

# Optimal Graph Design Using A Knowledge-driven Multi-objective Evolutionary Graph Algorithm

Christos A. Nicolaou, Christos Kannas, and Constantinos S. Pattichis, *Member, IEEE*

**Abstract**—Designing appropriate graphs is a problem frequently occurring in several common applications ranging from designing communication and transportation networks to discovering new drugs. More often than not the graphs to be designed need to satisfy multiple, sometimes conflicting, objectives e.g. total length, cost, complexity or other shape and property limitations. In this paper we present our approach to solving the multi-objective graph design problem and obtaining a set of multiple equivalent compromising solutions. Our method uses multi-objective evolutionary graphs, a graph-specific meta-heuristic optimization method that combines evolutionary algorithms with graph theory and local search techniques exploiting domain-specific knowledge. In the experimental section we present results obtained for the problem of designing molecules satisfying multiple pharmaceutically relevant objectives. The results suggest that the proposed method can provide a variety of valid solutions.

## I. INTRODUCTION

OPTIMAL Graph Design (OGD) deals with the discovery of graph structures that satisfy certain predefined criteria while conforming to a set of specifications concerning the building blocks and the rules available for the construction of the graph. The problem is characterized by its combinatorial nature and the potentially large size of the search space defined by the number of feasible graph structures. As is the case in most real life problems, OGD is typically multi-objective, i.e. the graphs need to satisfy more than one criterion and thus, it is essentially a multi-objective combinatorial optimization problem. OGD includes a wide variety of problems including communication network design [1], route planning [2] and molecular design [3] among others.

The aim of this paper is to introduce a newer development of our recently proposed multi-objective optimization algorithm specifically designed to evolve graphs using a pool of existing building blocks and exploiting problem

domain specific knowledge. The implementation of the algorithm, as well as the results presented, focus on the specific problem of molecular/drug design, also known as de novo drug design (DND). The rest of the paper is organized as follows. The next two sections of the paper briefly describe fundamental graph theory elements and their application to molecular graphs and, multi-objective optimization. Section IV of the paper introduces the proposed algorithm while section V describes the general DND problem and presents some experimental results from the application of the algorithm. The final section summarizes our conclusions and outlines some directions for future research.

## II. BACKGROUND

A graph  $G = (V, E)$  consists of a set of vertices  $V(G)$  and a set of edges  $E(G)$ . In the case of labeled graphs both vertices and edges have identifiers, i.e. each vertex and edge has a label drawn from a predefined set of vertex and edge labels. Graphs can be directed or undirected. In directed graphs edges are ordered pairs of the vertices they connect where, in undirected, edges simply list the pair of vertices they connect. Two vertices  $V_I, V_J$  of graph  $G$  are connected, or adjacent, if there is an edge  $E_{IJ} = (V_I, V_J) \in E(G)$ . If there is a path  $P = (E_1, E_2, \dots, E_n)$  between every pair of vertices in a graph  $G$ , then  $G$  is a connected graph. A graph  $S = (V_S, E_S)$  is a subgraph of  $G = (V, E)$  if and only if  $V_S \subseteq V$  and  $E_S \subseteq E$ . If  $E_S$  contains all edges in  $E$  connecting the vertices in  $V_S$  then  $S$  is an induced subgraph of  $G$ . A common induced subgraph between  $G_1$  and  $G_2$  is a graph  $CS$  that is an induced subgraph of both  $G_1$  and  $G_2$ . The largest induced subgraph between  $G_1$  and  $G_2$  is known as the Maximum Common Induced Subgraph (MCIS). The largest contiguous common substructure is known as the Maximum Common Substructure (MCS).

Chemical structures are typically represented as labeled, undirected graphs where atoms correspond to vertices and chemical bonds are

Manuscript received August 17, 2009.

C. A. Nicolaou is with the Cyprus Institute, Nicosia, Cyprus, the University of Cyprus, Nicosia, Cyprus and Noesis Chemoinformatics, Nicosia, Cyprus. (phone: +357-22208644; fax: +357-22208625; e-mail: c.nicolaou@cyi.ac.cy).

C. Kannas, is with Noesis Chemoinformatics and the University of Cyprus, Nicosia, Cyprus (e-mail: ckannas@noesisinformatics.com).

