Determination of Structural Similarity by Quantitative Comparisons of Wiswesser Line Notation Entries

THOMAS F. KALTENBACH and GARY W. SMALL*

Department of Chemistry, University of Iowa, Iowa City, Iowa 52242

Received September 19, 1989

A means for computing a quantitative measure of similarity between structures encoded in the Wiswesser line notation (WLN) is presented. The WLN string for each compound being compared is first converted to a generalized form. This form consists of three separate strings for each compound and is designed to be more amenable to a character-by-character comparison of two strings. A similarity score is computed on the basis of the best matching of the largest substrings common to both structures. The comparison algorithm is evaluated through the use of five sets of example structures.

INTRODUCTION

Chemical structural databases have come into wide use as a means for cataloging a variety of physical, chemical, and biological data. In most cases, these databases are interrogated by substructure searching. For a target substructure, this procedure produces a list of the compounds in the database containing the target structural unit. A substructure search has no capability for ranking structures in terms of their similarity, however. The database may contain no compounds corresponding to the target substructure, yet contain many compounds with a related substructure. Algorithms for determining structural similarity have been reported for several specialized applications,[1-5] although no generalized algorithm for use in searching a database and computing similarity scores has been reported.

To formulate a quantitative index of structural similarity, a database of structures is required, stored in some encoded format. A widely used scheme for encoding chemical structures is the Wiswesser line notation (WLN), a method of representing each chemical structure in a unique and compact form.[6,7] This notation is stored in the form of a character string (i.e., a string of alphanumeric and punctuation symbols) and can easily represent a large compound with relatively few characters. This character string format is readily suited to automated comparison. Earlier work investigated some WLN searching techniques, but did not involve quantitative comparisons.[8-11] In this paper, a generalized algorithm is introduced for comparing two WLN-encoded structures, producing a quantitative measure of structural similarity. This algorithm is examined in detail and evaluated through the use of five sets of example structures.

EXPERIMENTAL PROCEDURES

The algorithms outlined in the following section were implemented in computer software on a PRIME 9955 minicomputer operating in the Gerard P. Weeg Computing Center at the University of Iowa. The software consists of approximately 1200 lines of PRIME FORTRAN 77 source code (excluding comment lines) and approximately 2100 lines of PRIME Pascal source code (excluding comments and blank lines). The Pascal constructs used are essentially standard Pascal, but the FORTRAN 77 code takes advantage of recursive subroutines, a PRIME extension to the FORTRAN 77 language. For this reason, the general portability of the software in the current implementation is limited to computer systems that (a) allow FORTRAN and Pascal code to be linked together into an executable module and (b) have a FORTRAN compiler which allows recursion. The possibility of a second version, running under MS-DOS on IBM PC (and compatible) microcomputers, is currently being investigated.

RESULTS AND DISCUSSION

Wiswesser Line Notation. The Wisesser line notation (WLN) was developed by William J. Wiswesser in 1949 and consists of numeric digits to represent carbon chains or ring sizes and various alphabetic characters to represent heteroatoms and common molecular substructures such as functional groups or benzene rings.[12,13] These characters are used in a formalized system to encode each chemical structure in a unique and compact manner. Examples of WLN structural representation are shown in Figure 1. Portions of the following discussion refer to an older version of the Wiswesser line notation.[6] This is unavoidable due to its widespread use in certain areas of chemistry. The current standard WLN version is defined by Smith and Barker.[7]

Ring systems are encoded under the WLN system through the use of two sets of delimiters: the L,J pair and the T,J pair. The L,J pair is used when denoting a carbon-based ring system, while the T,J pair is used when denoting a heterocyclic ring system. Examples of this formulation are shown in Figure 2. In the WLN for cyclohexane, shown in Figure 2, the 6 indicates a six-membered ring, and the presence of the T before the terminating J indicates that the ring is saturated. In the WLN for naphthalene, the 66 indicates that there are two six-membered rings present which are fused together, and the absence of any T's before the J delimiter indicates that both the rings are unsaturated. Substituents on these ring systems are denoted by a space followed by a letter A–W, which encodes the position of the locant to which the substituent is attached. Thus, the WLN L6TJ AQ encodes the structure of cyclohexanol, a hydroxyl group (Q) attached to the A position of the cyclohexane ring.

The Wiswesser notation, due to its compactness and long existence, is by far the most common line notation in use. This fact alone makes it worthwhile to pursue the use of the WLN system in developing a structural similarity index.

Processing Overview. Since the Wiswesser line notation is based on alphanumeric characters, the most direct method for comparing two WLN-encoded structures is on a character-by-character basis. The characters or sequences of characters that are common to the two WLN strings should represent the chemical substructures that are common to the two compounds. Those characters that did not appear in both WLN strings could be equated to dissimilarity between the two structures. A numerical score could be readily computed on the basis of the matching and mismatching characters in such

0095-2338/90/1630-0073$02.50/0

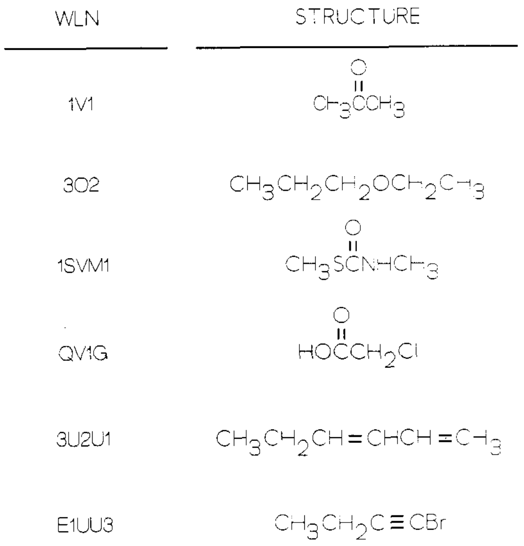

Figure 1. Examples of some simple structures encoded in the Wiswesser line notation.

