Tunable Machine Vision-Based Strategy for Automated Annotation of Chemical Databases

Jungkap Park,[†] Gus R. Rosania,[‡,§] and Kazuhiro Saitou*[,†]

Departments of Mechanical Engineering and Pharmaceutical Sciences, University of Michigan, Ann Arbor, Michigan 48109

Received January 21, 2009

We present a tunable, machine vision-based strategy for automated annotation of virtual small molecule databases. The proposed strategy is based on the use of a machine vision-based tool for extracting structure diagrams in research articles and converting them into connection tables, a virtual "Chemical Expert" system for screening the converted structures based on the adjustable levels of estimated conversion accuracy, and a fragment-based measure for calculating intermolecular similarity. For annotation, calculated chemical similarity between the converted structures and entries in a virtual small molecule database is used to establish the links. The overall annotation performances can be tuned by adjusting the cutoff threshold of the estimated conversion accuracy. We perform an annotation test which attempts to link 121 journal articles registered in PubMed to entries in PubChem which is the largest, publicly accessible chemical database. Two cases of tests are performed, and their results are compared to see how the overall annotation performances are affected by the different threshold levels of the estimated accuracy of the converted structure. Our work demonstrates that over 45% of the articles could have true positive links to entries in the PubChem database with promising recall and precision rates in both tests. Furthermore, we illustrate that the Chemical Expert system which can screen converted structures based on the adjustable levels of estimated conversion accuracy is a key factor impacting the overall annotation performance. We propose that this machine vision-based strategy can be incorporated with the text-mining approach to facilitate extraction of contextual scientific knowledge about a chemical structure, from the scientific literature.

INTRODUCTION

Nowadays, rather than a mere repository of molecular structure information, the chemical database is becoming an essential research tool as a comprehensive knowledge bank of molecules. For example, virtual collections of chemical compounds can be used to design and keep track of chemical synthesis of combinatorial libraries[1,2] as well as serving as a systematic repository for storing and sharing the various assay data and biological activities of chemical agents in chemical genomics and systems biology.[3,4] In addition, virtual libraries of small molecules can serve as the main source for *in silico* drug discovery applications, including molecular docking and QSAR prediction models.[5,6] It is thus hardly surprising that cheminformatics research has devoted much effort into developing techniques for the storage, retrieval, and processing of chemical databases in order to maximize the value of such an intellectual asset.[7]

In order to enrich the chemical database, many research and development organizations have made an effort not only to register new chemical structures but also to annotate database entries using related information such as method of synthesis, chemical and physical properties, or biological activities. The related information can be derived from the scientific literature, other public databases, and computational methods. Scientists have added experimental property data to the CAS Registry System which is the largest commercially accessible chemical database in the world monitoring the scientific literature.[8] In the case of PubChem (the largest, publicly available chemical database linked to the National Center for Biotechnology Information data warehouse), each chemical structure can have cross-reference links to related structures, bioassay data, and bioactivity description as well as scientific research articles.[9] There are many other databases focusing on more specific information of molecules, while these two very large databases are built on a broad range of chemical sources. For instance, the DrugBank database contains comprehensive information of both FDA-approved drug molecules and their associated targets.[10] As a collection of commercial vendor catalogs, the eMolecules database allows users to access the supplier information of commercially available compounds and to purchase them online.[11] There are also a few annotation services such as SciFinder,[12] IDdb3,[13] and SureChem[14] that enable users to retrieve patent documents or journal articles containing identical or similar chemical structure for the query chemical structure. Commonly, all these chemical database systems are cross-linked to each other so that users can explore chemical information distributed over chemical databases, scientific articles, and Web sites efficiently.

While many chemical information systems have attempted to integrate all chemical information published up-to-date, much time and resources are spent on exploring a vast amount of unstructured information sources such as journal

* Corresponding author e-mail: kazu@umich.edu.
[†] Department of Mechanical Engineering.
[‡] Department of Pharmaceutical Sciences.
[§] Co-corresponding author e-mail: grosania@umich.edu.

10.1021/ci900029v CCC: $40.75
Published on Web 07/21/2009

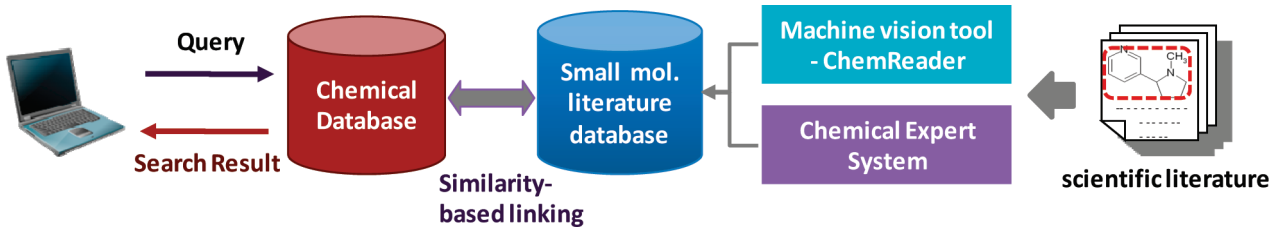

Figure 1. Annotation pipeline to a chemical database using a similarity-based linking method with a tunable level of estimated accuracy.

articles, patents, project reports, and books. In practice, it is a daunting task for chemical experts to compile all chemical information in the scientific literature published so far, and often such manual curation results in the high cost of access.[15,16] Therefore an automated system annotating chemical structures in the chemical database with one or more relevant links to the scientific literature is highly demanded.[17]