C. S. Pattichis is with the Computer Science Department, University of Cyprus, Nicosia, Cyprus (e-mail: pattichi@ucy.ac.cy).

represented by edges. In this context, molecular fragments, or substructures, are induced subgraphs of molecular graphs. Scaffolds are molecular fragments defined in association to well-defined sets of compounds. A scaffold derived from a compound set is a substructure characteristic of the compounds in the set. Often, a compound set scaffold is thought of as the MCS of that set. More informally scaffolds can be thought of as the common, distinguishing “core” of a compound set. Privileged substructures are scaffolds positively correlated with favorable behavior, i.e. scaffolds present preferentially in compounds with a desired biological profile. For a more thorough review of chemical substructure mining see [4].

### III. MULTI-OBJECTIVE OPTIMIZATION

A large class of optimization problems need to simultaneously achieve multiple, often conflicting objectives. In contrast to single-objective problems where optimization methods explore the feasible search space to find the single best solution, in multi-objective settings no best solution can be found that outperforms all others in every criterion [5]. Instead, multiple “best” solutions exist representing the range of possible compromises of the objectives [1]. These solutions, known as non-dominated, have no other solutions that are better than them in all of the objectives considered. The set of non-dominated solutions is also known as the Pareto-front or the tradeoff surface. Fig. 1 illustrates the concept of non-dominated solutions and the Pareto-front in a bi-objective problem.

Problems that require the accommodation of multiple objectives are widely known as multi-objective problems (MOP) or “vector” optimization problems [6]. Traditionally MOPs have been simplified, either by ignoring all objectives but one, or, by aggregating them. Multi-objective optimization (MOOP) methods follow a different approach that is founded on compromises and tradeoffs among the various objectives to be met. The aim of MOOP methods is to discover a set of satisfactory compromise solutions and, through them, the globally optimal solution(s) by optimizing numerous dependent properties [1]. A major benefit of MOOP methods is that local optima

corresponding to one objective can be avoided by consideration of all the objectives simultaneously, thereby escaping single objective dead-ends.

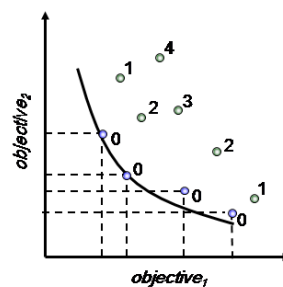


Fig. 1. A MOP with two minimization objectives and a set of solutions (circles). Non-dominated solutions are labeled ‘0’.

In recent years, Evolutionary Algorithms (EAs) have been used extensively for MOPs with several multi-objective optimization EAs (MOEA) cited in the literature [3], [7], [8]. MOEAs are particularly attractive since their population-based approach enables the simultaneous search of multiple search space regions and thus the identification of numerous Pareto-solutions in a single run. Additionally, since EAs impose no constraints on the morphology of the search space, they are suitable for complex, multimodal search spaces with various local optima such as the ones typically found in MOPs [9].

### IV. METHOD

Recently, we proposed the Multi-objective Evolutionary Graph Algorithm (MEGA) that combines evolutionary techniques with graph data structures to directly manipulate graphs and perform a global search for promising solutions [3]. MEGA has been designed to enable the use of problem-specific knowledge and local search techniques, to improve performance and scalability. This section briefly describes the algorithm and elaborates on some new features implemented. A more detailed description is found in [3].

MEGA operates on two population sets, the normal, working population and the secondary population or the Pareto-archive. The former population consists of the individuals subjected to objective performance calculation and obtained through evolution in a single iteration. The latter supports a form of elitism aimed at preserving promising solutions found throughout

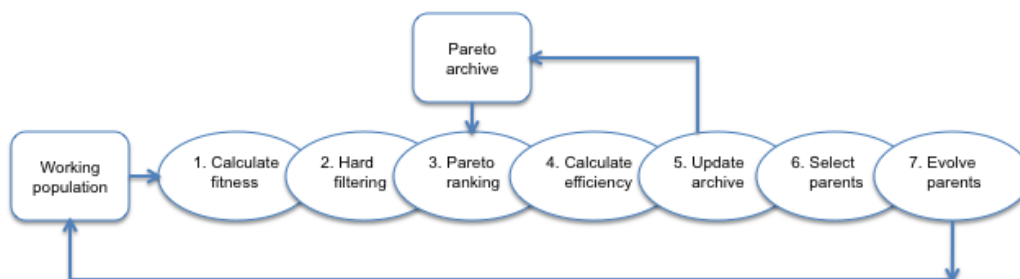


Fig. 2: A diagram of the MEGA algorithm

evolution and, ensuring that the final Pareto-approximation will contain the best solutions found.