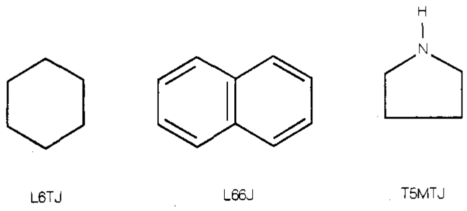

Figure 2. Examples of ring system notation under the WLN system.

a way that the score value reflected the degree of structural similarity of the two compounds.

The Wiswesser line notation, however, is designed to encode each chemical structure in a compact manner. To keep the size of the WLN string to a minimum, there are many cases in which a given substructure may be omitted. In these cases, the existence of the substructure is implied by the way in which the rest of the WLN string is specified. There are several cases where this string contraction can occur, including methyl groups omitted in ring and branching notations, repeating substructure units omitted in multiplier specifications, etc. The ability to leave symbols out of the WLN string greatly reduces the size of the string, but at the same time it means that the system does not readily lend itself to a character-by-character comparison. For such a comparison, the ideal string would have no omissions or contractions, but rather would have every structural unit explicitly expressed. Since each structural unit would be represented by some character(s), this ideal string would allow the character comparison to most accurately reflect a structural comparison. The goal of the work described here is to process a given WLN string into a generalized form that is amenable to character-by-character comparison.

The processing of WLN strings is divided into two basic stages: ring processing and nonring processing. Ring processing includes any cyclic or polycyclic structures (i.e., any simple ring or fused ring systems) present in the WLN notation. Nonring processing includes straight and branching chains, any ring substituents, or any atoms that bridge two ring systems. A given WLN string is first checked for the presence of a ring or ring system, and if one is present, the ring and nonring portions of the WLN are separated and processed independently. Figure 3 is a flowchart that describes this processing scheme.

Nonring Processing. The first stage in the processing of the nonring portion of the WLN string is to expand any multipliers into a longer form. A multiplier is defined as a single-digit number preceded by a space and indicates that the preceding string is to be repeated the specified number of times. A multiplier must be expanded so that each occurrence of the argument is present in the string and free to match with other characters in a comparison of two WLN strings. Each multiplier is processed by removing the multiplier notation and inserting the multiplied argument the appropriate number of times.

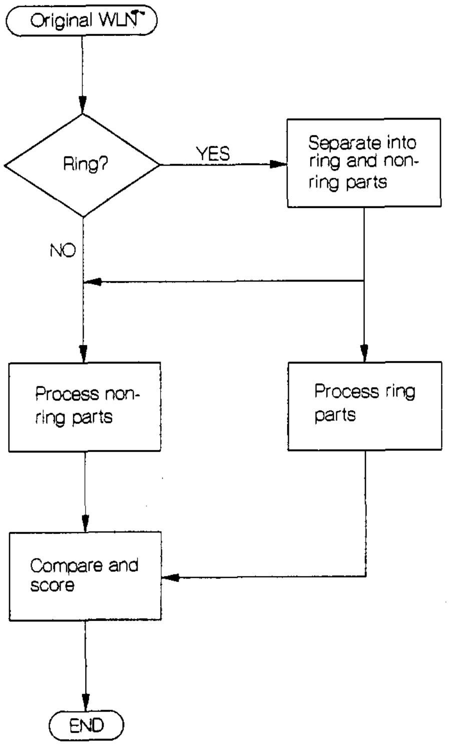

Figure 3. Overview of WLN string processing scheme.

Another stage in nonring processing is to expand any numeric digit (implying a carbon chain) into a string of 1's. This is analogous to writing $(CH_2)_3$ as $CH_2CH_2CH_2$, and such a step is essential for a character-by-character comparison of two carbon chains. For example, if a heptyl group (WLN: 7) were to be compared with an octyl group (WLN: 8), a total mismatch would occur because $7 \neq 8$, even though the two groups are structurally very similar. However, if these were converted to 1111111 and 11111111, respectively, seven of the characters would match, more accurately reflecting the similarity of the two structures.

In the WLN system, methyl groups can be omitted from the string if their presence is implied by the existing structure. These contracted methyls pose a problem in string comparisons, as described above for multipliers, and must be reinserted. This reinsertion step must be performed in the several situations in which methyl groups are omitted. Common occurrences of methyl contractions include the notation for branching chains and ring substituents.

Ring Processing. Ring processing occurs on two different levels: the *ring system* level and the *ring chain* level. The ring system level involves structures which consist of a single ring system and its substituents, that is, structures which are made up of one or more rings fused together, as in naphthalene (Figure 2). The ring chain level involves structures which are a chain of rings, that is, one or more ring systems linked by nonring structures. For this processing, a recursive technique is used to parse and separate the ring systems. These two types of processing produce two types of generalized WLN strings for later comparison. The ring chain string contains only the linking structures and a token character R for each individual

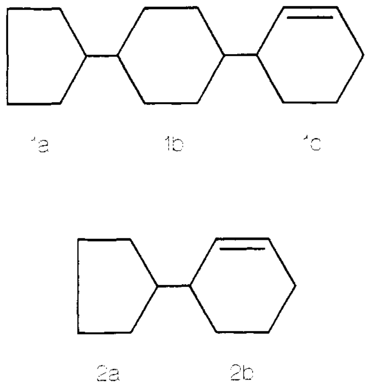

Figure 4. Diagram of best mapping of ring systems between two compounds. Best mapping here is system 1a with system 2a, system 1c with system 2b. System 1b is left unmapped.

ring system. The individual ring system processing produces a string containing the substituent information for that system and a string detailing the ring system structure. The following discussion of ring processing addresses first the ring chain processing and then the ring system processing—the order in which the software performs the actual operations.

