Extracting Chemical Information from Molecular Depictions

Amena Mahmoud
Computer Science Department
Faculty of Computers and Information,
Mansoura Univ.
Mansoura - Egypt
Email: Amena_mahmoud {at} mans.edu.eg

Magdy Zakaria
Computer Science Department
Faculty of Computers and
Information, Mansoura Univ.
Mansoura - Egypt

Taher Hamza
Computer Science Department
Faculty of Computers and
Information, Mansoura Univ.
Mansoura - Egypt

Abstract— **The Chemical Information contained in researches or on internet is often drawn as chemical formulas embedded in depictions. To digitally convert these chemical structure formulas into their equivalent computer representation, several software systems have been developed. This papers aims to provide critical reviews for these systems and discuss our current approach. The proposed approach is considered as an automated tool for converting chemical structure diagrams into standard chemical format. Basic algorithms for recognizing lines and letters representing bonds and atoms are used for extracting information from graphs.**

Keywords; chemical molecules; bonds; atoms; molecule recognition; regenration.

I. INTRODUCTION

Different document formats further complicate the problems of extracting chemical information from the literature. Although documents exist in two basic format types of 'text' or 'image', each type has many variations (e.g. text, rich text format, word document, hypertext markup language, portable document format, or graphical interface file and tagged image file formats). Journal articles are generally found in a text format embedded with images corresponding to figures and/or tables. Chemical entity recognition will require the ability to extract information form the text and images in all multiple variations.

Two examples of information resources linking chemical structures with biomedical targets, pathways and phenotypes are PubMed [1] – the database of the scientific literature corpus – and PubChem [2] – a publicly available database of over 19 million chemical structures, each of which can have a cross-reference link to similar structures and bio-activity descriptions.

In general, one can envision two ways to parse scientific articles for chemical information: by searching for names or structure diagrams of chemical agents. The chemical structure diagrams in scientific articles are typically drawn manually using a program such as ChemDraw [3], ISIS/ Draw [4], DrawIt [5], and ACD/ChemSketch [6]. Once a structure is drawn, the structural description can be translated into a computer readable format, such as ISIS, MOLfile or SMILES [7] formats, which describes the atoms, bond orders, and connectivity patterns of atoms in molecules.

However, the diagrams of chemical molecules in scientific journals and reference books are encoded as digitized images (e.g. JPG, PNG or GIF), which are embedded within lines of text in a form that is not readily translatable into a computer readable format. Therefore, most references to chemical agents in scientific research articles cannot be easily linked to other repositories of scientific knowledge, and are thus not amenable for analysis or searching using cheminformatic software [14].

An effective image searching capability would require converting the digital depictions of chemical diagrams into structured representations such as SMILE strings or atom connectivity tables in standard chemical file formats. Novel drug candidates or newly synthesized molecules are usually referenced by chemical structure diagrams rather than molecule names. In addition, a single molecule may have a number of synonyms such that it could be referenced by different names in different articles. Thus, the capability of exploring research articles or patents where the chemical structure or similar compounds are drawn would complement existing text-based search engines for chemical information.

The current research presents a novel application of machine vision methods for the identification of chemical composition diagrams from two-dimensional digital depictions. The method is based on the use of Optical Character Recognition algorithms for text extraction and Hough transform algorithm for lines extraction from digital depictions. We compare this method with previous approaches that transform such images to a computer readable format. The present approach is attempting to recognize features such as the

presence of a chemical representation in the original depictions. This information can be used for providing chemical information of the original image as part of an image classification process.

II. CRITICAL REVIEW

Although pattern recognition in images is not a new field, there exist only a few approaches dealing with the chemical structure recognition problem. In the 1990s, several software programs were developed that could extract chemical structure diagrams in scientific articles and convert them to structured representations. Recently, with the active development of cheminformatic tools for processing published chemical information, more software programs were launched and continue to be updated.

