Application of Genetic Algorithms in the Field of Constitutional Similarity[†]

Eric Fontain

Institute of Organic Chemistry, Technical University of Munich,
Lichtenbergstrasse 4, W-8046 Garching, Germany

Received June 12, 1992

Constitutional similarity is defined in terms of the minimum chemical distance (MCD) between isomeric ensembles of molecules (EM). Genetic algorithms are introduced as a new method for the calculation of MCD. These algorithms mimic the process of natural evolution within an artificial population of genome vectors. It is shown how the application of the concept of constitutional similarity can substantially increase the efficiency of reaction generation in the field of reaction mechanism elucidation with the program RAIN.

INTRODUCTION

Within chemistry a large variety of more or less well-defined similarity measures are in use.[1] The concept of chemical similarity is applied in the areas of drug design,[2] molecular shape analysis,[3] structure–activity studies,[4] database retrieval,[5] reaction mechanism elucidation,[6] etc. This paper will give answers to the following three questions:

(1) How can we define constitutional similarity?

(2) With a given constitutional description of the molecules in terms of connection tables, how can we measure constitutional similarity?

(3) What is a typical application of the measured value of constitutional similarity?

When looking for a measure of constitutional similarity, one must either look for structural features that the molecules have in common[7] or one must consider differences in the bonds of the structures. The Minimum Chemical Distance[8] (MCD) affords a mathematically based definition of constitutional similarity. It measures the constitutional differences of isomeric ensembles of molecules (EMs) by counting the number of electrons that have to be redistributed in order to convert EM *A* into EM *B*. When the EMs *A* and *B* are represented by their corresponding *BE* matrices,[8] the MCD is defined as the minimum sum of the absolute values of the difference matrix of *BE*(*A*) and *BE*(*B*):

$$\mathrm{MCD}_{A,B} = \min \sum_{ij} |BE(A)_{ij} - BE(B)_{\mathbf{p}(i)\mathbf{p}(j)}|$$

A permutation vector **p** is introduced that assigns the atoms of *A* onto the atoms of *B* such that a minimum value for the chemical distance is achieved. If we neglect the contribution of the free valence electrons, the MCD is defined as twice the sum of the minimum number of bonds to be *broken* and bonds to be *made* in the course of the transformation. A small chemical distance indicates two closely related structures, while very dissimilar structures correspond to a large chemical distance.

Figure 1 shows a reference structure **1** (ampicillin) together with a list of molecules with increasing MCDs. The first structure **1′** in the list is identical to the reference. Thus, it has a chemical distance of 0. The second molecule **2** is a tautomer and, thus, yields a relatively small chemical distance of 8, which indicates its close relationship to ampicillin. In structure **3** the thiazole ring has been replaced by a thiazolidine system with a chemical distance of 16. The molecules **4**–**7** in the list are more and more dissimilar to the reference structure. The last structure **7** in the list does not resemble ampicillin at all.

MINIMIZING THE CHEMICAL DISTANCE

Calculation of the MCD is a special case of an optimization. The structures that are to be compared are superimposed such that a maximum structural overlap is achieved. In mathematical terms this task is called a quadratic assignment problem and has the adverse property of NP-completeness. The arising combinatorial difficulties can easily be demonstrated by the example shown in Figure 2. The atoms of **8** can be mapped in up to 10^{49} different ways onto the atoms of **9**. However, only two of these assignments yield the minimum distance of 8. Naturally, it is impossible to scan all the mappings in a "brute force" approach, calculate for each mapping the number of redistributing electrons, and compare all these numbers, in order to find the minimum. We need a fast and robust algorithm that either scans the whole solution space very efficiently or has specific techniques that exploit the chemical nature of the underlying graphs.

