How To Add Chemical Abstracts Service Registry Numbers and Structures to Databases via Chemical Names Comparison

Heinz A. Krebs* and Ulrich Jordis

Institute of Organic Chemistry, Vienna University of Technology, Getreidemarkt 9/154, A-1060 Vienna, Austria

Received July 31, 1999

For the conversion of nonstructural chemical databases to structure databases, a series of algorithms to find the closest match between existing names to names in a reference database are described. On the basis of the best match, new fields such as the Chemical Abstracts Service Registry Number (CASRN) or structures were added to the database.

INTRODUCTION

Frequently, existing chemical databases or chemical inventory systems lack unequivocal structure information such as the Chemical Abstracts Service Registry Number (CASRN) or the structures themselves. Computer programs for the conversion of chemical substance names to connection tables (1), translation to other languages (2), validation of nomenclatures (3), interconversion (4) of nomenclatures, and computer generation of nomenclature from structures (5) have been described. For the migration of a chemical inventory system consisting of a single-user MS-DOS/dBase/Clipper program to a modern networkable multiuser application, we wanted to automatically add CASRNs to the existing database and use these generated CASRNs to link the records of the new database to other information sources including chemical structures and molecular safety data sheets (MSDS). This conversion would then open the possibility of (sub)-structure searches.

STRATEGY

The records of an old chemical inventory database included fields for the molecular formula and German chemical names. The principal idea was to develop algorithms for using these two fields to find the most similar compound in an established database. Matched entries would then allow the import of the CASRN. Based on the CASRN match, structures could be imported and further links to compound information like MSDS added.

To develop and validate the algorithms, an established German database was first compared to an established English database. The German version of the Fluka catalog on CD (FCD) and the Available Chemicals Directory (ACD) can be used, as FCD already contains the CASRN thus allowing the evaluation and validation of the success of the algorithms. Additionally, the inherent problems of comparing German with English names would have to be solved. The algorithms developed would then be used to convert the existing database (of >7000 chemicals) to the new database containing both structures and the CASRNs.

* Corresponding author. E-mail: hkrebs@mail.zserv.tuwien.ac.at.

ALGORITHMS FOR COMPARING CHEMICAL NAMES

1. One2One Comparison. This algorithm counts the number of equal characters at the same position in both names:

3-CHLOR-1-BUTANOL
2-CHLORBUTANOL

To obtain a better match, reverse counting is added:

3-CHLOR-1-BUTANOL
2-CHLORBUTANOL

The similarity is determined by the ratio of correct characters and the number of characters of the chemical name:

$$\text{similarity } r_0 = \text{correct/length}$$

in the last example $r_0 = 13/17 = 0.76$.

2. Shift Correlation. To avoid errors created by deviation of sequence of substituents and difference in endings of names (Trimethylamin vs Trimethylamine), matching characters are counted by sliding one name against the other:

```
                    2-CHLORONITROHEXANON
1-NITRO-2-CHLORHEXANONE                          (1st step)
 1-NITRO-2-CHLORHEXANONE                         (2nd step)
......
       1-NITRO-2-CHLORHEXANONE                   (8th step)
        1-NITRO-2-CHLORHEXANONE                  (9th step)
         1-NITRO-2-CHLORHEXANONE                 (10th step)
          1-NITRO-2-CHLORHEXANONE                (11th step)
......
                          1-NITRO-2-CHLORHEXANONE  (28th step)
                           1-NITRO-2-CHLORHEXANONE (29th step)
                            1-NITRO-2-CHLORHEXANONE (30th step)
...... end up with:
                                    1-NITRO-2-CHLORHEXANONE
```

3. ShiftBlock Correlation. Shift correlation results in a high correlation of different structural isomers. Therefore the algorithm was improved by searching equal blocks of three characters instead of one (in the following scheme only the first characters of the corresponding blocks are marked).

3-NITRO-2-CHLORBENZOL
2-NITRO-3-CHLOROBENZENE

10.1021/ci9902649 CCC: $19.00
Published on Web 03/03/2000

Table 1

algorithm	similarity	all characters	Only-Char	No-Voc
One2One	r_0 (%)	63	65	68
	r_{EXP} (%)	62	63	67
	r_{DIV} (%)	69	70	75
Shift	r_0 (%)	60	62	64
	r_{EXP} (%)	66	66	66
	r_{DIV} (%)	77	77	78
ShiftBlock	r_0 (%)	78	79	77
	r_{EXP} (%)	74	75	73
	r_{DIV} (%)	81	82	79
One2One	r_0 (%)	65	70	71
(with previous translation)	r_{EXP} (%)	64	69	70
	r_{DIV} (%)	71	74	76
Shift	r_0 (%)	64	64	68
(with previous translation)	r_{EXP} (%)	69	69	69
	r_{DIV} (%)	80	80	80
ShiftBlock	r_0 (%)	78	78	78
(with previous translation)	r_{EXP} (%)	75	76	74
	r_{DIV} (%)	81	83	79

gives a similarity $r_0 = 11/21 = 0.52$ versus

2-NITRO-3-CHLORBENZOL

2-NITRO-3-CHLOROBENZENE

which gives a similarity $r_0 = 15/21 = 0.71$.

