Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution

Igor V. Filippov*,† and Marc C. Nicklaus‡

Laboratory of Medicinal Chemistry, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, and Laboratory of Medicinal Chemistry, NCI, NIH, DHHS, NCI-Frederick, Frederick, Maryland 21702

Received February 22, 2008

Until recently most scientific and patent documents dealing with chemistry have described molecular structures either with systematic names or with graphical images of Kekulé structures. The latter method poses inherent problems in the automated processing that is needed when the number of documents ranges in the hundreds of thousands or even millions since graphical representations cannot be directly interpreted by a computer. To recover this structural information, which is otherwise all but lost, we have built an optical structure recognition application based on modern advances in image processing implemented in open source tools, OSRA. OSRA can read documents in over 90 graphical formats including GIF, JPEG, PNG, TIFF, PDF, and PS, automatically recognizes and extracts the graphical information representing chemical structures in such documents, and generates the SMILES or SD representation of the encountered molecular structure images.

INTRODUCTION

Proliferation of computer technologies has brought forward the necessity of new data formats to exchange information in a machine-readable way within the context of a scientific publication. Such new formats suitable for representing chemical structural information have recently appeared, for example InChI, CML, etc.[1,2] However, the bulk of chemical literature that existed before these developments does not employ such well-defined and computer-parsable formats for the representation of molecular information. Some of the most common ways to describe a chemical structure are chemical nomenclature (IUPAC names) and graphical descriptions, images of Kekulé structures printed within scientific or patent documents. The task of later automatic extraction of such structural information has proved to be challenging enough that even though several software packages have been developed, none has achieved universal acceptance.[3−9] Our approach to recovery of chemical information from published material is to reuse to the fullest extent possible the existing software created by the open source community and to invite further development and participation by releasing our work as free and open source. To our knowledge, OSRA is the first open source program for optical structure recognition. OSRA has been designed with a wide range of applicability in mind: it does not rely on the document image being of any particular resolution, color depth, or having any particular font used. To manipulate images, OSRA employs the ImageMagick library[10] that allows parsing of over 90 different image formats, including the popular TIFF, JPEG, GIF, PNG, as well as Postscript and PDF (through the Ghostscript library),[11] formats. OSRA is implemented as a command-line utility, which users are welcome to download from the group's web server.[12] To demonstrate the capabilities (and limitations) of OSRA, we have also provided a web interface.[13] Participation in the further development of this open-source code project is highly encouraged. For this purpose, we have created a SourceForge project with an SVN repository.[14]

ALGORITHM

The basic work flow is similar to that used by the previous implementations:

1. grayscale and binarization
2. segmentation
3. anisotropic smoothing and thinning
4. vectorization and bond/node detection
5. atomic label and charge recognition
6. circle bond (for old style aromatic rings) recognition
7. double and triple bond detection
8. special bond detection: wedge and dash bonds
9. bridge bond detection
10. compilation of the connection table
11. confidence estimate

Grayscale and Binarization. A color image is first converted to grayscale via the following mechanism: for every pixel a color vector (R, G, B) is transformed into a gray-level vector (Gr, Gr, Gr), where $Gr = \min(R, G, B)$. This is different from the more common grayscale conversion methods where $Gr = (R + G + B)/3$ in that it allows for a better later binarization for light-colored portions of the image (such as yellow symbol for sulfur for example). A global threshold is used for binarization. Local (adaptive) thresholding has been tested but so far found unsatisfactory because of the appearances of artifacts in the threshold value-changing regions.

The image by default (unless it is a PDF or Postscript document) is processed at three different scales (resolutions): 72, 150, and 300 dpi. The scale affects the limits on the

* To whom correspondence should be addressed. E-mail: igorf@helix.nih.gov.
† SAIC-Frederick, Inc.
‡ NCI-Frederick.

10.1021/ci800067r CCC: $40.75
Published on Web 02/17/2009

maximum character size and overall molecular image size, as well as the choice for thinning and anisotropic smoothing. In case of a PDF or Postscript document, only a resolution of 150 dpi is used.

Segmentation. The rectangular areas containing the images of chemical structures are selected based on the following criteria:

i. ratio of the black pixels to the total area of the rectangle is between 0.0 and 0.2
ii. aspect (height to width ratio) is between 0.2 and 5.0
iii. the rectangle does not intersect with existing structure-containing rectangles
iv. the width and height are above the minimum values (currently 50 pixels) if the resolution is above 150 dpi
v. the width and height scaled to a resolution of 300 dpi is below a maximum value of 1000 pixels (if the resolution is above 150 dpi)

Anisotropic Smoothing and Thinning. After the selection of rectangular areas in the original image containing the chemical structures, we calculate a "noise factor" for each such area. A noise factor is defined here as a 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 image is too noisy, i.e. 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,[15] which implements a method for removal of small variations in pixel intensities while preserving global image features based on nonlinear multivalued diffusion partial differential equations. The next step is the application of a thinning function 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.[16] Currently anisotropic smoothing and thinning are only performed for images at a resolution of 300 dpi.