The traditional approach for automated knowledge extraction from the scientific literature is based on processing raw text information. In fact, various applications using text-mining and natural-language processing (NLP) technology have been developed to integrate unstructured data in the biological and biomedical literature into biological databases.[18] For example, the identification of biological entities such as genes, proteins, or diseases to facilitate the retrieval of relevant documents has been an area of interest in NLP for many years.[19−21] In the case of the chemical document processing, instead of sequences representing genes or proteins within a document, chemical named entities should be identified first. For this purpose, document segmentation and machine learning techniques have been successfully applied.[22,23] Since a chemical compound might be expressed in various ways including generic name, IUPAC systematic nomenclature, abbreviations, and index number (e.g., CAS registry numbers, EINECS and Beilstein registry numbers), extracted chemical named entities need to be converted into their chemical structure. There already are several name-to-structure converting tools such as OPSIN,[24] Lexichem,[25] ACD/Name to Structure,[26] or Name=Struct.[27] A demonstration of this approach can be found in the IBM Chemical Search alpha site[28] which identifies and indexes over 3.6 million chemical structures in the U.S. patent corpus from 1976−2005 using text-mining techniques and Name=Struct software.[29]

Another way to link entries in a chemical structure database with the scientific literature is to relate chemical structure diagrams embedded in the text of a scientific article to the corresponding structure entry in the database. Since novel chemical structures are usually referenced by chemical structure diagrams rather than chemical names in published articles and patents, this approach can provide a distinct advantage compared to the text-based approach mentioned above.[30] There are two essential stages in recognizing chemical structure diagrams from documents: identification of a chemical structure diagram and conversion of a diagram to a connection table. In a similar way as the text-mining approaches, the chemical structure diagrams in a digitized document can be identified using document processing and machine learning techniques.[31,32] Also, in order to translate raster images of the chemical diagrams into a standard, machine-readable chemical file format, several machine vision-based tools are available, including Kekule,[33] IBM OROCS,[34] CLiDE,[35,36] chemoCR,[37,38] OSRA,[39] and ChemReader.[40] However until now the annotation of chemical databases using a machine vision-based approach has not been directly examined.

Here, we demonstrate a machine vision-based approach for automated annotation by linking published journal articles to entries in a chemical database, PubChem. For chemical structure extraction we used ChemReader - a software tool for converting chemical structure diagrams into the connection-table, which outperformed other available software like OSRA V1.01 and CLiDE V.2.1 for all sets of images collected from different sources such as Web sites and real journal articles in our previous study.[40] In particular, it was observed that ChemReader kept its performance at the test images embedded in journal articles, while other software dropped their performance significantly. As a next step of the ChemReader project, we have designed and examined an annotation strategy which is capable of linking published real journal articles to entries in the chemical database (Figure 1). In the following sections, we describe how we addressed our annotation test and the test result as well as enhanced algorithms of ChemReader.

MATERIALS AND METHODS

Machine Vision Tool - ChemReader. ChemReader is a fully automated, machine vision-based tool for extracting chemical structure diagrams in research articles and translating them into standard, machine-readable chemical file formats. Figure 2 shows the essential recognition steps of a chemical structure diagram in ChemReader. The chemical structure diagram digital image consists of a long sequence of bits that give pixel-by-pixel values. In the first step, the pixels are grouped into components based on pixel connectivity. These connected components are then classified as text or graphic objects. Text objects are transferred to a character recognition algorithm and converted to character symbols. Since the results can contain nonexisting chemical symbols or valences, to detect and correct these errors, a chemical "spell checker", a recovery process similar to conventional OCR error correction, confirms the final chemical symbols. Graphical objects representing bond connectivity are analyzed using the (Generalized) Hough Transformation, Corner Detection algorithm, and a few other geometric operations. Finally, from recognized chemical symbols and bonds, the whole of the structural information is assembled, and a connection-table is generated, which can be converted into a standard chemical file format. The detailed description of the ChemReader algorithm can be found in our previous report.[40]

Virtual "Chemical Expert" System. Any chemical OCR systems including ChemReader, no matter how accurate they

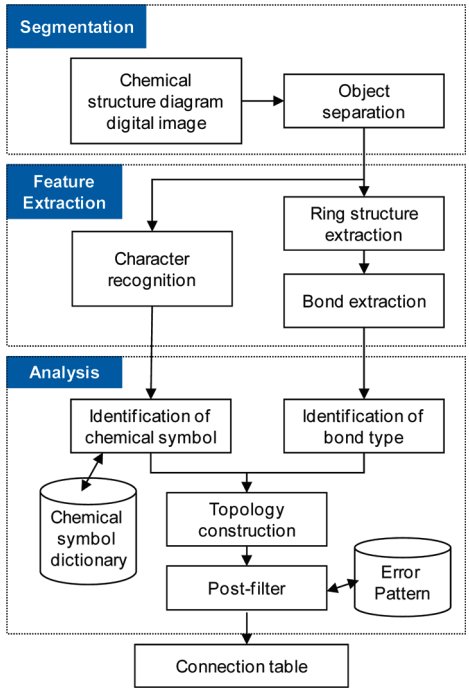

Figure 2. Chemical structure recognition process in ChemReader.

become in the future, will never be completely error free since there will always be chemical structure diagrams with low resolution, high noise level, and/or unconventional notations, which can disguise even the most sophisticated machine-vision algorithms. One strategy to deal with these errors is to avoid annotation with output structures that are likely to lead to false-positive links. By extension, since the accuracy of the output structure produced by a machine-vision tool is related to the relevance of annotated information, it would be possible to tune the accuracy of the annotation system by estimating a confidence in the recognition result and using it as a parameter for linking. Thus we have developed a virtual "Chemical Expert" system which can estimate the accuracy of recognized structures by examining a few main types of recognition errors described below.

