6th annual IEEE Conference on Automation Science and Engineering
Marriott Eaton Centre Hotel
Toronto, Ontario, Canada, August 21-24, 2010

Image-based Automated Chemical Database Annotation with Ensemble of Machine-Vision Classifiers

Jungkap Park, Kazuhiro Saitou, *Senior Member, IEEE*, and Gus Rosania

Abstract—**This paper presents an image-based annotation strategy for automated annotation of chemical databases. The proposed strategy is based on the use of a machine vision-based classifier for extracting a 2D chemical structure diagram in research articles and converting them into standard chemical file formats, a virtual "Chemical Expert" system for screening the converted structures based on the level of estimated conversion accuracy, and a fragment-based measure for calculation intermolecular similarity. In particular, in order to overcome limited accuracies of individual machine-vision classifier, inspired by ensemble methods in machine learning, it is attempted to use of the ensemble of machine-vision classifiers. For annotation, calculated chemical similarity between the converted structures and entries in a virtual small molecule database is used to establish the links. Annotation test to link 121 journal articles to entries in PubChem database demonstrates that ensemble approach increases the coverage of annotation, while keeping the annotation quality (e.g., recall and precision rates) comparable to using a single machine-vision classifier.**

I. Introduction

To search for chemical information in the scientific literature, chemical entities and their related information such as method of synthesis, chemical and biophysical properties, or biological activities need to be compiled in a structured form. With the aid of computer and informational techniques, cheminformatics research has devoted much effort into developing techniques for the storage, retrieval, and processing of chemical information in order to maximize the availability of chemical information published so far [1]. For example, scientists have registered new chemical structures with experimental properties to the CAS Registry System which is the largest commercially accessible chemical database in the world monitoring the scientific literature [2]. In the case of PubChem (the largest, publicly available chemical database integrated to the National Center for Biotechnology Information data warehouse), each chemical entries can have cross-reference links to related structures, bioassay data, and bioactivity description as well as relevant scientific articles. Thus, nowadays these chemical databases are, rather than a mere repository of molecular structure information, essential research tools allowing people can explore chemical information distributed over the world efficiently [3, 4].

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 articles, patents, project reports, and books. In practice, it is a very daunting task for chemical experts to compile chemical information in the scientific literature, and often such manual curation results in the high cost of access [5]. Therefore an automated system annotating chemical structures in the chemical database with one or more relevant links to the scientific literature is highly demanded.

The traditional approach for data mining 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 [6]. 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. Since a chemical compound might be expressed in various ways including generic name, IUPAC systematic nomenclature, abbreviations, and database index number, extracted chemical named entities need to be converted into their chemical structure by name-to-structure converting tools [7-10]. A demonstration of this approach can be found in the IBM Chemical Search alpha site [11].

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 alone, this *image-based* annotation approach can complement the text-based approach mentioned above [12]. Basically, there are three essential stages in the image-based annotation: identification of a chemical structure diagrams from documents, conversion of a diagram to a chemical file format, and linking the converted structure to relevant entries in a chemical database. In order to extract raster images of the chemical diagrams and convert them into a standard, machine-readable chemical file format, several machine vision-based classifier tools, so called chemical OCR systems such as Kekule [13], IBM OROCS[14], CLiDE [15], chemoCR [16], OSRA[17], and ChemReader [18] have been developed.

In our previous works, we proposed an image-based annotation strategy in cooperation with ChemReader which

Manuscript received February 12, 2010. This work was supported in part by Polyergic LLC.

J. Park and K. Saitou are with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109-2125 USA (e-mail: {jungkap,kazu}@umich.edu). G. Rosania is with the Department of Pharmaceutical Sciences, University of Michigan, Ann Arbor, MI 48109-1065 USA (e-mail: grosania@umich.edu).

978-1-4244-5449-5/10/$26.00 ©2010 IEEE

has been developed in our lab [19]. As a case study, the proposed annotation scheme was tested by attempting to link chemical structures in real journal articles to entries in the PubChem database. Even though our ChemReader outperformed other available software like OSRA V1.0.1 and CLiDE V2.1 in a recognition test [18], many of chemical structures were discarded before the annotation task due to limited accuracy of machine vision algorithms. In fact, almost half of target articles couldn't be linked to any entries in the PubChem database.

