Computers and Chemical Engineering 78 (2015) 79–84

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Note

Generic mathematical programming formulation and solution for computer-aided molecular design

Lei Zhang, Stefano Cignitti, Rafiqul Gani*

CAPEC-PROCESS, Department of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads, Building 229, DK-2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 15 April 2015
Accepted 22 April 2015
Available online 30 April 2015

Keywords:
Molecular design
CAMD
Chemical structure
Group contribution
MILP/MINLP

ABSTRACT

This short communication presents a generic mathematical programming formulation for computer-aided molecular design (CAMD). A given CAMD problem, based on target properties, is formulated as a mixed integer linear/non-linear program (MILP/MINLP). The mathematical programming model presented here, which is formulated as an MILP/MINLP problem, considers first-order and second-order molecular groups for molecular structure representation and property estimation. It is shown that various CAMD problems can be formulated and solved through this model.

1. Introduction

Computer-aided molecular design (CAMD) is a method to design molecules with desired properties. That is, through CAMD, it is possible to generate molecules that match a specified set of target properties. CAMD has attracted much attention in recent years due to its ability to design novel as well as known molecules with desired properties. The attention is in particular targeted at the design of chemical based products, such as solvents, refrigerants, active pharmaceutical ingredients, polymers, surfactants, lubricants, and more (Gani, 2004).

Property prediction methods are needed in molecular design, as they enable the prediction of the target properties of the candidate molecules. Here, CAMD methods can be regarded as the reverse engineering approach to property prediction, as the target properties are known while the molecules that match them need to be determined. Typically, almost all CAMD methods use group contribution (GC) based property prediction methods (from Franklin, 1949 to Hukkerikar et al., 2012) to evaluate the generated compound with respect to the specified set of desirable target properties (Harper et al., 1999). The GC-based methods belong to a class known as additive methods (Hukkerikar et al., 2012).

$$F(p) = \omega_1 \sum_i N_i C_i + \omega_2 \sum_j M_j D_j + \omega_3 \sum_k O_k E_k + \cdots \quad (1)$$

In Eq. (1), p is the desirable property, C_i is the contribution of first-order group i, N_i is the number of occurrences of first-order group i; D_i is the contribution of second-order group i, M_i is the number of occurrences of second-order group i; E_i is the contribution of third-order group i, O_i is the number of occurrences of third-order group i; ω_1, ω_2, ω_3 are weights that may be imposed on each of the additive terms. From a practical point of view, the highest order of Eq. (1) is three (Marrero and Gani, 2001). Second and third order additive methods are able to distinguish some isomeric molecular structures in CAMD problems. In this paper, only first and second order groups are considered. Third order groups can also be considered using this new model, but it is not necessary for most CAMD problems.

With the advent of connectivity-based prediction methods, several researchers have developed new strategies for embedding it with CAMD method. Constantinou et al. (1996) proposed a systematic strategy for generating isomers from a set of groups. Harper et al. (1999) proposed a framework for CAMD method, where the pre-design phase defines the basic needs, the design phase determines the feasible candidates (generates molecules and tests for desired properties) and the post-design phase performs higher level analysis of the molecular structure and the final selection of

* Corresponding author. Tel.: +45 45 252882; fax: +45 45 882258.
E-mail address: rag@kt.dtu.dk (R. Gani).

http://dx.doi.org/10.1016/j.compchemeng.2015.04.022
0098-1354/

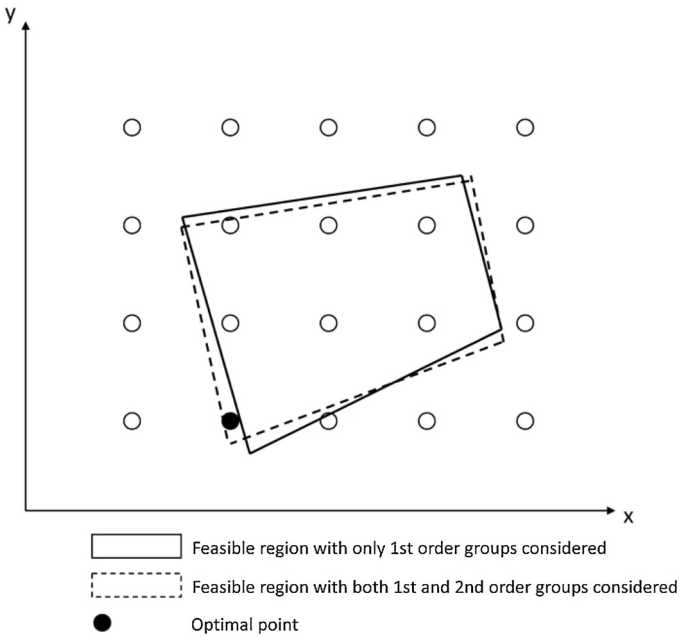

Fig. 1. Feasible region of CAMD problems using different modeling approach.