• Number of molecules: ChemReader assumes that every input image contains only one chemical structure diagram. Thus multiple molecules in ChemReader output indicate that there are some errors which occurred in the recognition process. For example, if ChemReader misses a bond in the line detection or fails to locate a correct node position in the topology construction, the final molecular structure is often separated by several fragments. To identify these broken structures, the number of connected components in the graph of output structures is computed. If there are more than one connected components, the output will be regarded as wrongly recognized.

• Bond length: The chemical structure diagram drawn in the "standard" two-dimensional format keeps bond length uniform over the entire structure. Too short or long lines in the output structure might not correspond to true bonds in the original input structure. Thus divergence of extracted bond length could be an indication of errors which occurred in the recognition process. Here the most populated bond length in pixel is regarded as the representative length of the output structure. Based on the representative length, the Chemical Expert system sets an acceptable range for bond length to the given structure, which can be defined as $L_{MP} - T1 < Acceptablelength < L_{MP} + T2$, where L_{MP} denotes the representative length and both T1 and T2 are adjustable thresholds. Thus if a bond length is out of the range, the output structure will be filtered out.

• Bond angle: Since chains or ring systems which frequently appear in the chemical structure diagram are usually drawn by a fixed angle, specific bond angles such as 60°, 90°, 108°, 120°, and 180° are likely to be dominant in the bond angles of most chemical structures. Figure 3 is an example showing how a bond angle histogram could be different in correctly and wrongly recognized structures. It can be observed that bond angles around 120° are dominant in both histograms, but a wrongly recognized structure has more distributed bond angles. Given the ranges of acceptable bond angles, the Chemical Expert system detects odd bond angles in the output structure and then will filter it out if the number of odd angles is larger than a certain threshold.

• Chemical symbol: Each chemical symbol in the output structure has its confidence score given by a likelihood function implemented in the "Chemical Spell Checker" (see ref 40 for details). This confidence score can be also used to estimate the probability of existence of errors in the output structure. A chemical symbol with the likelihood value less than a certain threshold is regarded as a misrecognized symbol. If the number of misrecognized symbols exceeds half the number of chemical symbols in the structure, then the output structure will be filtered out.

• Distance of nonbonded atoms: It is very rare that a group of nodes are located close to each other in the 2D chemical structure diagram. In ChemReader, many types of recognition errors finally induce an irregular node distribution in the 2D space. For checking this, the number of node pairs that are within half the representative bond length from each other is calculated in the Chemical Expert system. If this number of adjacent node pairs is over a certain threshold, then the output structure is likely to have some recognition errors.

Similarity-Based Linking Method. A useful database annotation scheme does not necessarily require perfect, exact matches between database entries and scientific articles. In fact, the ability to link to similar but not identical structures may be important when the intent is to synthesize drug leads that are not identical to the molecule in question and to identify related compounds in the scientific literature. Such similar but not identical molecules, having been synthesized in other drug development projects, could provide some new ideas for developing a derivate for given virtual ligand candidate molecules. Thus, for the purpose of retrieving similar molecules from a chemical database, many different chemical-similarity search methods which use substructure keys, atom pairs, or other molecular properties have been developed and widely used.[41,42] The similarity between two molecules can be quantified by computing chemical coefficients such as the Tanimoto coefficient or Euclidean distance coefficient on the basis of their selected properties. As the number of chemical structures in a chemical database is explosively increasing, the similarity calculation should not be unnecessarily computationally heavy. Therefore, the Tanimoto coefficient in conjunction with the PubChem

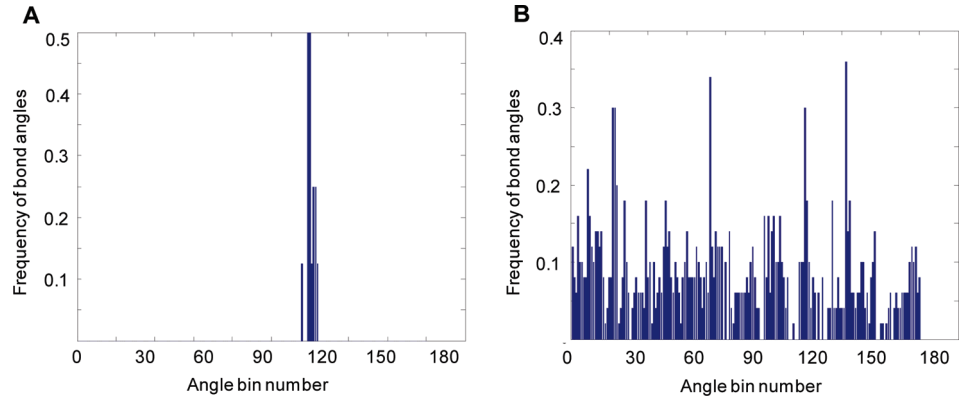

Figure 3. Bond angle histograms for ChemReader's output structure containing (A) no recognition error and (B) recognition errors.

Table 1. Article Sets for an Annotation Test[a]

journal title	no. of articles	no. of chemical structure diagrams
J. Am. Chem. Soc.	23	104
Angew. Chem., Int. Ed. Engl.	15	105
J. Med. Chem.	36	187
Chem. Commun. (Cambridge, U. K.)	13	61
Chem. Biol.	14	64
J. Biol. Chem.	14	58
Tetrahedron Lett.	6	30
total	121	609

[a] Articles are collected from the PubMed journal database.

Table 2. Contingency Table

		relevant[a]	
		yes	no
linked[b]	yes	True Positive (TP)	False Positive (FP)
	no	False Negative (FN)	True Negative (TN)

[a] Relevant structures are PubChem compounds having Tanimoto coefficients over 90% of the original (output) structure in this test. [b] Linked structures are PubChem compounds having Tanimoto coefficients over 90% of the original (output) structure in this test.

binary fingerprint[43] allowing a rapid evaluation of chemical similarity is employed in this test.

