

Image-to-Structure Task by ChemReader

Jungkap Park¹, Ye Li², Gus R. Rosania³, and Kazuhiro Saitou¹

¹Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan, ²Shapiro Science Library, University of Michigan, Ann Arbor, Michigan, and ³Department of Pharmaceutical Sciences, University of Michigan, Ann Arbor, Michigan

1. BACKGROUND

Chemical structure recognition software aims to extract raster images of 2D chemical structure diagrams and convert them into a standard, machine-readable chemical file format. Such software, so called chemical OCR can be used for mining chemical entities appeared in scientific literature. Since traditional text-based mining methods haven't attempt to utilize image data in documents yet, chemical OCR software will pave a new way for the development of chemical literature mining [1, 2].

This year, the TREC Chemical IR campaign has launched a new topic called "Image-to-Structure (I2S) task" where participants are asked to process given images and recognize chemical structures in the images. While the immediate objective of this task would be to evaluate the existing chemical OCR software, it ultimately aims to create a platform to see how information in image data can be incorporated with existing text-mining approach to facilitate further development of chemical IR techniques.

We developed a chemical OCR software, ChemReader which specifically tailored to a chemical database annotation scheme [3, 4]. The recognition algorithms are optimized to achieve high accuracy and robust performance sufficient for fully automated processing of scientific articles. In our previous reports, ChemReader was able to outperform other chemical OCR software on several sets of sample images from diverse sources in terms of the rate of correct outputs and the accuracy on extracting molecular substructure patterns. Since then, other existing tools have been continuously updated, and new chemical OCR tools also have been released. Thus this task is a good opportunity for chemical OCR developers to evaluate their algorithms against common image set, and see strengths and weaknesses of their own comparing them to others.

Here we report how we performed the I2S task during May-July, 2011. We first briefly present the recognition algorithms in ChemReader, followed by the updates during the training. Then, we will show the results of its evaluation on test set and discuss major errors of ChemReader on the test. Most importantly, on the basis of the lessons learned from this task, we will discuss issues and insights in the chemical OCR development.

2. CHEMICAL STRUCTURE RECOGNITION TOOL - ChemReader

The algorithms of our chemical OCR tool, ChemReader, were already reported in our previous paper [4]. Since then, the overall structure of the algorithm has been changed with additional steps and enhanced algorithms. The latest version of ChemReader, consists of the following steps:

1. Identification of a cluster of pixels that contain a chemical structure

Assuming there is a single chemical structure diagram in given image, ChemReader collects set of pixels that are likely to be part of chemical structure by employing region growing method. First of all, pixels of input image are clustered based on pixel connectivity. After this pixel-based segmentation, it picks the largest connected component as initial seed, and examines neighboring connected components to determine whether neighbor components should be added to the region of chemical structure. This process is iterated until there is no feasible neighbor to be added. The decision is made based on a certain distance threshold. In general, the distance threshold is determined by considering distance between the initial seed and its nearest neighbor. Components that not are included in the region will not be processed in the next steps.

2. Preprocessing: re-sizing, de-noising, and bond length estimation

It is often necessary to resize and de-noise the input image so that the chemical structure diagram within the input image has optimally adjusted bond lengths and character sizes to ChemReader's recognition algorithms. With the first run of line detection as explained below, the length of single bonds is estimated. If the estimated bond length is shorter than a certain threshold (currently 25 pixel), the image is resized to ChemReader's preferred size.

3. Identification of Text(character)

Connected components that have similar heights and areas are labeled as characters. The most popular area/height combination is assumed to be those of text components. In order to distinguish the small isolated lines or circles representing bonds from the text components, the relative location and horizontal/vertical run profile of each component are also checked.

4. Identification of circles within benzene rings

Among non-character components, if pixels of components are distributed with almost same distances from the center of the component, those are regarded as circles for representing aromatic bonds.

