Computer-aided molecular design with combined molecular modeling and group contribution

Peter M. Harper a, Rafiqul Gani a, Petr Kolar b, Takeshi Ishikawa b

a CAPEC, Department of Chemical Engineering, Building 229, Technical University of Denmark, 2800 Lyngby, Denmark
b Mitsubishi Chemical, Mizushima Plant, 3-10 Ushidori, Karashiki 712, Japan

Received 2 April 1998; accepted 21 October 1998

Abstract

Computer-aided molecular design (CAMD) provides a means for determining molecules or mixtures of molecules (CAMMD) having a desirable set of physicochemical properties. The application range of CAMD is restricted due to limitations on the complexity of the generated molecular structures and on the availability of suitable models for property prediction. A new CAMD methodology that addresses this issue by combining molecular modeling techniques with a traditional CAMD approach is presented. The new method includes a new molecular/atomic structure generation algorithm, a large collection of property estimation methods, and, a link to molecular modelling tools. Application of the new CAMD method is highlighted through two industrial examples. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Group contribution; Molecular modeling; CAMD; Property prediction

1. Introduction

In recent years, CAMD has attracted much attention because of its ability to design and/or find compounds (or mixtures of compounds) with desirable properties. In principle, CAMD-based techniques can be employed for a very large range of problems as diverse as solvent design/selection, CFC substitutes, alternative process fluids, polymer design, and drug design. The limiting steps in any CAMD-based technique are generation of a large range of feasible molecular structures and the reliable estimation of the necessary physicochemical properties. Various CAMD-based techniques have been proposed in recent years and they can be classified as (a) Database search [1], (b) Generate and Test [2,3], and (c) Mathematical programming and genetic algorithm [4,5].
Typically, almost all CAMD methods have used group contribution based property prediction methods to evaluate the generated compound with respect to the specified set of desirable properties. Group contribution methods are simple, have acceptable accuracy for many properties and are predictive in nature. Also, group information is used by most CAMD methods to create the molecular structure. While these property prediction methods have proved themselves an indispensable tool they do not provide all the answers needed to fully evaluate a generated compound. This is because group contribution methods do not exist for all necessary properties, and, reliability of prediction is often questionable for large, complex molecules. Menuai and Newsham [6] have proposed an alternative approach to group contribution within the specific CAMD application area of solvent design and selection. Menuai and Newsham employ molecular modeling instead of group contribution for the evaluation of design parameters. It should be noted however, that the evaluation/calculation of properties using molecular modeling techniques is of significantly higher computational complexity and limits the size of the possible search space using conventional computational resources. In this paper, we have extended our CAMD (Generate and Test) approach [3] to include improved compound selection features (so that database searches can also be carried out), improved generation of molecular structures and property prediction and a link to molecular modeling (to overcome some of the limitations related to complexity of molecular structures and to provide better structure analysis options). The result is a hybrid overall methodology employing different estimation/validation methods in various steps depending on the size of the problem and available information. A commercial package, Chem3D [7], has been used to illustrate the link between CAMD and molecular modeling techniques.

2. Requirements for a CAMD method

In its simplest form, a CAMD problem can be formulated as the task of finding all compounds matching a specified set of properties. The important steps in a CAMD (generate and test) approach are, selection of building blocks (such as functional groups), combination of groups into chemically feasible molecules, estimation of the specified set of properties for the generated molecules, selecting as candidate compounds, those that match the specified set of properties (see Fig. 1).