When a WLN containing one or more ring systems is processed, the first task to be performed is a search for the first ring system. When the first ring system is found, an R character is placed in the ring chain string. The substituent information of this system is then parsed to determine if a second ring is present. If a secondary ring system is found, any structure that links the two ring systems is removed from the WLN being processed and inserted into the ring chain string, followed by the R that indicates the second ring system. Thus, two phenyl rings joined by a methylene would be represented R1R. The secondary ring system (including its substituents) is then removed from the current WLN string and placed into a second system string. The processing then starts from the beginning on the secondary ring system (this stage marks the beginning of the recursion). The secondary ring system is scanned for substituents as above, and the occurrence of another ring system invokes another level of recursive processing. When no additional ring systems are found, the processing of the current ring system completes and control returns to the processing of the previous ring system. The result is a single ring chain string, and one ring system string for each ring system present in the structure.

Next, processing begins at the ring system level. Each ring system string is ultimately converted into two strings: one generalized ring system string and one substituent string. To perform this conversion, information concerning the ring system must first be extracted from the WLN string. This information includes the number, size, and saturation of all (fused) rings; the location, saturation, and atom type of each locant in the system; etc. From this information, a ring string can be generated that represents any ring system in a generalized manner. In this string, tertiary and quaternary atoms are represented by their corresponding WLN symbols: Y and X for carbons, N and K for nitrogens. For example, a naphthalene could be encoded YCUCCUCYCUCCUC, which represents a chain of atoms with alternating single and double bonds (the U character represents a double bond), beginning with a tertiary carbon and moving away from it. In certain cases, substituents on a given ring system can influence which locant is designated as the first locant in the locant path. To prevent this from interfering with a comparison of the ring portions of two structures, the generalized string is generated from all possible starting positions. Each generated string from the first structure is compared to all generated strings from the second structure by using both forward and reverse matching, assuring the best possible comparison. When multiple ring systems are present, the ring systems are processed and scored in order from largest to smallest, and again each generated string for the first structure is compared with all possible generated strings from the second structure, assuring the best possible score.

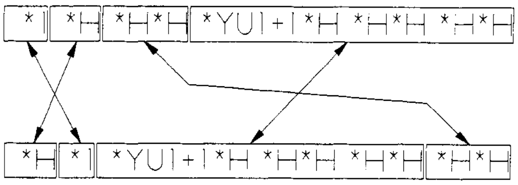

Figure 5. Matching of common substrings within a substituent string.

Table I. Processed WLN Strings

original WLN		generated strings
L6UTJ_A_	chain	R
DYU1	system	CUCCCCC
	subst	_*1_*H_*H*H_*YU1&1*H_*H*H_*H*H
L6UTJ_B_	chain	R
CYU1	system	CUCCCCC
	subst	_*H_*1_*YU1&1*H_*H*H_*H*H_*H*H

Substituent information for a given ring system is enhanced by inserting the implied hydrogens that are bonded to ring locants. In addition, the substituent position indicators in WLN are removed, and the position of each substituent is represented relative to other substituents by its position in the string. The new format for each substituent is ⟨space⟩ ⟨*⟩ ⟨substituent 1⟩ ⟨*⟩ ⟨substituent 2⟩. Thus, for a secondary ring carbon, a " *H*H" substituent would be generated. If this carbon had a methyl substituent, the corresponding string segment would be " *1*H". Each substituent on the same ring locant is separated by an asterisk (*), and substituents on adjacent ring locants are separated by a space and an asterisk.

Scoring. After the processing of the WLN strings is complete, both the ring and nonring portions of the structures are compared on a character-by-character basis, and a numerical score is computed to represent the degree of structural similarity between two given WLN strings. The comparison algorithm employed here is based on finding the strings or substrings that are common to both structures; these portions of the string represent the units within the compounds that are structurally similar. All other (nonmatching) portions of the strings represent structural dissimilarity. From these matching and nonmatching units, a numerical score can be computed that reflects the degree of structural similarity. A number of string similarity measures exist in the literature, primarily for use with spelling correction algorithms.[14–17] For two main reasons, these similarity measures are not readily adaptable to determining the structural similarity between two WLN-encoded compounds: (1) structural comparisons require both forward and reverse matching; (2) WLN symbols require weighting coefficients to rank their chemical significance.

The specific comparison algorithm developed here is called the largest common substring algorithm. This approach involves finding all possible substrings in each of the two WLN strings and then comparing the two sets of substrings for the best matchings. The best matching is obtained by first sorting the two sets of substrings on the basis of substring length and then comparing the strings starting with the longest substrings and ending with the shortest. As matching substrings are identified, the positions of the matching characters in the two original strings are flagged as matched and are subsequently excluded from further comparisons. In some cases, reverse matching is also performed; that is, the substring comparison will return a match if one string has the same sequence of