5. Hatched bond detection

From this point, we can expect that residual pixels are part of normal, wedge or hatched bonds. Among residual connected components, ChemReader attempts to find short line segments having uniform length and interval, as well as being collinear in the direction perpendicular to

the direction of the short line segments. Components whose diagonal length is shorter than the estimated single bond length can be candidate segments for hatched bonds.

6. Skeletonization

Left-over pixels are assembled together in a bitmap space and then skeletonized. This step makes the ring structure detection and line detection algorithms run faster as well as minimizes the effect of variation in line thickness on line detection algorithm.

7. Hexagonal or Pentagonal ring structure detection

The pentagonal and hexagonal ring structures are directly identified using Generalized Hough Transformation (GHT). This step is specially designed for low resolution images because line detection algorithm often fails to detect all the lines of ring structure correctly. The direct extraction of ring structure helps to construct the topology of small organic molecules more accurately and efficiently.

8. Line detection

ChemReader employs a modified Hough Transformation (HT) and a corner detection algorithm for extracting lines. While the HT provides information about the location and the direction of detectable lines, the corner detection algorithm allows determining the end points of them.

9. Filtering out collected lines

Sometimes text components or noise components are not clearly identified before entering the line detection step. So short line segments are filtered out and examined again in the next step.

10. Identification of Text(character) over {unidentified pixels & identified pixels in step 3}

Entering this step, there are unidentified small fragments including filtered line segments or left-over pixels. Such fragments are examined again whether there are text components. From original image, pixels corresponding to those small components are collected. Those pixels are assembled with pixels of identified text components in an empty bitmap space. From this image, text components are detected and finally confirmed. In this manner, characters that are glued to graphic components can be separated successfully.

11. Character recognition

Text components are sent to character recognition engine. ChemReader runs two different character recognition engines: GOCR open source library and its own engine. While GOCR is based on template matching of character features, the latter engine is based on the calculation of Euclidean distance between pixels of the character and of the character models of several fonts. Confidence scores for candidate characters is sum of two confidence scores given by those two engines. Total 10 candidate characters are assigned to each character.

12. Chemical spell checker

Given candidate characters and their confidence scores, the chemical spell checker tries to find a most likely chemical word based on a predefined, frequently-used chemical symbol table which contains 770 chemical abbreviations and fundamental chemical rules on molecular formula containing nonmetal and hydrogen atoms.

13. Repeat step 8-9 over unidentified pixels

Among unidentified components, there can be lines to be detected. So step 8 and 9 are repeated over left-over pixels. With less strict line-filtering criteria, every possible line segments are detected. That is, even if a line segment is significantly shorter than the estimated single bond length, ChemReader accepts it once the short line can be arranged together with previously detected bonds.

14. Merging or breaking lines

Often one bond is fragmented by multiple line segments. So if the end points of two lines are very close and their relative angle is about 180 degree, they need to be merged into one line segment. In contrast, if there should be junction nodes in the middle of a line segment, the line segment is split into multiple segments.

15. Graph construction

First, every end point of the identified bonds and center points of the identified chemical symbols are labeled as a node. Next, among these nodes, the ones located within a certain distance are merged into a single node.

16. Identification of connected components in the graph data structure

Since it is assumed that there is only one chemical structure in the input image, we need to pick only one connected component of the constructed graph. If there are more than one connected component, ChemReader selects the largest component and discard other components.

17. Output the connection table

3. TREC-CHEM 2011 Image2Structure Task

3.1 Image-sets

Two sets of images, training-set and test-set were given for the I2S Task. Each set consists of 1000 images in TIF format, each of which has only one chemical structure. According to the TREC-CHEM 2011 guideline, the set of images and reference structure have been chosen to satisfy the following criteria:

- No "M..." records except for "M END" - this gets rid of brackets (polymers), as well as charges and isotopes and other less common things
- Only one fragment in connection table
- The only allowed elements are C N O S F Cl Br I P and H; no r-groups
- No "wavy" bonds
- Valid InChI can be created
- Larger than 6 atoms
- Smaller than molecular weight 1000

3.2 Evaluation

"The evaluation is done automatically based on an existing set of corresponding SD files for each image using a structure based comparison algorithm (exact matching)."