The CAMD method described above needs to address a number of additional issues related to molecular structure (number, size and complexity), property prediction (how to estimate the specified set of properties for the candidate compounds) and analysis (how to screen/select the candidate compounds). In addition, any computer-aided system must also be robust, flexible and computationally efficient. A CAMD method must be able to generate the structures of large and complex molecules as well as the structures of isomers. In order to achieve this, the molecular structure generation method needs to consider higher order groups in addition to the traditional first order groups [3] that are often used for the molecular structure generation step. Similarly, if the first order groups alone cannot be used for the generation of isomers and large/complex molecules, their properties cannot also be estimated through contributions of these groups. For the test step, the CAMD method needs to look beyond group contribution based methods for property prediction. This means that additional structural information (atomic representation, connectivity index, etc.) needs to be generated from the molecular structure representation in terms of groups (first- and second-order). Generally, most methods available for prediction of properties depend (by definition) on the structure
of the generated compounds either explicitly or implicitly. Many of these methods need more than the molecular structural information. For example, representation in terms of fragments, atoms, connectivity index and so on. Other methods or correlations may depend on the molecular structure in addition to a set of properties (for example, vapor pressures). Therefore, in order to use these methods, the compound structure also needs an atomic representation. Note that the same atomic representation is also needed to establish a link to molecular modeling program packages such as Chem3D. In order to address the issues of robustness, flexibility and computational efficiency, any CAMD method based on the generate and test approach must avoid the problem of combinatorial explosion. One way to avoid combinatorial explosion is to employ a structured (multilevel) generate and test approach, where, every level generates and tests structures with the lower levels using molecular representation and the higher levels using atomic representations. This structured approach also provides a natural ordering of the use of the property prediction methods.

3. CAMD—extended version

3.1. New multi-level CAMD method

A multilevel approach is used with each level having its own generation and test steps. The results from each level are ‘promoted’ as input for the next level. This ensures that the size of the combinatorial problem is continuously held at a level where it can be handled effectively without ‘combinatorial explosion’. Furthermore, by having the most time consuming operations at the higher levels where the remaining candidates are the most promising, efficiency with respect to execution time is naturally obtained. Levels 1 and 2 employ macroscopic representation of molecules while levels 3 and 4 employ microscopic representation of molecules.

3.1.1. Level 1

In the generation step of this level, using an updated set of combination rules proposed earlier by Constantinou et al. [3], a collection of unique ‘vectors’ of fragments is generated. A fragment could
be viewed as a building block and a fragment vector represents a complete structure. Here, only first order groups are employed and the test step involves the prediction of specified properties by the group contribution approach only. The results from this level have two important qualities: (i) Any fragment vector will generate at least one feasible compound; (ii) All fragment vectors satisfy the specified property constraints.

3.1.2. Level 2

The generation step of this level combines elements of the individual fragment sets from level 1 to form molecular structures. This is done by recursively adding groups to a starting group, forming all possible combinations as illustrated by the (partial) generation tree in Fig. 2. Since the generation step in level 1 only guarantees that it is possible to combine the fragments to one or more molecular structures, it is necessary to impose rules for the generation step in this level to ensure that only feasible chemical compounds are generated. In each sub-level of this building process the intermediate results are screened and if an intermediate product is a finished compound (i.e., has no remaining free connections) it is removed. Similarly all intermediate products are compared and duplicates are removed. This procedure of repeated pruning (as illustrated in Fig. 2 by an ×) of non-productive endpoints (duplicates or premature finished structures) in the generation tree is vital in order to control the size of the problem. Apart from the trivial rules (all fragments in a set must be used, the valency constraint, etc.), additional restrictions on which groups can be connected is imposed in order to

Fig. 2. Isomer generation (level 2) and selected results for application example 1. (1) Shows instances where a complete compound has been formed without using all fragments. (2) Illustrates the duplicate removal process performed on each sublevel.
guarantee the generation of feasible compounds (e.g., the combination of a Cl group and a OH group is not allowed). These rules are derived from the basic set of the rules for level 1 [3] and reformulated to apply to individual groups.

Note that in level 2, second-order groups are considered. This means that some isomer structures (those permitted by the combination rules related to the second-order groups) are also generated. The results from the generation step of level 2 are molecules described by groups and connections between these as illustrated in Figs. 1 and 2.

The test stage of level 2, uses the additional information available (knowledge of the basic structure of the molecules) to refine predictions from level 1 and to test for properties that could not be estimated with the information available in level 1. Examples of such properties are methods (biodegradation probability estimation method by Boethling et al. [8]) using fragments larger than the ones used in level 1. For such methods, a structure needs to be defined in order to identify the fragments in question.

