CABIOS

Vol 13 no. 5 1997
Pages 537-543

Estimation of equilibrium constants using automated group contribution methods

Ronald G. Forsythe, Jr[1,4], Peter D. Karp[2] and Michael L. Mavrovouniotis[3]

[1]Department of Engineering, University of Maryland Eastern Shore, Princess Anne, MD 21853, USA, [2]Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, EJ229, Menlo Park, CA 94025, USA and [3]Chemical Engineering Department, Northwestern University, Evanston, IL 60208, USA

Received on February 26, 1997, accepted on April 28, 1997

Abstract

Motivation: *Group contribution methods are frequently used for estimating physical properties of compounds from their molecular structures. An algorithm for estimating Gibbs energies of formation through group contribution methods has been automated in an object-oriented framework. The algorithm decomposes compound structures according to a basis set of groups. It permits the use of wildcards and is able to distinguish between ring groups and chain groups that use similar search structures. Past methods relied on manual decomposition of compounds into constituent groups.*
Results: *The software is written in Common LISP and requires <2 min to estimate Gibbs energies of formation for a database of 780 species of varying size and complexity. The software allows rapid expansion to incorporate different basis sets and to estimate a variety of other physical properties.*
Availability: *Available on request from the authors.*
Contact: *E-mail: ronjr@erika.umd.edu; pkarp@ai.sri.com; mlmavro@nwu.edu*

Introduction

Researchers who use equilibrium constants for thermodynamic analysis of biotransformations frequently encounter reaction pathways for which equilibrium constants are not readily accessible. Without approximate values for the equilibrium constants, it is impossible to assess even the direction and feasibility of a biochemical reaction or pathway, or to uncover bottlenecks (Mavrovouniotis, 1993, 1996). The equilibrium constants are related to the Gibbs energies of biotransformation, which can be estimated using group contribution methods (Mavrovouniotis, 1991). Group contribution methods are frequently used for estimating physical properties of compounds from their molecular structures (Reid *et al.*, 1987). This paper presents a completely automated algorithm for decomposing compounds into their constituent groups, where past methods relied on manual group decomposition. Since compound structures are routinely stored in databases, rough estimates for the equilibrium constants can be generated very quickly; the time-consuming task of obtaining experimental values is reserved for cases where more accurate values are needed.

System and methods

Software has been developed using Macintosh Common LISP (MCL) 2.0.1 and runs on Motorola 608x0-based Macintosh systems. The code is not platform specific and has been compiled on Silicon Graphics systems running AllegroCL 4.3 and Sun/Sparc systems running Lucid Common LISP v.4.0 with the Common LISP Object System (CLOS) module installed.

Algorithms and implementation

The problem of estimating equilibrium constants for reactions can be transformed to one of estimating Gibbs energies of formation for all reactants and products of those reactions. This is accomplished by relating the standard Gibbs energy of change for the reaction, $\Delta G'_{\mathrm{rxn}}$, to the equilibrium constant, K', using

$$\Delta G'_{\mathrm{rxn}} = -RT\ln K' \qquad (1)$$

and by determining the standard Gibbs energy of a biotransformation from the standard Gibbs energies of formation of its species through

$$\Delta G'_{\mathrm{rxn}} = \Sigma\, \alpha_i \Delta G_i \qquad (2)$$

where α_i and ΔG_i are the stoichiometric coefficient and standard Gibbs energy of formation for species i in the reaction ($\alpha_i < 0$ for reactants and $\alpha_i > 0$ for products), respectively.

Mavrovouniotis (1991) presented a method for estimating Gibbs energies of formation for compounds by decomposing

