Fast Evaluation of Chemical Distance by Tabu Search Algorithm[†]

Vladimír Kvasnička* and Jiří Pospíchal

Department of Mathematics, Slovak Technical University, 812 37 Bratislava, Slovakia

Received January 15, 1994[⊛]

The concept of chemical distance is a principal formal tool of mathematical chemistry oriented to organic synthesis design problem. Applications of standard combinatorial methods (like the backtrack algorithm) to the evaluation of chemical distance are ineffective for larger molecular graphs. A tabu search algorithm is applied for the evaluation of chemical distance. For smaller molecular graphs (say, up to 10–12 vertices) it provides results that are closely (or even precisely) related to their exact values and moreover obtained in surprisingly short CPU time.

1. INTRODUCTION

The concept of *chemical distance* has been developed by Dugundji and Ugi[1–3] (see also refs 4–9 and for topological-chemical distance[10]) as a proper measure of similarity between two molecular graphs. Recently, this concept has become very important[3,6,11–13] as the main driving force for the construction of the so-called *reaction network* composed of such intermediate molecular graphs that are constituents of a reaction mechanism of the reaction: educt → product.

Let G_1 and G_2 be two molecular graphs composed of the same number N of vertices—atoms (we say that G_1 and G_2 are *isomeric* or *stoichiometric*). The corresponding adjacency matrices are denoted by $\mathbf{A}_1 = (A_{ij})$ and $\mathbf{A}_2 = (A'_{ij})$, respectively. Let $P = (p_1,p_2,\ldots,p_N)$ be a permutation of N objects. The *chemical distance* between G_1 and G_2 is defined as a minimum value of the L_1 norm of a "difference" of adjacency matrices for all possible permutations[1]

$$\mathrm{CD}(G_1,G_2) = \min_P \sum_{\substack{i,j=1\\(i<j)}}^{N} |A_{ij} - A'_{p_ip_j}| \tag{1}$$

We emphasize that if G_1 and G_2 are composed of atoms of different chemical nature, then the permutation relates only vertices of G_1 and G_2 that are of the same chemical nature. An alternative graph-theoretical definition of the chemical distance is[4–7]

$$\mathrm{CD}(G_1,G_2) = |E(G_1)| + |E(G_2)| - 2|E(G_1 \cap G_2)| \tag{2}$$

where $|E(G)|$ denotes the cardinality of the edge set of a graph G, and $G_1 \cap G_2$ is the maximum common subgraph (not induced!) of G_1 and G_2. Both these definitions are equivalent and it was proved[4–7] by making use of (2) that the chemical distance is a metric (it is nonnegative, zero only for isomorphic graphs, symmetric and satisfies the triangular inequality).

The evaluation of chemical distance belongs to the set of NP complete problems, i.e., its correct evaluation needs N! (an upper bound) matchings of vertices between G_1 and G_2. Many different exact methods[14–20] have been suggested for the evaluation of chemical distance. All these methods are unable to surmount the barrier of NP completeness, i.e., for larger graphs (say $N > 12$–15) they are ineffective. Recently, a genetic algorithm[21,22] and simulated annealing[23] were successfully applied for suboptimal evaluation of the chemical distance.

The purpose of the present communication is to describe an application of the tabu search algorithm[24,25] for an extremely fast suboptimal evaluation of chemical distance between smaller molecular graphs (up to 10–12 vertices). Although the effectiveness of the tabu search approach cannot be directly related to the effectiveness of the above mentioned two stochastic optimizations for larger graphs, the tabu search offers a very effective method for suboptimal evaluations of chemical distances giving results in fractions of CPU time of any other methods.

2. TABU SEARCH ALGORITHM

The *tabu search*[24,25] is an interesting approach which is recently receiving significant attention as an effective heuristic suitable for optimization of combinatorial problems. Its effectiveness has been demonstrated by many problems of operations research and graph theory (see ref 25 for extensive list of applications).