3.1.3. Level 3

After levels 1 and 2, the number of candidates has diminished due to their prediction/screening steps. The remaining compounds are unique molecules but the detailed connectivity is only implicitly described. The first step towards a detailed “microscopic” description and representation of the generated molecules within the new CAMD method is to clearly define the connectivity of the molecules. This is achieved using the algorithm outlined below:

1. Starting point: represent results from level 2 using a symmetrical matrix (compound definition matrix) describing the connectivity of the various groups.
2. Introduce atomic connection matrices for each group (building blocks used in level 1). These matrices describe a group/building block in terms of the atoms it consists of and how these atoms are interconnected.
3. Replace each group description in the compound definition matrix with the atomic description from step 2.
4. Re-map the intergroup connections and verify that the correct atom–atom bonds are described (i.e., after the expansion, the original connection information between groups is changed to connection information between individual atoms of the original groups).
5. Result: a matrix describing the compound connectivity at the atomic level is obtained.

Fig. 3 shows the starting point (in terms of structural information) and the result (atomic description matrix) from such a conversion. Furthermore, by having the detailed connectivity it is possible to utilise the vast number of estimation methods based on connectivity indices [9] in the test step of this level.

3.1.4. Level 4

With the connectivity established from level 3, the starting point at this level is a set of compounds each described by their atomic description matrix. Note that the set of compounds passed to this level are the remaining compounds surviving the screening procedures of levels 1, 2 and 3. Therefore, the confidence in these compounds is sufficiently high to justify a more detailed and time-consuming next step.

By taking the information of connectivity and atom types from level 3, a 3-dimensional graph (or molecule model) is created by applying a set of standard or default bond lengths and angles for the
various types of connections. By doing this, one obtains the true molecular models of the compounds which are directly transferable into molecular modeling programs such as MOPAC, Chem3D or SYBYL. A structure ready for further optimisation is shown in Fig. 3.

One important aspect of the generation of 3-dimensional models is the question of conformers. Note that since a set of ‘standard’ bond lengths and angles and random torsion angles is used, there is no guarantee that the resulting structure represents the prevalent conformer or is at an energy minimum. As a first and simple solution, the following action is proposed: perform an energy minimisation using the MM2 force field method to bring the structure, at least, to a (local) energy minimum.

3.1.5. Further analysis

Beyond level 4, many directions can be taken. One very interesting possibility is the generation and distinction of cis/trans and R/S isomers, something that is not possible in the previous levels. Another important aspect is the determination of the lowest energy conformer using stochastic global optimisation (e.g., the Boltzmann Jump algorithm). This can be further extended into the evaluation of properties based on a distribution of conformers rather than a single conformer. Such an approach would reflect the true nature of a bulk chemical more accurately.

3.2. Important features of the new CAMD method

3.2.1. Property prediction

In addition to the purely group contribution based methods, other methods based on connectivity indices (for estimation of properties such as refractive index), fragments (for estimation of properties
such as biodegradation probability), correlation based methods (for estimation of properties such as solubility parameters) and mixed methods (for estimation of properties such as vapour pressures and activity coefficients) have been incorporated. A total of 41 properties have so far been used successfully in the extended CAMD method.

3.2.2. Molecular modeling

Once the molecular structure is generated, the corresponding connectivity index and fragments information is also generated. For every generated molecule, the structure is further analyzed through molecular modeling with a link to Chem3D. This link is established by converting the molecular structure definition in terms of groups to an atomic description (the necessary input for Chem3D). Fig. 3 shows how the link between CAMD and Chem3D has been established. Inside Chem3D, the molecular structure is optimized in terms of energy and the corresponding properties are estimated. This link also provides a further generation of new molecules (or isomers). This feature, therefore, provides a powerful tool for analysis of the generated compounds. It also provides property estimation options (for compounds where the group contribution methods cannot be applied). One such option is to estimate the dipole moment or enthalpy of formation and then use correlation/connectivity index methods to estimate other properties (such as octanol–water partition coefficient, boiling point, etc.).

3.2.3. Compound selection

The link to molecular modeling also contributes to the final selection (or screening of tested compounds) because the stability of the molecular structure can be analyzed. In principle, the link to molecular modeling is employed to analyze compounds that are not present in any database. A database (of approximately 3000 known compounds with experimental data) is also linked to the CAMD method. In this way, all tested compounds are checked (optionally) in levels 2 and above against this database. If they are found in the database, then in addition to checking for their properties, they can also be analyzed with respect to known uses of the compound (this provides a knowledge base for the selection step of CAMD). For example, in solvent selection, all tested compounds present in the database, can be analyzed with respect to their known solvent properties. Note that the mixture properties need to be estimated and are usually not available in the database. For this reason, the search is performed after the generation step and not before (as in type–a CAMD methods based on database search). The search engine in the extended CAMD method is more efficient since only those compounds that satisfy the property constraints are identified and further analyzed.