Article Sets for Annotation Test. The annotation test was performed on a total of 121 journal papers from seven different journals in the fields of biomedical and molecular biology, each of which has at least one chemical structure diagram. The papers in the portable document file (PDF) format are downloaded via links in the PubMed journals database,[44] and then embedded images are extracted by parsing the document file according to the PDF specification.[45] Images containing nonchemical structures are discarded by hand. In general, the figures in the journal papers contain not only chemical structure diagrams but also simple symbols (e.g., reaction symbols) and text for the additional description. Since the current version of ChemReader assumes that there is only one chemical structure diagram within an input image, components not related to the chemical structure are removed manually using an image editor. Also, an image file is broken into pieces of an image in case the image file contains multiple chemical structures. Table 1 shows the title of journals, number of sampled articles, and number of extracted structure diagrams. Among the 609 structure diagrams in the testing set, 38 structures are duplicated, but those are present in different articles or drawn differently in an article. For the validation of our annotation strategy, we obtain original connection tables for testing chemical structures by drawing structures manually using ChemDraw software.[46]

Chemical Database for Annotation Test. The target database for our annotation test is the Pubchem database[47] which is the largest, publicly accessible chemical structure database, encompassing a collection of 19 million unique structures that have been chemically synthesized or isolated and are therefore known to exist. As integrated with other components in the NCBI Entrez data warehouse, a structure in the PubChem database can have cross-reference links to related structures, bioassay data, bioactivity description, and literature related to the structure. However, since the majority of the entries in the PubChem database have been obtained from disparate sources such as commercial vendors, reference catalogues, and existing small molecule collections, current PubChem entries do not possess much information about the synthesis method of the molecules, their properties, or their biological activities.[48] Therefore the PubChem database might be one of the target databases which our annotation scheme can enrich.

Error Analysis. As a measurement of the chemical database's annotation performance, the recall and precision rates are used. Precision is the ratio of linked structures that are relevant, whereas recall is the ratio of relevant structures that are linked. Once a structure diagram s_i is processed by ChemReader and then linked to entries in PubChem, precision $P(s_i)$ and recall $R(s_i)$ rates of the structure diagram can be computed as follows

$$P(s_i) = \begin{cases} 1.0, & \text{if } |TP(s_i)| + |FP(s_i)| = 0 \\ \dfrac{|TP(s_i)|}{|TP(s_i)| + |FP(s_i)|}, & \text{otherwise} \end{cases}$$

$$R(s_i) = \begin{cases} 1.0, & \text{if } |TP(s_i)| + |FN(s_i)| = 0 \\ \dfrac{|TP(s_i)|}{|TP(s_i)| + |FN(s_i)|}, & \text{otherwise} \end{cases}$$

where $\mathrm{TP}(s_i)$, $\mathrm{FP}(s_i)$, and $\mathrm{FN}(s_i)$ mean respectively the set of true positive links, the set of false positive links, and the set of false negative links to the structure, s_i. Table 2 is the contingency table describing those four notions. The averaged precision and recall rates over an output set also can be defined as

$$\bar{P}(S) = \frac{1}{|S|}\sum_{s_i \in S} P(s_i) \text{ and } \bar{R}(S) = \frac{1}{|S|}\sum_{s_i \in S} R(s_i)$$

where S denotes the set of output structures. By looking at the distribution of precision and recall rates of processed structures, we would see how those two measures are correlated in our annotation scheme and also could discuss what features and errors occurred at the machine vision tool affect critically on annotation performance.

RESULTS AND DISCUSSION

ChemReader processed and converted the total 609 chemical structure diagrams to the associated connection tables (mol-files), which are then examined and filtered out by the Chemical Expert system. To demonstrate how the Chemical Expert system can be utilized to tune the overall annotation performance, we proceeded to work with two cases of a test with different conditions in the Chemical Expert system: Test I with tolerant constraints and Test II with strict ones. For tolerant conditions, the "bond angle" condition is turned off in Test I, while Test II has all the conditions turned on with a 10% smaller threshold for the "bond length" constraint than Test I. To see how the Chemical Expert system classifies output structures, the Tanimoto similarity coefficients are computed between original structures and recognized structures. For generating a PubChem fingerprint and computing Tanimoto similarity coefficient, an open-source code provided by the NIH Chemical Genomic Center (NCGC)[49] is used in conjunction with ChemAxon's JChem toolkits.[50]

In Test I, 212 output structures could survive, while only 145 structures satisfied the strict conditions of Test II. The Tanimoto similarity coefficients can be seen as the extent of correctly including chemically important features in the output structure. The more missed or misinterpreted PubChem substructure patterns the recognized structure has, the smaller the Tanimoto similarity coefficient becomes. Thus, in order to reduce wrong annotation effectively, the Chemical Expert system should be able to discriminate those wrong structures of small similarity coefficients from output structures. Figure 4 shows similarity histograms for both rejected and survived structures in Test I (Figure 4A) and Test II (Figure 4B). In both tests, we can observe that most of the wrong structures of small similarity coefficients are filtered out successfully. In particular, among structures of similarity coefficients less than 0.7, 82% and 92% of those structures are filtered out respectively in Tests I and II by the Chemical Expert system.

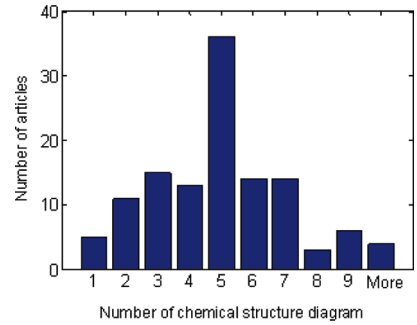

Figure 5. Histogram of the number of chemical structure diagrams in the sample articles. Most articles have multiple chemical structure diagrams.