Vectorization and Bond/Node Detection. The vectorization (bitmap to vector graphics conversion) is performed using the Potrace library by Peter Selinger.[17] We then attempt to find the positions of atoms and bonds using the vectorized form of the image. We examine each interval between the control points of a Bezier curve (a parametric curve, in this case cubic, commonly used in computer graphics to model smooth curves; the set of control points is the primary output of the Potrace library). The control point is flagged as an atom if any of the following conditions is fulfilled:

i. This control point is classified as a corner by the Potrace algorithm.
ii. The vector from this 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.
iii 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 bonds are then set as the vectors connecting the found atoms. Note the usage of normal component measures instead of the angles between two vectors (as is custom in previous implementations). It is difficult to come up with a general threshold for the angle between two bonds which would remain valid for a wide range of image drawing styles. Moreover angles are difficult to measure for smaller vectors in a pixilated environment, that is, when the size of a dot and thickness of a line are finite and nonzero. Measuring normal components instead allows for much more robust detection of bonds and nodes. Reliability is further improved by using the fact that the Potrace library generates control points for both sides of the same bond, and the skeletization procedure attempts to produce the best recreation of the bond structure by collapsing of the two sides of the same bond together.

Atomic Label and Charge Recognition. All connected sets of Bezier curves smaller in size than a maximum character height/width, or two characters aligned horizontally or vertically, are tested using GOCR[18] and OCRAD[19] (open source OCR tools) for being part of a heteroatom label or an abbreviation. All recognized characters are saved, and the corresponding Bezier curves removed from the list of detected bonds. The maximum height and width of the recognized characters is saved and used at a later stage to determine characters that are connected to the rest of the image, for example if a bond overlaps with an atom label. Small stand-alone bonds are either removed or recognized as lower case letters "l", "i", "r", etc., such as in the atomic label for chlorine "Cl", if they are found next to an upper-case character. Similarly the formal charges, the characters "−" and "+", are identified and assigned to the nearest atomic label.

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 average bond length away), and angles between the ring bonds and the vectors to the center of the circle being less than 90°. The current implementation fails when the inner circle touches the ring bonds.

Average Bond Length and Double/Triple Bond Detection. The average bond length is estimated in the following way: a sorted list of all 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 the 75th percentile instead of the more common 50th (the median) allows the program to avoid the bias toward 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 the 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.

Dashed and Wedge Bonds. Dashed bonds are identified as three or more "blobs" (of any shape as long as they are small enough) positioned within the average bond length from start to finish where a straight line can be drawn through the geometric centers of the "blobs". Wedge bonds are recognized by constructing a linear regression of thickness

versus position within the bond (least-squares estimate) and testing for a significant thickness increase or decrease along the bond.

Bridge Bonds. Bridge bonds are disambiguated based on the following simple rules: 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 (a) a difference in the number of fragments, (b) a difference in the number of rotatable bonds, or (c) a decrease in the number of 5- and 6-member rings by 2, then the atom is removed, and the intersection is presumed to be a bridge bond intersection. This simple rule, while not 100% foolproof, ensures that such disambiguation does not result in a molecule splitting into two or more fragments, the ends of the molecule flying apart, or the node being a connection atom between two rings (a spiro ring system).

Compilation of the Connection Table. OSRA currently is capable of using two different molecular backends: OpenBabel and RDKit. The chosen molecular backend is selected at the time of the compilation. 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 following superatom labels are recognized: MeO, MeS, MeN, CF, CF_3, F_3CN, CN, *n*Bu, EtO, O*i*Bu, *i*Pr, *t*Bu, COOH, Ac, AcO, NO_2, NO, SO_3 H, BzO, $N(OH)CH_3$, THPO. SMILES or SD format output is generated based on the resulting molecular object.