Here, to achieve higher chance of linking articles to structures in a chemical database, we address the annotation study again utilizing an ensemble of machine-vision classifiers, rather than depending on a single machine vision classifier. Proposed method is inspired by ensemble approaches in machine learning, which uses multiple models to obtain better predictive performance [20]. That is, a single chemical structure diagram is allowed to be processed by multiple machine-vision classifiers. Multiple interpretations of an input structure are then used for linking the original input structure to relevant entries in a chemical database.

II. Machine-vision classifier - Chemical OCR system

Chemical OCR systems extract a 2D chemical structure diagram from a document and convert it into a standard chemical file formats. Fig. 1 shows the essential recognition steps of a chemical structure diagram. The first step in chemical OCR systems is to identify all the individual chemical diagrams in a document, and segment these diagrams into atoms and bonds connected to form an individual molecule. Next, with the isolated chemical structure image which consists of a long sequence of bits that give pixel-by-pixel values, 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. Graphical objects representing bond connectivity are analyzed via the vectorization process [21] or the (Generalized) Hough Transformation [22, 23]. 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 chemical OCR systems can be found in our previous report [18].

ChemReader which is the one of chemical OCR systems employed here is a software developer toolkit tailored to a chemical database annotation scheme. The recognition algorithms are optimized to achieve high accuracy and robust performance sufficient for fully automated processing of research articles. In particular, for robust bond detection, ChemReader employs the Hough Transformation which is tolerant to noise. Also, for intelligent chemical symbol recognition, a chemical "spell checker," a recovery process similar to the conventional OCR error correction, is implemented in ChemReader. These features enable ChemReader to process complicated structures or symbol abbreviations in the low-resolution image.

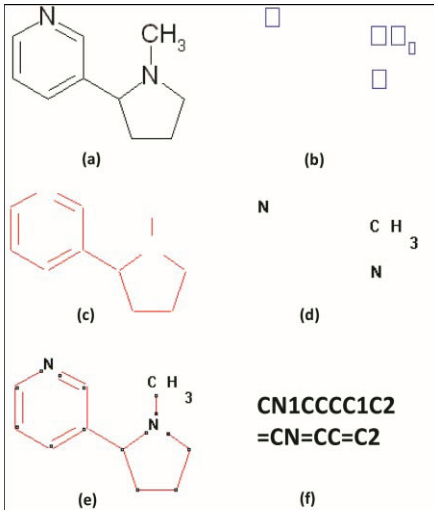

Fig. 1. General recognition steps of chemical structure diagram images. (a) input image, (b) character-line separation, (c) bond recognition, and (d) character recognition, (e) topology construction, and (f) data output in standard chemical file format.

Another chemical OCR program present in this report is OSRA-recently released, open source software written by the CADD group at the National Cancer Institute. Since most machine vision algorithms could yield quite different interpretations of the same input with a slightly different parameter setting, OSRA attempt to process a structure multiple times by varying parameters, and then picks one as an output based on its own empirical confidence function. This iterative processing of the same input could improve the overall ratio of correct outputs, so long as the confidence function is reliable enough. In our previous study, OSRA shows the most comparable accuracy to ChemReader. In this study, the latest version, OSRA V1.3.3 has been used (Recently, a new version, OSRA V1.3.4 was released).

III. Chemical expert system

Any chemical OCR systems including ChemReader and OSRA, no matter how accurate they 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 most sophisticated machine-vision algorithms. As a remedy, we have introduced a virtual "Chemical Expert" system which can estimate the reliability of extracted structure. Assuming that the reliability of output structure produced by chemical OCR systems is related to the relevance of annotated information, the Chemical Expert system examines a few main types of recognition errors and then judge if the output structure can be used further in the annotation pipeline. To estimate the reliability of output structure, following features are checked.

1) *Number of fragmentized molecules*: Assumes that input image contains only one chemical structure diagram, molecular fragments in the output structure indicate

recognition errors like missing bonds or wrong node conjunction.