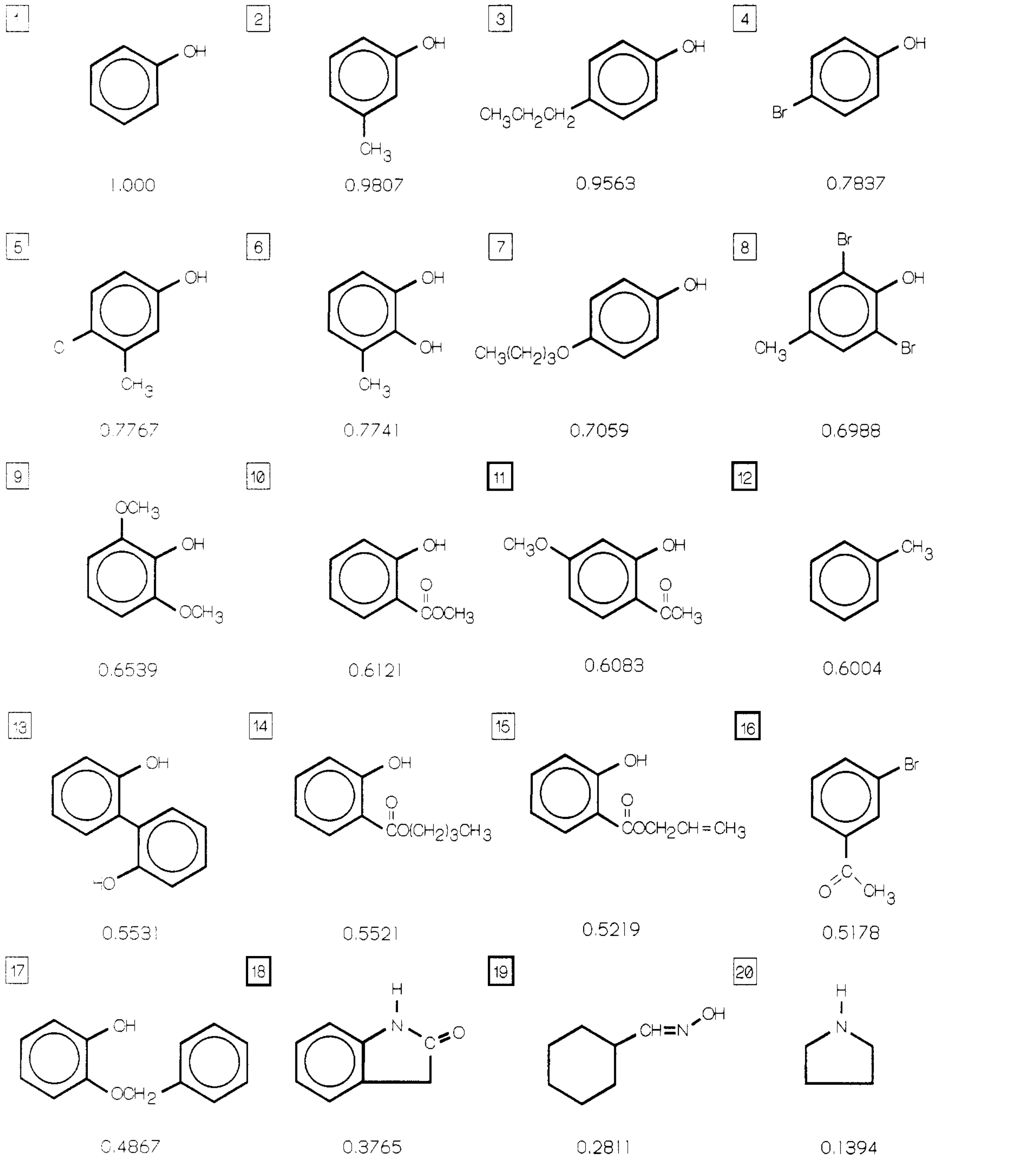

Figure 6. Test results for the target compound phenol compared with 19 compounds of varying structural similarity.

characters as the other, but in the reverse order. The result of these forward and reverse comparisons is a set of matching substrings, corresponding to matching structural subunits.

Once the largest common substrings have been obtained, the substrings must be ranked in terms of their structural significance. One simple method for obtaining a score for a matching substring is by length, since matching a large substring should be better than matching a shorter one. Unfortunately, this is sometimes not the case, since a single character in a WLN string may represent one atom, zero atoms, or more than one atom. For example, an O represents one oxygen atom, a U represents a double bond, but no atoms, and a V represents a carbonyl, C═O. Therefore, the scoring algorithm utilizes an atom count for each symbol in determining a numerical score. Another consideration is that structural units such as functional groups, while composed of few atoms, have great chemical significance. To give certain WLN symbols more influence in the final score, a weighting coefficient is introduced for each symbol. These factors are used to produce a numerical score for a substring

$$S = n + \sum_{i=1}^{n} c_i w_i \tag{1}$$

where S is the substring score, n is the number of characters in the substring, c_i is the atom count, and w_i is the weighting coefficient for the ith character in the string. The length of the substring is added to the sum to give some weight to those characters that, while structurally significant, have an atom count of zero.

The final score for a comparison of two WLN strings involves summing the scores of all the matching substrings and relating this score to the score of the unmatched portions in the WLN strings. A score that increases with increasing string similarity is produced