The basic entity that should be optimized by the tabu search is a permutation $P = (p_1,p_2,\ldots,p_N)$ of N objects. We assign to the permutation P a current chemical distance determined by the right-hand side of (1)

$$\mathrm{CD}(P) = \sum_{\substack{i,j=1\\(i<j)}}^{\mathrm{N}} |A_{ij} - A'_{p_ip_j}| \tag{3}$$

Then the exact chemical distance $\mathrm{CD}(G_1,G_2)$ is equal to the minimum of $\mathrm{CD}(P)$ over all ($N!$) permutations from the Nth degree symmetric group. A current permutation $P = (p_1,p_2,\ldots,p_N)$ is transformed into another permutation $P' = (p'_1,p'_2,\ldots,p'_N)$ by applying a transposition operator t_{ij} (for any $1 \le i < j \le N$)

$$\begin{aligned} t_{ij}(P) &= t_{ij}(p_1,\ldots,\overbrace{p_i,\ldots,p_j},\ldots,p_N) \\ &= (p_1,\ldots,p_j,\ldots,p_i,\ldots,p_N) \end{aligned} \tag{4}$$

Current chemical distances $\mathrm{CD}(P)$ an $\mathrm{CD}(P')$ are simply related by

[†] Dedicated to Professor Milan Kratochvíl on the occasion of his 70th birthday.

[⊛] Abstract published in *Advance ACS Abstracts,* August 15, 1994.

0095-2338/94/1634-1109$04.50/0

$$\mathrm{CD}(P') = \mathrm{CD}(P) + \sum_{\substack{k=1 \\ (k\neq i,j)}}^{N} [|A_{ik} - A'_{p_jp_k}| - |A_{ik} - A'_{p_ip_k}| + |A_{jk} - A'_{p_ip_k}| - |A_{jk} - A'_{p_jp_k}|] \quad (5)$$

All transposition operators form a set of feasible *transformations*

$$S = \{t_{12},t_{13},...,t_{N-1,N}\} \quad (6)$$

which contains $N(N-1)/2$ transformations. A *neighborhood* of the permutation P is composed of all permutations $P', P'',...$ that are created from P by all transformations $t\epsilon S$

$$U(P) = \{P'; \forall t\epsilon S\text{: } P' = t(P)\} \quad (7)$$

Its cardinality is the same as the cardinality of S, i.e., $|S| = |U(P)|$.

At this moment we are able to formulate the so-called *hill climbing* algorithm for finding a suboptimal solution of the problem (1) rewritten as follows

$$\mathrm{CD}(G_1,G_2) = \min_{P} \mathrm{CD}(P) \quad (8)$$

where the current chemical distance is determined by (3). For a randomly generated permutation P we look for a minimum of the current chemical distance not in the entire symmetric group but only in the neighborhood of the permutation P determined by (7)

$$\mathrm{CD}(P^*) = \min_{P'\epsilon U(P)} \mathrm{CD}(P') \quad (9)$$

The obtained permutation P^* is subsequently used in the next iteration (single hill climbing) as a "center" of new neighborhood $U(P^*)$. This process is repeated for a prescribed number of iterations, as seen in Figure 1. In general, the hill climbing algorithm provides a sequence of locally optimal current chemical distances that are oscillating. Its PASCAL pseudocode appears as follows

Hill climbing algorithm.

```
P:= randomly generated permutation;
time:=0; $CD_{min}$:=∞;
WHILE time<$time_{max}$ DO
BEGIN time:=time+1;
      $CD(P^*) = \min_{P'\epsilon U(P)} CD(P')$;
      IF $CD(P^*)<CD_{min}$ THEN
      BEGIN $CD_{min}$:=$CD(P^*)$;
            $P_{min}$:=$P^*$
      END;
      P:=$P^*$;
END;
```

The variables $\mathrm{CD}_{\min}$ and $P_{\min}$ record the best solution achieved in the whole run of the algorithm recurrently applied $t_{\max}$ times. The main shortcoming of this simple heuristic optimization algorithm is a *cycling* problem. After a finite number of iteration steps, the algorithm returns to a permutation already achieved as a local best solution in the previous iteration step. One of the very simple ways around this serious shortcoming of the hill climbing algorithm is the so-called *stochastic hill climbing*. After a prescribed number of iteration steps, the hill climbing is stopped, and it is again initiated by a randomly generated permutation. The best solution results as a record of all these independent performances of hill climbings.