2) *Bond length*: The chemical structure diagram drawn in the “standard” two-dimensional format keeps bond length being uniform over entire structure. Thus divergence of extracted bond length could be an indication of errors occurred in the recognition process.
3) *Bond angle*: Since chains or ring systems which are frequently appeared in the chemical structure diagram are usually drawn by 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.
4) *Non-existent chemical symbol*: Frequently, chemical OCR systems fail to interpret atomic symbols or chemical abbreviations in the chemical structure. Also, output symbols which do not make chemical sense are examined.
5) *Inconsistent distance between neighboring atoms*: It is very rare that a group of nodes are located close to each other in the 2D chemical structure diagram.

IV. Image-based chemical database annotation with ensemble approach

A. Similarity-based linking

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

B. Chemical database

The target database for our annotation test is the Pubchem database which is the largest, publicly accessible chemical structure database, encompassing a collection of 26 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 [25]. Therefore the PubChem database might be one of the target databases which our annotation scheme can enrich.

TABLE I
Article sets for an annotation test

Journal index	Journal title	# of articles	# of chemical structure diagrams
1	J. Am. Chem. Soc.	23	104
2	Angew. Chem., Int. Ed. Engl.	15	105
3	J. Med. Chem.	36	187
4	Chem. Commun.	13	61
5	Chem. Biol.	14	64
6	J. Biol. Chem.	14	58
7	Tetrahedron Lett.	6	30
	Total	121	609

TABLE II
Contingency table

		Relevant[a]	
		Yes	No
Linked[b]	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

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

C. Article set

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, and then embedded images are extracted by parsing the document file according to the PDF specification. 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

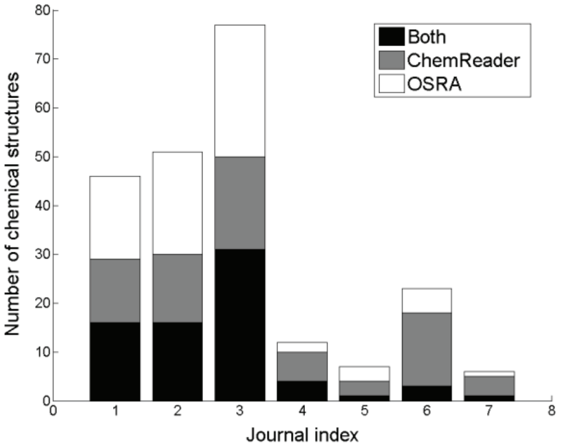

Fig. 2. Number of successful outputs produced by ChemReader or OSRA grouped by journal index. The successful output means that the output structure has Tanimoto similarity over 90% to the original structure.

validation of our annotation strategy, we obtain original connection tables for testing chemical structures by drawing structures manually using ChemDraw software [26].

D. Ensemble approach

Ensemble is a machine-learning technique for combining multiple models in attempt to obtain better predictive performance. To overcome weak accuracies of individual machine vision classifier (i.e. chemical OCR system), it would be possible to combine recognition results from multiple chemical OCR tools to improve the annotation performance. In our previous annotation test, a single output structure produced by only ChemReader was used to estimate the Tanimoto similarity coefficients between the associated input structure and entries in a chemical database. However, since different machine-vision algorithms could have different strengths in particular types of structures, multiple interpretation of the same input structure would increase the chance of including correct structure information.

In this study, two machine-vision classifiers, the ChemReader and OSRA have been employed to obtain two multiple output structures for the same input chemical structure diagram. That is, given an image of 2D chemical structure s_i, both ChemReader and OSRA produce their own output structures, $s_o^{ChemReader}$ and s_o^{OSRA} respectively. Since it is never easy to select one output structure which is likely to be correct for the annotation, the ensemble approach utilize both output structures together. Any PubChem entries having Tanimoto similairty over 90% to either $s_o^{ChemReader}$ or s_o^{OSRA} are linked to s_i. Thus if either $s_o^{ChemReader}$ or s_o^{OSRA} is correct output, all relevant entries are correctly annotated and then the number of true positive links is maximized. For the comparison, we performed the annotation test with individual machine vision classifier as well as the ensemble approach. Therefore there are three rounds of annotation tests, each of which utilizes one of following sets respectively.