Since the early 1980s there have been two approaches for overcoming the combinatorial problems in the calculation of minimum chemical distance. The first was published in 1980 by Clemens Jochum.[9] He tried to apply "traditional" methods taken from operations research, e.g., branch-and-bound techniques and perturbation calculations. The second approach was published in 1988 by Micaela Wochner.[10] Her algorithm was based on a modified relaxation procedure for the canonization of chemical structures. The basic concept was to locate common substructures, thus reducing the problem dimension, and then to find the minimum chemical distance by complete enumeration of the remaining assignments. Both approaches require an appreciable amount of computing time for larger molecules, especially when the similarity of the compared structures is rather small.

BASIC GENETIC ALGORITHM

When looking for a robust and efficient procedure that solves this quadratic assignment problem, we may take as a paradigm the process of natural evolution performing parallel and dynamic optimization of the genoms of an enormous number of creatures. Among the main mechanisms that drive this optimization process are

(1) selective reproduction that guarantees the "survival of the fittest"

[†] Presented May 28, 1992, at the 3rd Beilstein Workshop on Similarity in Organic Chemistry, Bolzano, Italy.

0095-2338/92/1632-0748$03.00/0

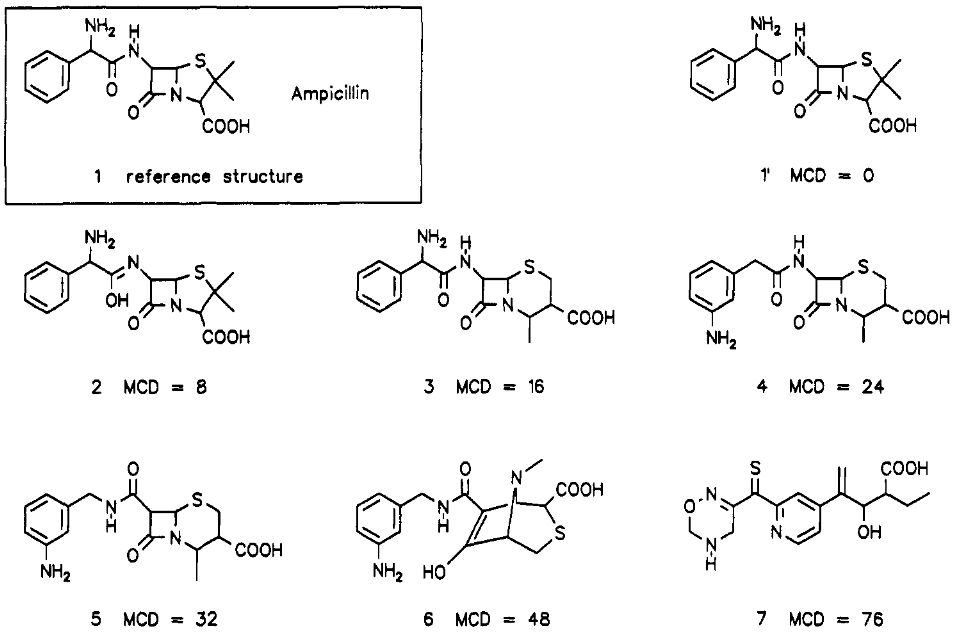


Figure 1. Reference structure ampicillin and a list of molecules that show increasing MCD, indicating their decreasing constitutional similarity to ampicillin.

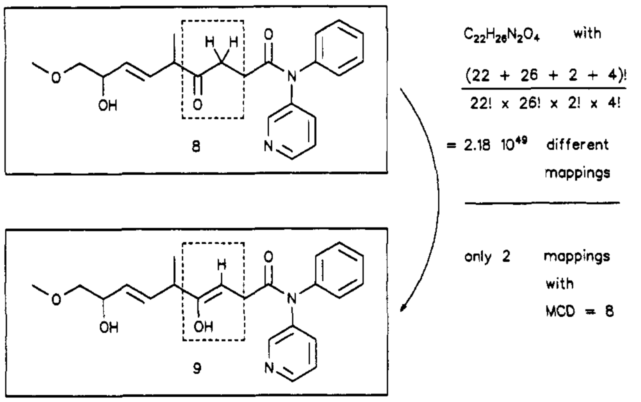

8
9
$C_{22}H_{26}N_2O_4$ with
$\frac{(22 + 26 + 2 + 4)!}{22! \times 26! \times 2! \times 4!}$
= 2.18 10^{49} different mappings
only 2 mappings with MCD = 8

Figure 2. Combinatorial problems in the atom mapping process.

