# The CombiUgi Project and Closing the Open Science Loop

June 19, 2007 update

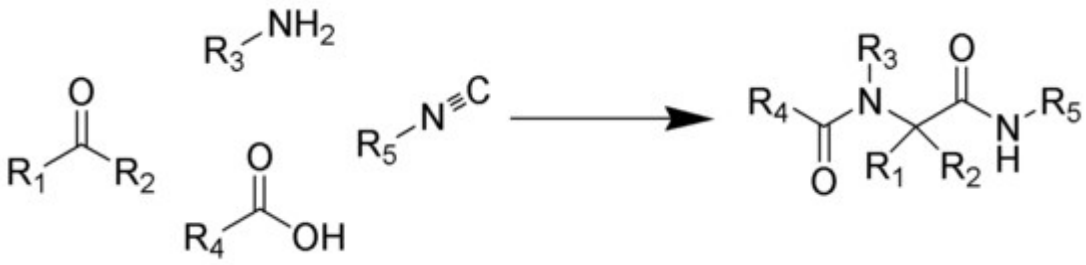
Jean-Claude Bradley

includes work from
Rikesh Parikh
Rajarshi Guha
Dan Zaharevitz

http://usefulchem.blogspot.com/2007/06/combiugi-says-order-2-naphthyl.html

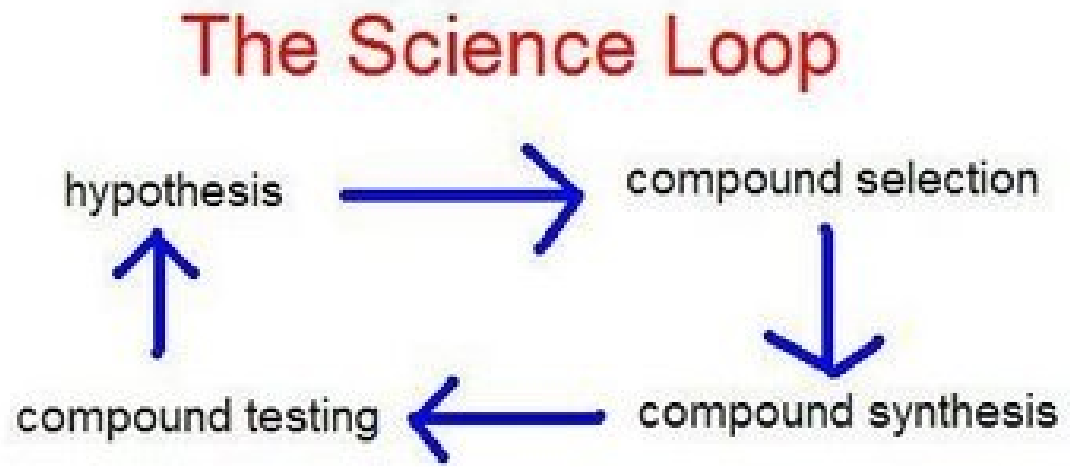
http://usefulchem.blogspot.com/2007/05/combiugi-and-closing-open-science-loop.html

# The Ugi Reaction




Scheme from http://en.wikipedia.org/wiki/Ugi_reaction

A few weeks ago I asked my undergraduate student Rikesh Parikh to kick off the CombiUgi project: to create lists of commercially available boc-protected amino acids, aldehydes, primary amines and isonitriles. He is now done and the links to purchase each compound is provided, in addition to the SMILES code.




By indexing these compounds in relevant search engines (I am working with Chemspider to make this happen) as **UsefulChem molecules available upon request** (and justification) we have an opportunity to close the loop on a practical Open Science project.

By the loop I mean a complete iteration from hypothesis to deciding which compounds to make to actually making them and getting testing results. These results will confirm or force a modification of the hypothesis and the cycle goes through another iteration hopefully closer to producing a useful outcome (a good drug lead compound for example).

I imagine that this loop operates in a lot of research groups. But doing the work under Open Science conditions lets it evolve in new ways. First of all, the **direction of progress is determined by the collaborators that elect to participate in the process, not necessarily scientific objectives**.

An example of that is our recent shift from the testing of our compounds as anti-malarial agents to testing them as tumor inhibitors simply because Dan Zaharevitz from the National Cancer Institute contacted me and suggested that we submit our compounds.

