Computers and Chemical Engineering 83 (2015) 267–277

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Evolutionary algorithm for de novo molecular design with multi-dimensional constraints

Robert H. Herring III, Mario R. Eden*

Department of Chemical Engineering, Auburn University, Auburn 36849, USA

ARTICLE INFO

Article history:
Received 17 October 2014
Received in revised form 23 June 2015
Accepted 27 June 2015
Available online 4 July 2015

Keywords:
Molecular design
Genetic algorithm
Descriptors

ABSTRACT

An evolutionary approach for solving molecular design problems with descriptors of varying dimensionality has been developed. Spatial fragment based descriptors are employed to generate candidate solutions within a population, which is evolved through the application of genetic operators toward an improved fitness. The candidate molecules are represented as graphs, and as such, customized operators of crossover and mutation have been developed to be compatible with this representation. The search space is conveniently represented through limitations on the occurrence of each fragment, as defined by the chosen data set, and the spatial capabilities of this space are captured through an initial conformational analysis. This spatial information is compressed and utilized to generate conformational space estimations throughout the algorithm, which expedites the search for solution graphs. The effect of various user determined input parameters is considered and exemplified through a case study involving the identification of solvents falling within a desired boiling point range, as estimated by a multi-dimensional property model.

1. Introduction

The area of computer-aided molecular design (CAMD) has greatly influenced the rate and cost at which novel chemicals with desired attributes have been identified. As such, great effort has been invested in new methodologies which allow for the solution of larger and more complex problems of this nature. The application of genetic algorithms (GAs) is one such technique which has shown early promise in the solution of large combinatorial, and highly non-linear molecular design problems (Venkatasubramanian et al., 1995). While there are many established stochastic methodologies capable of solving such problems, studies identifying an optimal approach for such problems have most often proven inconclusive based on the varying nature of individual problems and search spaces (Amaran et al., 2014). It is known, however, that GA's are simple to implement in parallel computing environments, which provides motivation for the development of such an approach that could be scaled up based on the problem size. Genetic algorithms (Holland, 1975) are a subclass of evolutionary algorithms, which mimic the process of natural selection, that encode the characteristics of an individual, in this case a potential candidate molecule, within a chromosome. The typical approach begins with a population of randomly generated individuals and the more fit members, i.e. those closest to possessing the desired characteristics, are selected to undergo computational analogs of natural recombination and mutation. This process is iterative until the resultant population possesses the desired attributes. The nonlinearities under consideration within this paper arise from the utilization of complex molecular descriptors, which cannot be linearly mapped into chemical space. Additionally, the combinatorial issues stem from the use of molecular fragments as building blocks for developing potential solutions.

Kawai et al. (2014) realized the importance of utilizing a fragment based approach for de novo molecular construction from previous studies in which infeasible molecular structures were regularly created through atom based mutations. Their evolutionary approach to searching for novel chemical structures began with a desirable initial structure, which was evolved by means of mutation and crossover operators while diversity was supplied by a fragment library created from molecules related to the target of interest. This study introduces the challenge of creating chemically feasible molecules, which were in this case verified though a measure of chemical similarity (e.g. Tanimoto coefficient). One limitation however is that the molecular fitness and evolution of structures were guided through this same measure of chemical similarity. It would be beneficial to introduce flexibility in terms of how molecular solutions are ranked, such as the ability to utilize characterization of any dimensionality. Such

* Corresponding author. Tel. +1 3348442064.
E-mail address: edenmar@auburn.edu (M.R. Eden).

http://dx.doi.org/10.1016/j.compchemeng.2015.06.012
0098-1354/

an approach has proven successful in providing measures of similarity between biologically active molecules (Nettles et al., 2006) as well as chemicals (Tseng et al., 2012). Accessibility and ease of use associated with modern molecular modeling software is allowing the utilization of spatial molecular characterization in a variety of fields. Additionally, combination of these higher order descriptors with the conventional topological and constitutional descriptors of the past is now much more feasible with the application of powerful variable selection techniques (Nicholls et al., 2004).