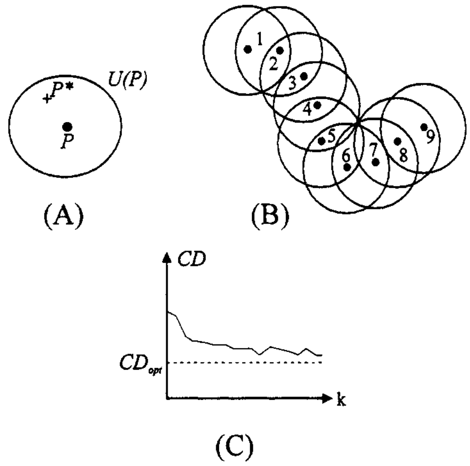

Figure 1. An outline of the hill climbing algorithm. (A) Local search within neighborhood $U(P)$, $P^*\epsilon U(P)$ corresponds to the best solution, see eq 9. (B) The hill climbing is a recurrent application of local searching. The best solution achieved in the kth iteration is used as a "center" of neighborhood in the next $(k + 1)$th iteration. (C) A plot of chemical distances vs number of iterations, chemical distances sharply decrease in the beginning, and at the end a cycling of solutions appears and the tail of the graph is either smooth or a zigzag line parallel with the horizontal axis.

The tabu search algorithm[24,25] offers another interesting possibility for overcoming the above mentioned obstacle of the hill climbing. The main limitation of hill climbing in combinatorial problem solving is that the local minimum obtained after a prescribed number of iterations, when no improving transformations are possible, is quite probably not a global minimum. Tabu search offers a heuristic to continue searching without returning back into a local minimum from which it previously emerged. The main concept for doing this is a *tabu list T* (called the *short-memory approach*), which stores temporarily some transformations already used in the previous iteration steps. The tabu list of transformations $t\epsilon S$ of the size k is constructed and updated during the performance of the algorithm. If a transformation t belongs to the tabu list, $t\epsilon T$, for a given iteration, it is not allowed to be used in local searching at that iteration. At the beginning of the algorithm the tabu list is empty. It is then constructed in k subsequent iterations and updated cyclically in the next iterations. An addition of each transformation thus erases the transformation recorded in the position k iterations ago. Numerical experiences indicate that a broad range of tabu sizes k exists for which the tabu list performs very effectively for driving the search beyond local minima and for obtaining progressively improved solutions in a nonmonotonic sequence. The size of the tabu list belongs to principal parameters of the tabu search. If it is too small, cycling will occur; if it is too big, it will restrict the search enough to skip the "deeper valleys" (better local minima) of the minimized objective function $\mathrm{CD}(P)$. The tabu list T is used for the construction of a modified neighborhood $U_T(P)$ of the current permutation P

$$U_T(P) = \{P'; \forall t\epsilon S\backslash T\text{: } P' = t(P)\} \quad (10)$$

It is composed of permutations P' that can be formed by making use of transformations from the set S but not belonging to the tabu list T. The local searching is performed for the modified neighborhood $U_T(P)$ but with an exception called the *aspiration level criterion*. It overcomes the tabu list restriction for a transformation $t \epsilon T$ at a given iteration if the resulting permutation $P' = t(P)$ gives a current chemical distance strictly smaller than smallest value CD_{min} obtained so far.

Tabu search algorithm

```
1    P:= randomly generated permutation;
2    time:=0; $CD_{min}$:=∞; T:=∅;
3    WHILE time<$time_{max}$ DO
4    BEGIN time:=time+1; $CD_{loc-min}$:=∞;
5           FOR t∈S DO
6           BEGIN P':=t(P);
7                  IF ((t∉T) and CD(P')<$CD_{loc-min}$) or
                      (CD(P')<$CD_{min}$) THEN
8                  BEGIN $P^*$:=P'; $t^*$:=t;
9                         $CD_{loc-min}$:=CD(P');
10                 END;
11          END;
11          IF $CD_{loc-min}$<$CD_{min}$ THEN
12          BEGIN $CD_{min}$:=$CD_{loc-min}$;
13                 $P_{min}$:=$P^*$;
14          END;
15          P:=$P^*$;
16          IF |T|<k THEN T:=T∪{$t^*$} ELSE
                                T:=T∪{$t^*$})\{$t_{oldest}$};
17   END;
```