The first commercial program to read and interpret digital depictions of chemical structures was *Kekule* [8], which has a built-in algorithm to fix character recognition errors using neural network for generating potential characters with scoring information estimating the likelihood that a specific character corresponds to a certain atom. Another program called *OROCS* [9], was developed which has an algorithm for isolating chemical structure diagrams from other elements, the document is segmented by a conventional connected components algorithm. If the size of a segment is larger than a threshold, it is potentially regarded as a chemical structure, and the polygonal shapes of chemical structure diagrams are used to make a final decision. Chemical Literature Data Extraction Project (*CLiDE*) is available commercially and not only aims at extracting chemical structures but also abstracting chemical information from text [10]. By employing the Documental Format Description Language (DFDL) which can describe logical relationships of objects and elements in a document, *CLiDE* builds logical associations between chemical structures and the text segments of document. Recently, a new program, called *chemOCR* [11], has been developed and made available. It adopted a chemical rule-based expert system for the extraction of chemical structure diagrams. The most interesting features, at the post-processing stage, is that *chemOCR* uses a graph-matching algorithm to select the best-matching chemical structure fragment against sub-graphs of chemical structures stored in a database. *OSRA* [12], another recently released program is free and open source software attempts to generate three output structures by varying parameters for the de-noising stage, and then picks one as an output based on its own empirical confidence function.

As well as *Kekule*, *OROCS* and *CLiDE* require at least a 300 dpi resolution in scanned images at the scanning step and manual correction at the post processing step to achieve reliable output. However, the drawn chemical structure diagrams are typically embedded in Word documents as GIF or JPG formats, whose the resolution is usually 72–96 dpi. Therefore, these software systems might be impractical tools for fully automated extraction of chemical structure information.

III. MOLECULE RECOGNITION PROCEDURE

The present research introduces a novel molecule recognition which takes a single molecule depiction either from a scanned image or a digitally generate done, and returns a graph structure representing that molecule. The scanned image may include some information as text or many chemical formulas. The proposed approach does not have an automated process to discriminate chemical formula diagram from graphical objects, so manual separation of chemical diagrams using other painting programs is needed to specify the working area that includes only the wanted chemical formula. The recognition is carried out by separating the diagram into parts and recognizes each one using a specific algorithm. The following is a schematic overview of single steps in the proposed procedure.

Algorithm 1 (Recognition Procedure)

INPUT:
An image of a chemical molecule
OUTPUT:
A graph G = (V;E) representing the molecule structure

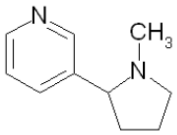
METHOD:
1. Image de-noising
2. Image segmentation
3. Character Extraction and Recognition
4. Line Endpoint Extraction
5. Creation of Graph Representation

A. De-noising

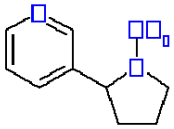
The first step in the recognition process involves an image processing de-noising. For this purpose, GREYCstoration [15], a free implementation of image regulation algorithm is used. GREYCstoration is an image regularization algorithm which is able to process a color image by locally removing small variations of pixel intensities while preserving significant global image features, such as edges and corners. The most direct application of image regularization is image de-noising.

B. Segmentation

In the proposed approach, the image was traversed from left to right, top to bottom searching for unlabelled black pixel. Once such a pixel has been found we label the entire connected component containing that pixel using a grass-fire algorithm [13] that recursively searches outwards for neighboring unlabelled black pixels, *fig.1*. Once the entire component has been labeled the systematic traversal continues.



(a) Original image



(b) After segmentation

Figure. 1. Chemical Structure Image Segmentation

C. Optical Character Recognition

Molecule images contain different kinds of string oriented chemical symbols, like atoms, SMILES strings and super atoms. For this reason optical character recognition (OCR) algorithms are required to correctly identify all contained character symbols. OCR is one of the most successful applications of automatic pattern recognition. It translates images of hand/type written text, usually captured by a scanner, into machine-editable text. These patterns are then recognized using template based recognition approach.