MEGA requires the supply of a set of molecular building blocks, the implemented objectives to be used for scoring the graphs and a set of attributes controlling mutation and crossover methods and probabilities, parent selection method, hard filters for solution elimination, etc. Optionally, a set of graphs to be used as the initial working population may be supplied as well. The supplied data is used to create graph-based chromosomes, to construct a list of building block objects and to initiate additional internal data structures required for algorithm execution. At this stage the external archive of solutions intended to store the secondary population is also created.

The first phase of the algorithm applies the objectives on the working population to obtain a list of scores for each individual. The list of scores may be used for the elimination of solutions with values outside the range allowed by the corresponding active hard filters. In the next step, the two populations, working and secondary, are merged and the individuals' list of scores is subjected to a Pareto-ranking procedure to set the rank of each individual. The combined population forms the new working population. The algorithm then proceeds to calculate an efficiency score for each individual using a novel methodology that operates both in parameter and objective space. The methodology employs an elaborate niching mechanism that performs diversity analysis of the population, based on the genotype, i.e. the chromosome graph structure, and subsequently assigns an efficiency score that takes into account both the Pareto-rank and the diversity analysis [3]. The efficiency score calculation technique can be fine tuned by user supplied parameters to favor the parameter space, the objective space or to

balance between the two. Traditional methods for efficiency calculation operating exclusively on the phenotype, i.e. the objective space, have also been implemented for performance assessment comparisons. The efficiency score of each individual is then used to update the Pareto-archive. The current Pareto-archive is replaced with a subset of the working population that favors individuals with high efficiency score, i.e. low domination rank and high chromosome graph diversity. Note that the size of the subset selected is limited by a user-supplied parameter.

Following the update of the Pareto-archive MEGA checks for the termination conditions; if satisfied the process terminates. However, if this is not the case the process moves to select the parent subset population from the combined population set using the "roulette" method [1] on the efficiency scores of the candidate solutions. The parents are then subjected to mutation and crossover according to the probabilities indicated by the user. The new working population is formed by merging the original working population and the newly produced mutants and crossover children. The process then iterates as shown in Fig. 2.

MEGA incorporates heuristics to enable the exploitation of existing problem specific knowledge. The heuristics involve the usage of the weights associated with the building blocks provided and result in favoring those with an increased weight. Additionally, the progress of the Pareto-front approximation is monitored to self-adapt certain attributes controlling the optimization process. Specifically, MEGA identifies underrepresented regions in the Pareto-front and adjusts the building blocks weights favoring the generation of solutions from certain regions of the space.

## V. APPLICATION

The validation of MEGA was performed

through application on the multi-objective DND problem. In the next sections DND is described and details about the application are provided.

### A. *De Novo Design*

De-novo drug (or ligand) design is an attempt to generate ligands from scratch based only on information about the pharmaceutical target site or known ligands [3]. Effectively, DND methods face the task of exploring a chemical search space estimated to be in the order of  $10^{60}$  [10]. Such space cannot possibly be fully enumerated and so powerful search methods need to be applied to detect the best possible solutions in a limited amount of time.

DND algorithms proposed in the literature typically use an evolutionary algorithm related technique for searching and a set of molecular fragments as building blocks. The use of a predefined collection of molecular fragments is sometimes combined with the identification of reaction points and synthetic rules that, when used, increase the synthetic feasibility potential of the designed chemical structures. Most methods are designed to accommodate a single objective either predicted binding affinity to a known protein target or similarity to a known ligand. MOOP-based methodologies are limited to [3] and [11] although the need for their greater adoption is gaining support within the drug discovery community [5] [9].

### B. *Experimental Design*

The experiments performed involved the design of selective Estrogen Receptors (ER), i.e. ligands that bind to ER- $\beta$  and not ER- $\alpha$ . All runs were performed on a computer equipped with an Intel Core 2 Duo 3 GHz processor and 2GB of memory. An account of the tests performed and the results obtained follows.

**Datasets.** Two datasets were used during the MEGA tests performed. Dataset 1, a set of well-known ER ligands, contains 7 compounds, 5 with increased selectivity to ER- $\beta$  and 2 with selectivity to ER- $\alpha$ . Dataset 2 is an ER-inhibitor dataset obtained from Pubchem [12]. The dataset consists of 86098 compounds tested on both ER- $\alpha$  (Bioassay 629) and ER- $\beta$  (Bioassay 633).

**Building Blocks.** A single collection of 51123 building blocks was used for all the tests performed. The building blocks were obtained via fragmentation of Dataset 2 described above

with the substructure mining tool provided by [13]. Building block weights were assigned according to the activity labels of the molecules containing it.