In the training, we used graph matching algorithm implemented in ChemAxon's JChem toolkits (<http://www.chemaxon.com/jchem/intro/index.html>). We also calculated Tanimoto similarity of PubChem fingerprints between output structures and of given reference structures.

3.3 Small training set

We selected 100 images from 1000 training images. Because all images in training set look very similar in terms of the image resolution and the complexity of chemical structure, we decided to use a small set of training images for quick evaluation and error analysis.

3.4 Training

We performed three rounds of major trainings. In each round, we categorized main types of errors and prioritized them according to the frequency of occurrence.

Figure 1 shows ChemReader's progress in the training. Each round of the training could increase the accuracy by ~15%. The first trial with untrained ChemReader gave us 57% accuracy. Most of chemical structures in the training set have uniform bond length, bond thickness and character size. Even we could hardly see unconventional notations in the training-set. So we decided to remove unnecessary heuristic algorithms in ChemReader which are actually required to recognize chemical structure diagrams with low resolution, high noise level, and/or unconventional notations. Followings are major changes that were made to ChemReader during the training.

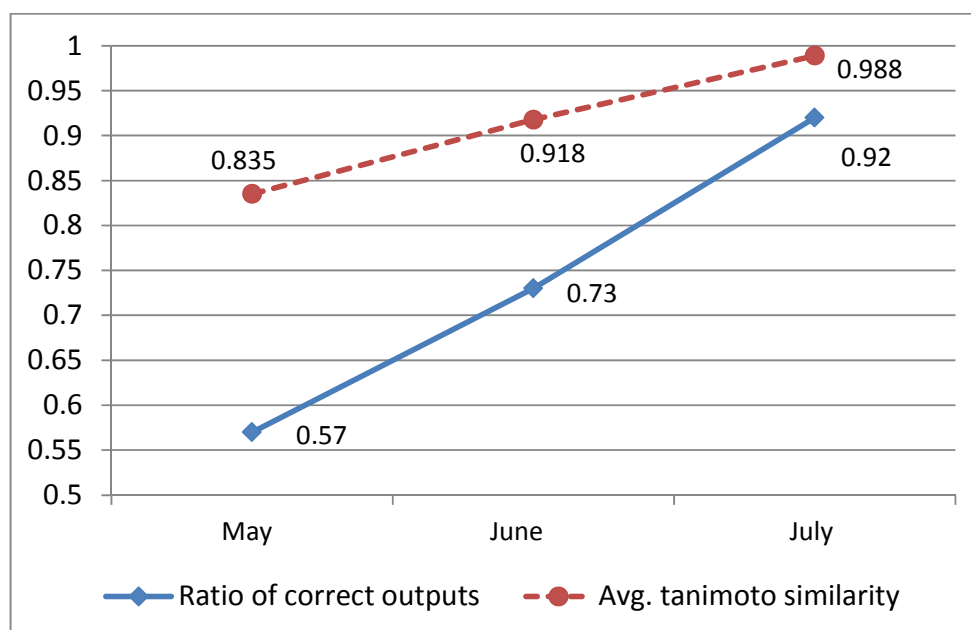


Figure 1. Training progress of ChemReader

- **Deactivated algorithms for I2S task**
 - Re-sizing and De-noising in step 2
 - Isolating subscripts which are likely glued to neighbor normal characters in step 11
 - Character type classification {upper, lower, subscript, superscript}
 - Merging lines in step 14
- **Limit possible characters:** { C N O S F I B r l P H h M e E t 1 2 3 4 5 6 7 8 9 () }
- **Set parameters to conservative values**
 - Minimum number of line segments representing hatched bond = 4
 - Minimum number of characters to estimate character height = 1
Possible character height = 17~22 pixels
- **Update chemical dictionary**
Only a few chemical abbreviations such as “Me”, “Ph” or “Et” appear in the training set. Most of symbols in the original dictionary are not necessary for this task. So we replaced the original dictionary with a light version of dictionary including abbreviated symbols appeared in the training set.

