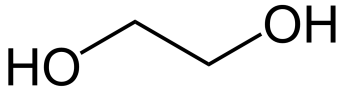
**Accelerated Materials Discovery Julia Rice (jrice@us.ibm.com)**

Advanced materials are at the heart of innovation, economic opportunities, and global competitiveness. They are the foundation in addressing challenges in clean energy, human welfare and national security. 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. We see 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 science, are the focus of many research fields such as semiconductors (e.g. low-k dielectrics, photolithography, and directed self-assembly), nanomaterials, polymeric drug delivery vehicles, desalination membranes, recyclable polymers for green chemistry etc.

At the IBM Almaden Research Center, we have considerable expertise in the experimental development and computational modeling of polymers. However, whenever we start a new project it is not easy to collate all the data in the literature. Ultimately, we would like to initiate a polymer design project by asking questions such as “What polymers make good drug delivery vehicles?”, “What are their relevant properties?”, and “What are their known problems?” However, there is no easy way to extract polymer information from patents and journal articles. One important component of this is extracting polymer structures from documents.

So, the project we propose would be to expand the open source tool OSRA [1,2] currently available for extracting chemical structures from text, to extracting polymer structures from text. The simplest polymers are chain polymers with a single “repeat” unit, specified by [..]n such as in the image of polyethylene glycol below. Today, OSRA can process the ethylene glycol molecule but not the polymer.





Ethylene glycol

Polyethylene glycol

In the case of the polymer, the processing would result in three chemical (or SD) files (as opposed to one SD file), the first file representing the chemical within the brackets (1), the second file representing the chemical entity outside the left bracket (2), and the third file representing the chemical entity outside the right bracket (3). Connectivity information of chemical entities (2) and (1), and (1) and (3) may also need to be generated. .

Polymers can have more than one repeat unit, for example,



Polymer structures can get progressively more involved, and there are variations in the way they are represented …



So, the goal would be to start with simple images and progress as far as we could in the time allotted. We would provide a selection of images and desired output files. OSRA is available under a GPLv2 license, and it is anticipated that any enhancements would be made similarly available. Estimate: 2-3 students; experience in C++; high school chemistry preferred.

[1*] J. C*hem. Inf. Model., **2009**, *49* (3), pp 740–743

[2] <http://sourceforge.net/apps/mediawiki/osra/>