One of the difficulties arising from the inclusion of spatial descriptors within a property model is that the conformational space, or potential energy surface, of each possible solution must be explored in order to estimate these spatial descriptors. This can become quite a task considering that the degrees of freedom associated with this space is $3N-6$, with N being the number of atoms in the molecule. Additionally, a molecule will most often have multiple accessible conformations, each of which can correspond to a local energy minimum on the potential energy surface, instead of a single conformation. A combination of these factors makes the consideration of one conformer, often times built from common bond lengths and angles, infeasible for the derivation of spatial descriptors as it is not representative of the set of accessible conformations often seen in bulk chemicals. One recent successful de novo molecular design study (Katarkar et al., 2015) was able to consider the structural, or two-dimensional (2D), and spatial, or three-dimensional (3D), properties in an evolutionary approach to discover inhibitors of the Hsp-47 heat shock protein. This study is based upon a previously introduced methodology (Douguet et al., 2005) that uses 3D molecular fragments to generate candidate molecules, which undergo the genetic operators of mutation and crossover to evolve toward an improved fitness. Each candidate molecule must undergo a conformational search in which several conformers are identified through an energy minimization. This step can become computationally intensive, especially for large search spaces, and an approach to estimate the conformational capabilities of candidate molecules within the chemical search space is introduced within the presented methodology. The expedited development of spatial properties is accompanied by a potential decrease in the quality of geometric information. This is the trade off in sacrificing a thorough conformational search for a more computationally efficient method of spatial development. The feasibility of such an approach lies in the sensitivity of the estimated properties/attributes with respect to spatial parameters, which may often fall within the error associated with the developed property model. The research gap filled by this methodology resides in the inability of current molecular design approaches to solve problems with multi-dimensional descriptors while considering the conformational space of potential solutions without an extensive conformational analysis performed through molecular simulations. The presented technique is able to estimate the conformational space 'on the fly' for candidate structures, along with facilitating the calculation of topological descriptors through its graph based representation.

The approach presented in this paper utilizes a fragment based descriptor, which is represented as a molecular graph, as building blocks to generate candidate solutions. These building blocks are generated from the set of molecules utilized in creation of the chosen property models such that the resultant region of chemical space searched has an increased likelihood of falling within the applicability domain of these models. The spatial characteristics of this space are captured through an extensive conformational analysis, after which the information generated is compressed to minimize redundancy. This compressed geometric information is then utilized throughout the algorithm to estimate the accessible potential energy surface for the candidate structures considered. Genetic operators of mutation and recombination have been developed to act on this graph based representation of molecules, ultimately guiding the population of candidate structures toward an improved fitness. The objective of this paper is to introduce this methodology and elucidate the effect of some user defined parameters. These effects are exemplified through a case study in which molecules are designed to fall within a certain boiling point range. The authors were invited to submit this paper as an extended version of a paper published in the proceedings of the 24th European Symposium on Computer-Aided Process Engineering (ESCAPE-24) (Herring and Eden, 2014).

2. Methodology

The methodology described herein will begin by extending the concept of signature descriptors to include spatial information. The development of this information through a conformational analysis, along with the subsequent compression of data, will also be discussed. This conformational analysis is necessary to develop spatial information about the molecular fragments, which is useful in estimating the subsequent whole molecule geometric information of candidate solutions. This data is compressed because much of the spatial information is overlapping and would cause combinatorial explosion if left untreated. Once the creation of spatial atomic signatures is introduced, the utilization of these fragments in reproducing molecular geometries representative of a more thorough conformational analysis will be considered with an example illustrating the accuracy and limitations of such a technique. Following this, the genetic algorithm designed to handle this graph based representation of potential solutions will be considered along with details for generation of a starting population, formulation/implementation of genetic operators, and utilization of fitness functions. The novelty of such an approach is that the genetic operators have been design to maintain and perpetuate spatial information within the candidate molecules created. Finally, the effect of some user defined variables will be discussed to help elucidate the decisions for these criteria. The generalized methodology can be visualized in Fig. 1.

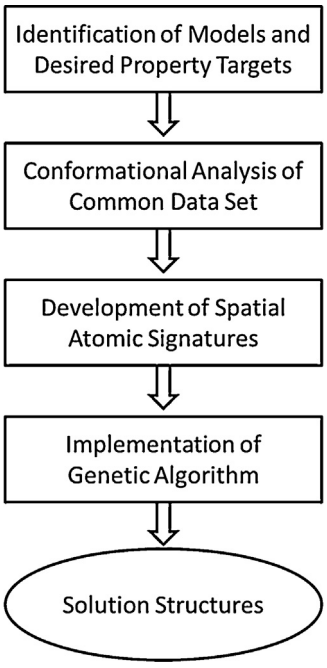

Fig. 1. Overview of CAMD methodology developed.