EXTRACT FROM
CHEMICAL INFORMATION AND COMPUTATION 2008, NUMBER TWO
236th ACS NATIONAL MEETING AND EXPOSITION
PHILADELPHIA, PA, AUGUST 17-21, 2008

Dr. Wendy A. Warr, Wendy Warr & Associates, 6 Berwick Court, Holmes Chapel, Cheshire CW4 7HZ, England, Tel/fax +44 (0)1477 533837, wendy@warr.com, http://www.warr.com

INTRODUCTION

The meeting attracted a total of 13,805 attendees, consisting of 8,684 full registrants, 3,086 students, 1,490 exhibitors and 545 exposition-only attendees. The exhibit consisted of 357 companies occupying 511 booths. More than 8,000 papers were presented, in 717 half-day oral sessions and 93 poster sessions. This report concerns mainly the Exhibition, and the technical sessions of CINF (the Division of Chemical Information) and COMP (the Division of Computers in Chemistry). To make the report more useful, I am including some news unrelated to the ACS meeting. I have used press releases dated from July 1 to December 31, 2008.

NEWS AND NEW PRODUCTS

Aureus Pharma

Aureus Pharma (http://www.aureus-pharma.com/) has released three new AurSCOPE Target knowledge databases. They include knowledge databases for nuclear receptor and protease drug targets, as well as the centralised product AurSCOPE Global Pharmacology Space (AurSCOPE GPS). AurSCOPE GPS comprises all of the pharmaceutically important therapeutic targets including G-Protein coupled receptors (GPCRs), kinases, ion channels, nuclear receptors and proteases.

The company has also released a new software solution, AurPROFILER, which enables researchers to evaluate target, cell and drug profiles rapidly using bio-assays derived from the activity data found in Aureus Pharma's AurSCOPE Target knowledge databases. AurPROFILER has been designed to work in conjunction with AurSCOPE GPS and with all Aureus' drug target focused knowledge databases. Compound profiling is routinely used by pharmaceutical researchers to evaluate drug target selectivity including off-target effects, polypharmacology and cytotoxicity. AurPROFILER offers easily interpretable graphics and rapidly provides *in silico* profiles. It was developed in collaboration with a pharmaceutical industry focus group consisting of several major pharmaceutical companies.

Aureus Pharma will participate in a consortium along with a number of industrial and academic research teams in a European Commission funded project entitled "preDiCT: Computational Prediction of Drug Cardiac Toxicity". The goal of the three year preDiCT project is to develop models to predict the effects of drugs on the heart. Computer models for cardiotoxicity will not only help to prioritise drug leads but will also ultimately avoid costly laboratory experimentation. Aureus' contribution to the project will be to use its expertise in ion channel/hERG knowledge databases and knowledge management to design, build and populate the platform required for storing the experimental data that will be used to build models, as well as the framework for collecting the knowledge generated in the modelling studies.

In addition to Aureus Pharma, other partners include F. Hoffmann-LaRoche, Fujitsu Laboratories of Europe, GlaxoSmithKline Research and Development, Novartis Pharma, University of Szeged, Universidad Politécnica de Valencia, and the CRS4, Centre for Advanced Studies Research & Development in Sardinia. The University of Oxford is the coordinator of the project. The project is part of a larger European Commission project linked to the Virtual Physiological Human Network of Excellence. More information related to preDiCT can be found at http://www.vph-predict.eu/.

Elsevier

ScienceDirect

The number of full-text articles on ScienceDirect (http://www.elsevier.com) surpassed nine million in July 2009. Elsevier is partnering with the Copyright Clearance Center (CCC) to use CCC's Rightslink service for marketing and monetising the rights to content. Rightslink is an online e-commerce service that allows content users to license content and order reprints online.

ScienceDirect has partnered with QUOSA to introduce the Document Download Manager, a new feature that enables researchers to download multiple full-text articles simultaneously. QUOSA, who specialise in building information management tools, are providers of the Download Manager technology with which researchers can initiate the download of up to 20 PDF versions of full-text research articles from any results list with many fewer clicks. The Download Manager also enables researchers to name downloaded full-text articles automatically, according to their own naming convention, and pre-select a preferred destination for downloads.

Elsevier has launched a new reference linking service enabling its reviewers to have access to referenced articles published by Elsevier, directly from the manuscript they are reviewing; by clicking on the hyperlinks listed alongside the referenced articles, reviewers are brought to the abstracts of those articles. Furthermore, depending on their personal or institutional subscription entitlements, reviewers can also link directly, *via* CrossRef's DOI service, through to the articles referenced in other publishers' journals. This functionality and the facilitating tools are integrated, through Elsevier's submission and peer review editorial system (EES). Elsevier's network of 7,000 journal editors and 500,000 reviewers using EES will have access to abstracts in Scopus and the full text of referenced articles in ScienceDirect, and will also benefit from the functionality of these systems, such as searching for related articles, author and citation information.

Illumin8

Elsevier has announced the newest upgrade to illumin8, its Web-based semantic text-mining solution. For end-users, illumin8 is now even "smarter" in that R&D professionals do not need to remember arcane search parameters or syntax. Instead, they can simply enter the phrase they are interested in and the system works out how to interpret it, allowing them to answer difficult questions. Enhanced features include:

- richer results providing a technology overview or landscape including organisations, products, approaches, benefits, and people associated with a technology or product category;
- smarter search logic that relieves the user of having to enter specialised syntax or parameters;
- explicit display of result types showing the semantic relationship of the result to the search input; and
- addition of nearly half a million pre-defined technology terms to the search guide.