### C. *Objectives*

The application of the algorithm relied on the encoding of two ligand-based objectives that measured the average similarity of a query molecule to known ligands. Similarity was calculated using the tool Fuzzee [14]. The specific method used operates on abstractions of molecular graphs that replace atoms with molecular features to produce the so-called feature graphs. The actual similarity is calculated in a pair-wise manner by first aligning the feature graphs of two molecules, identifying common features and then applying the Tanimoto similarity measure [15]. The two objectives encoded measured similarity of a given query molecule to a set of ER- $\alpha$  selective and ER- $\beta$  selective ligands. The experiments aimed at designing molecules selective to ER- $\beta$  and so the algorithm was set so as to maximize average similarity to the ER- $\beta$  ligand set and minimize the average similarity to the ER- $\alpha$  ligand set. A set of hard filters based on chemical structure objectives was also applied in order to remove potentially problematic designs from further consideration.

### D. *Algorithmic Settings*

The experiments performed aimed to measure the performance of the MEGA, MOGA and SPEA algorithms on the selectivity problem described previously. Tests using MOGA and SPEA served to assess the performance of MEGA in comparison to commonly used algorithms from the MOOP field. MOGA [8] was selected since it is one of the earliest, most commonly cited algorithms in the MOOP field. SPEA [7] is a more recent method that popularized the use of the Pareto-archive in MOEA algorithms. The experimental settings used population sizes 25, 50 and 100 and 100 generations. For each combination of input parameter settings 5 runs were performed, using different initial population sets. In each test case the initial population was selected from a user-defined data set. Runs were performed with the roulette parent selection mechanism using both

mutation and crossover. Mutation probability was set at 0.25 and crossover at 1.0. In the case of MEGA and SPEA the maximum Pareto-archive set was set to 1000. Performance assessment of the results from each run took place through a post-processing step that included the calculation of the Pareto-approximation set hypervolume [7] and the chromosome/structural diversity. The latter was calculated by averaging the Euclidean distances of each solution to all other solutions in the proposed set, using atom-pair descriptors [16] of the molecules involved. MEGA niching was set to balance between the diversity in parameter and objective space.

### E. Results

Experimental results indicate that MEGA can generate solution sets consisting of structurally diverse solutions characterized by increased similarity to ER- $\beta$  ligands over ER- $\alpha$  ligands. Throughout our tests MEGA solution sets compare favorably with those obtained using MOGA and SPEA.

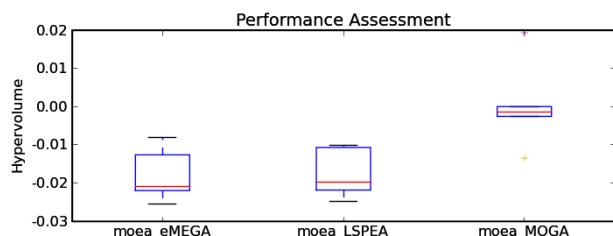


Fig. 3: Hypervolume performance for MEGA (eMEGA), SPEA (LSPEA), and MOGA for population 25. Low values correspond to better performance.

Fig. 3 presents the hypervolume measure box plots for MEGA, MOGA and SPEA. Pareto-approximations produced by MEGA and SPEA consistently perform better than those of MOGA. A similar trend has been observed for all population sizes examined.

Fig. 4 presents the results of the diversity calculations for the three algorithms both in parameter and objective space. It is clear from the plot that MEGA, with its unique niching mechanism specifically designed to promote structural diversity, can produce solution sets with increased diversity both in parameter and objective space. The sets proposed by MEGA and SPEA, while of comparable performance as measured by hypervolume differ in both diversity measures where MEGA is clearly more successful.

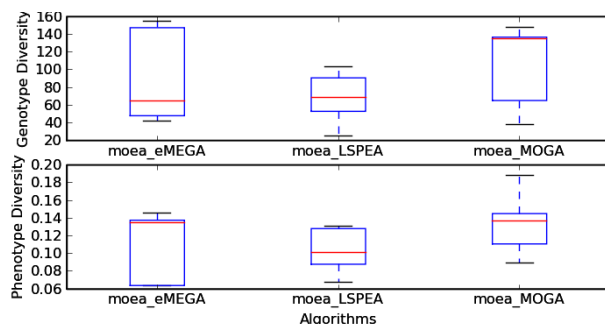


Fig. 4: Comparison of the diversity of the Pareto-approximations produced by MEGA, SPEA and MOGA in parameter (genotype) space and in objective (phenotype) space. Higher values correspond to greater diversity.