The tabu search algorithm is very similar to the above given hill climbing. The main difference here is a local searching combined with aspiration level criterion given in lines 5–11. In line 16 an updating of tabu list is expressed in simple set-theory formalism. If the cardinality of tabu list T is smaller than the tabu size k, then the tabu list is simply enlarged by the currently used transformation t^*. In the opposite case, if the cardinality of T is equal to k, then the addition of t^* to T is simultaneously accompanied by a subtraction of the oldest transformation t_{oldest}, which has been introduced to T just k iterations ago.

We have presented here the tabu search algorithm only in its basic form, further sophistications of the algorithm are extensively studied in the literature (cf. ref 25). In particular, a concept of the *long-term memory* is often used as a tool to intensify and diversify the search. The main principle in the approach is to endow the tabu search by a "memory" with a flexibility to reject the most frequently used transformations. An acceptance process in looking for the locally best solution is then based not only on objective function of CD(P) changes but also on the previous history of search. A very simple and straightforward realization of these general ideas is an application of transformation frequencies. Let f_{ij} be a frequency of the transformation t_{ij}. At the beginning of tabu search these entities are set equal to zero, then in the course of iteration process for each t^*_{ij} the corresponding frequency is updated, $f_{ij} \leftarrow f_{ij} + 1$. After a prescribed number of iterations these frequencies determine how many times single transformations t_{ij} have been used in the local search process. Then, the frequencies f_{ij} are used as penalty functions in looking for the locally best solution within a neighborhood $U_T(P)$. A permutation $P' = t_{ij}(P) \epsilon U_T(P)$, where $t_{ij} \epsilon S \backslash T$, is accepted as a temporarily locally best solution is the following condition is satisfied

$$CD(P') + \alpha f_{ij} < CD(P^*) \qquad (11)$$

where α is an empirically determined small positive constant. Thus the minimization described in lines 5–11 in the above tabu search algorithm actually minimizes the function CD(P') + αf_{ij}, but the recorded locally best value of the current chemical distance within $U_T(P)$ equals only CD(P'), CD(P^*):=CD(P'). This means that the most frequently used transformations are potentially rejected according to their high performance frequencies. The long term memory approach gives a chance to other transformations than to those already frequently used in the previous iteration steps.

Though this accelerating approaches appears on first sight very logical and although its implementation to tabu search algorithm is simple, according to our numerical experiences it did not affect substantially the numerical effectiveness of the tabu search for finding chemical distance between molecular graphs. Of course, this observation is not neccessarily true for other combinatorial problems with complicated and diverse searching domains, which is the case of almost all operations research problems.

3. APPLICATION AND SUMMARY

The tabu search algorithm has been extensively tested on randomly generated molecular graphs with prescribed numbers of vertices and edges. An exact upper bound of chemical distances between these randomly generated graphs has been constructed as follows: A graph G_1 with prescribed number of vertices and edges is generated. The second graph G_2 is created from G_1 by moving p randomly selected edges to unoccupied positions. Then, the chemical distance CD(G_1,G_2) is bounded from above by[6]

$$CD(G_1,G_2) \leq 2p \qquad (12)$$

where the relation "smaller than" is usually satisfied if G_1 or G_2 has a nontrivial automorphism. This means that we are able to evaluate an exact upper bound of chemical distance CD(G_1,G_2) for each pair of randomly generated graphs which differ in positions of p edges. In most cases, the correct value of chemical distance CD(G_1,G_2) will achieve the upper bound $2p$. In the performance runs of the algorithm, whenever the computed distance CD(P) is less than or equal to the upper bound $2p$, we shall treat the computed distance as correct, even though in some cases the actual distance will be slightly less than the computed distance. Tabu search parameters $time_{max}$ (maximum number of iterations) and k (size of tabu list) are $time_{max} = 300$ and $k = N(N-1)/4$, where N is the number of vertices of graphs.

