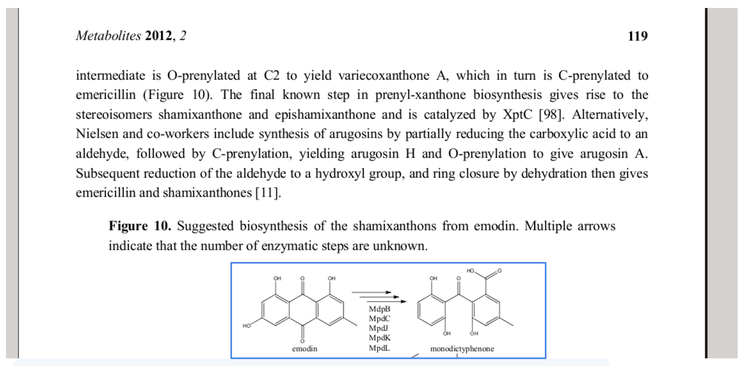
UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics

AMI2: High-throughput extraction of semantic chemistry from the scientific literature – the ChemistryVisitor

Andy Howlett, Mark Williamson, Peter Murray-Rust, Robert Glen
Unilever Centre, Cambridge

The AMI2 project aims to extract facts from the scientific literature, in particular from diagrams, which have been used less than text in the past. As part of this, the ChemistryVisitor is responsible for processing diagrams containing molecules and reactions / reaction schemes.



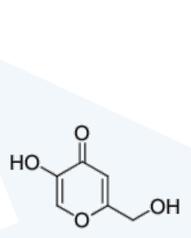
Metabolites **2012**, *2* **119**

intermediate is O-prenylated at C2 to yield variecoxanthone A, which in turn is C-prenylated to emericillin (Figure 10). The final known step in prenyl-xanthone biosynthesis gives rise to the stereoisomers shamixanthone and epishamixanthone and is catalyzed by XptC [98]. Alternatively, Nielsen and co-workers include synthesis of arugosins by partially reducing the carboxylic acid to an aldehyde, followed by C-prenylation, yielding arugosin H and O-prenylation to give arugosin A. Subsequent reduction of the aldehyde to a hydroxyl group, and ring closure by dehydration then gives emericillin and shamixanthones [11].

Figure 10. Suggested biosynthesis of the shamixanthons from emodin. Multiple arrows indicate that the number of enzymatic steps are unknown.

Example diagram from a paper

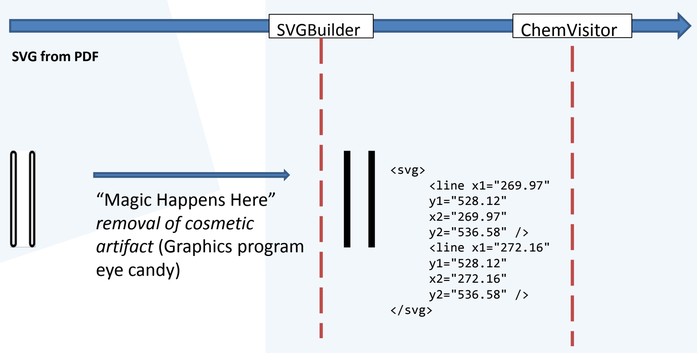
CML output for one of the molecules in the diagram at the bottom of the poster



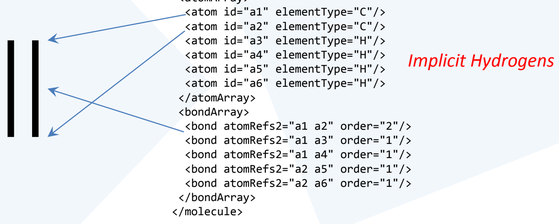
Kojic acid

```
- <atom id="a3" elementType="C" hydrogenCount="0">
    <label value="2" dictRef="cmlDict:locant"/>
    <label value="ortho" dictRef="cmlDict:locant"/>
  </atom>
- <atom id="a6" elementType="C">
    <label value="3" dictRef="cmlDict:locant"/>
    <label value="meta" dictRef="cmlDict:locant"/>
  </atom>
- <atom id="a7" elementType="C" hydrogenCount="0">
    <label value="4" dictRef="cmlDict:locant"/>
    <label value="para" dictRef="cmlDict:locant"/>
  </atom>
- <atom id="a8" elementType="C" hydrogenCount="0">
    <label value="5" dictRef="cmlDict:locant"/>
  </atom>
- <atom id="a9" elementType="C">
    <label value="6" dictRef="cmlDict:locant"/>
  </atom>
- <atom id="a11" elementType="O" hydrogenCount="0">
    <label value="O" dictRef="cmlDict:locant"/>
  </atom>
  <atom id="a16" elementType="H"/>
  <atom id="a17" elementType="H"/>
</atomArray>
<bondArray>
  <bond id="a1_a12" atomRefs2="a1 a12" order="S"/>
  <bond id="a1_a8" atomRefs2="a1 a8" order="S"/>
  <bond id="a2_a13" atomRefs2="a2 a13" order="S"/>
  <bond id="a2_a3" atomRefs2="a2 a3" order="S"/>
  <bond id="a3_a14" atomRefs2="a3 a14" order="S"/>
  <bond id="a3_a15" atomRefs2="a3 a15" order="S"/>
  <bond id="a3_a5" atomRefs2="a3 a5" order="S"/>
  <bond id="a4_a5" atomRefs2="a4 a5" order="S"/>
  <bond id="a5_a6" atomRefs2="a5 a6" order="D"/>
  <bond id="a6_a7" atomRefs2="a6 a7" order="S"/>
  <bond id="a7_a8" atomRefs2="a7 a8" order="S"/>
  <bond id="a8_a9" atomRefs2="a8 a9" order="D"/>
  <bond id="a4_a9" atomRefs2="a4 a9" order="S"/>
  <bond id="a7_a11" atomRefs2="a7 a11" order="D"/>
  <bond id="a6_a16" atomRefs2="a6 a16" order="S"/>
  <bond id="a9_a17" atomRefs2="a9 a17" order="S"/>
</bondArray>
</molecule>
</cml>
```