There is also loss in correctly recognized structures which could not satisfy the conditions in the Chemical Expert system. However, the fraction of loss is much smaller than the fraction of wrong structures being filtered out. In addition, as each article usually has multiple chemical structure diagrams, discarding a portion of outputs corresponding to an article does not necessarily mean that the article cannot be linked to structures in a chemical database. In case of our sample articles, one single article has five chemical structure diagrams on average (Figure 5). Thus the likelihood of linking each article to entries in the PubChem database could be higher than the ratio of survived structures. This can be verified by seeing the number of articles which have a survived structure diagram. For example, the survived structures in Test I are only 35% of the total input chemical structure diagrams (Figure 6A). However, since 35% of chemical structure diagrams are distributed from 63% of sample articles (Figure 6B), those 63% of articles could be linked to structures in the PubChem database. In addition, (23 + 27)% of articles have at least one chemical structure correctly processed by ChemReader. Therefore they would lead to true positive links to the chemical structures in the PubChem database.

In Test II, the rejection ratio increases to 76% (11% more than Test I) due to the strict rejection conditions, but the loss in articles which can lead to true positive links is only 5(=50−45)% (Figure 7). In details, the Chemical Expert system in Test II filters out 8(=18−10)% more wrong structures with 3(=17−14)% loss in correct structures than Test I. By these further rejected outputs, the percentage of

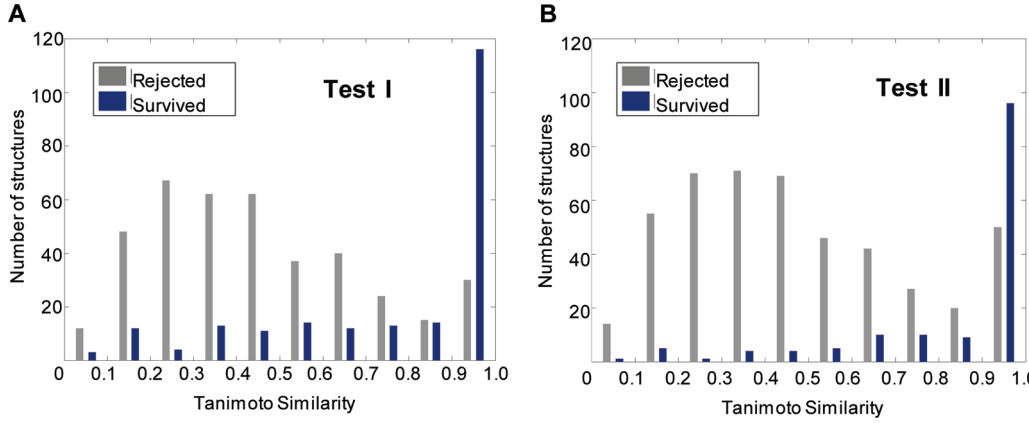

Figure 4. Tanimoto similarity histogram between original structures and ChemReader output structures: (A) Test I and (B) Test II.

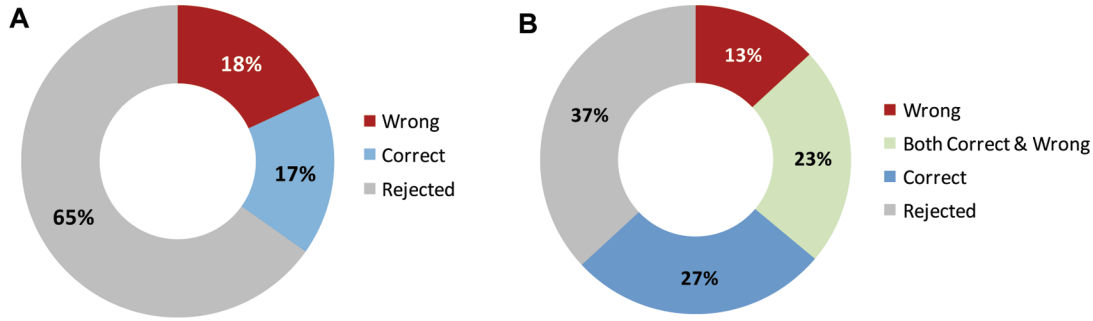

Figure 6. Percentages of rejected, correct, and wrong structures in Test I: (A) based on structures and (B) based on articles. 37%, 13%, and 27% of articles have no survived structures, only wrongly processed structures and correctly processed structures, respectively. Articles in the "Both Correct & Wrong" segment have both wrong structures and correct structures. Notice that 50% of the articles could be correctly annotated as shown in (B), even though 65% of the input images are rejected by ChemReader as shown in (A).

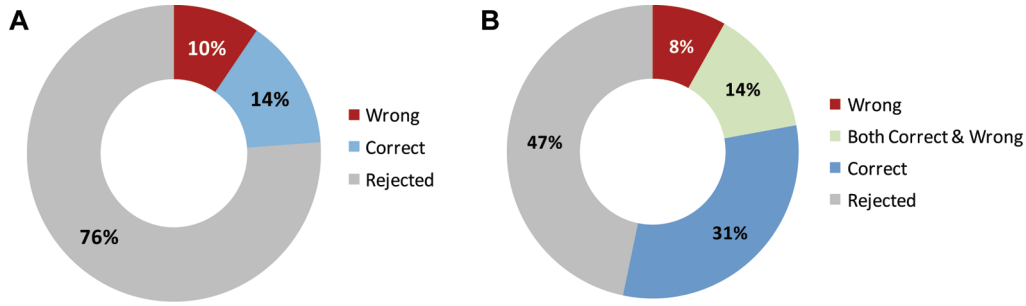

Figure 7. Percentages of rejected, correct, and wrong structures in Test II: (A) based on structures and (B) based on articles.