Improvements. The ShiftBlock correlation resulted in correct assignments of 60−78% (see Table 1). For improvements the chemical names of the source and the target database were altered before applying the algorithm.

4. OnlyChar. The style of chemical names (e.g., if a substituent is separated by a hyphen or not) differs in various vendor lists. For a higher similarity "meaningless" characters such as brackets and hyphens are deleted and only numbers and letters are used for comparison. Thus

2-(4-BROMPHENYL)-ETHANOL-AMINE

was converted to

24BROMPHENYLETHANOLAMINE

5. NoVoc. Taking into consideration that vowels do not contain much information in chemical names (like AMINE vs AMIN) but differ substantially in different languages, these characters are also deleted:

24BRMPHNYLTHNLMN

instead of

2-(4-BROMPHENYL)-ETHANOL-AMINE

Additionally for the comparison of databases based in different languages frequently occurring name fragments were exchanged: e.g., BENZOL by BENZENE, SÄURE by ACID, KALIUM by POTASSIUM, ESSIG by ACETIC, ...

Since the *deviation of size of names* should also be part of similarity, the following factors were used:

$$r_{DIV} = r_0 \frac{\text{len}(a) + \text{len}(b)}{2\,\text{len}(a)\,\text{len}(b)} \quad \text{and}$$

$$r_{EXP} = r_0 * e^{-|\text{len}(a)-\text{len}(b)|/\text{len}(a)}$$

A TEST SYSTEM FOR THE COMPARISON ALGORITHMS

The algorithms were evaluated by comparing the Fluka chemicals electronic catalog (FCD) to the ACD according to the following scheme:

1. Find all records in ACD with the matching molecular formula to the individual compounds from FCD.
2. Calculate the similarity for all hits and order them by their similarity.
3. Mark the match as successful if the CASRN of the most similar compound in the ACD corresponds to the CASRN in the FCD.

RESULTS

The simplest algorithm One2One with all characters gives an efficient recognition of correct names of about 63%. Incorporation of deviation of size increases the rate by 5%.

Shift correlation without incorporation of deviation makes the algorithm worse, in contrast to the r_{DIV} values, which are 7% better.

ShiftBlock comparison is the best algorithm with about 82% correct recognition.

Translation of names before comparing yields another 1−3%.

For the ShiftBlock comparison as the best algorithm, the probability that the correct name is contained in the best three hits is about 88−90%.

DISCUSSION OF ERRORS

Stereochemistry. About 5% of the errors are caused by different stereochemical isomers: e.g., (±)-butandiol is wrongly matched to (*S*)-(+)-1,3-butandiol.

Error in Molecular Formula. Identical chemicals have different molecular formulas in the databases used; e.g.

gallium(III) nitrat hydrat [63462-65-7]	$Ga(NO_3)_3$ in FCC
gallium(III) nitrate [13494-90-1]	$Ga(NO_3)_3$ in ACD
gallium(III) nitrate octahydrate [63462-65-7]	$Ga(NO_3)_3 \cdot 8H_2O$ in ACD

Different CASRNs. In some cases different CASRNs occur because the wrong isomers have been assigned in the vendor lists: e.g., (+)-fenchone [4695-62-9] in Fluka (which is correct) and [7787-20-4] in ACD.

SUMMARY

Using the ShiftBlock algorithm, correct CASRNs were correlated to ASCII-based databases with 85% fidelity. Although the results have to be inspected manually, this preselection allows us to simplify and speed up the final validation substantually. Using the CASRN as identifier, we were able to obtain more information from other databases, including the English name and chemical structure from the ACD as well as safety data from the Material Safety Data Sheets or physical properties.

REFERENCES AND NOTES

(1) Luque Ruiz, I.; Cruz Soto, J. L.; Gomez-Nieto, M. A. Error Detection, Recovery and Repair in the Translation of Inorganic Nomenclatures. 3. An Error Handler. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 483−90.
(2) Davydova, E. A.; Pappe, I.; Shevyakova, L. A. Automation of translation of systematic names of chemical organic compounds from Russian into German. *Vses. Inst. Nauch. Tekh. Inf., USSR* **1980**; *Chem. Abstr.* **1982**, *96*, 84715.
(3) Vander Stouw, Gerald G. Computer programs for editing and validation of chemical names. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 232−6.
(4) Kirby, G. H.; Polton, D. J. Systematic chemical nomenclatures in the computer age. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 560−3.
(5) Walker, S. B. J. Computer-generated chemical nomenclature. In *Chemical Nomenclature*; Thurlow, K., Ed.; Kluwer: Dordrecht, The Netherlands, 1998; pp 235−242.
CI9902649