4. Application examples

Two application examples are presented. The first example involves finding candidate solvents for the removal of phenol from process water prior to discharge to the sewer. The second example also involves finding candidate solvents but for separation of a close boiling mixture. The objectives of these examples are (1) to highlight the use of the new CAMD method to solve problems of interest to the industry and (2) to serve as proof of concept. Other examples not reported in this paper (but
details can be obtained from the authors) involve miscibility promotion, substitution of CFCs, candidate process fluids, azeotropic agents and many more.

4.1. Example 1: replacement of toluene as solvent for removal of phenol from a waste water stream

The process water contains 7% w/w of phenol. Toluene has earlier been used as a solvent for liquid–liquid extraction. However, because of current environmental regulations, toluene needs to be replaced. The desirable properties of the solvent, therefore, should be the same as that of toluene except for the environmental impact related properties. Butyl acetate has been reported as a candidate [10]. Figs. 1 and 2 show the generation of candidate molecules while Fig. 3 shows the optimized configuration of a candidate molecule. Considering alcohols, ethers, esters, ketones, acids and hydrocarbons, a total 5332 initial (level 1) descriptions were generated, 35 descriptions (fragment vectors) were selected, 288 molecules (level 2) were generated from these 35 descriptions and finally, 58 compounds satisfied all specified property constraints. In this example, level 3 did not include active screening but only the calculation steps serving as an aid in the assessment of potential candidates. Three of the candidate compounds are shown in Fig. 2. The performance of the candidate solvents has been compared with toluene and butyl acetate and a selected set of results are shown in Table 1. In Table 2 the estimation results from different levels are shown.

As an example of the use of molecular modelling tools (level 4), a manual conformer analysis around the C–O–C=O torsion angle has been performed for compound A using the MOPAC package included in Chem3D. In Table 3, the conformers corresponding to two different energy minima are shown along with their calculated enthalpy of formation and dipole moment. The results of the analysis illustrate one of the key application areas for molecular modelling techniques in CAMD. The two conformers have very similar energy levels (as indicated by the energy of formation) but very different dipole moments indicating that the actual dipole moment of the compound will depend heavily on the distribution of the conformers. This information is valuable when assessing the reliability of the estimated properties by using the simple group contribution based property estimation methods. If an estimated property or property function is known to be correlated with dipole moment, the molecular modeling results can be used to verify the correlation and the sensitivity of the dependence. This example is simplified significantly compared to real applications where the conformer analysis covers all torsion angles in the molecule (a nontrivial process best performed with special tools). However, this example serves as proof of concept. A detailed description and solution of the problem can be obtained from the authors.

Table 1

<table>
<thead>
<tr>
<th>Compound</th>
<th>Separation factor</th>
<th>Solvent loss</th>
<th>log P</th>
<th>Select</th>
<th>Capacity</th>
<th>ρ (g/cm³)</th>
<th>T\text{\textsubscript{flash}} (K)</th>
<th>T\text{\textsubscript{boil}} (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>125.9</td>
<td>3.2e–5</td>
<td>3.1</td>
<td>14.0</td>
<td>22.2</td>
<td>0.859</td>
<td>335</td>
<td>448.3</td>
</tr>
<tr>
<td>B</td>
<td>125.9</td>
<td>3.2e–5</td>
<td>3.1</td>
<td>14.0</td>
<td>22.2</td>
<td>0.859</td>
<td>335</td>
<td>448.3</td>
</tr>
<tr>
<td>C</td>
<td>125.9</td>
<td>3.2e–5</td>
<td>3.1</td>
<td>14.0</td>
<td>22.2</td>
<td>0.852</td>
<td>333</td>
<td>445.6</td>
</tr>
<tr>
<td>Toluene</td>
<td>305.1</td>
<td>8.3e–5</td>
<td>2.7</td>
<td>36.5</td>
<td>2.6</td>
<td>0.878</td>
<td>310</td>
<td>386.1</td>
</tr>
<tr>
<td>Butyl-acetate</td>
<td>105.0</td>
<td>0.0014</td>
<td>1.8</td>
<td>11.6</td>
<td>30.1</td>
<td>0.876</td>
<td>304</td>
<td>406.0</td>
</tr>
</tbody>
</table>
Table 2
Estimation results from different levels for example 1 (only results differing from the previous level are shown)