the product. Samudra and Sahinidis (2013) proposed a new optimization model using relaxed property targets and refined property targets with structural corrections. It is usually difficult to model and solve the MILP/MINLP problem with structure information considered due to the increased size of the mathematical problem and number of alternatives. Thus, alternative solution strategies have been proposed to ensure that solution can be found and that also a global optimum can be found. Harper et al. (1999) used a generate and test approach to decompose the CAMD problem; selection of building blocks (functional groups), combination of groups into chemically feasible molecules, estimation of the specified set of properties for the generated molecules, selection as candidate compounds, and finally, determination of those that match the specified set of properties. Samudra and Sahinidis (2013) decomposed the problem into three design steps: composition design, structure design and extended design. In composition design, a large number of compositions (molecules composed of groups) matching relaxed design criteria based on first-order property estimates are determined. Thus, the GC$^+$ property estimation model is relaxed (only considering the first-order groups) to obtain the building blocks, then the property model is refined with second-order groups (structure design information) based on the results of the first step. However, this may result in the possibility of an optimal solution being excluded. Second-order groups refine the property prediction and molecules that wrongfully lie outside the search space are neglected. As seen in Fig. 1, the solid line box is the feasible region of the decomposed model; the dash line box is the real feasible region of the CAMD problem. If decomposed approach method is used, the global optimal point is excluded from the feasible region. That means the optimal point obtained from the decomposed method might be a local optima. Samudra and Sahinidis (2013) used property relaxation method to avoid this situation. That is, instead of property interval $[X_k^L, X_k^U]$, they allow the property X_k to lie in the expanded interval $[0.9X_k{}^L, 1.1X_k{}^U]$. This relaxation is justified by the fact that the average errors in first-order property estimation of the GC$^+$ model rarely exceed 10% (Samudra and Sahinidis, 2013). But it is not always easy for the users to find the appropriate relaxations. On the other hand, the feasible region of the optimization problem will become larger when relaxations applied, which makes the solution of the problem harder.

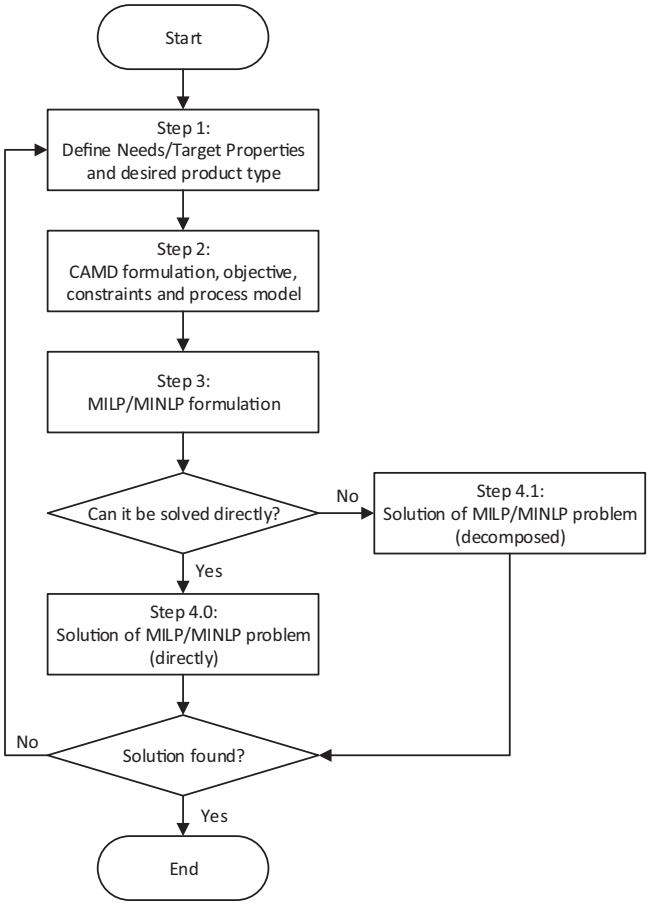

Fig. 2. Computer-aided molecular design framework.

In this short communication, a new model for CAMD problems is proposed. The models consider both first and second order groups simultaneously in the MILP/MINLP formulation, and the molecular structure is obtained from the solution of the adjacency matrix. This will avoid the possible situation in Fig. 1, where a possible optimal point may be excluded from the feasible region, and ensures the obtainability of a global optimal solution. This short communication is structured as follows. Section 2 gives a detailed description of the methodology and the mathematical formulation of CAMD problems with the proposed model; Section 3 gives three case studies; Section 4 draws some conclusions from the presented results.

2. Methodology

The computer-aided molecular design framework is presented here in Fig. 2. The framework has four steps (Cignitti et al., 2015), (1) problem definition: product needs, target properties and desired product type are defined here; (2) CAMD formulation: the needs, properties and product types are converted to a CAMD problem in which objective function and constraints related to molecular structure, product needs (property model) and process models are defined; (3) MILP/MINLP formulation: the CAMD problem from step two is set-up as a MILP/MINLP formulation; (4) solution of MILP/MINLP problem: the MILP/MINLP formulation is solved directly or through a decomposed strategy depending on the problem type, linearity and size.

If the needs and target properties of the designed molecule are defined in the design problem, the CAMD problem can be posed as a mathematical program in which the number of binary and continuous variables defines the search space.