Our numerical experiences with tabu search algorithm when applied for the calculation of chemical distances between randomly generated molecular graphs may be summarized as follows: The tabu search algorithm provides for smaller molecular graphs (up to 10–12 vertices), in the course of at most a few tens of iterations, correct values of chemical distances, see Table 1. A frequency of appearance of incorrect values within the prescribed 300 iterations was roughly 2%, in particular for pairs of isomorphic graphs (that is with $p = 0$ and CD = 0) and for a number of vertices greater than 10. For graphs composed of more than 15 vertices the effectiveness of the tabu search algorithm quickly decreases so that the failure of algorithm to provide correct chemical distances after 300 iterations is greater than 20%. Here, it is necessary to emphasize that these incorrect values are usually very close to the correct ones (in ~99% of the cases the incorrect values

Table 1. Median Values of a Number of Iterations of Tabu Search Algorithm for 50 Randomly Generated Graphs with Prescribed Numbers of Vertices, Edges, and Edge Shifts[a]

		number of edges				
no. of vertices	no. of shifted edges	7	8	9	10	11
8	0	10	8	11	11	7
	1	9	6	9	12	14
	2	3	3	4	4	4
	3	1	2	2	3	2
	4	1	1	1	1	2

		number of edges				
no. of vertices	no. of shifted edges	8	9	10	11	12
9	0	13	20	22	12	14
	1	15	11	21	19	20
	2	6	6	5	7	10
	3	2	3	3	4	4
	4	1	2	2	2	2

		number of edges				
no. of vertices	no. of shifted edges	9	10	11	12	13
10	0	23	13	19	15	24
	1	15	20	32	23	23
	2	8	11	14	20	19
	3	3	4	5	6	8
	4	2	2	3	3	3

[a] The algorithm ended after achieving correct chemical distance.

are only two units off from the correct results, i.e., instead CD = $2p$ we have obtained CD = $2(p + 1)$).

In summary, the tabu search algorithm represents for smaller molecular graphs an extremely fast numerical method for the calculation of chemical distances with very low incidence of incorrect results (smaller than 2%). Moreover, the incorrect results are usually closely related to the correct values. The tabu search algorithm opens new unexpected possibilities for the generation of reaction networks composed of intermediates of chemical reaction $G_1 \rightarrow G_2$, where G_1 and G_2 are educt and product molecular graphs, respectively.[11-18] Their construction for complex reactions (with many thousands of intermediates) has been up to now a prohibitive computational task mainly due to the low performance of known combinatorial algorithms[14-20] evaluating the chemical distance.

A sample TURBO PASCAL code of tabu search algorithm for calculation of chemical distance between two randomly generated graphs can be obtained via email (*kvasnic@cvt.stuba.sk* or *pospich@cvt.stuba.sk*) from the authors.

ACKNOWLEDGMENT

This reasearch was supported in part by the 1994 Chemical Structure Association Trust Award. We wish to express our thanks to Professor Fred Glover, University of Colorado, Boulder, for sending us preprints of his papers and other supplementary materials.

REFERENCES AND NOTES

(1) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.

(2) Jochum, C.; Gasteiger, J.; Ugi, I.; Dugundji, J. The Principle of Minimal Chemical Distance and the Principle of Minimal Structural Change. *Z. Naturforsch.* **1982**, *37B*, 1205–1215.

(3) Ugi, I.; Wochner, M. A.; Fontain, E.; Bauer, J.; Gruber, B.; Karl, R. Chemical Similarity, Chemical Distance, and Computer Assisted Formalized Reasoning by Analogy. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley: New York, 1990; pp 239–288.

(4) Johnson, M. A. Relating Metrics, Lines and Variables Defined on the Space of Graphs. In *Graph Theory and its Applications to Algorithms and Computer Science*; Alavi, Y., Chartrand, G., Lesniak, L., Wall, C., Eds.; Wiley: New York, 1985; pp 457–470.

(5) Baláž, V.; Koča, J.; Kvasnička, V.; Sekanina, M. A metric for Graphs. *Časopis Pěst. Matem.* **1986**, *111*, 431–433.