Scopus

Scopus has added "Articles-in-Press" (AiP) abstracts of accepted research papers published prior to being printed, from journals produced by Karger Medical and Scientific Publishers, and Nature Publishing Group (NPG). Later in 2008, AiPs from BioMed Central and IEEE will also be available. Scopus previously offered access to AiPs from Elsevier and Springer that included 2500 titles. This number will now rise to about 3000. Researchers will gain access to the full-text by linking from Scopus to the publishers' digital library. Scopus will be nearly doubling its Arts & Humanities (A&H) titles by April 2009.

Scopus has been chosen by the Organisation for Economic Co-operation and Development (OECD) to enhance the research performance analysis it conducts for member and non-member countries. The OECD will offer members insight on worldwide research productivity trends and potential opportunities for developing needed innovations. With the enhanced XML data provided by Scopus CustomData, such as linkage between authors and institutes, the OECD will be able to perform more in depth analysis than in the past. Scopus data will populate the OECD's print and online scorecards which members can use to assess and compare their country's research output to other nations. The OECD will also employ Scopus to help its members get a better view of co-authorship and cross-border research collaborations.

PharmaPendium

PharmaPendium has significantly expanded its content with a new release containing the European Medicines Agency's (EMEA) European Public Assessment Reports (EPARs). The addition of this content makes PharmaPendium the only source of consolidated, searchable access to US Food and Drug Administration (FDA) and EMEA drug approval documents on a single site, with a single search. EPAR documents cover medicines assessed by the Committee for Medicinal Products for Human Use (CHMP). They include efficacy, indication, safety and pharmacokinetics data and mode of action information. The EMEA database on PharmaPendium makes accessible approximately 80,000 pages of searchable documents for more than 300 active ingredients approved for the European market, some not approved in the US. It also presents preclinical and clinical toxicity and adverse effects data manually extracted from these documents. General product information documents are also included.

EMBASE

EMBASE has partnered with QUOSA to enable mutual customers to download and manage the full text of EMBASE.com search results *via* a newly established EMBASE.com channel in QUOSA Information Manager. The information manager boosts journal retrieval, search and management efficiency for individual users through post-search automated full article retrieval, article organisation and specialised full-text searching. In combination with QUOSA Virtual Library, the partnership enables expert searches on EMBASE.com to be integrated into an easy-to-use current awareness and literature access solution for larger groups. EMBASE.com combines EMBASE and MEDLINE on one platform; users can navigate from citations to full text from STM publishers.

Reaxys

Elsevier had a reception in Philadelphia to allow scientists to preview Reaxys (http://info.reaxys.com/index.php), a new workflow solution for synthetic chemists, which was released in January 2009. Reaxys is based on data from CrossFire Beilstein, CrossFire Gmelin and Patent Chemistry Database, now merged to provide harmonised content, additional functionalities and a redesigned interface. Reaxys also contains new features including a synthesis planner and advanced results handling to help scientists gather relevant information more efficiently.

Features include workflow and decision making support for synthesis design and planning, e.g. by comparing alternative synthesis routes and selecting the most relevant paths; quick access to data by displaying results in a tabulated overview of the most important information; and convenient and flexible output of data in most common formats. Reaxys has been developed with input and advice from development partners from corporate and academic research institutions around the world. In addition, the Web interface has been designed by experts in human-computer interaction. Search results reflect how chemists think and work, helping their workflow move from initial search and recognition of possible starting materials and their properties, through planning a strategy for the synthesis of a new compound, to proposing a detailed synthetic route. Most compounds in Reaxys can be identified by their CAS Registry Numbers.

Elsevier and CAS recently reached agreement to increase cooperation in information exchange, which will improve their respective operations and service offerings. Under that

agreement, Elsevier is providing enhanced electronic versions of its journal articles to CAS, and CAS in turn is helping Elsevier identify additional CAS Registry Numbers in certain of its databases. The CAS Registry Number identification aspect of the agreement is being carried out periodically through an automated, first level, “machine comparison”. That method is neither comprehensive nor detailed relative to many chemical structure conventions, and so this resulting addition of Registry numbers does not represent a complete comparison of Elsevier resources with the CAS Registry, the most comprehensive substance collection. Nonetheless, Elsevier anticipates that these added CAS Registry Numbers will be useful and helpful for search and identification of more Elsevier database substances. From time to time more CAS Registry Numbers will be added.

CINF AND COMP TECHNICAL PROGRAMMES

CINF and COMP Division technical programmes and abstracts are on the Web at http://www.acscinf.org/cinf/ and http://www.acscomp.org/.

Data Mining and Text Mining Approaches to Drug Discovery

Hierarchical clustering of chemical structures by maximum common substructures
Miklos Vargyas and Ferenc Csizmadia, ChemAxon Ltd, Maramaros koz 3/a, 1037 Budapest, Hungary, Fax: 361-453-2659 mvargyas@chemaxon.com

Vargyas gave his talk the subheading “Clustering made human”. Cluster analysis has been shown to be successful in the categorisation of physicochemical and biological properties of compounds. However, conventional approaches to clustering molecular structures, where chemical graphs are transformed into sequences of numbers, seldom meet chemists' expectations. Graph based techniques that cluster compounds with respect to common structural motifs are gaining in popularity as these can better mimic human categorisation.