- *ChemReader set*: a set of ChemReader output structures
- *OSRA set*: a set of OSRA output structures
- *Ensemble set*: the union of ChemReader set and OSRA set

E. Performance estimation

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 a chemical OCR system and then linked to entries in the PubChem, precision $P(s_i)$ and recall $R(s_i)$ rates of the structure diagram can be computed as follows.

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

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

where $TP(s_i)$, $FP(s_i)$ and $FN(s_i)$ 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 II is the contingency table describing those four notions. The averaged precision and recall rates over an output set also can be defined as

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

where S denotes the set of input structures.

V. Discussion

First of all, the recognition results show that successfully processed output structures by ChemReader or OSRA are not much overlapped. The Tanimoto similarity 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 becomes. Thus suppose output structures having Tanimoto similarity over 90% against corresponding input structures are successful outputs of chemical OCR systems, only 32% of successful outputs are commonly belong to both ChemReader set and OSRA set. Fig. 2 shows the number of successful outputs in each journal produced by ChemReader, OSRA, or Both. In every journal, the ratio of overlapped successful outputs by both systems is less than 40% of total successful outputs. This clearly implies that the ensemble set would induce more true positive links than could be obtained only from either ChemReader set or OSRA set.

The Chemical Expert system examines all output structures and removes output structures which do not satisfy a certain level of reliability. Fig. 3 and 4 show similarity histograms for both removed and accepted output

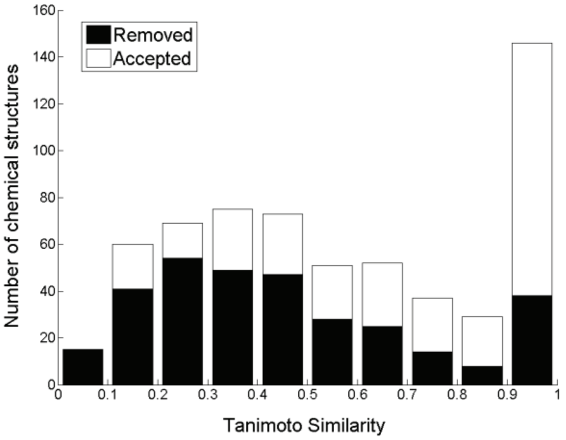

Fig. 3. Tanimoto similarity histogram between original structures and ChemReader output structures.

structures of ChemReader and OSRA, respectively. In both cases, most of the unsuccessful output structures of small similarities are desirably removed. Even though there is loss in successful structures which could not satisfy the conditions in the Chemical Expert system, the fraction of loss is much smaller than the fraction of unsuccessful structures being filtered out. Table III summarizes the final results of the Chemical Expert system in this test.

TABLE III
NUMBER OF REMOVED, ACCEPTED & CORRECT, AND ACCEPTED & WRONG STRUCTURES

		ChemReader	OSRA	Both[a]
Accepted	successful	85	61	29
	unsuccessful	203	192	136
Removed	successful	38	41	5
	unsuccessful	283	315	207

[a]Set of chemical structures commonly belong to ChemReader and OSRA outputs

Number of accepted chemical structures in ChemReader set, OSRA set, and Ensemble set is 288, 253, and 541, respectively. Note that 541 structures in Ensemble set corresponds to only 376 input chemical structures. Using a 90% Tanimoto similarity as a threshold for linking the structure in the articles with PubChem entries, 72,223, 88,100, and 101,691 PubChem compounds (unique structures) were identified as relevant entries to the molecules in ChemReader set, OSRA set, and Ensemble set, respectively. On the other hand, 43,577 PubChem entries via ChemReader set, 43,244 PubChem entries via OSRA set, and 69,469 PubChem entries via ensemble set were retrieved. All similarity searches are performed using the PUG SOAP interface [27] with a 90% Tanimoto similarity coefficient as a threshold.

Table IV shows the total number of TP, FP, and FN links in three tests. Since one PubChem entry can have multiple links to output structures, the sum of true and false positive links in Table IV is more than the number of retrieved unique PubChem entries. As expected, Ensemble set has much more TP links than any of ChemReader set or OSRA set. Table V shows the averaged recall and precision rates of three tests. Compared to ChemReader set and OSRA set, the ensemble set shows the highest recall rate as well as a moderate precision rate. In order word, without much decrease in precision rate, the ensemble approach could increase number of useful annotations.

