Summary of the meeting

Encoding molecular structure – requirements, methods, problems, and pitfalls

in Bielefeld, Germany, June 16-17, 2000.

The meeting was hosted by the *Graduiertenkolleg Strukturbildungsprozesse, Forschungsschwerpunkt Mathematisierung* at the University of Bielefeld and took place in the Mathematics Department there (cf. **www.mathematik.uni-bielefeld.de**).

Prof. **Andreas Dress** (Mathematics Department, University of Bielefeld; titular member of IUPAC's Commission II.2, Nomenclature of Inorganic Chemistry (CNIC); **dress@uni-bielefeld.de**) opened the meeting with some remarks to place it in the context of the restructuring IUPAC organization and the round table discussion on the future of chemical nomenclature which had taken place in Washington D.C. in March 2000.

A working group within CNIC had been considering aspects of computer-based nomenclature for some time and had actually set up a meeting in 1997, also in Bielefeld, where a framework had been proposed for a general encoding of chemical structures. In Berlin in August 1999, there had been contacts between the CNIC working group and CNOC (IUPAC's Commission III.1, Nomenclature of Organic Chemistry), and it had been agreed to arrange a separate meeting, again in Bielefeld, to explore common interests in computer-based nomenclature. In the meantime, the Washington meeting had been planned by upper IUPAC management, and it was decided that the most rational order would be to postpone the new Bielefeld meeting until after the Washington meeting. At the latter meeting, then, the idea of the chemical identifyer had been advanced, as can be seen from the minutes, which are accessible on the IUPAC home page. Dr. Stephen Heller (of NIST, see below) had been made responsible for conducting a feasability study in relation to the identifyer concept (see below) and was therefore a particularly relevant person to invite for the present meeting.

A number of talks were then given, with ample time for discussions between talks. Handouts from most of the talks are available upon request.

First, Prof. **Adalbert Kerber** (Mathematics Department, University of Bayreuth; **kerber@uni-bayreuth.de**) talked about ***Combinatorial chemistry and the need for canonical forms of molecules.*** He described applications of mathematical concepts and tools in combinatorial chemistry (such as double cosets for permutational isomer enumeration and generation, as developed by E. Ruch and coworkers) and explained the central problem of identifying molecules in real-world combinatorial libraries with molecules in virtual libraries created by computer programs.

The required comparison of structures calls for *canonical forms of graphs* at the topological level, and the canonization of additional structural descriptors if stereochemistry is also to be taken into account. The situation is similar to the nomenclaturists' problems of establishing unique names for structures within chemical nomenclature.

Combinatorial libraries are sizeable, as exemplified by prof. Kerber in one case with a tabulation of the number of isomers possible when reacting amino acids with the symmetrical core compound cubane-1,3,5,7-tetrakis(methanoyl chloride). If 10 amino acids are available, the virtual library counts 925 isomers; with 20 amino acids the size is 13700. Prof. Kerber has developed the software package *Molgen* (cf. **www.mathe2.uni-bayreuth.de/molgen4/**) together with the next speaker. This software may *e.g.* be used for enumerating combinatorial libraries. A further software package *Molcomb* actually generates libraries based on mathematical methods published in recent years as documented by prof. Kerber.

In the ensuing discussion, the analogies between, on one side, *data structures* in computers representing chemical structures and, on the other side, traditional *chemical nomenclature* were highlighted.

Prof. **Reinhard Laue** (also Mathematics Department, University of Bayreuth; **laue@uni-bayreuth.de**) then, in his talk entitled ***Normal forms***, further addressed the problems associated with the generation of unique representations of chemical structures and the general *recognition problem*, *i.e.*, being able to realize that seemingly different objects, for example in a database, are in fact equal. Graph-theoretical approaches due to G. Luks and B.D. McKay were presented, and the Morgan scheme proposed at Chemical Abstracts Services in the 1960's was also mentioned. Complications arise, however, due to the fact that molecular graphs are not sufficient for specification of real-world chemical entities, since the graphs do not, *e.g.*, take into account such phenomena as tautomerism, resonance structures, and stereochemical features, all phenomena which require the representations to exhibit a certain degree of *local flexibility*. The general problem is what a proper mathematical representation of a real-world object is.

The following discussion addressed, among other subjects, the question of efficiency in database searching.

Dr. **Stephen R. Heller** (National Institute of Standards and Technology (NIST), United States Department of Commerce, Gaithersburg, Maryland; **srheller@nist.gov**) then described the concept of the ***IUPAC Unique Chemical Identifyer*** (IUChI) and the feasability study reagarding this construct that he had taken responsibility for after the Washington meeting. The feasability study is ongoing now based on NIST resources and volunteer participation, and is due for completion and presentation to the IUPAC Bureau in September 2000.