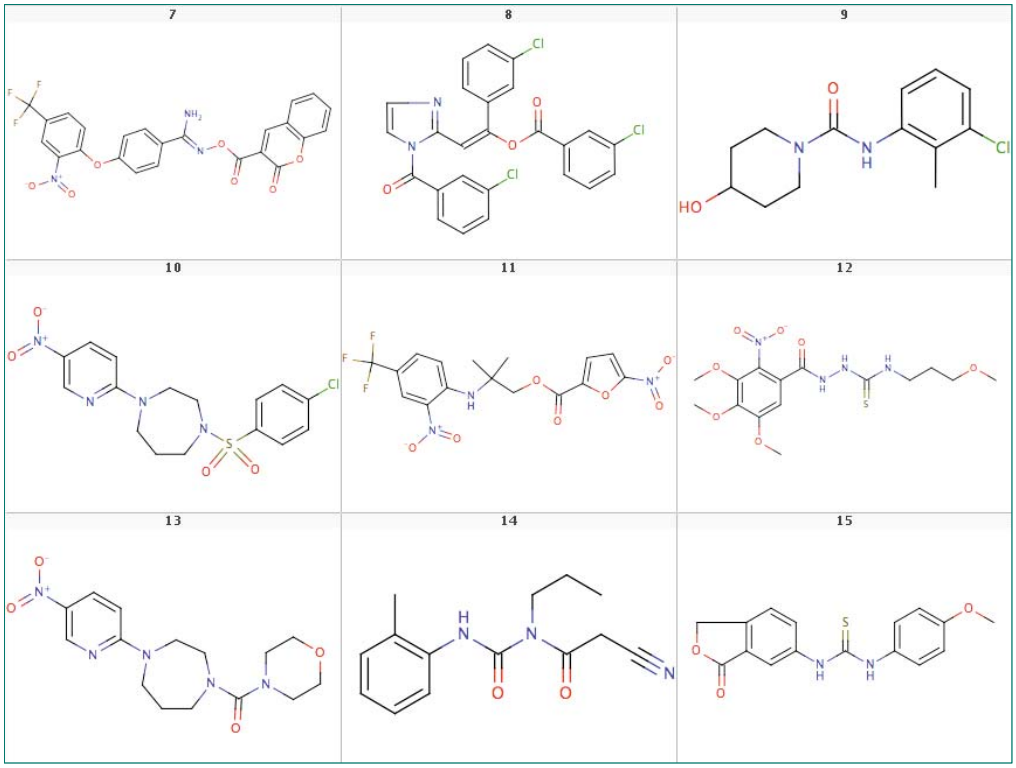
Clustering is carried out to reduce the number of objects with which one has to deal, to group subsets together, and to represent each group by one member of it. Conventional approaches are similarity-based and use fingerprints or other molecular descriptors, high dimensional, artificial chemical spaces, and similarity measures. Jarvis-Patrick clustering is often used, with Tanimoto similarity. It is non-hierarchical. It gives globular clusters, which is an advantage, but has a tendency to create large numbers of singletons, which is a disadvantage. Singletons are essentially “failures”.

ChemAxon does offer Jarvis-Patrick clustering. In one example, 999 objects formed two non-singleton clusters, and eight singletons. Average dissimilarity was 0.66208726, minimum dissimilarity, 0.0 and maximum dissimilarity 0.9411765. By tuning the parameters *t* and *c* this can be improved to 81 clusters and 115 singletons.

t	*c*	Number of clusters	Number of singletons
0.6	0.1	2	8
0.3	0.1	179	248
0.5	0.1	7	36
0.5	0.5	10	37
0.5	0.8	81	115

Parameter *t* denotes the maximum (allowed) dissimilarity of two compounds. That is, if they are more dissimilar than this upper bound then they are not considered to be members of the same cluster. Parameter *c* denotes the minimum (required) ratio of common neighbours of two compounds , that is, if the ratio of number of common neighbours (with respect to number of neighbours) is less then this lower bound then the two compounds being compared are not placed in the same cluster.

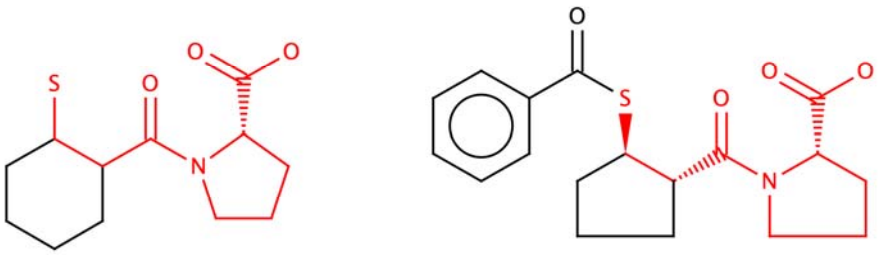
The most populated cluster, however, has structures that would surprise a chemist: the structures do not look alike.