Right after we started to submit our compounds, Dan left this message:

*The folks at Indiana have done a lot of cool stuff that is well worth looking at. One thing they have running in a preliminary form is a service that predicts acompound's activity in cell lines in the screen. This compound is predicted to be inactive in the cell lines in the prediction. I actually don't think that is a bad result. We probably should put up a place to discuss screens and screening strategy, but essentially a prediction tools such as this summarizes what is known. A compound that is predicted to be inactive, but turns out to be active is much more likely to show you something new and interesting than a compound that is predicted to be active and is active.*

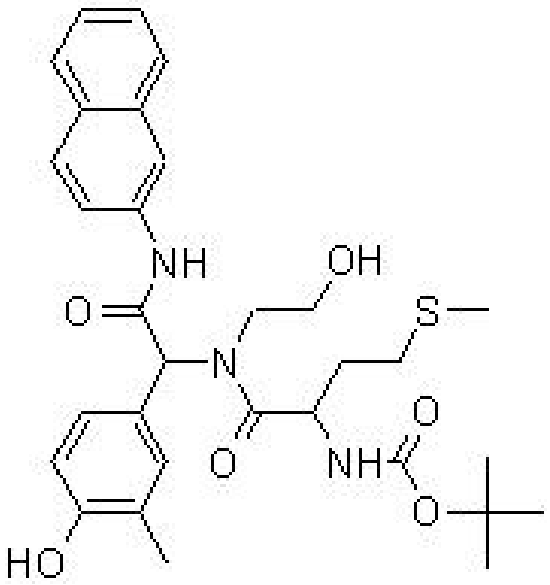
So that's the last piece that closes the loop. This web service will make a prediction about activity of the compounds generated by the CombiUgi algorithm and rank them. The flagged compounds will be identifed and synthesized then tested via NCI's assays for tumor cell inhibition.

My group' s core expertise is the synthetic component. As far as we are concerned the other 2 processes are black boxes. And for scientists involved in the computation and testing, our synthesis operation is probably a black box. But doing everything in the open, hopefully this will allow other researchers to propose other models and create derivative loops of their own.

We'd love to do the same for the anti-malarial assays but we have not found an established system in place like NCI that will do substrate screening routinely at no cost (except shipping of course).

**Is it becoming clearer why I think the scientific process can be automated in novel and useful ways with the progressive adoption of Open Science?**

Rajarshi really worked hard on getting an algorithm to create the Ugi product SMILE codes and passed them through his tumor cell inhibition program. Out of about 68,000 he identified a shortlist of 21 that showed the most activity ( see wiki for details). An example is shown below:



I find it very interesting that all the top hits involve 2-naphthyl isocyanide and over half involve boc-methionine. Is this real or even meaningful? We've been discussing these issues privately and I hope that Dan, Rajarshi and others continue the discussion openly.

## Top 21 Hits (SMILES format) for predicted anti-tumor activity

C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.OCC%91.Cc1cc(ccc1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCC%91.Cc1cc(ccc1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCC%91.Cc1cc(ccc1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCC%91.Cc1cc(ccc1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.CSCCC(NC(=O)OC(C)(C)C)%90.CCCCCCCC%91.Cc1cc(ccc1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCCCC%91.Oc1ccccc1%92.c%931ccc2ccccc2c1
C(=O)%90N%91C%92C(=O)N%93.Oc1ccc(CC(NC(=O)OC(C)(C)C)%90)cc1.CCCCCCCC%91.Cc1cccc(c1O)%92.c%931ccc2ccccc2c1

The point of this excercise is not so much to prove that this model is correct or that we have found a new anti-tumor lead (though that would be nice) but that we can close the scientific loop of hypothesis-synthesis-assay in a completely open and collaborative scientific environment.

I welcome suggestions of other compounds from our virtual library that might be worth making (for any disease-related target), as long as we have assays that someone can run.

We are also working with Tony Williams to see if ChemSpider can serve as a database to store and manage the virtual library, the predicted properties and the assay results. Hopefully then we could increase the library to several million molecules.

# InChI Tags

InChI=1/C11H7N/c1-12-11-7-6-9-4-2-3-5-10(9)8-11/h2-8H
2-naphthyl isocyanide

InChI=1/C5H11NO2S/c1-9-3-2-4(6)5(7)8/h4H,2-3,6H2,1H3,(H,7,8)
methionine