Template based recognition

This recognition strategy is structural based and involves template matching. The entire image is used as a feature; each symbol identified from the segmentation step is compared with all character templates of an alphabet. A suited distance function is used to compute a similarity measure between a character symbol and all templates. The template exhibits the highest similarity measure to the symbol, is identified. If this similarity is above a certain threshold, the character is assigned the character matched label. To encounter the invariant problem, different sizes of template characters may be used or the character symbols in the input image can be scaled to suit the template sizes, fig. 2

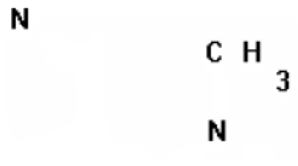


Fig. 2. Text Extraction

D. 2D Bond Extraction

Bonds, beside atoms, are the most frequently used symbols in chemical images. Among other things, they hold the information which atoms have to be connected. It is important to infer a relative accurate determination where the bond is situated in the picture. In addition to the connecting information, bond sets can symbolize themselves a collection of atoms. Although an aromatic ring system contains several carbon atoms, it is represented through a set of connected bonds. For the reconstruction of molecule depictions, it is therefore absolutely required to process these line drawings with an appropriate procedure, fig.3 shows types of bonds.

Fig. 3. Types of bonds

1) Line Detection Algorithm

Most bonds in a chemical structure drawing are simple straight lines. Therefore, a robust line detection algorithm is the key software component for extracting bond features from a chemical structure diagram. In digital image processing, the Hough Transform (HT) is a standard technique used for this purpose. It detects lines by mapping the image in the Cartesian space to the polar Hough space using the normal representation of a line in x-y space:

$$xi \cos \theta i + yi \sin \theta i = ri$$

The algorithm for the Hough transform can be expressed as follows:

1) Define the Hough Transform ρmin, ρmax, θmin, and θmax.
2) Quantify the ρ-θ plane into cells by formatting an accumulator cells array A(ρ,θ) where ρ is between ρmin and ρmax and θ is between θmin to θmax.
3) Initialize each element of an accumulator cell array A to zero.
4) For each black pixel in a binary image, perform the following:
    a) For each value of θ from θmin to θmax , calculate the corresponding ρ using equation:
    $\rho = x \cos \theta + y \sin \theta$
    b) Round off the ρ value to the nearest interval value.
    c) Increment the accumulator array element A (ρ, θ).
    d) Detect best line candidates as local maxima in an accumulator cell array.

Since a pixel corresponds to a sinusoidal curve in the Hough space, collinear pixels in the x-y space have intersecting sinusoidal lines. Therefore, all possible lines passing through every arbitrary pair of pixels in a chemical diagram image are

identified by checking the intersection points of curves in the Hough space, fig.4.

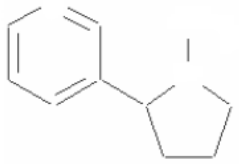


Fig. 4. Lines detection

2) Ring Structure Identification

Another interesting bond recognition problem occurs in aromatic systems, where a circle is often used to represent the conjugated electron system of the benzene ring. To identify these circles, an algorithm looks for the pixels of a connected component that are distributed with almost the same distance from the center of the component. With this algorithm, the presence of circular features can be detected by checking whether the standard deviation of distances from the center of an object is smaller than a certain threshold.

In low resolution images, it is often observed that a detected line have a different position, length or direction from the actual bond. This is especially the case for the bonds in a hexagonal or pentagonal ring structure because the pixels of the neighbor bonds can act as noise in the Hough Transform (HT). Accumulated errors of line detection around a ring structure would cause significant errors in constructing the topology of the chemical structure, fig. 4. This problem could be solved by detecting Pentagonal or Hexagonal ring structures directly using the Generalized Hough Transformation (GHT).

E. Topology Construction and Data Output

For data output, a graph representing the chemical structure is compiled based on the detected bonds and the recognized atomic or chemical symbols. 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. Based on this graph data structure, a node-edge connectivity-table is generated, which finally can be converted into a standard chemical file format (SMILE string).