(2) mutation introducing new genetic information by random

(3) crossover that guarantees the efficient mixing and outspreading of the genetic information

It was already in the early 1970s when John Holland[11] and his co-workers at the University of Michigan invented a basic concept for optimization procedures that completely rely on the rules of natural evolution. They called these types of procedures genetic algorithms.

A basic genetic algorithm describes rules for the selective reproduction and for mutation and crossover events within an artificial population of genome vectors. Each line within a box in Figure 3 represents one genome vector. A genome vector in most cases is a bit string that is constructed by a linear concatenation of all independent parameters that describe the problem to be optimized. In our case, a genome vector is a permutation vector **p** that assigns each atom of one structure to an atom of the other structure, thus defining an atom-to-atom mapping for one trial solution.

At first, the population of vectors is initialized with random sequences. Of course, attention has to be paid that the atoms are only assigned to atoms of the same element type.

The population then undergoes the process of selective reproduction. For each permutation vector, the corresponding number of electron movements is calculated. All these numbers within the population are compared to each other, and the permutation vectors are copied to a second population according to their relative "fitness". "Better" permutation vectors with large amounts of structural overlap receive more copies than permutation vectors that have a fitness below the average. "Very bad" permutation vectors may even die out, i.e., they are not copied to the second population. This new population differs from the first population only with regard to the numbers of the occurrences of the permutation vectors. The internal information of the individual permutation sequences remains constant during reproduction.

New genetic information is now introduced by a mutation operator. It randomly switches two positions within randomly selected permutation vectors. The mutation events occur with a specific probability, which, in most cases, is very low. The validity of the resulting vector is guaranteed, because only assignments of atoms that belong to the same element type may be switched.

A second reordering operator, which is activated more frequently, is the crossover operator. It randomly picks two permutation vectors from the population. These are both cut at two random positions. The equally sized middle pieces are exchanged. Then, within each of the two permutation vectors, homology has to be reestablished by changing the appropriate vector positions. This has to be done, because a valid permutation vector must contain each number exactly once.[12]

The second population now reenters the generation cycle. It is again subjected to the pressure of selective reproduction. Then mutation and crossover occur, and the third generation is complete. Within each generation cycle the best permutation vector is memorized. The average performance of all permutation vectors is constantly increasing, while the population scans the search space with implicit parallelism.

These genetic algorithms are able to find global optimums in huge search spaces very fast, and are, in most cases, insensitive to being trapped in local optimums. Of course, there are no criteria to decide whether the best found permutation vector represents the real global minimum of chemical distance. However, there exist many possible applications where it is sufficient to have a value near the global optimum.