Table 3. Total Number of TP, FP, and FN[a]

	TP	FP	FN
Test I	29,540	34,386	28,642
Test II	23,277	6845	7874

[a] The number of false positive and negative links in Test II dramatically decrease compared to Test I.

Table 4. Averaged Recall and Precision Rates over Structures in Tests I and II

	av recall	av precision
Test I	0.69	0.8
Test II	0.8	0.88

the articles that can be linked decreases to 53%, but 31% of the articles will be linked to entries in the PubChem database without any false positive or negative link. In fact, it is confirmed that a subset of articles that will be linked through both wrong and correct outputs in Test I become articles having only correct outputs in Test II. Also, 5% of the articles that will be linked through wrong structures in Test I disappear in Test II. It should be noted again that, although 76% of output structures are filtered out in Test II, (14 + 31)% of articles can be linked through correct structures that would have one or more true positive links to the structures in the PubChem database.

Next, we proceeded to look at how many PubChem entries could be correctly annotated using filtered output structures in both Tests I and II. At that time, there were 19,187,639 unique chemical structures in the PubChem compound database. Using a 90% Tanimoto similarity as a threshold for linking the structure in the articles with PubChem entries, 43,704 and 27,967 PubChem compounds (unique structures) were identified as relevant entries to the outputs in Tests I and II, respectively. On the other hand, using ChemReader's output, 39,593 PubChem entries for Test I and 27,597 PubChem entries for Test II were retrieved. Since one PubChem entry can have multiple links to output structures, the sum of true and false positive links in Table 3 is more than the number of retrieved unique entries in both tests. All similarity searches are performed using the PUG SOAP interface[51] with a 90% Tanimoto similarity coefficient as a threshold.

Table 3 shows the total number of TP, FP, and FN links in both tests. Interestingly, most false positive and false negative links are originating from several structures. For example, in Test I, 80% of FP (27,497) and of FN (20,244) links are involved only in 10 and 15 structures, respectively. This verifies that the use of the Chemical Expert system can be a key factor impacting the overall annotation performance. In fact, the Chemical Expert system rejects 8 and 11 out of those 10 and 15 structures at issue. Subsequently, the number of FP and FN links dramatically decrease in Test II as shown in Table 3. Furthermore, also in Test II, 80% of FP (5585) and of FN (6298) links are attributed only to 3 and 6 structures, respectively. Therefore the development of the Chemical Expert system should proceed such that it can perceive error types commonly found in a small number of those molecules and filter them out selectively. Such evolution of the expert system would enable a reduction in the number of FP and FN links with less loss of TP links.

The quality of annotations (links) is estimated by precision and recall rates as described in the 'Error Analysis' subsection. Table 4 shows the averaged recall and precision rates of Tests I and II. The overall recall and precision rates based on total TP, FP, and FN numbers in Table 3 can be different than the recall and precision rates of individual structures averaged over the testing set (Table 4) because, as mentioned above, a small fraction of the testing structures contributes

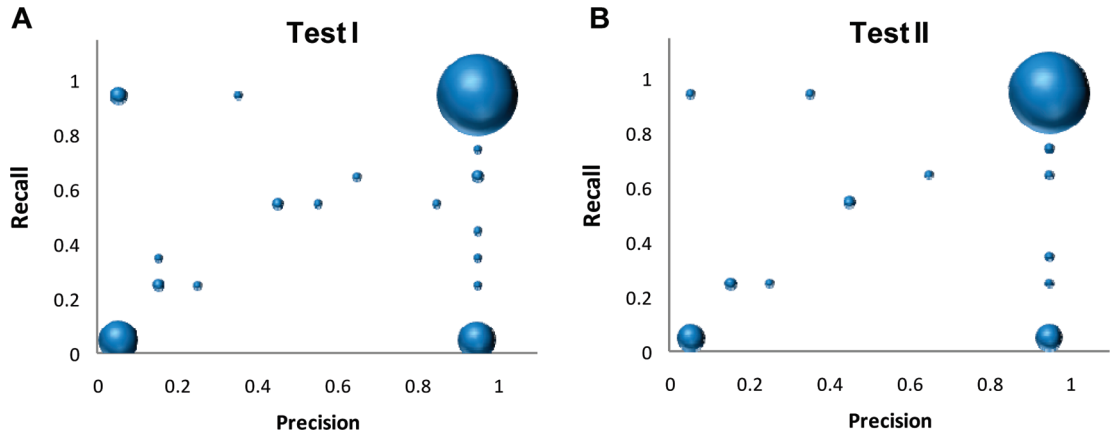

Figure 8. Distribution of recall and precision rates in Test I (A) and Test II (B). The size of the sphere is proportional to the number of structures corresponding to recall and precision rates.

to most of the FP or FN links. While 1.5 times more PubChem compounds could be annotated in Test I than Test II, both the averaged recall and precision rates of Test II are higher than those of Test I.

From the perspective of a chemical database user, the Chemical Expert system provides important information involving the reliability of the links. The relevancy of a link between a molecule in a chemical database and an extracted structure cannot be estimated with the Tanimoto similarity coefficient alone because of the possibility of recognition errors. The likelihood of a recognized structure corresponding to the original structure should be considered along with an intermolecular similarity. In fact, tolerant and strict conditions used in Tests I and II can be seen as certain levels of the estimated accuracy of the extracted structures. In this context, Table 4 illustrates a correlation between the stringency of the Chemical Expert system and the overall quality of annotations. This correlation indicates that, in a practical sense, the Chemical Expert system allows a chemical database user to request annotated information within a certain level of reliability. For a more practical application of the Chemical Expert system, it may employ a machine learning algorithm such as support vector machine or adaptive boosting in order to quantitatively estimate the reliability of the resulting annotations.