(6) Koča, J.; Kratochvíl, M.; Kvasnička, V.; Matyska, L.; Pospíchal, J.; *A Synthon Model of Organic Chemistry and Synthesis Design.* Lecture Notes in Chemistry 51; Springer Verlag: Berlin, 1989.

(7) Baláž, V.; Kvasnička, V.; Pospíchal, J. Dual Approach for Edge Distance Between Graphs. *Časopis Pěst. Matem.* **1989**, *114*, 155–159.

(8) Kvasnička, V.; Pospíchal, J. Two Metrics for a Graph-theoretical Model of Organic Chemistry. *J. Math. Chem.* **1989**, *3*, 161–191.

(9) Baláž, V.; Kvasnička, V.; Pospíchal, J. Two Metrics in a Graph Theory Modeling of Organic Chemistry. *Discrete Applied Mathematics* **1992**, *35*, 1–19.

(10) Balaban, A. T.; Bonchev, D.; Seitz, W. A. Topological/Chemical Distances and Graph Centers in Molecular Graphs with Multiple Bonds. *J. Molec. Structure* (*Theochem*) **1993**, *99*, 253–260.

(11) Kvasnička, V.; Pospíchal, J.; Baláž, V. Reaction and Chemical Distances and Reaction Graphs. *Theor. Chim. Acta* **1991**, *79*, 65–79.

(12) Fontain, E.; Bauer, J.; Ugi, I. Computer-Assisted Bilateral Generation of Reaction Networks from Educts and Products. *Chem. Lett.* **1987**, 37–40.

(13) Fontain, E.; Reitsam, K. The Generation of Reaction Networks with Rain. 1. The Reaction Generator. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 96–101.

(14) Pospíchal, J.; Kvasnička, V.; Gutman, I. Chemical Distance for Molecules Whose Classical Structural Formula is not Unique. *Collection of Scientific Papers of the Faculty of Science Kragujevac* **1992**, *12*, 109–125.

(15) McGregor, J. J. Backtrack Search Algorithm and the Maximal Common Subgraph Problem. *Soft. Pract. Exp.* **1982**, *12*, 23–34.

(16) Tsai, C.-C.; Nicholson, V.; Johnson, M.; Naim, M. A Subgraph Isomorphism Theorem for Molecular Graphs. In *Graph Theory and Topology in Chemistry*; Elsevier: Amsterdam, 1987; pp 226–230.

(17) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD). *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495–505.

(18) Wochner, M.; Brandt, J.; v.Scholey, A.; Ugi, I. Chemical Similarity, Chemical Distance and Its Exact Determination. *Chimia* **1988**, *42*, 217–225.

(19) Ugi, I.; Fontain, E.; Bauer, J. Transparent Formal Methods for Reducing the Combinatorial Wealth of Conceivable Solutions to a Chemical Problem-Computer Assisted Elucidation of Complex Reaction Mechanisms. *Anal. Chim. Acta* **1990**, *235*, 155–161.

(20) Kvasnička, V.; Pospíchal, J. Maximal Common Subgraphs of Molecular Graphs. *Reports in Molecular Theory* **1990**, *1*, 99–106.

(21) Fontain, E. The Problem of Atom-to-atom Mapping. An application of genetic algorithms. *Anal. Chim. Acta* **1992**, *265*, 227–232.

(22) Fontain, E. Application of Genetic Algorithms in the field of Constitutional Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 748–752.

(23) Pospíchal, J.; Kvasnička, V. Fast Evaluation of Chemical Distance by Simulated Annealing Algorithm. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 879–885.

(24) (a) Glover, F. Tabu Search—Part I. *ORSA J. Comp.* **1989**, *1*, 190–206; 2(1990)4. (b) Glover, F. Tabu Search—Part II. *ORSA J. Comp.* **1990**, *2*, 4–32.

(25) Glover, F.; Laguna, M. Tabu Search. In *Modern Heuristic Techniques for Combinatorial Problems*; Reeves, C. R., Ed.; Blackwell Scientific Publications: Oxford, UK, 1993; pp 70–141.