The *chemical identifyer* is envisaged as a construction comprising a user interface including graphical input options and a number of algorithms or encoding procedures leading, in principle, for each chemical compound to

(1) a chemical (IUPAC) name in the traditional sense

(2) a connectivity table (when applicable)

(3) a unique *IUPAC chemical identifyer number* (IChIN), a multidigit Arabic numeral, possibly prefixed by *e.g.* a letter to indicate general classification(s), such as organic compounds, inorganic compounds, minerals *etc*.

Whereas there are already commercial programs that provide feature (1) within specified classes of compounds, based on confidential and proprietary computer codes, features (2) and (3) would be new, and the idea would be to base them on open source codes. Thus, any user would, in principle, be able to derive the IChIN for any compound.

A string of computer characters (or a number of strings suitable for various purposes) could, in addition, be derived from the connectivity table by a hashing process or directly from the user input by a variety of algorithms.

The assignment of the IChI number would be irreversible in the sense that one could not, in general, derive the chemical structure from it.

The discussion following immediately after dr. Heller's presentation was concerned with the number of digits needed in the IChIN and with the missing reversibility of this representation. Some participants questioned whether it would be feasible or practical to insist on a fixed number of digits. It is at least clear that since the number of theoretically possible compounds is infinite, there can not be a limit to the number of digits needed to represent *all* of them. On the other hand, if 15 or 20 digits were sufficient to number all real-world compounds in a practical way, the IChIN would still retain some ease of handling, something similar to a bar code as used today.

The handouts provided by dr. Heller included an extensive listing of literature from the late 20th century dealing with unique numbering algorithms, structure representation and naming.

Dr. **Gunnar Brinkmann** (Mathematics Department, University of Bielefeld) in his talk ***How to generate and list all members of large classes of of large molecular structures*** advised to start any attempt at a general encoding of chemical structures *modestly* and create *an extensible system*. Together with PhD student **Sebastian Lisken,** he demonstrated the computer program CaGe (for 'carbon generator'; at present Unix-based, Windows version planned) which generates and enumerates generalized fullerenes, including nanotube structures. Because these structures are topologically planar, a linear-in-time algorithm for canonical numbering may be established. The full software setup offered by CaGe comprises a *structure generator*, an *embedder* creating a 2D or 3D realization of the structures and finally a *viewer* (*e.g. Rasmol* or *Geomview*).

Dr. **Janusz L. Wiśniewski** (Beilstein Informationssysteme GmbH, Frankfurt; titular member of CNOC; **jwisniewski@beilstein.com**) is the author of Beilstein's *Autonom* software for naming organic chemical structures and gave a talk about this system entitled ***Systematic nomenclature as fuzzy encoding of molecular structures:***

Autonom ***case study***.

The Beilstein database comprises chiefly organic compounds, which are selected according to frequency of mention in the literature (the 0.5 % most rarely mentioned compounds are excluded). *Autonom* was developed with the goal of being able to assign a unique IUPAC-acceptable name (in the "Beilstein dialect") to a majority of these compounds, subject to certain restrictions, particularly on molecule size. At the moment, its success rate is around 86 % at a speed of naming 30-40 structures/min (*i.e.*, in 14 % of cases, naming of the compound is, as dr. Wiśniewski put it, "politely refused", and a reason for not proposing a name is given to the client). Certain advanced features are included, such as the addition of Cahn-Ingold-Prelog chirality descriptors in the latest generalized version (compatible with the current version of the CNOC 'Preferred names' document).

The Power Point presentation used in this talk, which also contained details on the various versions of *Autonom* available for various types of computer hardware, may be obtained by contact to dr. Wiśniewski. Also, *Autonom* version 2.1 may be checked out at the Internet site **www.chemweb.com**.

A future challenge is to develop a 'reverse *Autonom*', *i.e.* a program which from a name derives a connectivity table.

Dr. **Ture Damhus** (Novo Nordisk A/S, Bagsværd, Denmark; titular member of CNIC; **damhus@teliamail.dk**), in a talk with the title (slightly changed relative to the announced title) ***Some requirements and opportunities for a global chemical identifyer system***, addressed the needs of the practising chemist and the practising nomenclaturist, as viewed by a member of an IUPAC nomenclature commission.

