Optical Structure Recognition Application entry in Image2Structure task

Igor V. Filippov[a], Dmitry Katsubo[b] and Marc C. Nicklaus[c]

[a] *Chemical Biology Laboratory, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland, 21702, USA*
E-mail:igorf@helix.nih.gov

[b] *Life Sciences Department, European Patent Office, The Hague, The Netherlands*
E-mail:dmitry.katsubo@gmail.com

[c] *Chemical Biology Laboratory, NCI, NIH, DHHS, NCI-Frederick, Frederick, Maryland, 21702, USA*
E-mail:mn1@helix.nih.gov

October 24, 2011

Abstract

We present Optical Structure Recognition Application (OSRA) as an entry into Image2Structure task of TREC-CHEM. OSRA is an open source utility to convert images of chemical structures to connection tables in an established computerized molecular format. There exists a large body of chemical information which has remained largely inaccessible to machine data mining techniques so far. One of the most common ways of describing a chemical structure in a journal publication or a patent document is by drawing a two-dimensional structure diagram which represents atoms and bonds of the molecule in a human-recognizable form. While easily interpreted by a human expert, such drawings are by themselves unsuitable for use in a computer database for applications such as virtual screening and computer aided drug development. OSRA allows recognition and conversion of such drawings into computer formats widely used by the chemoinformatics community.

1 Introduction

Optical Structure Recognition Application is an open source utility which automatically detects, extracts and converts images of chemical structures, such as those published in journal articles and patents, into machine readable formats such as SMILES (Simplified Molecular Input Line Entry Specification) or SDF (Structure Data File) formats.

The problem of extraction and identification of chemical structure depictions is distinctly different from the better-known tasks of optical character recognition or object detection. A chemical structure image contains much more information than a character in an alphabet — there are only a few dozen characters in most alphabets but millions of known chemical structures. It is easy to correct a misrecognized character when the options are limited to a hundred or so possibilities, but a misrecognized chemical structure might still be a valid molecule. Making the matter even more complicated, there are many ways to draw the same molecule, often in a such a way that only a trained chemist is able to recognize identical chemicals. For this reason, widely used techniques such as wavelet transforms or neural networks, used for example in face recognition, are not applicable here.

2 Overview of OSRA processing workflow

The general work-flow of OSRA is as follows.

2.1 Page segmentation

A page image is first segmented — all pairwise Chebyshev distances between connected components are calculated and a threshold value is computed such that chemical diagrams and corresponding characters forming atomic labels are grouped together, while text and linear vertical or horizontal page separators are removed. This threshold distance is estimated based on an emperical relationship between the size ratio of the connected components vs. their distance from each other.

Separators (connected components with a ratio of height to width above 100 or below 0.01 and size above 300 pixels — typically long lines) are identified and deleted. This allows the removal of linear vertical or horizontal separators early in the process and simplifies further page analysis. Table frames are also removed based on a similar procedure — a table is identified as a connected component which has an aspect ratio between 0.1 and 10, and at least 300 pixels of which are lying on the surrounding rectangle.

For each pair of segments, the area ratio r_{AB} is an integer computed as: $r_{AB} = [\frac{max(S_A,S_B)}{min(S_A,S_B)}]$, where S_A and S_B are the sizes (number of pixels) of the A and B segments, respectively. We define a feature matrix f as a matrix of dimensions $(1..max\_area\_ratio, 1..max\_distance)$ which contains the counts of the pairs of segments with a specific area ratio lying at a specific distance from each other, that is f_{ij} is equal to the number of pairs of segments A, B for which $i = r_{AB}$ and $j = d_{AB}$.

A crucial point which must be addressed in order for the algorithm to be as widely applicable as possible is the ability to distinguish between two different scenarios — whether we have an image of a page with multiple text blocks and/or drawings or a single structure drawing. In the former case, it is possible to apply statistical analysis to determine a threshold value for the distance which separates fragments that "belong together", and which should be treated as parts of the same chemical structure, from the segments that have no logical connection to the current structure depiction and will only complicate further processing. In the latter case such statistical analysis is likely to be impossible due to the low number of elements present in the image and an overzealous algorithm might erroneously discard vital pieces of information. The following step helps in making such a distinction:

For each row of the feature matrix f the number of cells containing zeros ($f_{ij} = 0$) is counted, then the entropy of the row is computed as

$E = -p \log p$, where $p = -\frac{z}{max\_distance-2}$, $z = \sum_{j=2}^{j<max\_distance} n_{ij}$, and $n_{ij} = 1$ if $f_{ij} = 0$ or 0 otherwise.

By empirical observation we have found that the row with the maximum entropy usually lies above 6 for pages with text and/or multiple graphics and is 3 or lower when only a single structure image is present. Therefore a threshold value of 4 was chosen to distinguish between the two types of images.