- **Loss of precision due to type conversion**

In ChemReader, type conversions were frequently occurred. During the training, we realized that several repeated conversions could make significant differences in the result. For example, object translation/rotation in bitmap space or transformation between Hough space and bitmap space. By using a round function, we could minimize the effect of the loss of precision due to type conversion.

- **Fragmentized characters**

In step 3, we assumed that each character corresponds to a connected component. However, we noticed that some characters in training set are fragmentized. In order to isolate such fragmentized characters correctly, we cluster pixels by region-growing approach. Thus a clustered segment has multiple connected components.

4. TEST RESULT

We submitted two sets of outputs for different parameter setting. Test I and Test II in Table 1 shows evaluation results for our submission. Both are significantly different from the estimated accuracy in the final training. There were two main causes of the accuracy decrease as follows:

- Stereo chemistry

As many chemical structures in image set has unidentified bond-stereo types, we set ChemReader to ignore bond stereo type. In addition, the ChemAxon's graph matching algorithm used in our training doesn't take bond stereo types into account. However, the evaluation method used in the actual test requires exact bond-stereo type matching. We found that the omission of bond stereo types could decrease the accuracy by ~10%.

- A bug in corner detection code

In addition to stereo chemistry issue, it turns out that our small training-set is an inappropriate sample. In fact, our corner detection code had a bug which causes to fail finding the end point of lines touching the boundary of the image. Unfortunately, we couldn't see the bug during the training because all chemical structures in the small training set are placed in the middle of image with a certain margin space. This bug also lowered the test accuracy by ~10%.

Since two issues above are not really related to algorithmic and/or parametric tunings, we re-ran ChemReader for test set after resolving above two issues in order to evaluate actual ChemReader's capability. Parameter values used in Test I was applied. For fair comparison, instead of our evaluation tool used in the training, we employed the evaluation tool used in the actual test. Test III in Table 1 shows the result. The ratio of correct outputs is 93% (930/1000) which is comparable to the final training accuracy. Even though our submitted result show very low accuracy compared to other participants, it is noted that Test III's accuracy is comparable to the highest accuracy (Figure 2).

Table 1. Image2Structure test result of ChemReader

	# of correct outputs	Avg. Tanimoto similarity
Test I	691	0.9769
Test II	689	0.9823
Test III	930	0.9913

5. Error analysis

We randomly selected 20 samples from Test III result, and categorized error type. Table 2 shows major recognition errors and their examples occurred in Test III. The most frequent error type is wrongly merged nodes which usually happen when two nodes are too close to be distinguished by a distance threshold. Secondly, while filtering out short line segments, we often miss normal bonds in structures. However, we believe that these two types of errors can be avoided when some chemical intelligence is incorporated into node merging and line filtering steps.

A nonstandard representation of chemical structures is one of factors causing recognition errors. For example, noise symbols which are not part of chemical structures would confuse ChemReader. Often wedged or hatched bonds drawn with a different style gives another challenge to chemical structure recognition algorithms. Also, current version of ChemReader is not capable to interpret 3D crossing bonds in structures. This might be another ability to be developed in the next version of ChemReader.