[4]*To whom correspondence should be addressed*

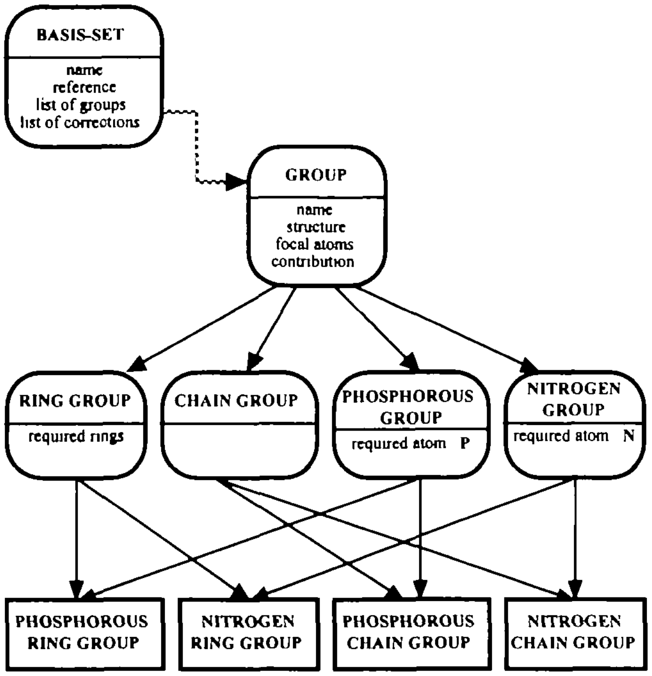

Fig. 1. Object classes and attributes that are used in the software. Each basis set comprises a set of groups. Each group inherently knows how to identify its own existence in a compound. Subclasses of group have been identified based on their key atoms and their participation in rings. These refined subclasses allow methods to be defined that reduce the apparent size of a basis set.

the compounds into constituent groups, and summing the energy contribution of each group through the relationship

$$\Delta G_i = \Sigma\, n_i g_i \qquad (3)$$

where n_i and g_i represent the number of occurrences and the contribution of group i in the compound, respectively. In past work, all group decompositions were determined manually. The reader is referred to Mavrovouniotis (1990, 1991) and Reid *et al.* (1987) for discussions regarding group contribution methods and their application to the estimation of equilibrium constants.

Object-oriented analysis of the estimation problem

In order to identify the objects, operations and attributes of a system, it is useful to write a brief summary of the problem statement. Typically, nouns indicate the objects and attributes of a system and verbs describe desired operations.

PROBLEM STATEMENT:

Estimate the Gibbs energy of transformation for a **reaction** by summing the Gibbs energies of formation of its **compounds**. *Estimate the Gibbs energy* of formation for a compound by *decomposing* it according to a specified **basis set** of **groups**.

```
ESTIMATE-GIBBS-ENERGY reaction
   for each compound in reaction
      ESTIMATE-GIBBS-ENERGY compound
   SUM contributions of compounds
   RETURN the estimated value.

ESTIMATE-GIBBS-ENERGY compound
   DECOMPOSE compound into groups
   SUM contributions of groups
   RETURN the estimated value.
```

Fig. 2. Different procedures are implemented when estimating-gibbs-energies on reactions and compounds. The determination of which procedures are to be followed is left to the object system.

A basis set is an ordered collection of groups that are used to decompose a compound (Figure 1). Associated with each group is the contribution that it makes to the overall property of a compound, as well as the structural fragment that defines the group. This information is stored within the attributes of the class group. The REACTION and COMPOUND classes also describe important objects in the system. Each reaction has slots (attributes) that record the compounds that participate in the conceptual reaction and each compound stores the atoms and bonds of its structure.

An important global operation for the system is the estimation of Gibbs energies. In object-oriented programs, the actual methods or procedures that are followed to meet the goal of an operation are dictated by the objects on which functions are called. For example, a single operation ESTIMATE-GIBBS-ENERGY can be called on reactions or on compounds; however, each of these classes implements different methods (Figure 2). When the function is called on a reaction, the system knows it should perform the operation on each of its compounds. However, when the function is called on a compound, the compound structure is decomposed into its constituent groups. The determination of which set of procedures is to be followed is delegated to the underlying object-oriented system—in this case, CLOS.

Substructure matcher

The group-decomposition algorithm employs a chemical substructure matcher to search out all occurrences of each group in the basis set within a compound. The matcher is called repeatedly for each group in the basis set. The substructure matcher we have developed is quite flexible. The structures it accepts as inputs can be specified as either SMILES expressions (Weininger, 1988), or as a list of atoms plus a list of bonds between those atoms. The matcher employs a backtracking search to locate the first occurrence of the query structure (the group) within the reference structure (the compound to be decomposed) (Figure 3).

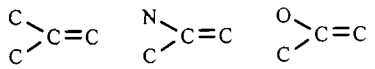

Fig. 3. Examples of substructures that must be distinguishable in compound structures.

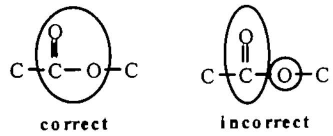

Fig. 4. In order to ensure proper decomposition of compounds, groups must be ordered according to their classification priority. Searching for '-COO-' before '-CO-' and '-O-' allows this decomposition to be performed correctly.

