Evolution of Molecules Using Genetic Operators on Reduced Molecules

Vladimir Sykora, Damjan Krstajic, Robert Leahy, David Leahy

Newcastle University

Introduction

New programming concepts have been applied in this work to the development of powerful and flexible genetic operators. The operators were designed to operate on reduced molecules. These genetic operators have been applied in a genetic algorithm with the objective of generating novel chemical scaffolds for HIV inhibition.

Context Molecule

Context molecules are useful data structures to characterise fragments of molecules created after bond disconnections. With them, we can trace back fragments that were generated by disconnecting bonds, and compute the whole molecule that originated it, without the need of using atom-indexing information.
A *context molecule* is a chemical graph [1] where nodes represent atoms and edges represent bonds. In addition to atomic information, the nodes in the *context molecule* contain context information that describes any former connection.
The context information of an atom that had a connection includes information about itself, and about the atom formerly connected to it.

Figure 1 shows an example of the resulting context molecules after a bond of a molecule has been broken.

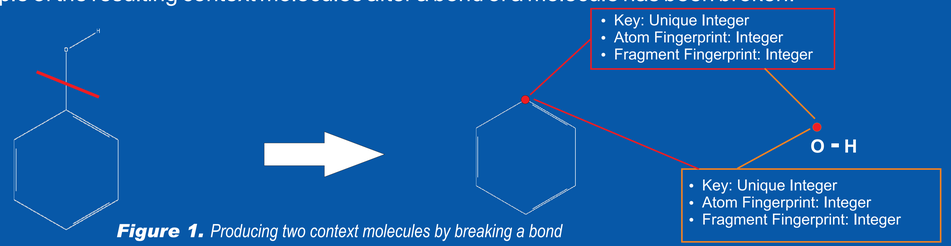


Figure 1. *Producing two context molecules by breaking a bond*

Reduced Molecule

A reduced molecule is a specialisation of a chemical graph, in which nodes contain a single *context molecule*. The *context molecules* contained in the nodes can in turn take any form, being whole fragments or single atoms.
A *reduced molecule* is created from a chemical graph using a set of fragments given by the user in the form of SMARTS [2] patterns. Each of these fragments is searched for in the initial chemical graph. A sub-group of fragments is selected from the found fragments with the objective of creating a set of fragments that are completely independent (no overlapping) and cover the majority of the initial chemical graph. A node in the resulting *reduced molecule* is created for each of the selected fragments, and for atoms that did not belong to any of the selected fragments. Two nodes in the *reduced molecule* are adjacent if the fragments or atoms were connected in the original chemical graph.
Context information is generated for atoms in the *reduced molecule* fragments which were connected to atoms in the original chemical graph that now belong to a different fragment.

Figure 2 shows the process of generating a reduced molecule from a simple chemical graph.

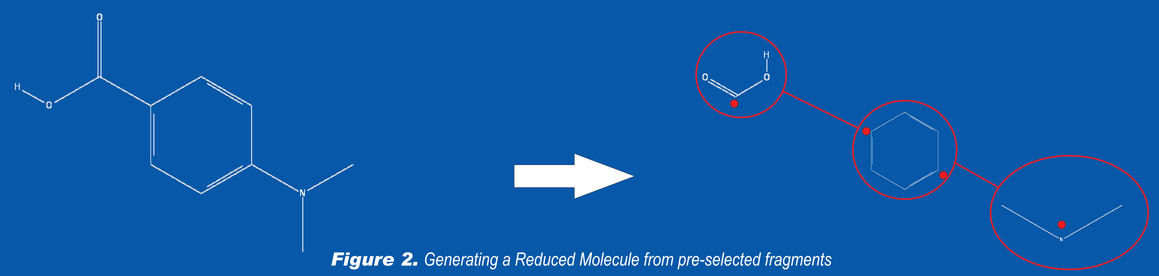


Figure 2. *Generating a Reduced Molecule from pre-selected fragments*

References

[1] Ivanciuc, O. Handbook of Cheminformatics. 13. Graph Theory In Chemistry. Gasteiger, J.; Engel, T. Editors. Wiley-VCH. 2003.
[2] Daylight Chemical Information Systems Inc. Theory Manuals.
[3] Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. “A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules”. J. Chem. Inf. Comp. Sci. , 44 (3), 1079 -1087, 2004.
[4]Schneider, G., Neidhart, W., Giller, T., Schmid G., "'Scaffold-Hopping' by Topological Pharmacophore Search: A Contribution to Virtual Screening". Communications of Angew. Chem. Int. Ed. 1999,38,No.19.
[5] Thomson Scientific. http://scientific.thomson.com/products/wdi/