Converting graphics primitives



SVG to CML



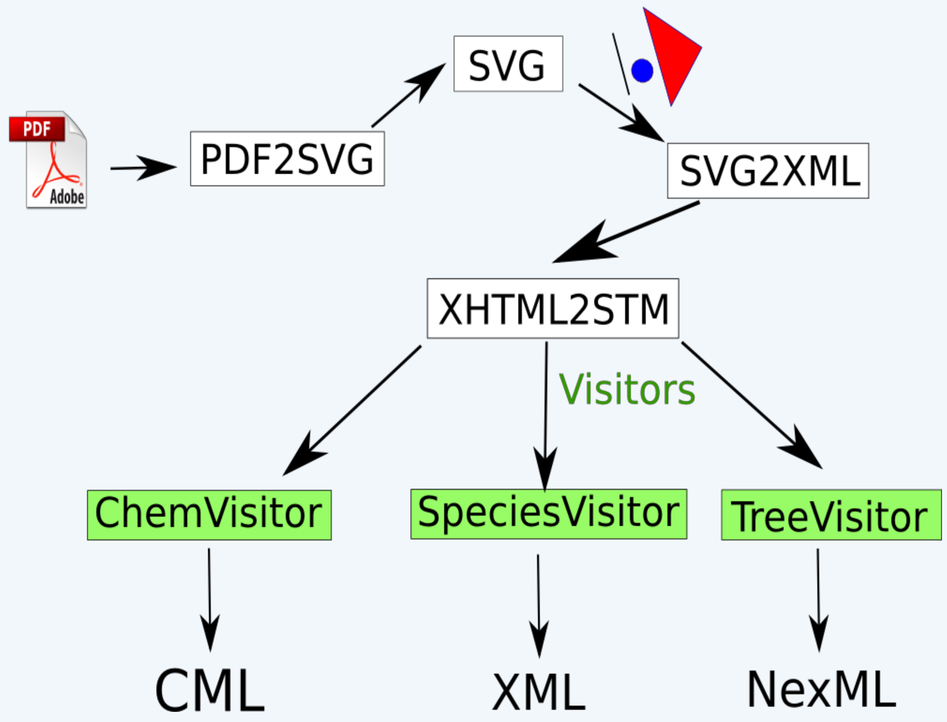
```
<?xml version="1.0"?>
<molecule xmlns="http://www.xml-cml.org/schema">
 <atomArray>
  <atom id="a1" elementType="C"/>
  <atom id="a2" elementType="C"/>
  <atom id="a3" elementType="H"/>
  <atom id="a4" elementType="H"/>
  <atom id="a5" elementType="H"/>
  <atom id="a6" elementType="H"/>
 </atomArray>
 <bondArray>
  <bond atomRefs2="a1 a2" order="2"/>
  <bond atomRefs2="a1 a3" order="1"/>
  <bond atomRefs2="a1 a4" order="1"/>
  <bond atomRefs2="a2 a5" order="1"/>
  <bond atomRefs2="a2 a6" order="1"/>
 </bondArray>
</molecule>
```

Implicit Hydrogens

Initial processing (currently the same for all visitors)

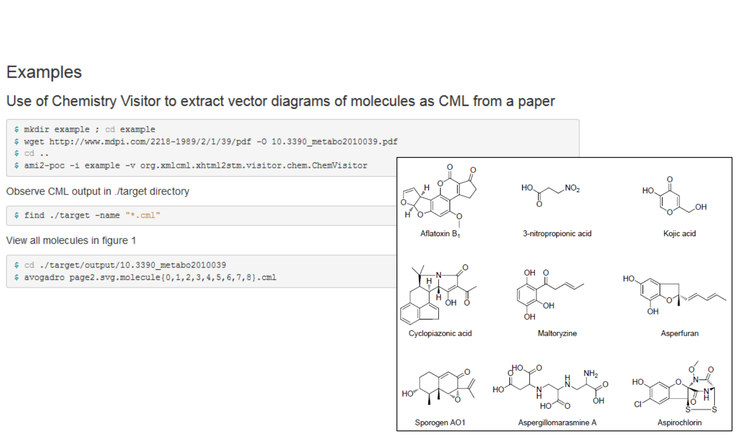
Chemistry is then determined by identifying which objects are joined to which, and what they are joined with

AMI2 architecture



Other visitors can e.g. extract species names from text and phylogenetic trees from diagrams

The program is available at
https://bitbucket.org/AndyHowlett/ami2-poc



Examples

Use of Chemistry Visitor to extract vector diagrams of molecules as CML from a paper

```
$ mkdir example ; cd example
$ wget http://www.mdpi.com/2218-1989/2/1/39/pdf -O 10.3390_metabo2010039.pdf
$ cd ..
$ ami2-poc -i example -v org.xmlcml.xhtml2stm.visitor.chem.ChemVisitor
```

Observe CML output in ./target directory

```
$ find ./target -name "*.cml"
```

View all molecules in figure 1

```
$ cd ./target/output/10.3390_metabo2010039
$ avogadro page2.svg.molecule{0,1,2,3,4,5,6,7,8}.cml
```

Aflatoxin B_1 · 3-nitropropionic acid · Kojic acid · Cyclopiazonic acid · Maltoryzine · Asperfuran · Sporogen AO1 · Aspergillomarasmine A · Aspirochlorin

We are currently working on applying the same principles to images (either digitally born or from photographs)