The substructure matcher includes several special capabilities to facilitate group decomposition. Every atom in the compound must be assigned to exactly one group, i.e. the positions of two groups within a compound cannot overlap. To provide this behavior, the matcher returns a description of atoms in the reference structure that matched the query structure. That description is used to form an additional input to the matcher in order to exclude previously assigned atoms from being reassigned in subsequent searches for the same query structure. In this way, all occurrences of the query structure are determined through iterative calls to the substructure matcher using updated descriptions, until no additional occurrences of the query structure are identified. Furthermore, these descriptions are used to ensure that occurrences of the active query structure do not overlap with atoms that have been assigned to previously matched groups.

The order in which groups are searched is significant. Figure 4 illustrates an example in which a -COO- group could be mistakenly characterized as a combination of -CO- and -O- groups. To avoid confusion as to which configuration is desired, the groups in the basis set are arranged in a partial order in which group G1 precedes group G2 if G2 is a substructure of G1.

Wildcards

Some groups specify a larger context of atoms in which the group must appear. For example, the group in Figure 5 contains a single carbon atom that must be connected to three non-hydrogen atoms as shown.

Wildcards significantly reduce the number of groups that have to be defined in a basis set. The bond arrangement in Figure 5 can be represented by the illustrated wildcard representation. This wildcard representation will match all structures that are shown in Figure 3 as well as other permutation of atoms that are bonded to the central carbon atom.

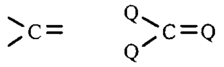

Fig. 5. Wildcards are used to match sets of atoms. They are especially useful for identifying specific bond combinations for central atoms. The wildcard Q is defined to match any atom except hydrogen.

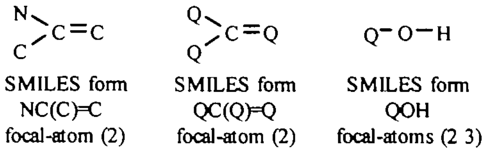

Fig. 6. Focal atoms allow a target structure to be identified by specifying a more complex structure that must be present. Focal atoms are optional for some groups, but are necessary for structures that implement wildcards.

The matcher allows user-defined wildcards that match any set of atoms. Generally, wildcards are defined to match all atoms except hydrogen. It is assumed that when hydrogen is an acceptable match in a group, it will be specifically listed in the query structure.

Focal atoms

We refer to the carbon in Figure 5 as a focal atom because it is part of the group proper, rather than the context for the group. Each group that includes a context must identify the focal atoms in the group. This approach allows those atoms in the reference structure that matched the non-focal atoms in a group to participate in matches with focal and non-focal atoms of other groups in subsequent calls to the matcher. On the other hand, focal atoms that have been assigned to one group will not be reassigned to focal atoms of later groups. This feature is also used for structures that do not use wildcards (Figure 6).

Finding all occurrences of a group in a compound requires the specification of an identifiable substructure. This substructure can incorporate wildcards and can specify focal atoms to target specific atoms and bonds, and returns the list of positions for each of the focal atoms in the compound. (Figure 7)

Ring identification and classification issues

Groups with similar structures can exhibit different levels of participation in rings (Figure 8). A limitation of the substructure matcher is that it is not able to determine efficiently whether a group is a member of a ring unless the ring structure is included in the query structure; in other words, the difficulty is that rings are non-local features while the

```
GET-OCCURRENCES-LIST of group in
    compound
  Search for the structure of group
    in compound
  For each occurrence of the
    structure, Return a list of
    positions corresponding to the
    focal atoms.
```

Fig. 7. It is necessary to identify the location of each focal atom for every occurrence of a group.

matcher works locally. Although it is possible to use wildcards to characterize rings of different sizes and configurations, this approach is computationally expensive because it can take longer to match or reject these larger structures. Furthermore, including all possible structures that must be used to match these ring structures would result in a substantial increase in the number of groups in the basis set.

Ring identification

Our software provides the locations and sizes of rings by finding the cross edges in a simple Breadth-First-Search. It does not include an algorithm that finds the Smallest Set of Smallest Rings (SSSR) because group contribution basis sets are only valid for compounds that are similar to those used to define the contributions for the groups (Gasteiger and Jochum, 1979). The basis sets that have been implemented cannot be extrapolated for use with compounds that contain the complex ring structures that SSSR algorithms are designed to locate. The types and locations of rings are stored in a list before the compound is decomposed. In this way, simpler query structures are used to identify the occurrences of ring groups in the compound. The descriptive information that accompanies these occurrences is then compared to the list of rings in order to verify that the appropriate ring participation requirements are met.