He started by recalling that the minutes from the Washington meeting had seemed to him so exhaustive in terms of viewpoints expressed and addressed that it felt as if there was not much more to say, but that the discussion at this meeting so far had provided motivation for nevertheless making a number of points. Well aware that some of his suggestions might be overly optimistic, he mentioned the following requirements for a chemical identifyer system and made the remark that not all of these are met by present IUPAC nomenclature, so that they could also be viewed as *opportunities* for an oncoming chemical identifyer system:

A chemical identifyer system

- must also be able to handle completely asymmetric structures (example shown of fullerene with a hole torn in it)

- *parent structures* (and substructures) *must be recognizable* (and new substructures/parent structures keep coming; furthermore, the same compound may have several sets of substructures relevant to different researchers)

- *particularities of the bonding scheme and stereochemistry must be searchable*

- the definition of 'a compound' must be flexible [cases of isomers, tautomers, fuzzy stereochemistry mentioned; also composition may be unknown at the time of encoding (*extreme case:* enzymes) or undeterminable]

- must be able to handle the chaotic bonding schemes of organometallic chemistry (connectivity matrix poorly defined)

- must be able to handle renaming of building blocks (examples from current CNIC work given)

The point was finally made that traditional nomenclature will stay, even if many functions will be taken over by computer-encoded data structures. So it must be part of the game !

Dr. **Olaf Delgado** (Mathematics Department, University of Bielefeld) spoke, under the title ***Virtual crystallography***, about *combinatorial tiling theory*, which provides a means of classifying crystal structures. Prof. Dress and coworkers have developed a theory which assigns so-called *Delaney symbols* to periodical tilings. The Delaney symbols are codes or 'blueprints' which define the tilings uniquely up to equivalence. It is well known which Delaney symbols correspond to tilings of the Euclidean plane, and it is easy to test planar tilings for equivalence. Dr. Delgado, in joint work with D. Huson, has lately derived a number of results regarding tilings of Euclidean 3-space, and in particular he is now able to decide which Delaney symbols correspond to 3-dimensional tilings. He also reported about some recent progress regarding the problem of associating to an arbitrary crystal structure a tiling that appropriately reflects its 'topology' as conceived by the chemist/crystallographer, pointing out, however, that a completely satisfactory way of doing this has not yet been found .

Dr. Delgado has coauthored *RepTiles* and other programs which generate and represent tilings in 2 and 3 dimensions.

Prof. **Andreas Dress** (see above) had given his talk the title ***The wild animal park of polyoxometallates and other polyhedral structures***. He used a number of slides displaying polyoxometallate structures provided by prof. **Achim Müller** (Chemistry Department, University of Bielefeld), who was present in the room with several of his students.

Prof. Dress incidentally started by noting that the mathematician Johann Benedict Listing in 1847 coined the word *topology* in his Göttingen paper *Vorstudien zur Topologie*, in which he used the word in exactly the way it is normally used in chemistry today.

The latest structural wonders in the polyoxometallate world are analogs of the fullerenes and some display icosahedral symmetry (dubbed *keplerates* by Dress and Müller).

The subsequent discussion questioned whether such large structures (the keplerates are clusters in the nm range!) should be named at all. Members of prof. Müller's group (who unfortunately left soon after) vehemently advocated using the 3-D structure itself, pointing out that in these days, new structures are produced at a pace where naming and searching for structures via other representations are practically impossible.

Thus, the scene was appropriately set for the final round table discussion, which had been given the title ***(How) can mathematics and computer science support the IUPAC Chemical Identifyer Project***.

Main points made and questions asked:

- the identifyer system must be open to future extensions

- the problem of overlapping substructures must be addressed

- prof. Dress had a vision of a number of mappings of the real world onto sets of numerical strings being established over time, such that "the" identifyer would be successively extended

- ***Q:*** is a completely universal identifyer possible, *e.g.* based on tiling theory ? ***A*** (dr. Brinkmann)***:*** difficulty: very sensitive to small changes in atom positions, *i.e.* infinitesimal changes could lead to different tilings as descriptors; ***reply*** (prof. Dress)***:*** not necessarily a big problem

- language problems in connection with traditional nomenclature may be used in arguing for funding – the EU is spending vast resources on translating information such as nomenclature

In connection with a discussion of the encoding of very large structures, dr. Heller mentioned that he is in contact with the protein data banks on this issue.

The meeting was closed at lunch time on June 17. Dr. Heller thanked the participants for providing much input of relevance to his feasibility study regarding the chemical identifyer project, but also warned that this project is a huge endeavor, and that a stepwise establishment of the desired systems will probably be the way ahead. *E.g.*, to begin with, only compounds with a defined connectivity will be addressed.

Summarized by Ture Damhus, July 2000