$$S_{\text{total}} = M/(M + X) \tag{2}$$

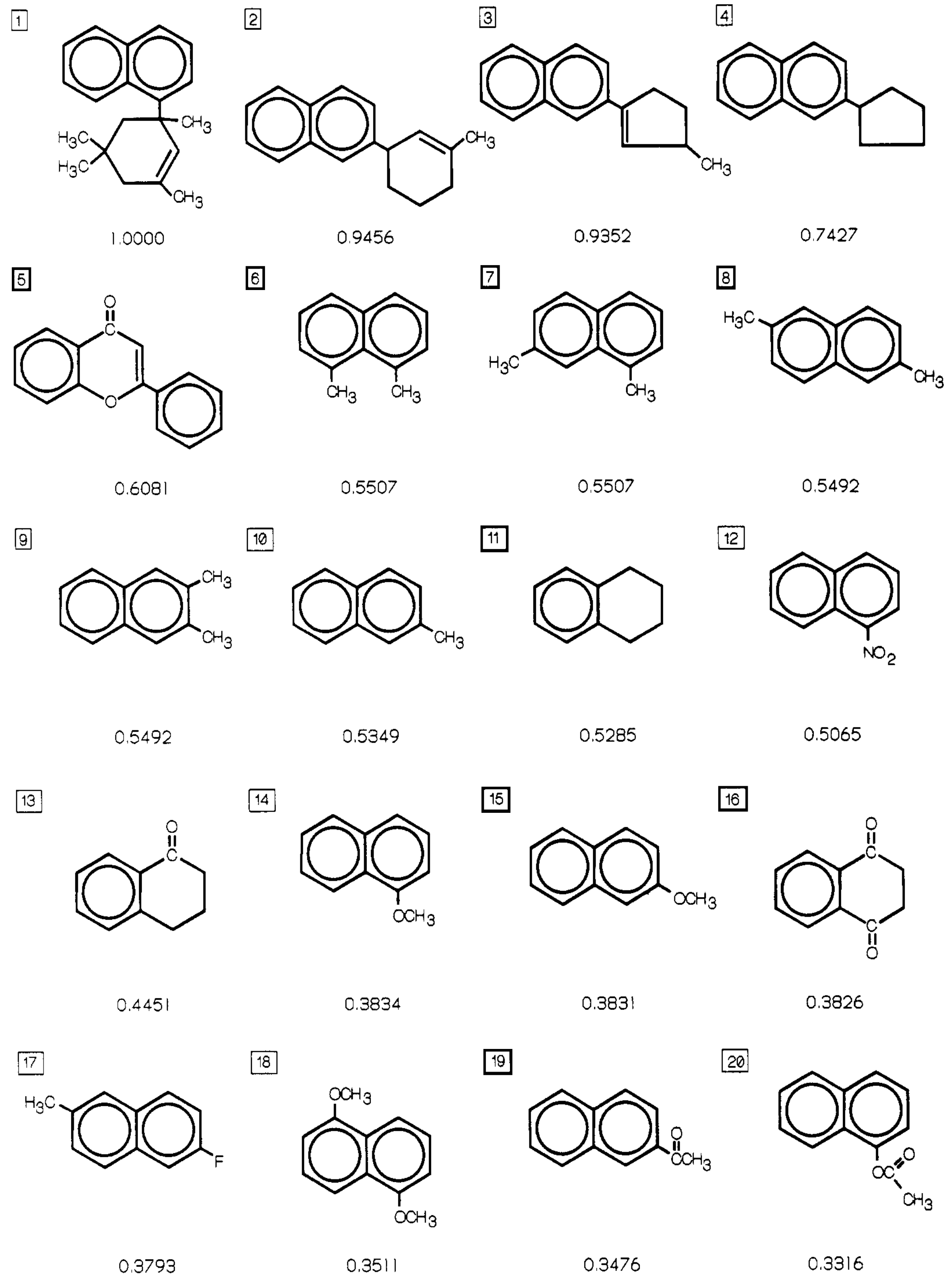

Figure 7. Test results for the target compound limonene.

where S_{total} is the largest common substring score, M is the sum of the scores of all the matching substrings, and X is the sum of the scores of all the mismatching substrings. However, simply summing the scores and computing the above ratio are not meaningful, because if all the pieces of a string match, say in three matching substrings, the sum of the scores for these substrings equals the score of the entire string. In this case, the score has failed to encode that the three matching structural fragments are not identically connected. For this reason, a *connectivity factor* is introduced to penalize a sum of scores based on the degree of fragmentation within the string, i.e., based on the number of substrings employed in the matching process. The connectivity factor is based on a relation between the number of substrings matched and the number of characters in the target string. Clearly, the smaller the number of substrings used in the matching, the higher the degree of connectivity. The ratio of the number of substrings to the length of the target string can be used to produce a score, S_{total}', in which connectivity is considered

$$S_{\text{total}}' = (1 - N/L)S_{\text{total}} \qquad (3)$$

where N is the number of matching substrings, L is the length of the target string, and S_{total} is defined in eq 2. The numerical scoring value computed, while linear in response, penalizes heavily against mismatches in smaller structures and increases slowly with increasing structural similarity. To achieve a more favorable response from the scoring metric, the computed score was mapped onto the radian scale from 0 to 1.0471976. The cosine of this value then becomes the connectivity factor, which ranges from 1.0 to 0.5. Thus, the maximum penalty for a match with many substrings is 0.5. The general similarity score is then

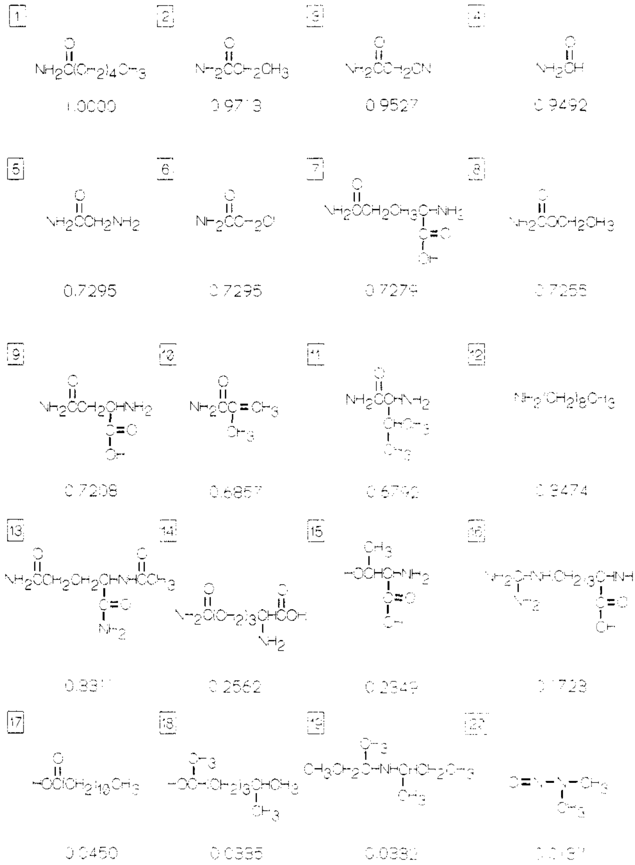

Figure 8. Test results for the target compound hexanamide.