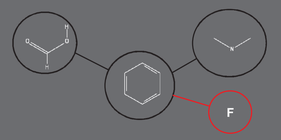
Genetic Operators

The genetic operators are the fundamental processes in genetic algorithms. These produce the variation in the genetic material necessary in the process of evolution. The two fundamental genetic operators are mutation and crossover. Mutation modifies the configuration of the chromosome, and the crossover combines the features of two chromosomes to generate two new single chromosomes.
In this work, genetic operators were developed to operate on reduced molecules. For that reason, our chromosome representation is a reduced molecule. This strategy is similar to the one presented by Brown [3].
These genetic operators are used in a genetic algorithm in order to evolve reduced molecules with desired characteristics.
The following operators were developed:

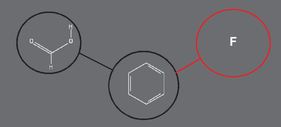
Mutation

The mutation operators are unary operators that transform the structure of the reduced molecule by modifying the configuration of the set of nodes, either by altering the contents or the topological configuration of nodes:

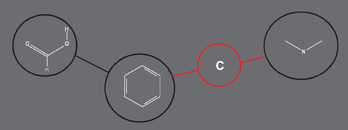
Append Adds a new vertex to the *reduced molecule* and connects it to an existing vertex



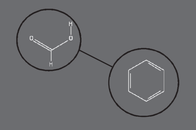
Replace Substitutes an existing vertex of the operand *reduced molecule* with a new one



Insert Selects an edge of the operand, breaks it, inserts a new node inbetween, reconnects



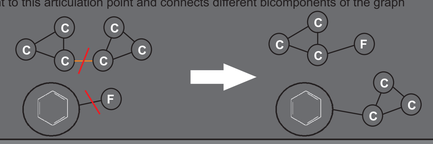
Delete Selects a node, deletes it, and reconnects the hanging edges



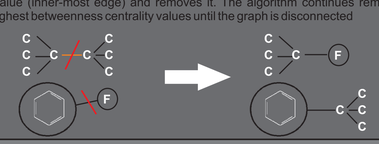
Crossover

Crossover is a binary operator that takes two parent reduced molecules, and produce two new reduced molecules that are combination of the original two. The process starts with the generation of two disconnected subgraphs for each of the initial reduced molecules. Following, the subgraphs are exchanged between parents, and then reconnected.

Articulation Points Selects an articulation point, and then removes an edge that is adjacent to this articulation point and connects different bicomponents of the graph



Betweenness Centrality Selects the edge with the highest betweenness centrality value (inner-most edge) and removes it. The algorithm continues removing edges of the following highest betweenness centrality values until the graph is disconnected



Results

Our genetic operators were applied in a genetic algorithm with the objective to evolve molecules with a pharmacophore distribution similar to one present in a potent HIV protease inhibitor. The pharmacophores were encoded using the CATS representation of Schneider [4]. This abstraction gives the flexibility of “scaffold hopping”, i.e. the identification of molecules with similar biological activity but with significantly different molecular backbones.The fitness function used in the genetic algorithm is the Tanimoto distance from the topological pharmacophores vector of the evolved structures to the pharmacophores of the template HIV inhibitor.
To generate the reduced molecule representations, we used a set of 550 SMARTS selected from fragments contained in the Sigma-Aldrich catalog of organic reagents.The initial population was a random set of 500 molecules obtained from the World Drug Index[5].
The genetic algorithm was controlled to have a minimum of 200 and a maximum of 1000 chromosomes per generation, and run until generation 50.

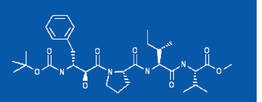


Figure 3. *HIV protease inhibitor template*

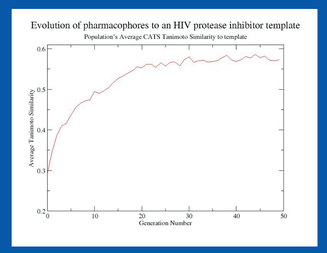


Figure 4. *Population Average Similarity*

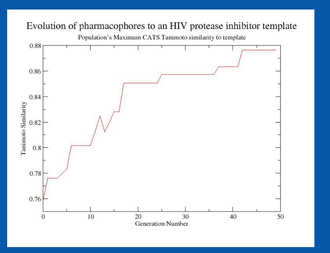


Figure 5. *Population Maximum Similarity*

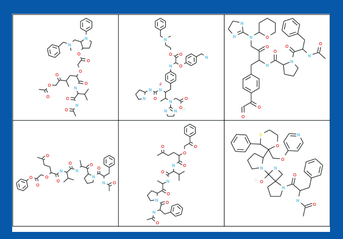


Figure 6. *Best structures in last population*

The Chemical Descriptors Library CDL: **http://cdelib.sf.net**