Overall, the results of MEGA demonstrate the ability of the algorithm to explore the chemical graph space given clear objectives to use for solution scoring. In comparison with commonly used MOOP algorithms MEGA consistently outperforms MOGA, the most commonly used technique in DND, and compares favorably with SPEA since it produces Pareto-approximation sets of similar quality but greater diversity.

Time requirements for the execution of the runs were sufficiently reasonable. A typical run of MEGA with population 50 and 100 iterations took approximately 40 minutes.

## VI. CONCLUSION

The research presented builds on our recent work on MEGA, a new class of hybrid multi-objective algorithms that have been designed specifically for the OGD problem. MEGA uses graphs for solution representation and exploits available knowledge through the use of privileged fragment building blocks. The algorithm uses a unique niching mechanism developed to preserve chromosome diversity and elitism in the form of a Pareto archive, to avoid loss of promising solutions. The results obtained from the application of MEGA to the significant and challenging problem of designing ER- $\beta$  selective ligands indicate that the algorithm produced solutions having significantly higher similarity to ER- $\beta$  ligands than ER- $\alpha$  ligands and that its results compare favorably with established MOOP methodologies.

Our future work plans focus mainly on the inclusion of additional knowledge into the algorithm. We plan to add further domain knowledge exploitation capabilities to MEGA in the form of rules, and local search techniques.

An additional, ongoing direction of research, investigates the parallelization of the algorithm implementation to enable the application of the system on large scale distributed systems.

#### ACKNOWLEDGMENT

The authors would like to thank Dr. J. Apostolakis for providing access to the Chil chemoinformatics software suite and Prof. E. Mikros for supplying his expert opinion on estrogen receptors and their inhibitors.

#### REFERENCES

- [1] Y. Colette, P. Siarry, *Multiobjective Optimization: Principles and Case Studies*; Springer-Verlag: Berlin, Germany, 2004.
- [2] N. Jozefowicz, F. Semet and E-G.Talbi, An evolutionary algorithm for the vehicle routing problem with route balancing, *Europ. J. of Operat. Res.*, vol. 195, no. 3, pp 761-769, 2009.
- [3] C. A. Nicolaou, J. Apostolakis, and C. S. Pattichis, "De Novo Drug Design Using Multiobjective Evolutionary Graphs", *J. Chem. Inf. Model.*, vol. 49(2), pp. 295-307, 2009.
- [4] C. A. Nicolaou, C. S. Pattichis. Molecular Substructure Mining Approaches for Computer-Aided Drug Discovery: A Review. In *Proc. of ITAB*, Ioannina, Greece, October 26-28, 2006.
- [5] K.-H. Baringhaus, H. Matter, "Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters", in *Chemoinformatics in Drug Discovery*, T. Oprea, Ed., Weinheim: Wiley, 2004; pp. 333-379.
- [6] C. A. Coello Coello, "Evolutionary Multiobjective Optimization: A Critical Review", in *Evolutionary Optimization*, R. Sarker, M. Mohammadian, X. Yao Eds., v. 48, New York: Springer, 2002, pp117-146.
- [7] E. Zitzler, L. Thiele, Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **1999**, 3, 257-271.
- [8] C. M. Fonseca, P. J. Fleming, Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. In *Proceedings of the Fifth International Conference on Genetic Algorithms*; S. Forrest, Ed.; Morgan Kaufmann: San Mateo, CA, 1993; pp 416-423.
- [9] C. A. Nicolaou, N. Brown, C. S. Pattichis, Molecular optimization using computational multiobjective methods. *Curr. Opin. Drug Discovery Dev.* **2007**, 10, 316-324.
- [10] R. S. Bohacek, C. Martin, W. C. Guida, The Art and Practice of Structure-Based Drug Design: A Molecular Modelling Approach. *Med. Res. Rev.* **1996**, 16, 3-50.
- [11] N. Brown, B. McKay, F. Gilardoni, J. Gasteiger, A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1079-1087.
- [12] D. L. Wheeler, et. al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2006**, 34, D173-D180.
- [13] NoesisChemoinformatics, Ltd. Nicosia, Cyprus. <http://www.noesisinformatics.com> (accessed May 14, 2009).
- [14] <http://www.chil2.de> (accessed April 3, 2009).
- [15] P. Willet, J. M. Barnard, G. M. Downs, Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 39, 983-996.
- [16] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 118-127.