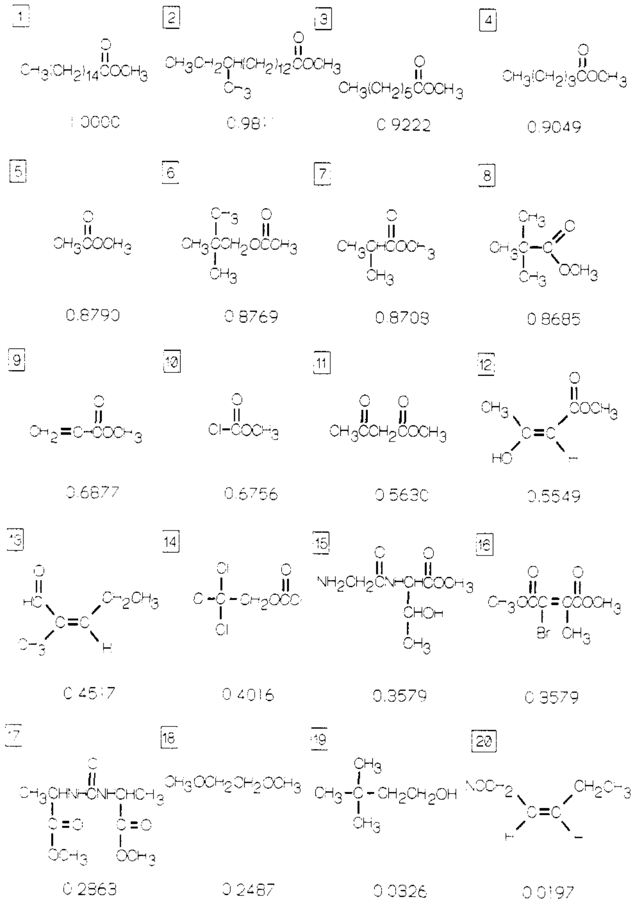

Figure 9. Test results for the target compound methyl hexadecanoate.

$$S_{\text{total}}' = \cos\,[1.0471976(N-1)/(L-1)]S_{\text{total}} \qquad (4)$$

The cosine function was chosen to moderate the rapid linear drop in the $(1 - N/L)$ term because it has a more desirable response function.

Scoring Multiple Ring Systems. In situations where one or both of the structures contains more than one ring system, a best mapping of the ring systems must be performed. In Figure 4, the best mapping of the individual ring systems within the structures is 1a matched with 2a and 1c with 2b. This assures that the comparison score penalizes only the six-membered ring (1b) that is dissimilar between the two structures. The best mapping is determined by generating ring strings and scoring all possible combinations of ring system pairs, retaining those with the highest scores. Once the best mapping is determined, all subsequent scoring of substituents is based partially on this best mapping. The subtituents for each of the ring systems are entered into the substituent scoring string in the order of the best mapping of the systems. This ensures that the sequence of substituents is correctly matched with the structural features of the ring systems.

Overall Comparison Score. To generate a single score to reflect the total structural similarity of two compounds, the three subscores (ring chain, ring system, and substituent/nonring) must be combined. This is accomplished by computing a weighted mean of the three scores, where the weights on the individual terms reflect the total number of atoms involved in the comparison. The formula for S_{combined}, the weighted mean, is

$$S_{\text{combined}} = (SC_1 + XC_2 + GC_3)/(C_1 + C_2 + C_3) \quad (5)$$

where S is the system score, X is the substituent/nonring score, G is the ring chain score, and C_1, C_2, and C_3 are the numbers of atoms associated with the three scores, respectively. If neither structure has a ring component, then the overall comparison score simplifies into simply the nonring/substituent score. In the case in which one compound is a ring and the other is a nonring, the score is arbitrarily chosen to be zero.

Example of Processing. The function of the above processing stages can best be shown by an example. The processing, comparing, and scoring of two sample WLN strings is described below. The two compounds chosen for this example are the first two compounds shown later in Figure 7. The original WLN strings and the generated strings are shown in Table I (blank spaces are shown as underlines for clarity). The ring chain string for both compounds contains just the character R, since each compound contains only one ring system. The ring system string for each compound is also identical and represents the cyclohexene ring as a chain of carbon atoms, with a double bond between the first and second carbons in the chain. The differences between the two compounds are apparent in an inspection of the substituent strings. Although the two compounds have the same substituents, the substituent positions are not the same, producing a difference in the substituent string.

The comparison of the three sets of strings is as follows. For the ring chain string and the ring system string, the strings match exactly. This produces a score of 1.0 for each of those two matches by applying eqs 1, 2, and 4. The substituent strings match in four substrings, with no unmatched characters, as shown in Figure 5. Applying eqs 1 and 2 produces a score of 1.0, since all characters in both strings have been matched. Applying eq 4 penalizes for the number of substrings used in the matching, reducing the score to 0.9937. Then applying eq 5 computes the final comparison score as a weighted mean of the three string scores, producing the value of 0.9981.