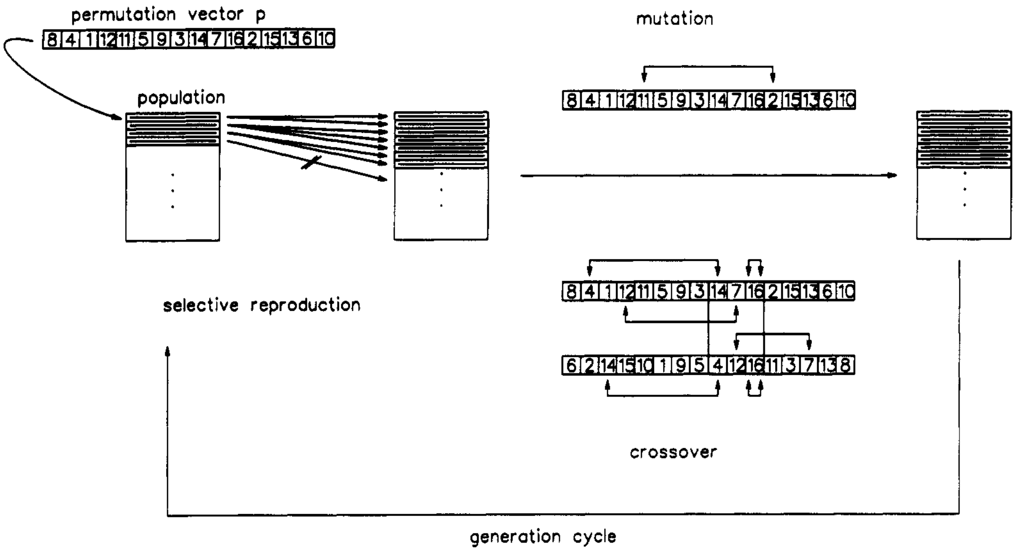


Figure 3. Basic genetic algorithm.

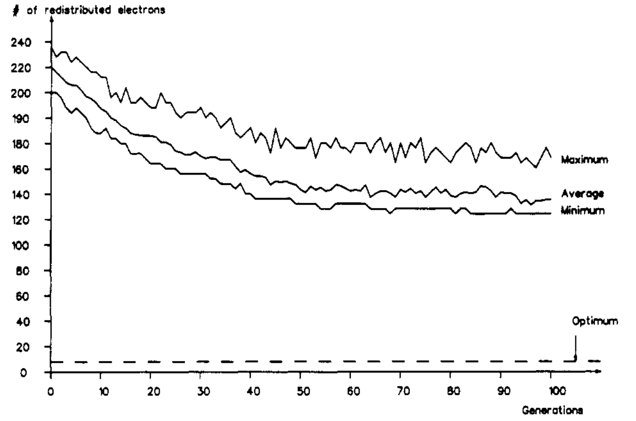


Figure 4. Result of basic genetic algorithm applied on problem in Figure 2.

The first experiments[13] with genetic algorithms in the minimization of chemical distance were very encouraging. With molecular structures of sizes up to approximately 30 atoms, the optimum values were found within less than 100 generation cycles. The afforded computing times were low (about several seconds on an IBM PC), but when we tried larger structures with about 50–100 atoms, the efficiency decreased as can be seen in Figure 4. The experimental setup contained a population of 100 permutations vectors. In each generation cycle, every vector was subjected to crossover with a probability of 60% and to mutation with a probability of 30%.

Figure 4 shows a typical result of a genetic algorithm experiment. The chosen structure mapping problem for this experiment was that in Figure 2. The curves indicate for each generation the lowest, the highest, and the average value of the chemical distance within the whole population. The optimum value of 8 for the minimum chemical distance is indicated by a dashed line. We can see a rapid decrease of all values within the first 30–40 generations, which is a result of the increasing fitness of the population. This demonstrates that the mechanisms of reproduction and genes mixing work very well. Nevertheless, this result is disappointing. The slope of the curves is getting less and less. Long before it reaches the optimum value, the whole population drops into an indifferent stage on a rather high level. However, as a result of the properties of genetic algorithms, there is no doubt that the optimum will be reached, but the number of required generation cycles would be much too large.

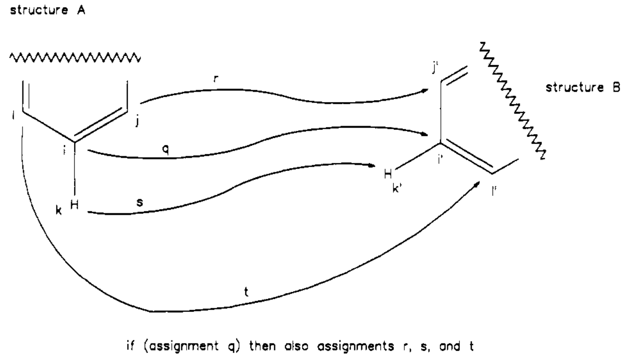


Figure 5. Reordering operator that preserves the neighborhood of an atom during assignment.