If it is determined that the page contains text as well as graphics it is advantageous to remove text blocks before processing the chemical structure images. To do so, first the characteristic distance between the text characters is determined by taking the 1st row of the matrix f_{ij} and locating the first local minimum (d) which occurs after the first local maximum (m): $f_{1m-1} < f_{1m}$, $f_{1m+1} < f_{1m}$, and $f_{1d-1} > f_{1d}$, $f_{1d+1} > f_{1d}$, $d > m$.

All segments within distance d from each other are then grouped together. If such a group contains more than the threshold number of connected segments (8 in our case), and the fill ratio (number of pixels divided by the area of the rectangle that a segment occupies) or the aspect ratio (width/height) are above preset thresholds (0.2 and 10, respectively), the group of segments is deemed to be a text block and removed from further processing.

All remaining segments are then grouped according to their pairwise distance — the threshold is chosen to be twice the value of d found in the previous step; or, for a single image page, an arbitrary high number — 100 pixels. Each group of segments — a perspective chemical structure depiction — is subjected to the following filtering criteria: the fill ratio has to be below 0.2, the aspect ratio between 0.1 and 10, both height and width should exceed a characteristic single character height and width, and at least one of the dimensions, either height or width, should exceed double the font height (or width). A characteristic font height and width are set to be 22 and 21 pixels respectively at a resolution of 150 dpi and scaled with the resolution accordingly.

2.2 Binarization, Thining and Anisotropic Smoothing.

A grayscale image is obtained by converting a color vector (R, G, B) into a gray-level vector (Gr, Gr, Gr) where $Gr = min(R, G, B)$. Note that this is different from the more common grayscale conversion methods where gray level intensity is a linear combination of red, green, and blue intensities.

Four different resolutions (or scales) are typically used by default on non-PDF files — the first three are 72, 150, 300 dpi, and the fourth resolution is determined dynamically in the range of 500-1200 dpi. The scale affects the limits on the maximum character size as well as the parameters for thinning and anisotropic smoothing. Trying the processing at different scales allows for a certain degree of independence from the scan resolution that was used when the document in question was scanned (or produced by some other means).

To detect the need for a smoothing procedure at higher resolutions, a quantity we call a noise factor is calculated. The noise factor is defined here as the ratio of the number of linear pixel segments (vertical or horizontal) with a length of 2 pixels to the number of line segments with a length of 3 pixels. If the noise factor is between 0.5 and 1.0, an anisotropic smoothing procedure is performed. Noise removal and anisotropic scaling are achieved using the GREYCstoration anisotropic smoothing library[1].

A thinning function is required to normalize all lines to be 1 pixel wide. Image thinning is done rapidly by the subroutine from the article “Efficient Binary Image Thinning using Neighborhood Maps” by Joseph M. Cychosz[9].

2.3 Vectorization.

The Potrace library[6] is used for vectorization. Atoms are recognized as the control points of Bezier curves where any one of the following conditions is met:

- The control point is classified as a corner by the Potrace algorithm.
- The vector from the control point to the next represents a change of direction with a normal component of at least 2 pixels as compared to the vector from the last atom to this control point.
- The distance from the last atom to the next control point is less than the distance from the last atom to the current control point.

The vectors connecting the found atoms are recognized as bonds. Note the usage of normal component measures instead of angles between pairs of vectors.

2.4 Atomic label recognition.

GOCR[3] and OCRAD[2] are used to test all connected sets of Bezier curves smaller in size than a maximum character height and width, or two characters aligned horizontally or vertically. Recognized characters are then assembled to build atomic labels. Optionally Tesseract[4] and cuneiform[5] libraries can also be used for OCR processing, however in our experiments their use did not result in an increase in recognition.

2.5 Circle bond recognition.

If a circle of sufficiently large diameter is found inside of a ring, the ring is flagged as aromatic. Additional conditions include the ring atoms being sufficiently close to the circle (not more than half of the average bond length away), and angles between the ring bonds and the vectors to the center of the circle being less than 90 degrees. The current implementation fails when the inner circle touches the ring bonds.

2.6 Average bond length and double/triple bond detection.

The average bond length is estimated in the following way: a sorted list of all the bond lengths is created, and the “average” bond length is taken to be the value at the 75th percentile by rank within this list. Choosing 75th percentile instead of the more common 50th (the median) eliminates the bias towards smaller bond lengths which is very common during the initial stages of processing, while also discarding longer than usual bonds which might appear in some structure depictions. The average bond length is re-evaluated several times throughout the processing of the image as more structural elements are being identified. Similar mechanisms are used for measuring distance within the bond pairs comprising double bonds and average bond thickness. The double and triple bonds are then identified as bond pairs (triples) which a) are parallel to each other, b) are within the double bond pair distance of each other, and c) are within each other's "shadow" — that is, the bonds of the bond pair are not separated too far along the line parallel to them.