It is tedious to use this traditional method, tuning the parameters in a trial-and-error fashion. Also, the results are not interpretable (the algorithm does not provide an explanation) and they often do not meet chemists' expectations. It is hard to reproduce on the computer what the chemist does when he scatters structures around manually.

The astronomer's job is easier. Gravity does the job for star clusters. Vargyas showed a picture of Andromeda, the largest galaxy of the Local Group. The Andromeda Galaxy, the Milky Way Galaxy, the Triangulum Galaxy, and about 30 other smaller galaxies contain one trillion (10^{12}) stars. Andromeda is known to harbour a dense and compact star cluster at its very centre. Clustering stars is easy because stars have a visible spatial arrangement and the distance between stars defines clusters, galaxies, etc. Clustering molecules is hard because of the lack of innate spatial arrangement. There are infinite types of chemical spaces for artificial arrangement, and various "distance metrics", but it is usually hard to visualise the results because of high dimensionality. There are various approaches, of which none is superior. The "best method" depends on the application area, and on the actual data.

We need a method that requires little or no tuning and gives an easy to understand, simple "explanation". A novel approach is structure based clustering based on maximum common substructure (MCS). The MCS is largest substructure shared by two molecules, e.g.,



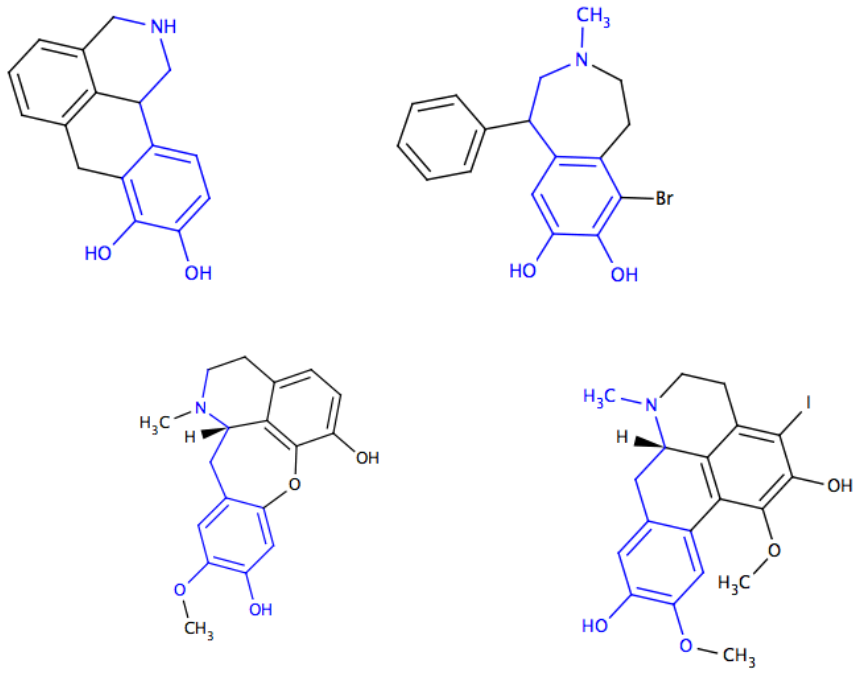
It is a simple concept, more "human", and visual, yet it is hard (i.e., expensive and slow) to compute. In substructure searching the query structure is known; it "simply" has to be found as part of the target structure. This is known as subgraph isomorphism. Graph isomorphism is

even “simpler”, yet it is NP-hard. Finding the answer can take a long time. The algorithm scales exponentially with respect to the number of graph vertices, in the worst case, but validating an answer is fast. Finding an MCS is even more complex: the “query” structure is not known and all possible substructures need to be checked. Even the number of substructures is exponential. Vargyas tabulated the features of the two approaches that have been used for finding an MCS. As far as he is concerned, speed is not the major factor, so he favours backtracking.

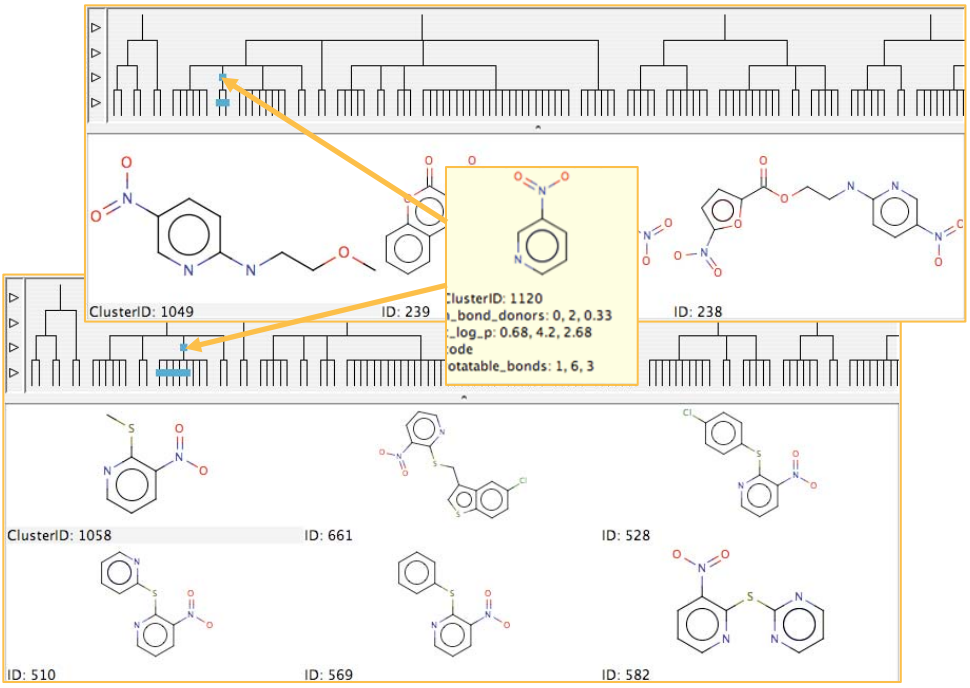
Backtracking	Clique detection
Ad hoc	High mathematical elegance
Average complexity is better than worst case	Average complexity is same as worst case
Dynamic heuristics	Static (initial) heuristics
Colouring is easy	Colouring is hard*
Fuzzy matching	Fussy matching