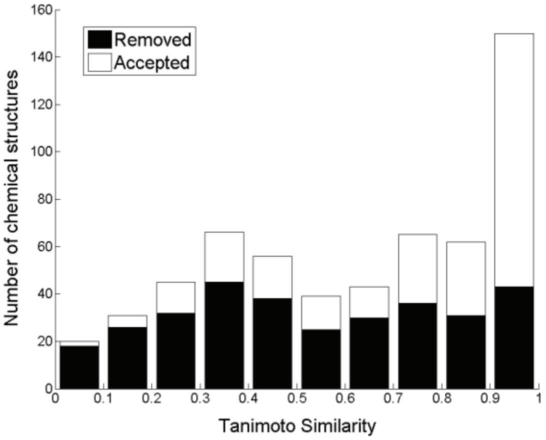

Fig. 4. Tanimoto similarity histogram between original structures and OSRA 1.3.3 output structures.

TABLE IV
TOTAL NUMBER OF TP, FP, AND FN LINKS

	TP	FP	FN
ChemReader	24592	30844	47631
OSRA	33105	21067	54995
Ensemble	45707	51535	55984

TABLE V
AVERAGED RECALL AND PRECISION RATES OVER STRUCTURES

	Avg. Precision	Avg. Recall
ChemReader	0.563	0.569
OSRA	0.491	0.568
Ensemble	0.544	0.619

It should be noted that our current ensemble approach does not much improve in the recall and precision rates while the number of true positive links increase significantly. The reason is attributed to the manner constructing the ensemble set. Currently, we simply join the ChemReader set and the OSRA set for the ensemble set. However this simple union causes inherently the increase in FP and FN links as well as TP links. To improve the quality of annotations significantly as well as in the coverage of annotation with the ensemble approach, it would be required that the ensemble set includes only the most reliable output structure from multiple outputs produced by different chemical OCR tools. This might become possible as making the Chemical Expert system be able to estimate the reliability of output structures from different chemical OCR tools.

For the future work, 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 [5]. 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 [28], 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.

VI. Conclusion

This paper presents an image-based annotation strategy for automated annotation of chemical databases. The proposed strategy is based on the use of a machine vision-based classifier for extracting a 2D chemical structure diagram in research articles and converting them into standard chemical file formats, a virtual "Chemical Expert" system for screening the converted structures based on the level of estimated conversion accuracy, and a fragment-based measure for calculation intermolecular similarity. In particular, in order to overcome limited accuracies of individual machine-vision classifier, inspired by ensemble methods in machine learning, it is attempted to use of the ensemble of machine-vision classifiers. Annotation test to link 121 journal articles to entries in PubChem database demonstrates that ensemble approach increases the coverage of annotation, while keeping the annotation quality comparable to using a single machine-vision classifier.

References

[1] G. R. Rosania, G. Crippen, P. Woolf, D. States, and K. Shedden, "A cheminformatic toolkit for mining biomedical knowledge," *Pharmaceutical Research,* vol. 24, pp. 1791-1802, Oct 2007.

[2] D. W. Weisgerber, "Chemical Abstracts Service Chemical Registry System: History, scope, and impacts," *Journal of the American Society for Information Science,* vol. 48, pp. 349-360, Apr 1997.

[3] D. K. Agrafiotis, V. S. Lobanov, and F. R. Salemme, "Combinatorial informatics in the post-genomics era," *Nature Reviews Drug Discovery,* vol. 1, pp. 337-346, May 2002.

[4] M. A. Miller, "Chemical database techniques in drug discovery," *Nature Reviews Drug Discovery,* vol. 1, pp. 220-227, Mar 2002.

[5] M. Baker, "Open-access chemistry databases evolving slowly but not surely," *Nature Reviews Drug Discovery,* vol. 5, pp. 707-708, Sep 2006.

[6] M. Krallinger, R. A. A. Erhardt, and A. Valencia, "Text mining approaches in molecular biology and biomedicine," *Drug Discovery Today,* vol. 10, pp. 439-445, Mar 2005.