2.7 Dashed and wedge bonds.

Wedge and dashed bonds represent bonds that are directed "out of" or "into" the page to convey 3D information about a non-flat chemical structure. Dashed bonds are three or more "blobs" of any shape as long as they are 1) small enough, 2) positioned within the average bond length from start to finish, and 3) a straight line can be drawn through their geometric centers. Wedge bonds are recognized by testing for a significant thickness increase or decrease along the bond.

2.8 Bridge bonds.

Bridge bonds are bonds which visibly intersect on the structure diagram but are not actually connected at the point of intersection. To detect a bridge bond versus an actual connection between four bonds we use the following algorithm: if an atom is connected to four pairwise collinear single bonds (none of which is a terminal bond) and this atom node removal does not result in:

- Difference in the number of fragments
- Difference in the number of rotatable bonds
- Decrease in the number of 5- and 6-membered rings by 2

the atom is removed and the intersection is presumed to be a bridge bond intersection.

Table 1: Image2Structure training and challenge set results

Run Options	default	300 dpi
Training set	84.3%	86.1%
Challenge Set	84.8%	85.6%

2.9 Confidence estimate.

To find the best structure resolution among several possible choices we use the following "confidence function":

$confidence = 0.316030 - 0.016315N_C + 0.034336N_N + 0.066810N_O + 0.035674N_F + 0.065504N_S + 0.04N_{Cl} + 0.066811N_{Br} + 0.01N_R - 0.02N_{X_x} - 0.212739N_{rings} + 0.071300N_{aromatic} + 0.329922N_{rings5} + 0.342865N_{rings6} - 0.037796N_{fragments}$

where N_C is the number of carbon atoms, N_N is the number of nitrogen atoms and so on, N_{rings} is the total number of rings, $N_{aromatic}$ is the number of aromatic rings, N_{rings5} is the number of 5-membered rings, N_{rings6} is the number of 6-membered rings, and the number of fragments is $N_{fragments}$.

2.10 Compilation of the connection table.

OSRA uses the OpenBabel[10] chemoinformatics library for conversion into SMILES or SDF. A molecular object is constructed based on the connectivity information along with the stereo- and aromaticity flags. Fragments based on superatoms are added at this stage as well. The superatom dictionary can be modified by a user at run-time without recompilation.

3 Image2Structure Task

For our submission to Image2Structure task we have used the latest released version of OSRA (1.3.8) without any modifications. We have submitted two runs — one with the default settings and the second with the resolution fixed at 300 dpi — thus eliminating the automatic scale selection. The recall rates for the training runs and the final results for the challenge set are presented in Table 1.

4 Conclusion

We have presented Optical Structure Recognition Application as an entry into the Image2Structure challenge. OSRA is a free and open source solu-

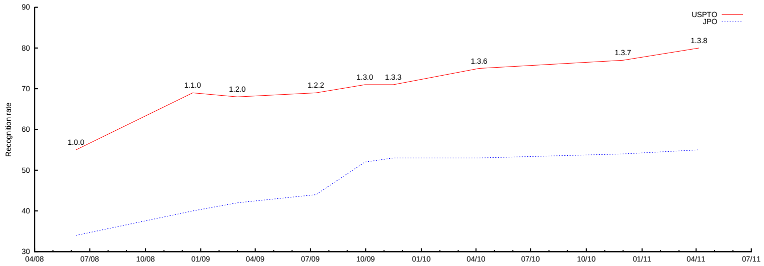

Figure 1: Recognition rate improvements

tion which has been in development for about three years and is now being used by many academic researchers, commercial companies and governement agencies to extract chemical structure information from patents and scholarly articles. The progress in the recognition rate improvement over the years based on patent documents from the USPTO[7] and JPO[8] is presented in Figure 1.

Funding Disclaimer:

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

[1] *GREYCstoration.* http://www.greyc.ensicaen.fr/ dtschump/greycstoration/

[2] *Ocrad — GNU Project — Free Software Foundation (FSF).* http://www.gnu.org/software/ocrad/ocrad.html

[3] *Optical Character Recognition (GOCR).* http://sourceforge.net/projects/jocr/

[4] *tesseract-ocr.* http://code.google.com/p/tesseract-ocr/

[5] *Linux port of Cuneiform.* https://launchpad.net/cuneiform-linux

[6] *Peter Selinger: Potrace.* http://potrace.sourceforge.net/

[7] *SourceForge.net: OSRA.* http://osra.sourceforge.net

[8] *Chem-Infty Dataset: A ground-truthed dataset of Chemical Structure Images.* http://www.iapr-tc11.org/mediawiki/index.php/ChemInfty_Dataset:_A_ground-truthed_dataset_of_Chemical_Structure_Images

[9] J. M. Cychosz. Efficient binary image thinning using neighborhood maps. In *Graphics gems IV*, pages 465–473. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[10] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 2011, 3:33.

[11] 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*, 49(3):740–743, MAR 2009.

[12] I. V. Filippov and M. C. Nicklaus Extracting chemical structure information: Optical structure recognition application. In *Proceedings of the Eight IAPR International Workshop on Graphics Recognition*, pages 133–142, 2009.