Polymer Discovery

Thomas Goddard, Konstantin Litovskiy, Nathan Nichols-Roy, Matthew Reed, Igor Shvartsner, Nicholas Smith, David Zeppa

Abstract

The recognition of polymer images in the literature and patents is key for automatic understanding of the wealth of polymer data already known. While tools such as OSRA (Optical Structure Recognition Application) [1] exist to identify and interpret chemical structure images, these tools do not yet work for polymers. Our tool, P-OSRA (OSRA-P) extends OSRA’s capabilities by being able to recognize brackets and parentheses in chemical diagrams. To date P-OSRA pre-scans the image; records and alters the image by removing brackets and parentheses from the diagram; calls OSRA, and then collects and edits the resulting SMILES string from OpenBabel3 to reflect changes needed for describing a polymer image. P-OSRA then populates a data-model to allow for subsequent querying of polymer structures (such as repeat units).

Introduction

Accelerated materials discovery is at the core of innovation, economic opportunities, and global competitiveness. The research process is responsible for bringing new materials to market. In 2011, President Obama launched the U.S. Material Genome Initiative (MGI) and challenged researchers, policy makers, and business leaders to reduce the time and resources needed to bring new materials to market—a process that today can take 20 years or more. There is great potential in leveraging modern data mining, big data analytics techniques, and physics based modeling (high performance computing) to significantly shorten the Research & Development cycle in material sciences. Polymers, as an important part of material sciences, are the focus of much research for applications in the areas of semiconductors, nanomaterials, polymeric drug delivery vehicles, desalination membranes, and recyclable polymers for green chemistry. In order to accelerate polymer discovery, the first challenge is to automate the retrieval of all polymers made to date in the literature and in patents. Although there are standardized nomenclatures for small molecules, standardized nomenclature for polymers is not straightforward forward and it is often more informative to represent these molecules as molecular structure diagrams or acronyms (e.g. PEG). Thus, this project’s goal is to identify polymer images in published research, analyze them and output the structure to a data-model, that allows the substructures to be queried.

Background

Optical Structure Recognition Application (OSRA)

P-OSRA extends OSRA, a tool for parsing images of chemical structures in documents, to support polymer diagrams that use bracket and parenthesis notation in their representations. P-OSRA also offers a data interface to allow storage and querying of repeat units and end groups (sub-components of the polymers). This is important because “Polification” 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. Our approach to recovery of chemical information from published material is to reuse 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 an open source application. 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 ImageMagic library [4] that allows parsing of over 80 different image formats, including the popular TIFF, JPEG, GIF, PNG, as well as PostScript and PDF... [5, page 1].

Acknowledgement

Many thanks and appreciation for IBM Researchers Dr. Julia Rice, Dr. Hans Horn, and Dr. Amanda Engler; creator of OSRA, Dr. Igor Filippov; IBM Research Scott Spangler; and Professor Linda Werner, UCSC.

Results

High Level Process Diagram of P-OSRA

At first, we took the approach of using point density data from a vector conversion of a structure image to pinpoint segments of the diagram that could be brackets. This proved mostly inaccurate and difficult to implement. Next, we tried an algorithm to detect endpoints in the diagram and found this to be promising. This algorithm works by checking each pixel’s surrounding adjacent pixels to see if it is an endpoint. We then look for pairs of endpoints that properly align, check for reasonable spacing and bond interaction, and tentatively store the coordinates of these endpoints as the location of a single bracket. From these coordinates, we draw a bounding box around the bracket and color all pixels inside the box white, removing the bracket, and then redraw the bond without the bracket. Figure 2 is the resultant image after tracing, in green, endpoints that are believed to be brackets. Figure 3 depicts the polymer polystyrene and the process in which we alter the image. After detecting the endpoints of the image and finding the brackets, we are able to box out the brackets entirely and replace them with polymer atom (we chose the bi-valent Po atom as it is highly unlikely to appear in any polymer structure).

Creating a Data Model

We have designed a data model in Object Relational Model format (ORML) [6] so that we can store and query the images. With a queryable database, researchers can search for particular substructures or “similar” repeat units that they may desire to include in a new polymer. Retrieval of these units will allow them to rapidly assemble polymer structures. Further, P-OSRA lends itself to parallelization allowing batch processing to be much more efficient. With our additions, we hope to preserve the same level of performance and plan to better incorporate OpenMP in our image pre-processing stage to maximize efficiency.

Conclusions

We have created a  extension to OSRA, making it more comprehensive by the ability to parse polymer chemical diagrams. Our plans include the addition of more comprehensive and detailed documentation and development tools to OSRA to make it easier for other researchers to extend OSRA’s capabilities. POSRA is written in C++ and is currently supported on Windows and Linux systems, both 32 bit and 64 bit. We plan to release POSRA as open source software under the public domain.

References