Ring classification

The algorithm currently distinguishes between benzene, heteroaromatic and non-aromatic rings as defined in the Mavrovouniotis (1991) basis set. It is assumed that aromatic bonds are explicitly labeled in the electronic form of the compound structure.

The algorithm for decomposing compounds into their constituent groups is developed by piecing together the substructure matcher and the ring location and classification algorithms. The filter function tests for appropriate ring involvements and ensures that occurrences do not overlap previously matched atoms. If there are unmatched atoms after all groups have been processed, then the basis set cannot be used to estimate properties for the compound.

decompose returns a list of all groups and the number of occurrences of each. Each group stores its contribution to the total property.

Efficiency issues

A key factor in determining the time required to decompose a compound is the number of groups in the basis set that must be searched. Using preliminary ring searches and implementing wildcards results in a significant reduction in the number of groups that must be defined. It is also possible to reduce the apparent size of the basis set by searching for

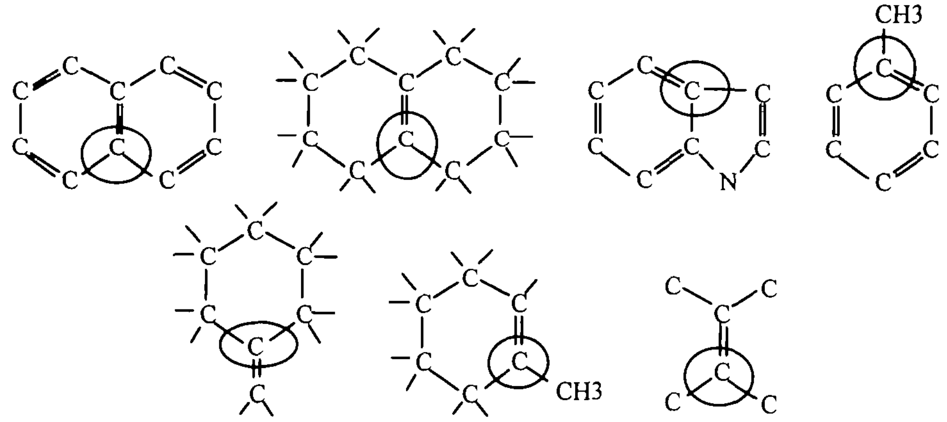

Fig. 8. The basis set defined by Mavrovouniotis (1991) uses the structure >C= to find occurrences of seven different groups. The groups are distinguished based on participation in rings: (**a**) participates in two fused benzene rings; (**b**) participates in two fused non-benzene rings; (**c**) participates in two fused rings: one benzene and one non-benzene; (**d**) the formal single bond and the formal double bond participate in a benzene ring; (**e**) two single bonds participate in a non-benzene ring; (**f**) the double bond and the single bond participate in a non-benzene ring; (**g**) does not participate in a ring.

```
CLASSIFY rings
   for each item in rings
   if all atoms in item are carbons
      with aromatic bonds and the size
         of item is 6,
   then item is a benzene ring.
   if all atoms in item have aromatic
      bonds,
   then item is a heteroaromatic ring.
   Label remaining items as
      nonaromatic.
```

Fig. 9. Rings are found using a Breadth-First-Search and are then classified as being benzene, heteroaromatic or non-aromatic.

```
DECOMPOSE compound
   FIND and CLASSIFY rings in compound
   loop for each group in basis set
      GET-ALL-OCCURRENCES of group in
         compound
      FILTER occurrences
      ADD occurrences to list of found
         atoms list
   If any unmatched atoms->can't use
      basis set on this compound

FILTER occurrences
   If the occurrence overlaps with a
      found atom, remove it.
   If a ring-group occurrence does not
      coincide with the appropriate
      ring involvements, discard it
```

Fig. 10. Compounds are decomposed by finding all occurrences of each group in the compound. The occurrences that are found are filtered according to ring participation and the list of previously matched atoms.

```
SCREEN ring-group
   Skip this group if no rings exist
      in the compound

SCREEN chain-group
   Skip this group if all atoms
      participate in rings.

SCREEN phosphorous-group
   Skip this group if P does not occur
      in the compound.

SCREEN nitrogen-group
   Skip this group if N does not occur
      in the compound.

SCREEN nitrogen-ring-group
   Skip this group if no rings exist
      in the compound.
   Skip this group if N does not occur
      in the compound.
```

Fig. 11. The procedures for screening groups are defined through multiple inheritance.

groups only when there is a chance of finding them in a compound.