*Colouring in this context means the association of chemical properties to atom centres, such as atom type, charge, hybridisation, etc.

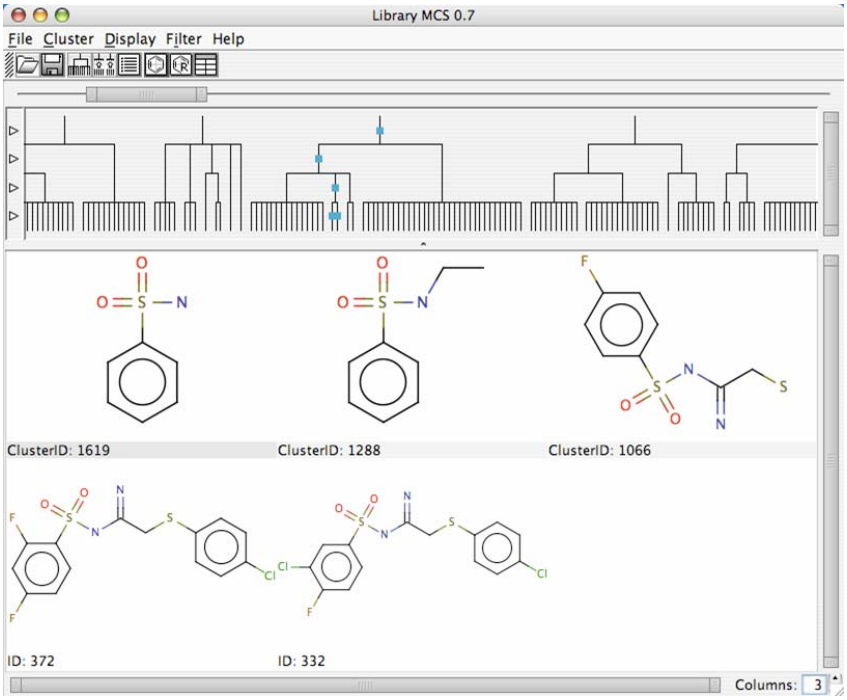
Another problem for the MCS algorithm is that lots of structures need to be clustered, not just two, e.g.,



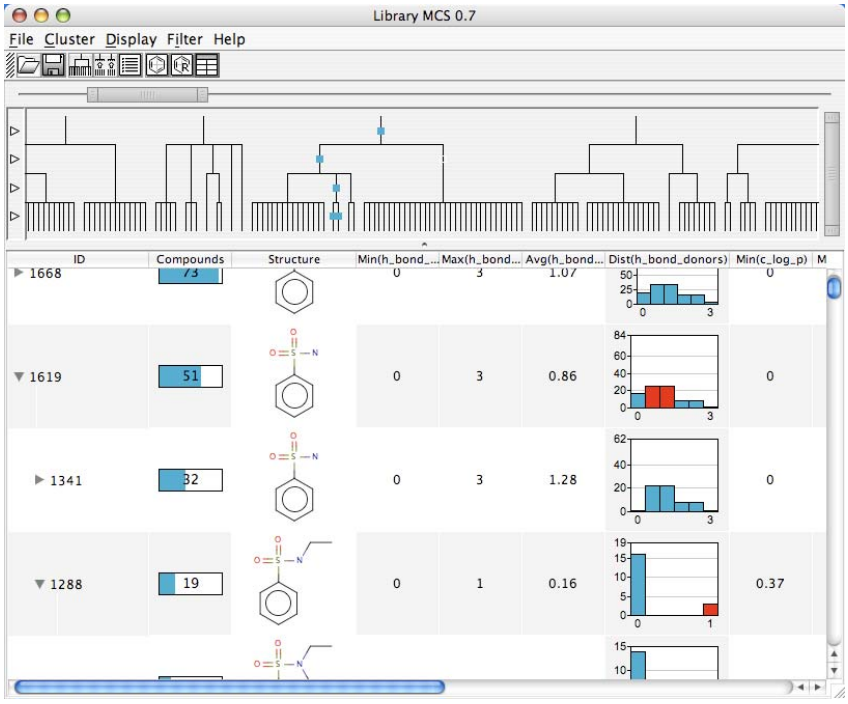
ChemAxon’s LibraryMCS is a hierarchical clustering program. Unlike some other graph based clustering methods, it neither involves a similarity based pre-clustering step nor relies on predefined fragments.