1) Molecule Formula Generation

Extracting the molecule formula amounts to an enumeration of the single atoms occurring in the graph structure. In addition we might need to add hydrogen atoms for single unsaturated Carbon atoms in the molecule. The algorithm proceeds as follows:

Algorithm 2 (Molecule Formula Generation)
INPUT:
Molecule graph G = (V;E)
OUTPUT:
The molecule formula
METHOD:
1. Initialize empty counter for the molecule formula
2. For each $v \in$ V do
(a) If *v* has label C then, add c to the formula
(b) elseIf, *v* has 6 labels C, add H to the formula
(b) else add the label of *v* to the molecule formula.

For example, in the molecule of Fig. 1, add the vertices N and NCH3 directly to the formula. For the 6vertices we add one additional Hydrogen atom H for each of the four vertices that have valence 3 and none for the two that have valence 4. This results in the final molecule formula of C10H7N2.

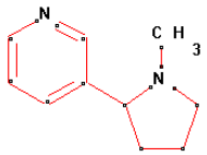


Fig. 5. Final Image Reconstruction

IV. CONCLUSION

Access to relevant information and knowledge is essential for all steps of the drug discovery process. Computer-aided information extraction (IE) systems have been developed to support the work of scientists by extracting relevant information from scientific publications and presenting it in an aggregated, condensed form. In this review, our proposed approach will be used on current information extraction strategies in the life sciences with a special focus on biological entity recognition and more recent developments towards the identification and extraction of chemical compound structures.

The proposed approach will be as the following:

For a given Image ***I*** :

- Defining the connected components (any pixels are adjacent if they have a common edge or a common corner).
- Decompose the graph into parts depending on its connectivity.
- Making a database of characters.
- Comparing the extracted component of figure with the stored characters of the database according to template based recognition algorithm.
- Storing each extracted character in a matrix with its (x,y) position.

- Decide the priority of structure reconstruction according to pharmaceutical rules.
- Defining vectors according to vecorization algorithms.
- For vectors connected edges which have no assigned atom, a carbon atom is added.
- For a ring structure , 6 carbon atoms are assigned.
- Finally, converting the matrix to a Chemical Format (SMILE).

Many improvements, corrections and professional quality output still await implementation. Using the de-noising algorithm, the proposed approach could process about 65-70 images out of 100 successfully.

REFERENCES

[1] PubMed [http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html#Introduction]
[2]PubChem [http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_Overview]
[3] ChemDraw [http://www.cambridgesoft.com/software/ChemDraw/]
[4] ISIS/Draw [http://www.symyx.com/products/software/decisionsupport/isis-draw/index.jsp]
[5] DrawIt [http://www.chemwindow.com]
[6] ACD/ChemSketch [http://www.acdlabs.com/products/chem_dsn_lab/chemsketch/]

[7] SMILES; [www.daylight.com]

[8] McDaniel JR, Balmuth JR: Kekule: OCR – Optical Chemical(Structure) Recognition. J Chem Inf Comput Sci 1992, 32:373-378.
[9] Casey R, Boyer S, Healey P, Miller A, Oudot B, Zilles K: OpticalRecognition of Chemical Graphics. In Proceedings of the SecondInternational Conference on Document Analysis and Recognition: Japan; 1993:627-632.
[10] Ibison P, Jacquot M, Kam F, Neville AG, Simpson RW, Tonnelier C, Venczel T, Johnson AP: Chemical Literature Data Extraction:The CLiDE Project. J Chem Inf Comput Sci 1993, 33:338-334.
[11] Algorri ME, Zimmermann M, Friedrich CM, Akle S, Hofmann-Apititus M: Reconstruction of Chemical Molecules from Images. In Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS): 23–26 , 2007 .
[12] OSRA: Optical Structure Recognition [http://cactus.nci.nih.gov/osra/]
[13] I. Pitas. Digital Image Processing Algorithms. Prentice Hall, 1993
[14] Gkoutos GV, Rzepa H, Clark RM, Adjei O, Johal H. Chemical Machine Vision: Automated Extraction of Chemical Metadata from Raster Image. J Chem Inf Comput Sci. 2003;43:1342–1355.
[15] GREYCstoration: open source algorithms for image denoising and interpolation [http://cimg.sourceforge.net/greycstoration/]