Two tests that are routinely used are:

(i) Do not search for a group if a key atom of the group does not exist in the compound;

and,

(ii) Do not search for a group if it requires ring participation and there are no rings in the compound.

These are easily implemented in an object-oriented program by defining subclasses for GROUP that emphasize these relevant differences. RING-GROUP and CHAIN-GROUP subclasses distinguish between groups based on their participation in rings. PHOSPHOROUS-GROUP, SULFUR-GROUP, NITROGEN-GROUP, OXYGEN-GROUP and CARBON-GROUP are based on differences in key atom types.

Some groups exhibit properties of more than one of these classes. Multiple inheritance has been used to define those groups that exhibit properties of more than one superclass. Subclasses have been limited to combinations of one atom type with either a ring or a chain superclass (Figure 1) (Figure 9) (Figure 10).

These new classes allow the definition of an operation called SCREEN that allows an object to determine whether or not it has a chance of being found in a compound before it is submitted to GET-ALL-OCCURRENCES. Only those groups that pass the screen tests are actually sought in the compound structure. The methods that are used to SCREEN GROUPS are delegated to the appropriate subclasses that have the information on how to perform this test (Figure 11).

Compounds are decomposed by identifying and classifying the rings of the structure, and by searching for every occurrence of each group in the structure. As each atom in the compound is assigned to a group, it is marked as being matched. Those occurrences that overlap with previously matched atoms or that do not have the appropriate ring and chain participation are filtered from further consideration. Decomposition time is reduced by screening groups before they are actually searched. The final form of DECOMPOSE that includes the screening operation is provided in Figure 12.

Results and discussion

The program was tested on a knowledge base of 780 compound structures (mostly small metabolites) varying in size from 8 to 155 atoms (Karp, 1992). Ring complexity varied from straight- and branched-chain compounds to species containing four fused rings. A Macintosh LC475 running at 25 MHz took 45 min to estimate Gibbs energies for the entire database using the basis set provided by Mavrovouniotis (1991). These times reflect compiled code with the ring and atom screening tests activated. Similar times on Sun IPX,

```
DECOMPOSE compound with basis-set
   FIND and CLASSIFY rings in compound
   loop for each group in basis-set
      SCREEN group
      GET-ALL-OCCURRENCES of group in
         compound
      FILTER occurrences
      ADD atoms to foundations-list
   RETURN error if there are unmatched
      atoms
```

Fig. 12. The procedures for DECOMPOSE have been amended to incorporate the screen test.

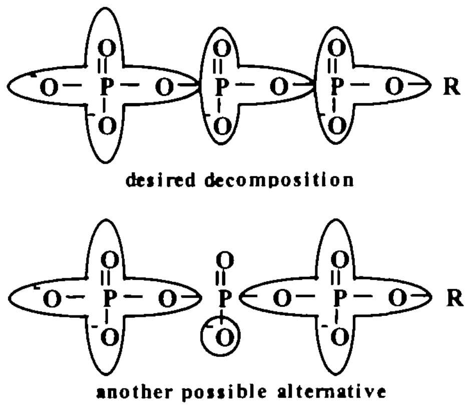

Fig. 13. There is more than one way to decompose phosphorous groups. Humans naturally infer the correct decomposition; however, the automated procedures do not naturally make the correct distinction. Additional groups have been added to the basis set to ensure the correct decomposition.

Phosphate groups	value
-O-PO$_2^{-1}$- (participating in a ring)	14.8
-CO-OPO$_3^{-2}$	-72.5
pentaphosphate	-50.3
tetraphosphate	-45.1
triphosphate	-39.9
diphosphate	-34.7
-OPO$_3^{-2}$ (primary)	-29.5
-OPO$_3^{-2}$ (secondary)	-30.0
-OPO$_3^{-2}$ (tertiary)	-25.7
-PO$_3^{-2}$	9.5
-O-PO$_2^{-1}$-O-PO$_2^{-1}$-O-	-35.0
-O-PO$_2^{-1}$-O-	-29.8
-O-PO$_2^{-1}$-	-5.2

Fig. 14. New groups have been added to the basis set to ensure accurate decomposition of compounds containing phosphorus.