Results. The performance of the structural comparison and scoring routines can be seen by the following examples, shown in Figures 6–10. In each example, a target structure is compared to a set of 20 compounds of varying structural

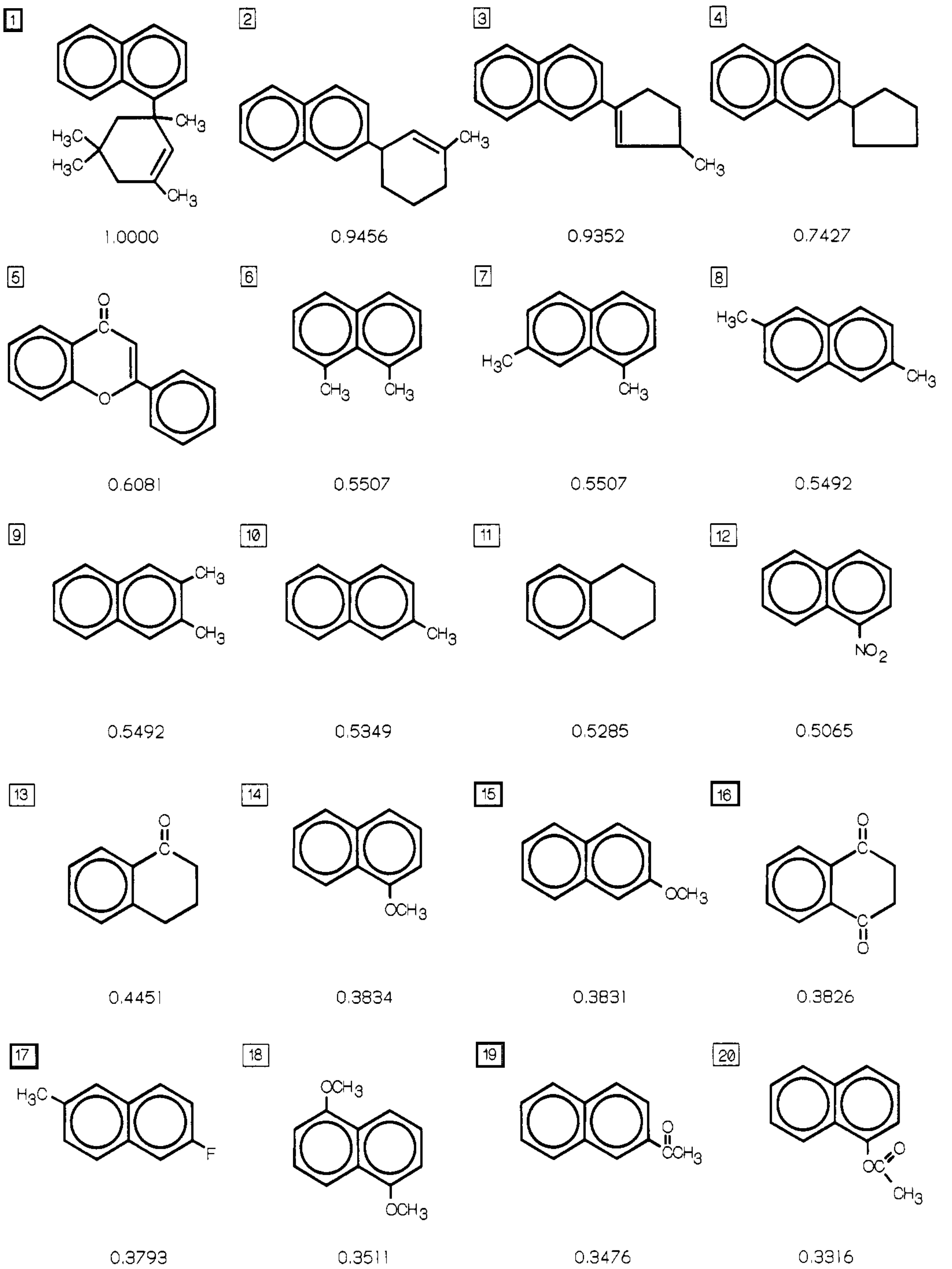

Figure 10. Test results for the target compound 1-(1,3,5,5-tetramethyl-3-cyclohexen-1-yl)naphthalene.

similarity. Each set of 20 was chosen to contain the target structure and 19 other compounds. The figures show each set of 20 as they are ranked by the scoring routines. Figures 6 and 7 are examples of ring processing, Figures 8 and 9 show nonring compounds, and Figure 10 depicts an example of processing compounds with multiple ring systems.

In the results for the target compound phenol, shown in Figure 6, it can be seen that the scoring algorithm does rank the set of compounds in a general order of decreasing structural similarity. Most similar to the target compound are the alkyl-substituted phenols, compounds **2** and **3**, then a halogenated phenol, and so on. The scores for compounds **2** and **3** are above 0.95, indicating a high degree of similarity. It could be argued that perhaps compound **6**, a methyl-substituted diol, should be ranked above a halogenated phenol, and regardless of whether or not this argument is valid, it does point out a weakness in the current scoring system. The scoring algorithm does not have the intelligence to determine the principal functional group of a compound and pentalize additional occurrences of this functional group less than occurrences of a different functional group. Stated differently, the presence of two hydroxyl substituents should be penalized less than the presence of one hydroxyl and one bromide substituent. Such "intelligence" could be incorporated into a future revision of the software.

The target compound in Figure 7 is limonene, and in general, the ranking results of this set of compounds are quite good. The principal point of interest is that changing the position of the methyl and vinyl substituents on the cyclohexene ring gives an appropriately ranked score, as seen in compounds **2–5**. One disturbing result is that a compound with a large branching substituent (compound **7**) is ranked above a methyl-substituted cyclohexene (as in compound **9**). One reason for this ranking is that methyl and methylene groups are represented identically (with a 1) in the generalized WLN string. This means that the vinyl group is matching with a segment of the alkyl chain, in particular, with the segment beginning with the unsaturated tertiary carbon. The resultant ranking should be more appropriate if methyl and methylene groups were represented by different symbols.

