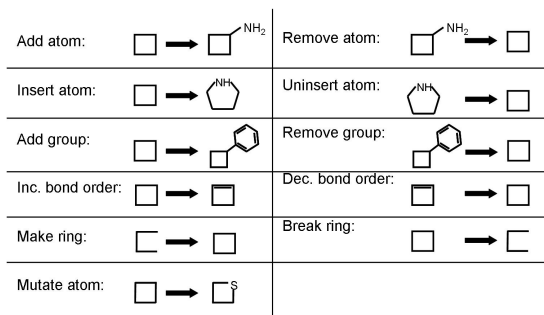


Figure 2: Mutation operators.

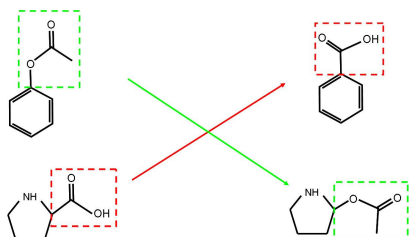


Figure 3: Recombination operator.

3. AUTOMATED EVOLUTIONARY SEARCH

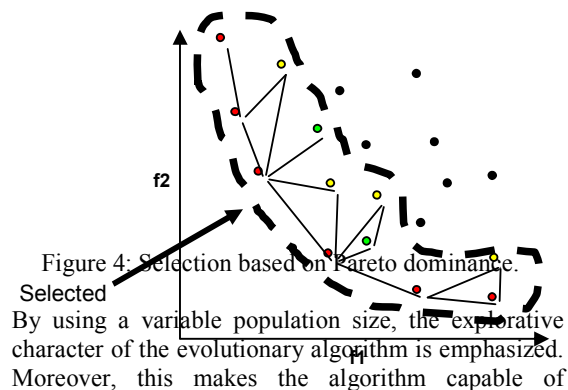
For the new, automated design of molecular structures, a multi-objective evolutionary algorithm is used that bases its fitness on the closeness of solutions with respect to a pre-defined set of very strict targets for a number of physicochemical properties.

The automated evolutionary algorithm uses the concept of Pareto dominance to introduce a pre-order on the set of candidate solutions and uses this as a selection criterion. The design of this algorithm is motivated by the base algorithm VV in Rudolph and Agapie [4], for which convergence to the efficient set has been proven, given appropriate variation operators. For the variation operators used in our study, however, only convergence to a set of locally efficient points in the operator-induced landscape is guaranteed.

The initial population of molecule structures consists of μ solutions that are randomly created by random recombination of fragments of known drugs from the database of the National Cancer Institute (NCI) in the USA.

In the evolution loop first λ ($=10$) solutions are generated. Molecules are either created by mutation or recombination. These operators operate on TreeSmiles encodings of the molecules and can be visualized using bond-graph notation (Figures 3 and 4). The candidate solutions are evaluated by means of different property computing/prediction algorithms, resulting in a nine dimensional vector for each solution.

The selection (Figure 4) draws all non-dominated solutions from the union of the parent and offspring population. Whenever the cardinality of this set is below a threshold μ_{\min} ($=40$) additional solutions are selected. The additional solutions are determined by non-dominated sorting, which results in subpopulations (partitions) of rank 1, 2, ... Partitions are added to the set of solutions until the minimal population size is exceeded. In order to keep the diversity of the solution set high, duplicates are eliminated from the selected population.



covering a large portion of non-dominated solutions without a need for an archive. The experimentally observed behaviour shows that the population size stays within a manageable range. This makes the selection strategy practical from a computational point of view, i.e. truncation of the population is not needed.

4. EXPERIMENTS

Experiments were based on the inverse design of six well known molecules from pharmacy. For each test-molecule a target was set to find five molecular structures with similar properties as the test-molecules. With a maximum of 1000 generations the algorithm found five solutions in 53 of the 60 test-runs. The running time of the algorithm varied between ten minutes for the simple molecular structures to one hour for more complex molecular structures. With this we show the applicability of the automated approach. The results of one test-run are displayed in Figure 5. Here the algorithm found five molecules with properties that were similar to the drug Zolofit.

5. CONCLUSIONS AND OUTLOOK

Extending the Molecule Evaluator™ with this automated evolutionary algorithm makes it applicable for a broad range of purposes. The chemist can exploit the advantages of both interactive evolution and automated evolution, and choose the most suitable method for the problem at hand.

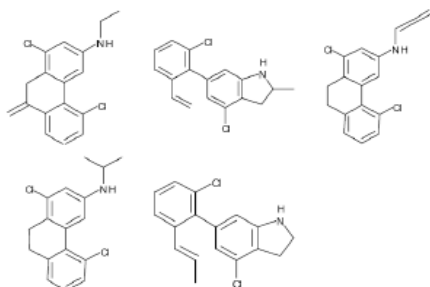


Figure 5: Results of a test-run for the design of molecular structures with the same properties as Zolofit.

Future research will focus on niching techniques, landscape analysis, molecule stability, molecule synthesizability, and applications in real-world problems.

REFERENCES

- [1] Lameijer E.-W., Kok J.N., Bäck T., and IJzerman A.P., The Molecule Evaluator. An interactive evolutionary algorithm for the design of drug-like molecules, *Journal of Chemical Information and Modeling*, 46(2):545 – 552, 2006

- [2] Sedwell, A.N.; Parmee, I .C.:Techniques for the design of molecules and combinatorial chemical libraries, In. *Proc. IEEE Congress on Evolutionary Computation*, Singapore (CEC2007), IEEE Press, Piscataway, NY, Page(s): 2435-2442. 2007
- [3] Kruisselbrink J.W., Bäck T., IJzerman A.P., and van der Horst E., *Evolutionary Algorithms for automated drug design towards target molecule properties*. GECCO 2008, Atlanta, GA (*in print*).
- [4] G. Rudolph and A. Agapie: [Convergence Properties of Some Multi-Objective Evolutionary Algorithms](#), pp. 1010-1016 in A. Zalzala et al. (eds.): *Proceedings of the 2000 Congress on Evolutionary Computation (CEC 2000)*, Vol. 2, IEEE Press, Piscataway (NJ) 2000.