Confidence Estimate. By default, OSRA attempts processing at three different resolutions (scales); therefore it may have up to three different perspective outputs. To automatically decide on the best variant we employ the following "confidence function":

$$\begin{aligned}\text{confidence} = {} & 0.316030 - 0.016315N_{\text{c}} + 0.034336N_{\text{N}} + \\ & 0.066810N_{\text{O}} + 0.035674N_{\text{F}} + 0.065504N_{\text{S}} + \\ & 0.198795N_{\text{Cl}} - 0.212739N_{\text{rings}} + 0.071300N_{\text{aromatic}} + \\ & 0.329922N_{\text{rings5}} + 0.342865N_{\text{rings6}} - 0.037796N_{\text{fragments}}\end{aligned}$$

The function was generated by performing linear regression analysis on the Tanimoto similarity between the real structures and the approximations generated by OSRA at various resolution levels using various simple molecular properties: element counts (N_{c}, number of carbon atoms; N_{N}, number of nitrogen atoms; and so on), ring counts (N_{rings}, total number of rings; N_{aromatic}, number of aromatic rings; N_{rings5}, number of 5-member rings; N_{rings6}, number of 6-member rings), and the number of fragments ($N_{\text{fragments}}$) as regressors. A correlation of 0.89 was achieved with about 40 structures used and over 100 corresponding approximations. While it is meaningless to compare the value of this confidence function for different structures, it proved to be a simple and effective way to choose the most appropriate version between several variants of the same structure. The reason for that is easy to see: a scale with the most heteroatoms recognized is more likely to be the correct one of the three, the same goes for the counts of 5- and 6-member rings, and the reverse is true for the number of fragments and nodes (which are taken to be carbon atoms).

DISCUSSION

We present the first open source optical structure recognition application. The source code can be downloaded from the group's web site, a compiled version for Microsoft Windows is also available. The web-based interface allows for interactive testing and visualization, as well as further lookup of the recognized structures in the Chemical Structure Lookup Service,[20] and conversion to various other molecular structure formats such as MOL, PDB, etc.

Table 1. Recognition Rate Comparison between OSRA and CLiDE, 42 Structures Total

	Perfect by InChI	Average Tanimoto	$T > 85\%$	$T > 90\%$	$T > 95\%$	uuuuu
OSRA	26	95	39	37	33	28
CLiDE	11	87	26	20	17	12

One of the most common questions asked about the optical structure recognition problem in general and OSRA in particular is how good the recognition rates are. To address this question, one has to define a measure of the accuracy of recognition. There have been several such definitions proposed in the past. For example, a structure is considered recognized if it takes less than 30 s for a human expert to correct it[5] or a recognized structure has no more than one error,[8] etc. The former seems subjective by today's standards, but even the latter leaves a lot of room open for interpretation, such as, if a double bond is missing, is this counted as one error or two? And what if a double bond is mis-categorized as a single bond? And this does not even begin to address the question of applicability domain: do we consider only high quality black-and-white images, or are we attempting to parse noisy or color images at a lower resolution as well? We therefore propose a different method for measuring the accuracy of a recognition engine, one that we think is more objective and usable in practical applications. One of the natural applications of a chemical structure recognition engine would be, for example, to attempt to look up the recognized structures from a document in a database of available chemicals. Therefore a natural accuracy measure would be a similarity index between the output of the program and the actual structure. On our web site, we output a Tanimoto similarity index based on CACTVS[21] fingerprints between the structure as corrected by the user (presumably yielding the correct structure) and the structure that OSRA had produced. While no well-known fingerprint proved to be an ideal choice for this project because all of them tend to punish easily correctable errors severely and assign structures with large discrepancies good mutual similarity indices, it is a quantitative measure that is easily implemented and understood. A better suited fingerprint would be one which produces results that are more intuitive, that is, closer to the accuracy measurements proposed in the past but more automated and rigorous, which however goes beyond the scope of this study.

To compare the recognition rates with the only other commercially available optical structure recognition program today, CLiDE, we used the so-called "small test set" kindly provided by Simbiosys Inc. (Table 1). We selected 11 files from this set. Of the remaining three, one file does not contain a structure image, one is poorly segmented by both OSRA and CLiDE, and the last file contains larger molecules, which neither OSRA nor the authors are prepared to handle at the moment. Of the resulting output set one structure was deemed to be a false positive and was removed from both OSRA and CLiDE output.

Table 2. Recognition Rate on the Internal Test Set, 215 Structures

	Perfect by InChI	Average Tanimoto	$T > 85\%$	$T > 90\%$	$T > 95\%$	uuuuu
OSRA	107	93	182	167	147	122

Here, "perfect" means the number of structures identical to the human-curated version according to InChI; the average Tanimoto similarity is between the human-curated set and the computer-processed set; "$T > 85\%$" indicates the number of structure pairs with Tanimoto similarity above 85% etc. The last column is the number of structures that have an identical "uuuuu" identifier, an identifier developed at the NCI CADD group, which is indifferent to the stereochemistry, tautomerism, charge, and isotope information and takes into account only the largest fragment.[20] The differences with the previously reported results[22] are because the most recent version of OSRA now supports SD format output and the comparison can now be made made using SD files for both CLiDE and OSRA. We used OSRA version 1.1.0 with OpenBabel backend, SD file format output.