KNOWLEDGE-AUGMENTED GENETIC ALGORITHM

We must take into account that the basic genetic algorithm that was applied in the last section did not use any information about the two structures to be matched. However, we should remember that finding an atom-to-atom mapping that *minimizes* the structural differences automatically implies a search for a *maximum* structural overlap. This, in most cases, results in larger structural entities that can directly be mapped onto each other without change. A modified, so-called knowledge-augmented genetic algorithm can exploit this fact.

For this purpose a new operator is introduced that preserves the neighborhood of an atom of structure **A**, when it is assigned to an atom of structure **B** (see Figure 5). This operators works as follows: if an atom i in structure **A** is mapped onto an atom i′ in structure **B** according to assignment *q*, then the assignments of its neighboring atoms j, k, and l are reordered such that these are mapped onto neighbor atoms of i′ in structure **B**. Thus an assignment does not only map a single atom, it rather maps a complete sphere of radius 1 around an atom. The bond orders are not considered, and within each remapping process, the neighbors of an atom are randomly ordered, thus preserving the overall stochastic nature of the algorithm. This reordering operator is invoked in the same way as the mutation operator and the crossover operator, that means it affects a permutation vector with a certain probability in each generation cycle.

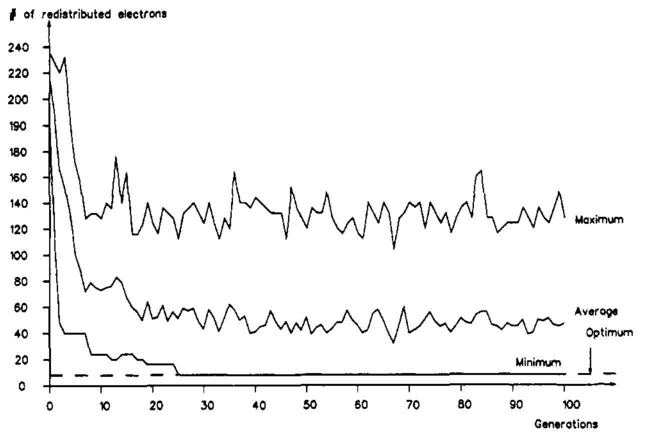


Figure 6. Result of knowledge-augmented genetic algorithm applied on problem in Figure 2.

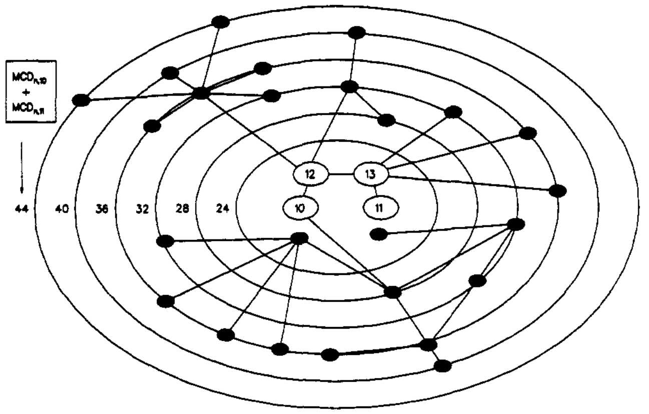


Figure 7. Reaction tree for **10** → **11** as generated by RAIN. Molecules are positioned according to their sums of MCDs to **10** and **11**.