<table>
<thead>
<tr>
<th>Level</th>
<th>Compound</th>
<th>Separation factor</th>
<th>Solvent loss</th>
<th>T_{flash} (K)</th>
<th>Select</th>
<th>Capacity</th>
<th>( \rho ) (g/cm(^3))</th>
<th>log ( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>A</td>
<td>126</td>
<td>3.2( \times ) 5</td>
<td>336</td>
<td>14</td>
<td>22.2</td>
<td>0.865</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>126</td>
<td>3.2( \times ) 5</td>
<td>336</td>
<td>14</td>
<td>22.2</td>
<td>0.865</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>126</td>
<td>3.2( \times ) 5</td>
<td>336</td>
<td>14</td>
<td>22.2</td>
<td>0.865</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>A</td>
<td></td>
<td>334</td>
<td></td>
<td></td>
<td></td>
<td>0.859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td>334</td>
<td></td>
<td></td>
<td></td>
<td>0.859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td>333</td>
<td></td>
<td></td>
<td></td>
<td>0.852</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.1</td>
<td></td>
</tr>
</tbody>
</table>

4.2. Example 2: design of extractive distillation agent

In this example, a binary mixture of two close boiling (boiling points close to 400 K and having four carbon atoms) organic acids is to be separated by extractive distillation. The problem definition in terms of solvent properties are, therefore, different from example 1. Also, while in example 1, aromatic compounds were not desirable as solvents, in this example, we are looking for aromatic solvents. This separation problem comes from the chemical industry and due to confidentiality agreement, the identity of the compounds cannot be revealed. The desirable solvent properties can however be revealed (as shown in Fig. 4). The results obtained at the end of level 2 is shown in terms of vapour–liquid phase diagrams (solvent-free basis) for one of the candidate solvents. The phase diagrams are made as a function of solvent to feed ratio. It can be seen that the solvent satisfies the primary requirement of its ability to extract solute A from the mixture. Application of levels 3 and 4 further confirms this candidate as well as gives more insight to its structural properties. Also, other structures based on the atomic representation of the candidate solvent can be generated in level 4.

Table 3
Structure and calculated properties for different conformers (example 1)

<table>
<thead>
<tr>
<th>Structure</th>
<th>Conformer 1</th>
<th>Conformer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_f )</td>
<td>-122.2 kcal/mole</td>
<td>-121.3 kcal/mole</td>
</tr>
<tr>
<td>Dipol Moment</td>
<td>1.790 Debye</td>
<td>4.142 Debye</td>
</tr>
</tbody>
</table>
More details on the solvent atomic structure and molecular modeling results can be obtained from the authors.

5. Conclusion

A new multilevel generate and test CAMD method has been presented. This method offers the possibility of creating a bridge between the traditional group oriented approach typically used in chemical engineering and the more detailed molecular modeling approach used by chemists without sacrificing the strong points of either of these methods. In the most prevalent use of CAMD in chemical engineering, solvent design, the additional information on the generated compounds together with the property estimation and visualisation features offer useful assistance in the final selection procedure. Note that the purely group contribution based CAMD methods, at best, give only a list of possible candidates which need to be further investigated. Link to molecular modeling/simulation tools provide the necessary additional insight and analysis of the performance aspects of candidate solvents. For the specialist the key task of manipulating and examining compounds on a detailed molecular level is unaltered and existing methods are still available. An area where the new CAMD method is of interest is in the identification of the starting point for such detailed studies. Almost any task of designing a compound still need some basic (simple) criteria to be fulfilled. These can be as simple as restrictions on type/number of functional groups and molecular weight or more complex like density at a given temperature, boiling point or miscibility with a specific solvent. The new CAMD method not only handles such constraints but at the same time, generates more structures, tests the generated structures with a large variety of property estimation methods, and, finally, delivers the result in a form ‘ready to use’ by the molecular modeling/design specialist.

References