To verify the results on a larger and more diverse set we performed the same analysis on our internal test set (Table 2). This set is composed of 66 images of various resolutions and color depths (black-and-white, gray scale, and color) and contains a wide variety of drawing styles. With this set, OSRA had seven false positives and missed three structures because of the imperfect segmentation. The total number of valid recognized structures in the 66 images and documents is 215.

The results appear to be consistent and very competitive. While the fraction of perfectly recognized structures is not yet very high, one has to remember that a molecular structure image contains much more information than for example a single character in a text subjected to an OCR procedure and the space of known chemicals (tens of millions) is much greater than the space of characters in any alphabet, so the direct comparison to regular OCR is not valid in this case. Still, the fact that a large portion of the structures has been recognized at a Tanimoto similarity level of 85% or above gives hope that the automatic recognition might be useful for example for locating a structure within a large database of known chemicals. The main sources of errors come from the imperfect segmentation, OCR mistakes, and noise in the scanned image. By releasing OSRA as an open source program, we hope to attract interested parties to participate in the further development of what we hope will be a useful addition to the set of publicly available chemoinformatics tools.

ACKNOWLEDGMENT

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 AND NOTES

(1) Heller, S.; Stein, S.; Tchekhovskoi, D. InChI: Open access/open source and the IUPAC international chemical identifier. *Abstr. Pap. Am. Chem. Soc.* **2005**, *230*, 1025–1026.
(2) Murray-Rust, P.; Rzepa, H. Chemical markup, XML, and the worldwide web. 1. Basic principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928–942.
(3) Contreras, M.; Allendes, C.; Alvarez, L.; Rozas, R. Computational perception and recognition of digitized molecular structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 302–307.
(4) Casey, R.; Boyer, S.; Healey, P.; Miller, A.; Oudot, B.; Zilles, K. Optical Recognition of Chemical Graphics. *Proceedings of the 2nd International Conference on Document Analysis and Recognition*; Tsukiba Science City, Japan, October 1993; IEEE: Piscataway, NJ, 1993; 627−632.
(5) McDaniel, J.; Balmuth, J. Kekule—OCR optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373–378.
(6) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A.; Simpson, R.; Tonnelier, C.; Venczel, T.; Johnson, A. Chemical Literature Data Extraction—The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 338–344.
(7) Zimmermann, M.; Thi, L.; Hofmann, M. Combating illiteracy in chemistry: Towards computer-based chemical structure reconstruction. *ERCIM News* **2005**, *60*, 40–41.
(8) Zimmermann, M. Large scale evaluation of chemical structure recognition. *Proceedings of the 4th Text Mining Symposium in Life Sciences*; October 9−10, 2006; Fraunhofer SCAI: Germany, 2006.
(9) Michigan Alliance for Cheminformatic Exploration. http://www-personal.umich.edu/grosania/MACE071806OSANIA.ppt. (accessed September 26, 2008).
(10) ImageMagick: Convert, Edit, and Compose Images. http://www.imagemagick.org/script/index.php. (accessed September 26, 2008).
(11) Ghostscript, Ghostview and GSView. http://pages.cs.wisc.edu/ghost/. (accessed September 26, 2008).
(12) OSRA: Optical Structure Recognition. http://cactus.nci.nih.gov/osra. (accessed September 26, 2008).
(13) OSRA: Optical Structure Recognition. http://cactus.nci.nih.gov/cgi-bin/osra/index.cgi. (accessed September 26, 2008).
(14) SourceForge.net: OSRA. http://sourceforge.net/projects/osra/. (accessed September 26, 2008).
(15) GREYCstoration. http://www.greyc.ensicaen.fr/dtschump/greycstoration/. (accessed September 26, 2008).
(16) Cychosz, J. M. Efficient binary image thinning using neighborhood maps. In *Graphics gems IV*; Academic Press Professional, Inc.: San Diego, CA, 1994; pp 465−473.
(17) Peter Selinger: Potrace. http://potrace.sourceforge.net/. (accessed September 26, 2008).
(18) Optical Character Recognition (GOCR). http://sourceforge.net/projects/jocr/. (accessed September 26, 2008).
(19) Ocrad—GNU Project—Free Software Foundation (FSF). http://www.gnu.org/software/ocrad/ocrad.html. (accessed September 26, 2008).
(20) Sitzmann, M.; Filippov, I. V.; Nicklaus, M. C. Internet resources integrating many small-molecule databases. *SAR QSAR Environ. Res.* **2008**, *19*, 1–9.
(21) Xemistry Chemoinformatics. http://www.xemistry.com/. (accessed September 26, 2008).
(22) Filippov, I. V.; Nicklaus, M. C. Optical structure recognition application. *Proceedings of the 236th ACS National Meeting*; Philadelphia, PA, August 17−21, 2008; American Chemical Society: Washington, DC, 2008.

CI800067R