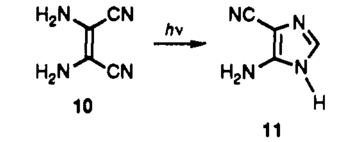
Figure 6 shows a typical result of a knowledge-augmented genetic algorithm run. The effect of the reordering operator in comparison to a simple genetic algorithm without knowledge about the problem structure is obvious. The optimum value, that means one specific assignment within a solution space of 10^{49}, is reached with less than 30 generation cycles. This result is qualitatively reproducible, but it should be emphasized that a genetic algorithm always is stochastic and nondeterministic in nature, and thus all results have some degree of uncertainty.

As a result of these experiments, it can be said that knowledge-augmented genetic algorithms can serve very well in the calculation of minimum chemical distance and, thus, provide a measure of constitutional similarity, especially in cases where other methods are less favorable.

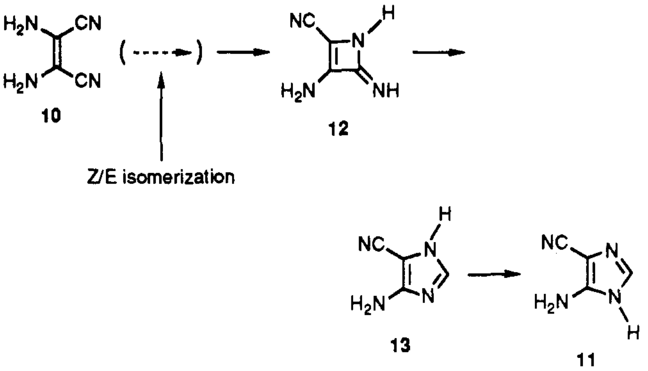
APPLICATION

The last part of this paper presents an example for the application of constitutional similarity in the field of reaction mechanism elucidation. Since the mid-1980s, we developed the PC program RAIN (Reactions And Intermediates Networks)[14] that produces reaction paths using a formal reaction generator,[15] together with a reaction network management system. The possible applications of RAIN are numerous, but the most interesting one is the elaboration of reaction mechanisms. In this operating mode, RAIN produces all conceivable reaction pathways that connect given starting materials with given products of a chemical reaction. The reaction generator is controlled by formal constraints only and does not use any reaction library. Furthermore, no calculations of molecular energies or reaction enthalpies are performed. RAIN takes into account only the constitutional aspects of the molecules. Stereochemistry and explicit molecular geometry are not considered.

We take as an example the photochemically induced conversion of an enaminonitrile **10** to an imidazole **11**:



This reaction was investigated by J. P. Ferris,[16] and it is believed to be a key step within the prebiotic formation of adenine from hydrogen cyanide. If we use RAIN in order to get all conceivable mechanistic pathways for this reaction, we shall receive the following answer:



This reaction path was found to be in full agreement with Ferris' experimental results. The Z/E isomerization step is not generated by RAIN because, up to now, the program does not consider stereochemistry. Figure 7 shows the complete outcome of the monolateral reaction generation, that means that RAIN produced all formally conceivable reaction pathways which emerge from the starting material **10**. The diverse reaction paths were produced by simply redistributing electrons in bonds and free valencies according to a set of formal rules. The reaction tree is rapidly growing and contains 29 different structures after only three reaction steps. The molecules are positioned on concentric circles. The increasing radii of the circles indicate decreasing structural similarity to the enaminonitrile **10** and the imidazole **11**. The molecules are placed on the circles according to the sum of their MCDs to **10** and **11**. It can easily be seen that the intermediates that are part of the connecting route between educt and product (molecules **12** and **13**) are located near the center. All the other generated reaction paths lead to molecules that are more and more dissimilar to the educt and the product. These molecules are irrelevant with regard to the reaction mechanism. Constitutional similarity provides here the means to cut off those branches whose structures are too dissimilar to the starting material and the product of the investigated reaction.