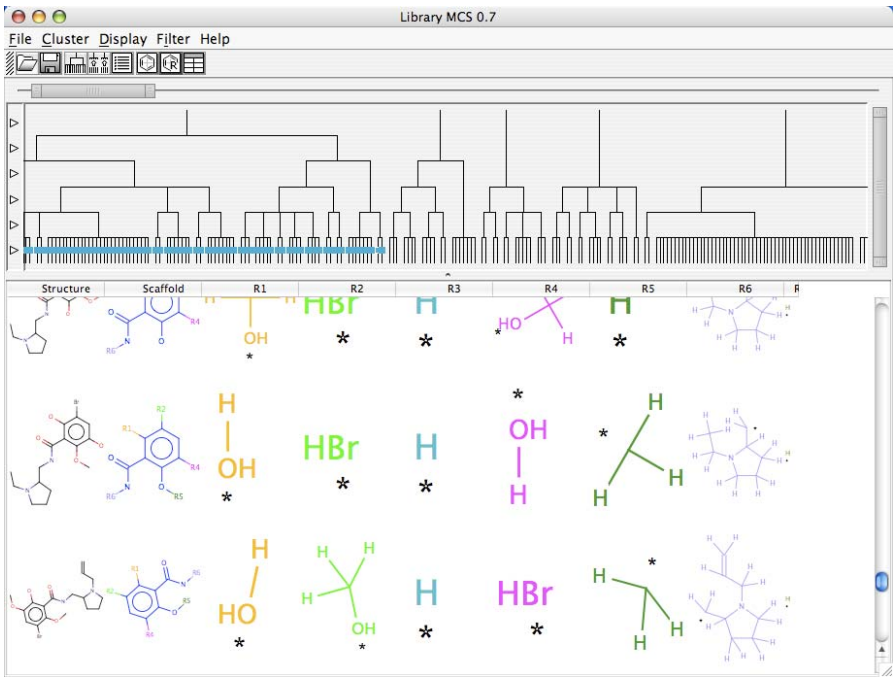


It offers intuitive visualisation.

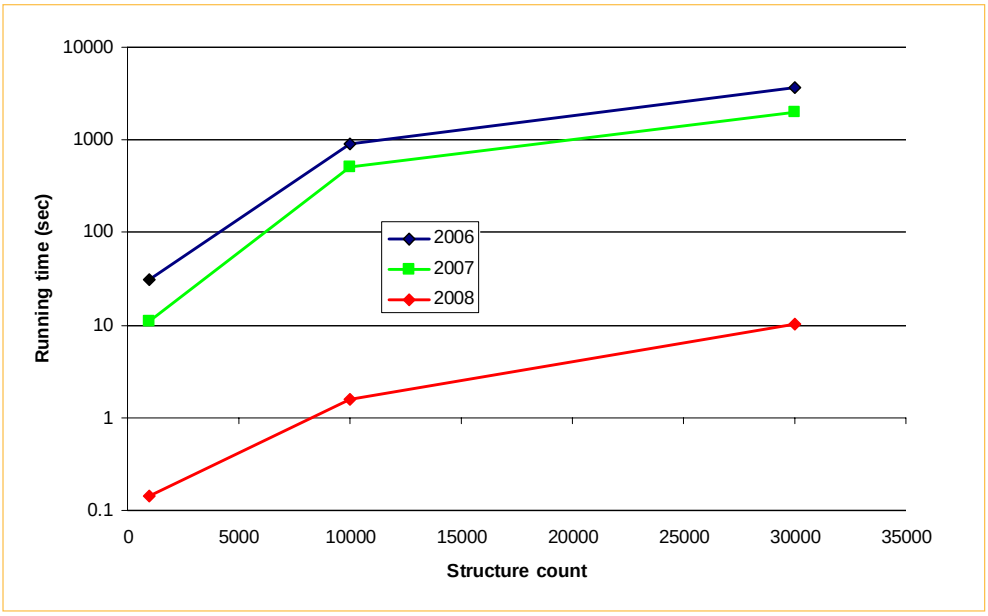


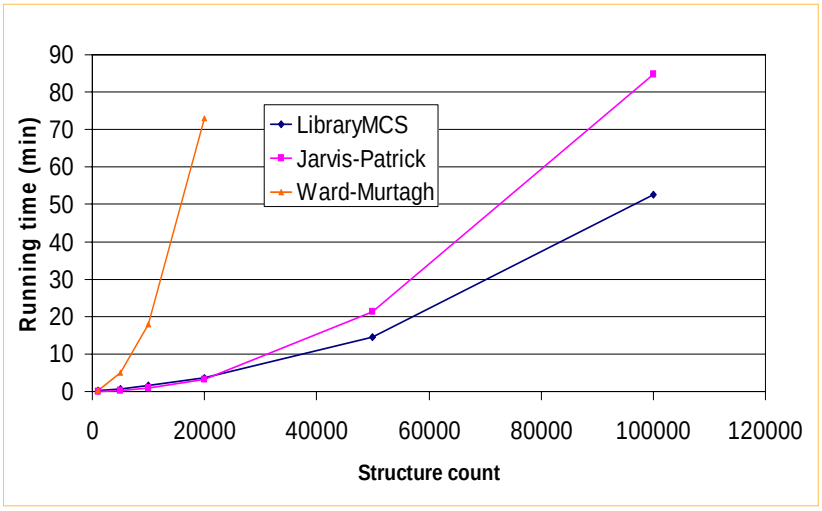
It also features an SAR-table view and R-group decomposition.