[7] OpenEye Scientific software, "Lexichem," OpenEye Scientific software, Nov. 2009. [Online]. Available: http://www.eyesopen.com/products/toolkits/lexichem-tk_ogham-tk.html

[8] P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," *Computational Life Sciences Ii, Proceedings,* vol. 4216, pp. 107-118, 2006.

[9] Advanced Chemistry Development, "ACD/Name to Structure Batch," Advanced Chemistry Development, Dec. 2009. [Online]. Available: http://www.acdlabs.com/products/name_lab/rename/batch.html

[10] J. Brecher, "Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature," *Journal of Chemical Information and Modeling,* vol. 39, pp. 943-950, 1999.

[11] J. Rhodes, S. Boyer, J. Kreulen, Y. Chen, and P. Ordonez, "Mining patents using molecular similarity search," in *Pacific Symposium on Biocomputing 2007*, 2007, pp. 304-315.

[12] M. Zimmermann, J. Fluck, C. M. Friedrich, and M. Hofmann-Apitius, "A Critical Review of Information Extraction Technologies in Chemistry," in *The International Conference in Trends for Scientific Information Professionals*, 2007.

[13] J. R. McDaniel and J. R. Balmuth, "Kekule: OCR-optical chemical (structure) recognition," *Journal of Chemical Information and Computer Sciences,* vol. 32, pp. 373-378, 1992.

[14] R. Casey, S. Boyer, P. Healey, A. Miller, B. Oudot, and K. Zilles, "Optical recognition of chemical graphics," in *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, 1993, pp. 627-631.

[15] A. T. Valko and A. P. Johnson, "CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition," *Journal of Chemical Information and Modeling,* vol. 49, pp. 780-787, Apr 2009.

[16] M. E. Algorri, M. Zimmermann, and M. Hofmann-Apitius, "Automatic Recognition of Chemical Images," in *Current Trends in Computer Science, 2007. ENC 2007. Eighth Mexican International Conference on*, 2007, pp. 41-46.

[17] I. V. Filippov and M. C. Nicklaus, "Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution," *Journal of Chemical Information and Modeling,* vol. 49, pp. 740-743, Mar 2009.

[18] J. Park, G. R. Rosania, K. A. Shedden, M. Nguyen, N. Lyu, and K. Saitou, "Automated extraction of chemical structure information from digital raster images," in *Chemistry Central Journal*. vol. 3, 2009.

[19] J. Park, G. R. Rosania, and K. Saitou, "Tunable Machine Vision-Based Strategy for Automated Annotation of Chemical Databases," *Journal of Chemical Information and Modeling,* vol. 49, pp. 1993-2001, Aug 2009.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. vol. 1857, 2000, pp. 1-15.

[21] X. Hilaire and K. Tombre, "Robust and accurate vectorization of line drawings," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 28, pp. 890-904, 2006.

[22] M. C. K. Yang, L. Jong-Sen, L. Cheng-Chang, and H. Chung-Lin, "Hough transform modified by line connectivity and line thickness," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 19, pp. 905-910, 1997.

[23] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition,* vol. 13, pp. 111-122, 1981.

[24] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of Chemical Information and Computer Sciences,* vol. 38, pp. 983-996, July 1998.

[25] Y. Zhou, B. Zhou, K. Chen, S. F. Yan, F. J. King, S. Jiang, and E. A. Winzeler, "Large-Scale Annotation of Small-Molecule Libraries Using Public Databases," *Journal of Chemical Information and Modeling,* vol. 47, pp. 1386-1394, 2007

[26] CambridgeSoft, "ChemDraw," CambridgeSoft, Dec. 2007. [Online]. Available: http://www.cambridgesoft.com/software/ChemDraw/

[27] NCBI PubChem, "PUG SOAP," NCBI PubChem, [Online]. Available: http://pubchem.ncbi.nlm.nih.gov/pug_soap/pug_soap_help.html

[28] S. B. Singh, R. D. Hull, and E. M. Fluder, "Text Influenced Molecular Indexing (TIMI): A Literature Database Mining Approach that Handles Text and Chemistry," *Journal of Chemical Information and Computer Sciences,* vol. 43, pp. 743-752, 2003.