If an arbitrary limit for the sum of the MCDs is set to 28, which is about twice the MCD between **10** and **11**, RAIN can prune these misleading branches. This substantially reduces the size of the reaction tree to be generated. The reaction tree with a depth of three reaction steps is reduced to six molecules, without losing any of the intermediates that are participating in the reaction mechanism.

CONCLUSION

The striking effect of the use of constitutional similarity on the efficiency of reaction generation in the elucidation of reaction mechanisms is demonstrated. With larger problems, the effects may be even more important. Since RAIN is designed to be operated on a PC, the substantial decrease of the size of the reaction trees enables the user to investigate reactions that could not be handled otherwise because of their combinatorial aspects. This paper demonstrates the usefulness of the concept of constitutional similarity based on minimum chemical distance and the power that lies in the genetic algorithms, which makes them a useful tool for many optimization problems in chemistry.

REFERENCES AND NOTES

(1) Rouvray, D. H. The Evolution of the Concept of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990, pp 15–42.

(2) Dean, P. M. Molecular Recognition: The Measurement and Search for Molecular Similarity in Ligand–Receptor Interaction. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990, pp 211–238.

(3) (a) Mezey, P. G. Three-Dimensional Topological Aspects of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990, pp 321–368. (b) Hopfinger, A. J.; Burke, B. J. Molecular Shape Analysis: A Formalism to Quantitively Establish Spatial Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990; pp 173–209.

(4) Randic, M. Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990; pp 77–145.

(5) Willet, P. Algorithms for the Calculation of Similarity in Chemical Databases. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990, pp 43–63.

(6) Ugi, I.; Wochner, M.; 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.; Wiley: New York, 1990, pp 239–288.

(7) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–159.

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

(9) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD). *Angew. Chem.* **1980**, *92*, 503–513; *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495–505.

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

(11) Holland, J. H. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press: Ann Arbor, 1975.

(12) (a) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989. (b) Goldberg, D. E.; Lingle, R. Alleles, Loci, and the Traveling Salesman Problem. In *Proceedings of an International Conference on Genetic Algorithms and Their Applications*; Grefenstetle, J. J., Ed.; Carnegie Mellon University: Pittsburgh, 1985; pp 154–159.

(13) Fontain, E. The Problem of Atom-to-Atom Mapping. An Application of Genetic Algorithms. *Anal. Chim. Acta* **1992**, in press.

(14) (a) Fontain, E.; Bauer, J.; Ugi, I. Computer Assisted Bilateral Generation of Reaction Networks from Educts and Products. *Chem. Lett.* **1987**, 37–40. (b) Fontain, E.; Bauer, J.; Ugi, I. Computer-assisted Mechanistic Analysis of the Streith Reaction Using the Program RAIN. *Z. Naturforsch.* **1987**, *42B*, 889–891. (c) Ugi, I.; Fontain, E.; Bauer, J. Transparent Formal Methods for Reducing the Combinatorial Abundance of Conceivable Solutions to a Chemical Problem—Computer-Assisted Elucidation of Complex Reaction Mechanisms. *Anal. Chim. Acta* **1990**, *235*, 155–161.

(15) Fontain, E.; Reitsam, K. The Generation of Reaction Networks with RAIN. 1. The Reaction Generator. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 97–101.

(16) (a) Ferris, J. P.; Kuder, J. E. Chemical Evolution. III. The Photochemical Conversion of Enaminonitriles to Imidazoles. *J. Am. Chem. Soc.* **1970**, *92*, 2527–2533. (b) Ferris, J. P.; Orgel, L. E. An Unusual Photochemical Rearrangement in the Synthesis of Adenine from Hydrogen Cyanide. *J. Am. Chem. Soc.* **1966**, *88*, 1074. (c) Ferris, J. P.; Trimmer, R. W. Photochemical Conversion of Enaminonitriles to Imidazoles. Scope and Mechanism. *J. Org. Chem.* **1976**, *41*, 19–24.

Registry No. 1, 69-53-4.