Minutes of the IUPAC IChI meeting, Columbus, OH
June 30, 2002

Attendees:

Steve Bachrach (sbachrach@trinity.edu)
Jonathan Brecher (jsb@camsoft.com)
John Brennan (jbrennan@epo.org)
Steve Heller (srheller@nist.gov)
Sandy Lawson (alawson@mdli.com)
Alan McNaught (mcnaughta@rsc.org)
Peter Murray-Rust (pm286@hermes.cam.ac.uk)
Warren Powell (wpowell2@juno.com)
Henry Rzepa (h.rzepa@ic.ac.uk)
Steve Stein (steve.stein@nist.gov)
Dmitrii Tchekhovskoi (dmitrii.tchekhovskoi@nist.gov)
Matt Toussant (mtoussant@cas.org)
Bill Town (bill.town@chemweb.com)
Tony Williams (tony@acdlabs.com)

Summary:

Steve Stein reviewed the progress made by NIST in developing the test version of the IUPAC Chemical Identifier – the IChI. The test version handles simple organic molecules. To date, in all of the testing (almost 70 copies have been distributed) there are no known examples of chemicals that the program does not handle. A number of suggestions (described below) were made regarding testing and output. The overall view was that the project is progressing considerably faster than expected. A lecture by Steve Stein on the project was given the following day at the CAS/IUPAC Conference on Chemical identifiers and XML for Chemistry and a copy of the slides presented can be viewed at:

http://www.hellers.com/steve/pub-talks/columbus-702/frame.htm

Background on the project can be found at:

http://www.iupac.org/projects/2000/2000-025-1-800.html

Steve Stein presented the results of the work being undertaken at NIST in support of the IUPAC Chemical Identifier (IChI) project, which is the result of the programming efforts of Dmitrii Tchekhovskoi. He started by reviewing how NIST evolved the algorithm, which was based on existing work of others and did not involve any new principles. A beta-version of the algorithm was distributed starting in March 2002. While almost 70 people have requested and received the beta version, the feedback has been minimal. The most useful comment from a beta tester has been a suggestion to consider being able to represent molecules with both defined and relative stereochemistry centers. (Articles in *Nature* and *The Alchemist* have stimulated some requests for the test version, but have not resulted in any feedback to date.) No examples of chemical structures which the program cannot handle have yet been found. The testing has been done on a few databases, such as the NIST and NCI structure databases. A large file (perhaps 1 million) structures from MDL is expected to be tested shortly. While testing millions of structures is possible, there still is the question of needing to actually examine the output to be sure it is what was input. With a 700 MHz PC, average processing time per structure is 2 milliseconds, but faster PCs will reduce this time.

Steve Stein indicated there were two major problems found: Chemists and Chemicals. Chemists are a problem as they have different ideas on how to represent chemicals. This is a human problem not likely to be resolved. Chemicals are a problem since the chemical structure depends on conditions – such as temperature, pH, and so on. This is also a problem not likely to be resolved.

The assumptions being made for the IChI algorithm are:

1. Throw away all electron density
2. Free rotation around all single bonds
3. Always a basic connectivity layer, then an isotopic layer, then a stereochemistry layer (*Z/E* and sp^3), and then a tautomer layer.

Issues not yet resolved in the current beta version include stereogenic centers (an atom and its bonding partners which is not superimposable on its mirror image) and zwitterions. NIST expects to have all normalization rules defined and programmed by the end of 2002. A second beta version is expected to be released at that time. It was suggested that a feature to compare multiple structures and show their differences be included in the next test release.

The issue of what the final and actual output should look like was discussed. At present the plan is to have a number of parts to the output:

1. Molecular formula
2. Connectivity listing (i.e., connections between atoms)
3. Isotopic and stereochemistry

It was suggested that Sandy Lawson look into creating some hash-code identifier with some degree of chemical intelligence. Sandy agreed to consider this project.

Jonathan Goodman, Cambridge, (who was unable to attend) is working on a Java version of the IChI. The best way NIST has found to test the IChI is to renumber the structures and see if the results are the same.

The meeting attendees were all very pleased with progress made at NIST. The hope was expressed that others could be brought into the project for expert help in deciding on how to handle proteins, polymers, and other chemical classes of compounds, so that the IChI could be a true, pure, and complete chemical identifier.

Steve Heller, Secretary
July 8, 2002

Appendix: Agenda for Meeting

Progress in the Development of the IUPAC Chemical Identifier

A project review meeting to be held in the Executive Board Room (Pfahl 102) of
the Blackwell, 2110 Tuttle Park Place, Columbus, Ohio 43210, on Sunday, June 30, at
9.00 am

Agenda

I. Technical Progress Review

1) Up to last review (August, 2000 to July 4, 2001)

2) From last review to Beta-test release (March 2002)
   - XML structured output
   - Separate into sub-layers: *Z/E* and *sp*^{3} stereo
   - Added 5-membering ring tautomers

3) Since Beta test
   - Identify true stereocenters
   - Efficiency improvements
   - Tautomer/stereo perception (underway)

4) Beta Test Results
   - 66 copies requested
     - Four technical comments
     - No negatives
   - Technical issues discussed
     - Output format
     - Tautomer definition
     - Reversibility
   - Implications of Low Response

5) Plans
   - Algorithm
     - Stereo for conjugated networks/zwitterions
     - Finalize tautomer definition
   - More testing
     - Data Sources

   1) NIST
   2) NIH
   3) 1M from MDL
   4) chemical catalogs

   - Establish output format

Presently favor Formula – Connectivity format
Represent multiple species for a single compound.
Relation to CML, others
Include additional data for 'reversibility'
Coordinates, bond and charge positions
Additional ambiguity/error detection
For chemistry 'spell checker'
Write guidelines for structure input
Need help.

6) Timeframe
Beta testing done.
Final test version – December, 2002

II. Promotion

Articles in *Nature* and *The Alchemist*

http://www.rsc.org/IUPAC8/attachments/IChI-Nature502.pdf
http://www.chemweb.com/alchem/articles/1015947904091.htm

Others?

Write paper(s).

Implement in NIST reference databases

Set up distribution/information website (IUPAC?).

General Discussion

What can be done now to prepare.
Find volunteers/partners
Encourage Announcements/Articles

Build list of potential applications
XML chemical identifier
Digital Object Identifier
Structure validity checker
Establish a baseline structure convention
Organizing compounds/Inventories
Communications
Technical Documents
Automated Processing

III. After Version 1 – Has our view changed?

Organometallics next?
Polymers
Markush
Proteins
Conformations and additional stereo (extend organics)