Recent evaluation by different research groups has indicated that LibraryMCS is capable of producing high quality clusters agreeing with human categorisation within a practicable time (approximately 1000 structures/second). LibraryMCS scales linearly and has been getting faster every year, although the actual timings depend on the size and diversity of the structures clustered. It performs better than Jarvis-Patrick clustering and is much faster than Ward's. The following results were produced from the Maybridge collection.





LibraryMCS can be used to cluster virtual high throughput screening hit sets, focused libraries, combinatorial chemistry libraries, diverse sets and, in the future, even corporate libraries. Typical uses are virtual high throughput screening hit set profiling, R-group decomposition by learned scaffolds, perception of novel scaffolds, reverse engineering of combinatorial libraries, diversity assessment of large chemical libraries and compound acquisition.

Exhaustive MCS search is slow but it can be made faster by introducing exact or inexact heuristics. In some cases the program misses the MCS but this does not badly affect the clustering. The use of use of heuristics, on average leads to less than 10% misclassifications. The method is useful for obtaining a bird's eye view of larger or more diverse sets. Vargyas demonstrated exact and inexact heuristics. LibraryMCS is still under development but it can be tried online at http://www.chemaxon.com/shared/libMCS.

In the development pipeline are integration with Instant JChem, Spotfire integration, and a new dynamic viewer. Enhancements being planned are disconnected MCS, multiple class members, and the use of Bemis-Murcko and other frameworks. In the long term it would be nice to extend the method to libraries of 1 million compounds, or more, and to implement additive clustering.

ADME Informatics. Converting Raw Data to Useful Knowledge for Drug Discovery

Processing drug discovery raw data collaboratively and openly using Open Notebook Science

Jean-Claude Bradley, bradlejc@drexel.edu, Department of Chemistry, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, Rajarshi Guha, rguha@indiana.edu, School of Informatics, Indiana University, Bloomington, IN 47406, and Phillip Rosenthal, San Francisco Division of Infectious Diseases, University of California, San Francisco, CA 94143

Recently there has been a movement towards making the scientific process more open. Bradley described a spectrum of research tools from closed to open, from the traditional, unpublished laboratory notebook, to the traditional journal article (which still does not include a lot of the elements such as the failed experiments), to an open access journal article, to a fully transparent Open Notebook source. The spectrum in teaching runs from a traditional paper text book and face-to-face lectures, to public lecture notes, to public assigned problems, to archived public lectures and free online textbooks.

The UsefulChem project in chemistry (http://usefulchem.wikispaces.com) is an open source science project, using Web 2.0 tools, led by the Bradley Laboratory at Drexel University. Since all laboratory experimental results are made public, the work is also described as Open Notebook Science. Bradley reported on a project designed publicly to report ongoing research within a research group working on the development of anti-malarial agents. The project makes use of free hosted tools as much as possible so that the infrastructure can be easily replicated by other research groups. A great many different tools are used.

IUPAC International Chemical Identifiers (InChIs), InChIKeys and compound names are used as tags on blog and wiki pages to facilitate indexing on common search engines. The handling of large libraries and interfacing with online databases is generally accomplished with SMILES lists. Substructure searching and annotation are handled by ChemSpider. JSpecView is used to manipulate JCAMP-DX spectra over a browser interface. Other technologies used include Collaborative Drug Discovery (CDD, http://www.collaborativedrug.com/), Google Docs, and mailing lists.

One of the greatest benefits to Bradley of doing Open Notebook Science has been to find some excellent collaborators: Rajarshi Guha of Indiana University has been doing docking for Bradley's synthetic chemistry team, and other workers have tested compounds. For example, Phil Rosenthal of the University of California, San Francisco (UCSF) has been testing compounds for anti-malarial activity.

The experimental section of a typical journal article has lots of information missing. In contrast, Bradley showed a screen shot of work targeting the enzyme falcipain-2 (Phil Rosenthal is testing compounds in this field) and he clicked on a link to EXP150, going to the wiki (http://usefulchem.wikispaces.com/Exp150) that is the laboratory notebook of how those compounds are actually made. It looks like a standard laboratory notebook but it makes more information available. Bradley clicked on a link and went to ChemSpider which calculates properties, molecular weight, SMILES and InChI. A user can do a substructure search in ChemSpider without having software on his or her own server.

The notebook wiki also links to Rajarshi Guha's docking procedure. By clicking on the results links, Bradley ended up with a Google Doc that had a list of SMILES in order of docking score *versus* falcipain-2. The procedure section can be pasted into a user's own publication. The spectrum has raw data behind it and, using JCAMP combined with JSpecView, peaks can be expanded, and coupling constants calculated. Compare this with the sort of less useful spectrum stored as a PDF in a supplementary information section. Bradley is using ChemSpider to archive compound characterisation information, e.g., to store NMR spectra and images. The spectra can be expanded in ChemSpider, which allows a predicted NMR spectrum to be found for comparison with an experimental one. It is nice for students to be able to get a predicted spectrum.