The results from the target compounds in Figures 8 and 9 show good results for acyclic compounds. The methyl/methylene problem can be seen again in Figure 9, where compound **6** is ranked among the methyl esters because of the presence of 1OV1 in the processed WLN. This problem should be resolved by the solution presented above. In the results shown in Figure 10, it can be seen that the compounds with two ring systems are ranked above those with only one, and within the dual-system compounds, the ranking is appropriate. Compounds **6** and **7** and compounds **8** and **9**, are dimethylnaphthalenes with similar scores, as is appropriate (those with a methyl in the B position on the ring are ranked above those without).

Computational Speed. The execution time of the software depends on the complexity of the structures being analyzed. Current execution time on the Prime 9955 systems is on average 0.2 CPU s/comparison for noncyclic compounds, 1.2 CPU s/comparison for monocyclic systems, and 5.0 CPU s/comparison for polycyclic systems for the five sets of test compounds shown above. The comparison process has not yet been optimized for speed. It is expected that greater efficiency can be obtained by enhancing certain modules. The largest common substring routine, for example, currently employs an inefficient algorithm based on generating all possible substrings for every string compared. This routine could be replaced by a routine utilizing a fast string-matching algorithm.[18-20] Enhancement of key routines should both reduce the memory requirements and increase the speed of comparisons.

Limitations. In the current implementation, some of the more advanced features of the Wiswesser line notation system are not supported. For example, perifused ring systems (a ring system with a locant common to three or more rings) are not supported, due to a limitation in the ring-walking algorithm employed in determining the location of ring locants in the ring system. This algorithm is used in both the processing and scoring of ring systems. It may be possible to overcome this limitation with further work. In addition to this limitation, nested multipliers are not supported, although processing of nonnested multipliers is implemented in the current version. Other features such as the ring-of-rings contraction and organometallic compounds are also not supported.

CONCLUSION

In general, the results described above are appropriate and show a logical ranking based on structural similarity. While it is true that judging which structures are most similar to a given target structure is a somewhat subjective procedure, the ranking obtained here seems justifiable. There is also some flexibility built into this approach, as individual WLN symbols can be given different weights in the scoring, thus making it possible to tailor the system to suit particular needs.

One potential extension of this work is the development of a similar structural similarity measure based on structures stored in the form of connection tables. It should be fairly straightforward to generate generalized characters strings like those described above for quantitative comparison. The ability to accept structures stored in connection table format would greatly enhance the utility and applicability of this system.

ACKNOWLEDGMENT

William J. Wiswesser of the U.S. Department of Agriculture, Frederick, MD, is gratefully acknowledged for supplying a machine-readable copy of the WLN-encoded herbicide database of 31 000 compounds, which was used in testing this system.[21]

REFERENCES AND NOTES

(1) Varkony, T.; Shiloach, Y.; Smith, D. H. Computer-Assisted Examination of Chemical Compounds for Structural Similarities. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 104–111.

(2) Randic, M.; Wilkins, C. Graph Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31–37.

(3) Adamson, G.; Bush, J. Comparison of the Performance of Some similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.

(4) Johnson, M.; Naim, M.; Nicholson, V.; Tsai, C. C. Comparing the Substructure Metric to Some Fragment-Based Measures of Intermolecular Structural Similarity. *Pharmacochem. Libr.* **1987**, *10*, 67–69.

(5) Willet, P.; Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity: Measures of Inter-Molecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18–25.

(6) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*; McGraw-Hill: New York, 1968.

(7) Smith, E. G.; Barker, P. A. *The Wiswesser Line-Formula Chemical Notation*, 3rd ed.; Chemical Information Management: Cherry Hill: NJ, 1976.

(8) Crowe, J.; Leggate, P.; Rossiter, B.; Rowland, F. The Searching of Wiswesser Line Notations by Means of a Character-Matching Serial Search. *J. Chem. Doc.* **1973**, *13*, 85–92.

(9) Ofer, K. A Computer Program To Index or Search Linear Notations. *J. Chem. Doc.* **1968**, *8*, 128–129.

(10) Bowman, C.; Landee, F.; Lee, N.; Reslock, M.; Smith, B. A Chemically Oriented Information Storage and Retrieval System. III. Search a Wiswesser Line Notation File. *J. Chem. Doc.* **1970**, *10*, 50–54.

(11) Eakin, D. R.; Hyde, E.; Parker, G. Use of Computers with Chemical Structural Information. ICI CROSSBOW system. *Pestic. Sci.* **1974**, *5*, 319–326.

(12) Wiswesser, W. J. How the WLN Began in 1949 and How It Might Be in 1999. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 88–93.

(13) Vollmer, J. Wiswesser Line Notation: An Introduction. *J. Chem. Educ.* **1983**, *60*, 192–196.

(14) Davidson, L. Retrieval of Misspelled Names in an Airline's Passenger Record System. *Commun. ACM* **1962**, *5*, 169–171.

(15) Heckel, P. A Technique for Isolating the Differences Between Files. *Commun. ACM* **1978**, *21*, 264–268.

(16) Hall, P.; Dowling, G. Approximate String Matching. *ACM Comput. Surveys* **1980**, *12*, 381–402.

(17) Lowrance, R.; Wagner, R. An Extension of the String-to-String Correction Problem. *J. ACM* **1975**, *22*, 177–183.

(18) Boyer, R. S.; Moore, J. S. A Fast String Searching Algorithm. *Commun. ACM* **1977**, *10*, 762–772.

(19) Hunt, J. W.; Szymanski, T. G. A Fast Algorithm for Computing Longest Common Subsequences. *Commun. ACM* **1977**, *20*, 350–353.

(20) Knuth, D.; Morris, J.; Pratt, V. Fast Pattern Matching in Strings.

(21) Creager, R.; Wiswesser, W. Summary of Results on 31 000 Compounds Evaluated for Herbicidal Activity. Technical Bulletin No. 1721; U.S. Department of Agriculture, U.S. Government Printing Office: Washington, DC, 1987.