Sparc10 and Silicon Graphics O2 systems were 13, 8 and 2 min, respectively. Besides faster hardware, the difference in the execution times between the Macintosh and the workstations may also be partly attributed to better compilers for the Lucid Common LISP and the AllegroCL4.3 used on the Sun and Silicon Graphics systems.

An unexpected benefit that the software provided was in the area of identifying errors in compound structures that were stored in the database. The software identified 10 structures as 'not decomposable' that were later found to contain errors. This is a limited benefit since some errors will result in decomposable structures and still remain undetected.

It should be noted that the Mavrovouniotis (1991) basis set for computing Gibbs energies has been amended because it was found that the published definitions for phosphate groups can result in erroneous decompositions. Although it has been established that the order in which groups are searched is important, it was later discovered that the order in which atoms are identified is equally important. For example, there is more than one way to decompose chains of phosphate groups (Figure 13).

Although humans can infer the correct decomposition, the initial computer implementation could not. To address this difficulty, the basis set was amended by defining larger groups that eliminate the confusion by matching the specific configurations that were intended. The new groups use appropriate sums of the smaller groups as their contribution terms (Figure 14).

The entire program has been designed to be reusable. The same code has already been used to implement two additional basis sets that calculate eight different physical properties. These include the Joback group contributions for critical properties (P_c, T_c and V_c), the normal boiling point and the freezing point (Joback, 1984); and the Joback method for ideal-gas properties [$\Delta G_f°$(298K), $\Delta H_f°$(298K) and $C_p°$] (Joback, 1984). The current implementation can use additional basis sets, provided that these group sets satisfy the following requirements:

(i) All non-overlapping occurrences of a group within a reference compound must be identified. Conversely, every atom in the compound is assigned to exactly one group.

(ii) Wildcards may be used to indicate when the atom to which a group is attached is not important. This occurs in basis sets that search for specific bond arrangements of atoms.

(iii) When a larger substructure is used to provide context, focal atoms may be used to specify the target structure of the group.

(iv) Ring groups and chain groups that use similar query structures must be distinguishable.

Acknowledgements

This work has been supported in part by the National Institutes of Health (National Library of Medicine grant LM 05278, National Center for Research Resources grant 1-R01-RR07861-01 and by the National Science Foundation grant BES-9531787).

References

Benson,S.W. (1968) *Thermochemical Kinetics*. John Wiley & Sons, New York.Benson,S.W. (1968) *Thermochemical Kinetics*. John Wiley & Sons, New York.

Forsythe,R.G., Prickett,S.E. and Mavrovouniotis,M.L. (1992) *An Introduction to Object-Oriented Programming in Process Engineering*. CACHE Corp., Austin, TX, 112 pp.

Gasteiger,J. and Jochum,J. (1979) An algonthm for the perceptıon of synthetically important rings. *J. Chem. Inf. Comput. Scı.*, **19**, 43–48.

Joback,K.G. (1984) A unified approach to physical property estimation using multivariate statistical techniques. MS Thesis ın chemical engineering, MIT, Cambridge, MA.

Karp,P. (1992) A knowledgebase of the chemical compounds of intermediary metabolism. *Comput Applic. Biosci.*, **8**, 347–357.

Mavrovouniotis,M.L. (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.*, **36**, 1070–1082.

Mavrovouniotis,M.L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol Chem.*, **266**, 14440–14445.

Mavrovouniotis,M.L. (1993) Identification of localized and distributed bottlenecks in metabolic pathways. In *Proceedings of the Fırst International Conference on Intelligent Systems for Molecular Biology (ISMB-93)*. AAAI Press, Menlo Park, CA, pp. 275–283.

Mavrovounıotis,M.L. (1996) Duality theory for thermodynamic bottlenecks in bioreaction pathways. *Chem. Eng. Sci.*, **51**, 1495–1507.

Reed,T.A. (1988) *An Introduction to Algorithm Design and Structured Programming*. Prentıce Hall, New York.

Reid,R.C., Prausnitz,J.M. and Poling,B.E. (1987) *The Properties of Gases and Liquids*, 4th edn. McGraw-Hill, New York.

Weinınger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf Comput. Sci.*, **28**, 31–36.