All the details of the experiment are on one very long page, but there has to be a log, so that the researchers can actually construct the rest of the experiment based on the log. A proper log is absolutely critical. When Bradley says that his team's experiments are in real-time, what he means is that the log has to be online by the end of the day. The other sections of the experiment may take weeks to be uploaded. Readers do not have to take Bradley's word for it that a reaction yield was 59%: they can go back and reinvestigate every single aspect and the arguments that were made.

Experiments can be accessed in various ways; for example, by the table of contents or by using Google search. At the very bottom of each blog or wiki page, there are tags. Bradley uses InChIs as tags, but for very large molecules, such as his Ugi products, the InChIs are not indexed properly by Google. So, recently, Bradley has started to use InChIKeys. Using Google Custom Search, all of Bradley's blogs, and wikis and all the pages that his team has generated can be searched in a system that will look only at Bradley's approved pages. All this is available free of charge for anyone to do.

How are people finding Bradley's experiments? Sitemeter shows that most people find them through a Google search. They could be looking for specific compounds, for example looking for the NMR of TFA. Or they might input a molecular formula. They could be asking "tell me everything that you know about guanidine" and they would certainly find hits in ChemSpider. They might be looking for experimental conditions. Searching for side reactions of amines, say, in the traditional literature is not likely to be useful, yet a typical laboratory notebook is almost all failures. A user searching for kinetics of Boc deprotection in UsefulChem will be able to find lots of kinetics analysis. For teachers, the site also offers chemistry videos for free downloading and 3D periodic tables. Other people are looking for bigger pictures, like lysomal targets, cheminformatics, and project proposals. Experiments can also be searched by the traditional table of contents file.

Bradley gave a more specific example of how Open Notebook Science could be used. Someone searches for purification of phenylacetaldehyde by distillation and finds an experiment carried out by one of Bradley's students. The results show the procedure and a picture of his experimental set-up, with its flasks, thermometer etc. Some problems with the NMR spectrum were recorded. In the log section all the details are recorded and, in a discussion, phenylacetaldehyde was reported to boil at 195º .The first fraction was collected at 162º. Perhaps the boiling point was wrong or perhaps the experiment was wrong? In the notebook the user can actually see what is wrong, i.e., see droplets floating in the flask, and conclude that the experiment must be redone in an inert atmosphere.

Details such as these are not available in traditional notebooks. The wiki shows the history of each page. The statement that "phenylacetaldehyde was purified" has a third party time stamp of October 27, 2006: This was changed in a later version. Red is used for deleted text and green for added text. Bradley added his own comments on January 3, 2007: there was a conversation with the student concerning whether there was water in the condenser or not (the tubes were disconnected before the photograph was taken).

Bradley showed an introductory page for one student ("My name is Shannon") where all her experiments were listed. There were 200 Ugi experiments. These are tricky to compare in Google, so Bradley has summaries to compare them in Google Docs. He has made a YouTube video demonstrating the experimental set-up. For example, you might want to see where the thermometer was. Five thousand people have viewed one of Bradley's videos.

Guha, Bradley and Rosenthal have used a mailing list for inter-group collaboration. This is good for debugging. The results of their collaboration are ready for publication. Bradley has given the structures of 71,000 compounds (from Ugi reactions) to Rajarshi Guha for docking with the enzyme falcipain 2. He showed micromolar inhibition of both the enzyme and red blood cells through binding site V1 of the enzyme, and another group of compounds in the V2 site which inhibits the enzyme but does not stop the infection of red blood cells. Anyone can drop in on the experiment and see what is going on: this is a good way to get collaborators.

Collaborators can use CDD and report assay results; they can put X-ray structures into eCrystals; and they can talk to a spectrum in Second Life. Bradley and his colleagues have

been having meetings in Second Life. Because JCAMP is used, macros can be used to calculate kinetics and a reaction profile can be plotted automatically. Andy Lang is building 3D structures in Second Life. Gus Rosania's group at the University of Michigan is collaborating with Bradley; he soon made his own wiki and 10 of his students upload their notebooks every day. Cameron Neylon at Southampton has also been doing Open Notebook Science, using a modified blog instead of a wiki.

Obviously people are good at inputting results but there are advantages to getting machines to do some of the data manipulation, so Bradley has been working on getting results into machine friendly format. A log can be translated into a machine readable format. InChIs can be scraped and ported into a database. Mettler Toledo has lent Bradley a robot which records the log of experiments and makes XML files. A MiniMapper system was then built for optimisation of the Ugi reaction. Google Docs is used to program and report. Bradley showed a 5D plot of the optimisation. Different shapes corresponded to different solvents. A 50% yield was increased to 68% by this optimisation.

We are getting to the stage where human beings can actually collaborate with machines if the human beings choose to make information available to them. Eventually we will get to the point where machines can actually do real science, formulating hypotheses, testing them, analysing the results, and then planning the next experiment. If this is to happen, we need to have free services; we need to have a possibility that anybody in the world can write a script that will try to process information and produce something useful. In conclusion, we should communicate first and standardise second. In a Science 2.0 world, redundancy is king. People used to put everything into one paper and hope that everyone would cite it. Nowadays the data is much more distributed. You can still publish your results in traditional journals, but you can put the data into multiple places.