An analysis of the distribution of recall and precision rates indicates how the current annotation performance can be improved. Figure 8 shows the distribution recall and precision rates per structure in Test I (Figure 8A) and Test II (Figure 8B). The size of sphere is proportional to the number of structures, of which precision and recall rates are within a circle having as the center the center of the bubble and a radius of 0.05. The percentages of structures of which both precision and recall rates are under 0.5 are only 16.5% for Test I and 10.3% for Test II, with most of these having a zero recall or precision. So we could expect that annotation performance would dramatically increase without much loss of true positive links if the ChemReader's algorithm is enhanced such that those structures that lead to zero precision and recall rates are processed correctly.

Another point that we can address from the precision and recall distribution is that a wrong structure is likely to have either zero or one as the precision rate. Two big bubbles at the bottom-left and bottom-right in Figure 8 indicate these two groups of wrong structures. Structures having zero recall and 1.0 precision are those that could not be linked to PubChem entries even though PubChem contains relevant structures. 11.8% of Test I and 9.0% of Test II belong to that case. By visual inspection, we observed that a common feature of those structures is that they contain some user-defined chemical symbols (such as e.g. R, X, or Y) which ChemReader cannot interpret. As the chemical meanings of such symbols are usually described in the figure captions or text, by allowing ChemReader to access to the figure caption or text information around the chemical structure diagram, the recall rate would increase.

Based on this result, we plan to combine the existing functionality with text-mining and NLP technologies to use information in figure captions and the body of the manuscript for increasing the accuracy of the annotations. In traditional text-mining approaches, the article is indexed by several keywords including chemical names extracted from the title or the abstract section. For example, the National Library of Medicine (NLM) added chemical names into MeSH data so that articles in the PubMed database could be searchable by the chemical name.[52,53] Similarly, we propose that chemical structure diagrams in a scientific article can be used for MeSH indexing of articles. As demonstrated at the TIMI system,[54] such integration of both chemical and textual descriptors enables linking the article with the chemical structure, which can uncover the contextual scientific knowledge sought by the pharmaceutical, biological, and medicinal chemistry research community.

CONCLUSIONS

We have elaborated a tunable, similarity-based annotation strategy for linking molecules in a chemical database with scientific research articles, using a machine vision tool for translating images of molecules to atom and bond connectivity files. The proposed annotation strategy enables linking chemical structure diagrams within the scientific literature to chemically related structures in a chemical database in a practical manner. In particular, by using the Chemical Expert system, the reliability of the links can be tuned, and thus the accuracy of the annotations can be quantitatively assessed. For the validation, chemical structure diagrams in a total 121 journal articles were processed by ChemReader and then linked to entries in the PubChem compound database. The results show that ChemReader could process chemical

structure diagrams distributed over more than 45% of the articles, and those articles could be linked to PubChem entries with promising precision and recall rates. In addition, by adjusting the stringency of the conditions used in the Chemical Expert system, the overall performance of annotation could be tuned. Based on observations on wrongly processed structures leading to false positive/negative links during annotation, it is expected that the annotation performance would increase significantly by improving the accuracy of converting molecules that include user-defined, nonstandard chemical symbols as part of the drawing.