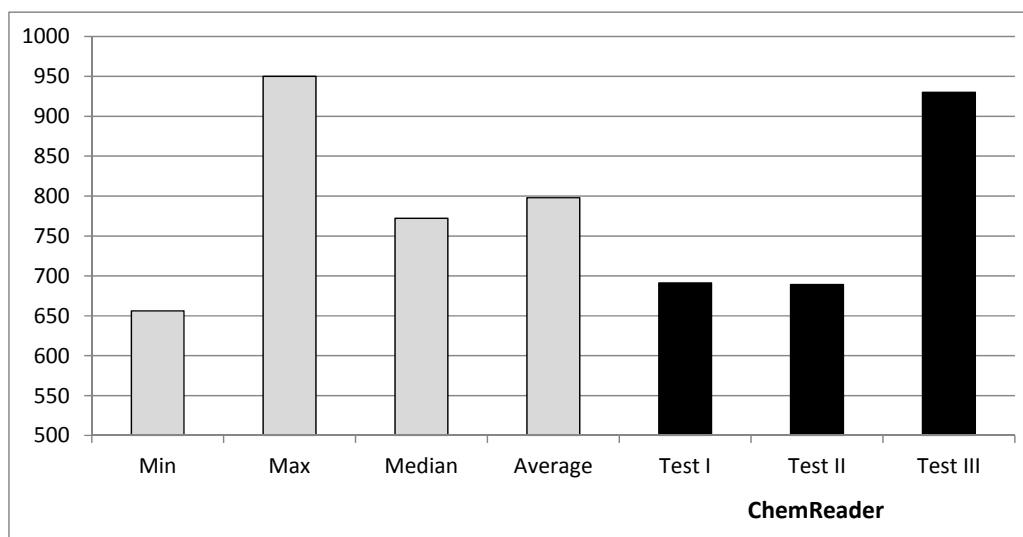
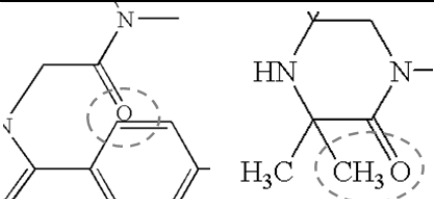
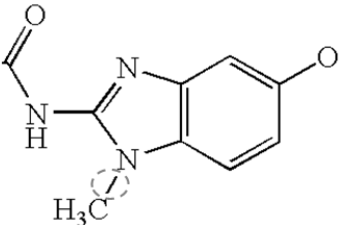
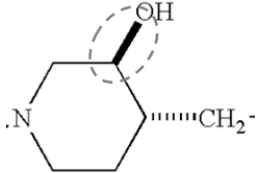
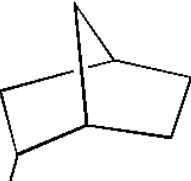
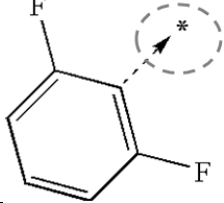


Figure 2. 2011 TREC-CHEM Image2Structure task result

6. Conclusion

We had the first round of the I2S task. Because of the simple omission of the stereo chemistry in the output structures and inappropriate training samples, ChemReader scored lower accuracy than expected. However, after resolving the stereo bond issue and a bug in the corner detection, ChemReader could obtain 93% accuracy. The error analysis on the final test indicates that ChemReader needs to incorporate more chemical intelligence in its algorithms.

Table 2. Error types, frequency and examples in Test III (20 samples)

Error type	Frequency	Examples in the test set
Wrongly merged nodes	6 (30%)	
Missed bonds	4 (20%)	
Incorrect bond stereo type	3 (15%)	
3D crossing bonds	3 (15%)	
Noise symbols around structure	2 (10%)	
Etc.	2 (10%)	

REFERENCES

- [1] D.K. Agrafiotis, V.S. Lobanov, and F.R. Salemme, "Combinatorial informatics in the post-genomics era," *Nat. Rev. Drug Discov.*, vol. 1, (no. 5), pp. 337-346, May 2002.
- [2] G.R. Rosania, G. Crippen, P. Woolf, D. States, and K. Shedden, "A cheminformatic toolkit for mining biomedical knowledge," *Pharmaceutical Research*, vol. 24, (no. 10), pp. 1791-1802, Oct 2007.
- [3] 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, (no. 8), pp. 1993-2001, Aug 2009.
- [4] 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," *Journal*, [online], vol. 3, (Date 2009), Available <Go to ISI>://000264395400001.