ACKNOWLEDGMENT

This work has been funded in part by NIH grant P20 HG003890-01 to G.R.R. We would like to thank Peter Dresslar (TorreyPath, Inc.) and Khalid B. Kunji for assisting with the manual processing of the chemical structures.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug Discovery* **2002**, *1*, 337–346.
(2) Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964–1969.
(3) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.
(4) Schreiber, S. L. The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chem. Eng. News* **2003**, *81*, 51–61.
(5) Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 220–227.
(6) Jonsdottir, S. O.; Jorgensen, F. S.; Burnak, S. Prediction methods and databases within cheminformatics: emphasis on drugs and drug candidates. *Bioinformatics* **2005**, *21*, 2145–2160.
(7) Rosania, G. R.; Crippen, G.; Woolf, P.; States, D.; Shedden, K. A Cheminformatic Toolkit for Mining Biomedical Knowledge. *Pharm. Res.* **2007**, *24*, 1791–1802.
(8) Weisgerber, D. W. Chemical Abstracts Service Chemical Registry System: History, Scope, and Impacts. *J. Am. Soc. Inf. Sci.* **1997**, *48*, 349–360.
(9) Kaiser, J. Chemists want NIH to curtail database. *Science* **2005**, *308*, 774.
(10) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2007**, *34*, 668–672.
(11) eMolecules. http://www.emolecules.com (accessed May 12, 2009).
(12) SciFinder. http://www.cas.org/SCIFINDER/SCHOLAR/ (accessed May 12, 2009).
(13) IDdb3. http://www.iddb3.com (accessed May 12, 2009).
(14) SureChem. http://www.surechem.org (accessed May 12, 2009).
(15) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133–4139.
(16) Baker, M. Open-access chemistry databases evolving slowly but not surely. *Nat. Rev. Drug Discovery* **2006**, *5*, 707–708.
(17) Banville, D. L. Mining chemical structural information from the drug literature. *Drug Discovery Today* **2006**, *11*, 35–42.
(18) Krallinger, M.; Erhardt, R. A.; Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today* **2005**, *10*, 439–445.
(19) Fukuda, K.; Tamura, A.; Tsunoda, T.; Takagi, T. Toward information extraction: identifying protein names from biological papers. *Proceedings of Pacific Symposium on Biocomputing*, Hawaii, U.S.A., Jan 4−9, 1998; pp 707−718.
(20) Krauthammer, M.; Rzhetsky, A.; Morozov, P.; Friedman, C. Using BLAST for identifying gene and protein names in journal articles. *Gene* **2000**, *259*, 245–252.
(21) Tanabe, L.; Wilbur, W. J. Tagging gene and protein names in biomedical text. *Bioinformatics* **2002**, *18*, 1124–1132.
(22) Kemp, N.; Lynch, M. Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 544–551.
(23) Wilbur, W. J.; Hazard G. F.; Divita, G.; Mork, J. G.; Aronson, A. R.; Browne, A. C. Analysis of Biomedical Text for Chemical Names: a comparison of Three Methods. *Proceedings of the AMIA Symposium,* Washington, D.C., U.S.A., Nov 7−11, 1999; pp176−180.
(24) Corbett, P.; Murray-Rust, P. High-Throughput Identification of Chemistry in Life Science Texts. *CompLife* **2006**, 107–118, LNBI 4216.
(25) Lexichem. http://www.eyesopen.com/products/toolkits/lexichem.html (accessed May 12, 2009).
(26) ACD/Name to Structure. http://www.acdlabs.com/products/name_lab/rename/batch.html (accessed May 12, 2009).
(27) Brecher, J. Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 943–950.
(28) IBM Chemical Search Alpha. https://chemsearch.almaden.ibm.com/chemsearch/SearchServlet (accessed May 12, 2009).
(29) Rhodes, J.; Boyer, S.; Kreulen, J.; Chen, Y.; Ordonez, P. Mining Patents Using Molecular Similarity Search. *Proceedings of Pacific Symposium on Biocomputing*, Hawaii, U.S.A., Jan 3−7, 2007; pp 304−315.
(30) Zimmermann, M.; Fluck, J.; Friedrich, C. M.; Hofmann-Apitius, M. A Critical Review of Information Extraction Technologies in Chemistry. Presented at The International Conference in Trends for Scientific Information Professionals [Online], Barcelona, Spain, Oct 21−24, 2007. Fraunhofer SCAI Web site. http://www.scai.fraunhofer.de/fileadmin/images/bio/chemoCR/Talks/Martin_Hofmann-Apitius_ICIC_talk.pdf (accessed May 20, 2008).
(31) Simon, A.; Johnson, P. Recent Advances in the CLiDE Project: Logical Layout Analysis of Chemical Documents. *J. Chem. Inf. Comput. Sci.* **1997**, *32*, 109–116.
(32) Gkoutos, G. V.; Rzepa, H.; Clark, R. M.; Adjei, O.; Johal, H. Chemical Machine Vision: Automated Extraction of Chemical Metadata from Raster Images. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1342–1355.
(33) McDaniel, J. R.; Balmuth, J. R. Kekule: OCR - Optical Chemical (Structure) Recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373–378.
(34) Casey, R.; Boyer, S.; Healey, P.; Miller, A.; Oudot, B.; Zilles, K. Optical Recognition of Chemical Graphics Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, Oct 20−22, 1993; pp 627−632.
(35) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A. G.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 338–334.
(36) Valko, A. T.; Johnson, A. P. CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition. *J. Chem. Inf. Comput. Sci.* **2009**, *49* (4), 780–787.
(37) Algorri, M. E.; Zimmermann, M.; Hofmann-Apitius, M. Automatic Recognition of Chemical Images. In *ENC 2007: Eighth Mexican International Conference on Current Trends in Computer Science*, Morelia, Mexico, Sep 24−28, 2007; IEEE Computer Soc.: Los Alamitos, California, U.S.A., 2007; pp 41−46.
(38) Algorri, M. E.; Zimmermann, M.; Hofmann-Apitius, M. Reconstruction of Chemical Molecules from Images. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, Aug 22−26, 2007; IEEE: New York, 2007; pp 4609−4612.
(39) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740–743.
(40) Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; Saitou, K. Automated Extraction of Chemical Structure Information from Digital Raster Images. *Chem. Cent. J. [Online]* **2009**, *3*, Article 4. http://journal.chemistrycentral.com/content/3/1/4 (accessed Mar 22, 2009).
(41) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(42) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.
(43) PubChem Substructure Fingerprint V1.2, 2007. NCBI PubChem Web Site. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt (accessed Mar 22, 2009).
(44) PubMed. http://www.ncbi.nlm.nih.gov/pubmed/ (accessed May 11, 2009).
(45) *Adobe PDF*. http://partners.adobe.com/public/developer/pdf/index_reference.html (accessed May 11, 2009).
(46) *ChemDraw Software*. http://www.cambridgesoft.com/software/ChemDraw/ (accessed May 11, 2009).
(47) PubChem database. http://pubchem.ncbi.nlm.nih.gov/ (accessed May 11, 2009).
(48) Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzeler, E. A. Large-Scale Annotation of Small-Molecule Librar-

ies Using Public Databases. *J. Chem. Inf. Model.* **2007**, *47*, 1386–1394.
(49) Open source code for generating PubChem Fingerprint provided by the NIH Chemical Genomics Center. http://ncgc.nih.gov/pub/openhts/code/NCGC_PubChemFP.java.txt (accessed May 11, 2009).
(50) ChemAxon's JChem toolkits. http://www.chemaxon.com/jchem/intro/index.html (accessed May 11, 2009).
(51) Pub SOAP for PubChem. http://pubchem.ncbi.nlm.nih.gov/pug_soap/pug_soap_help.html (accessed May 11, 2009).
(52) Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/ (accessed May 11, 2009).
(53) Baker, M. Open-access chemistry databases evolving slowly but not surely. *Nat. Rev. Drug Discovery* **2006**, *5*, 707–708.
(54) Singh, S. B.; Hull, R. D.; Fluder, E. M. Text Influenced Molecular Indexing (TIMI): A Literature Database Mining Approach that Handles Text and Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 743–752.

CI900029V