From Big Data to Chemical Information

RSC Chemical Information and Computer Applications Group (CICAG) and
Dial-a-Molecule Grand Challenge Network
22nd April 2015, Burlington House

We kindly thank Dr Colin Bird (University of Southampton) for preparing and providing this report

Introduction

The event was co-sponsored by RSC Chemical Information and Computer Applications Group (CICAG) and the Dial-a-Molecule Grand Challenge Network, which is funded by the EPSRC. The meeting brought together a diverse group of attendees interested in the challenges presented by "big data" and whether the chemistry situation might be different in any way. The morning session was devoted to talks assessing the scope and effect of "big data" from a chemistry perspective, while the afternoon session comprised talks about various approaches to managing "big data" and exploiting the opportunities that it presents. The day concluded with a keynote by Tony Williams, covering a notably wide range of topics related to RSC activities with large datasets.

Emergent themes

While it was to be expected that our speakers would offer different perspectives on the definition of "big data", it was perhaps less obvious that several of them would suggest that chemical data is not necessarily "big" data. Issues related to data integration arose in most of the talks and the need to capture metadata at source was one of the themes that arose more than once. A number of our speakers urged us to look first at what was already available before creating a new resource such as an ontology. While "big data" does present chemistry with diverse challenges, the tone of the meeting was optimistic: there are opportunities to meet those challenges.

Rise and Impact of Big Data

Big Data and the Dial-a-Molecule Grand Challenge

Richard Whitby

As the Principal Investigator for Dial-a-Molecule, Richard presented a range of big numbers: quantities that underpin the challenges associated with making novel molecules quickly. The real challenge for the synthetic organic chemist lies in deciding how to plan a synthesis so that you know it will work. Consequently, organic synthesis has to change from its current compound-driven approach to being a data-driven discipline. It is the data that has the lasting value.

Richard suggested that the predictability of a reaction sequence was analogous to a chess problem, except that we do not know enough about the weightings, owing to a lack of data. To improve that situation, we need to capture data at source, especially for reactions that we deem to have failed. Big numbers are involved, the largest being the estimated upper limit of the reaction space in terms of connections between molecules (10^18 to 10^200). Although we can to some extent reduce the reaction space with techniques such as functional group approximation, the space

remains huge; it is difficult to say where we are, as the amount of information is still restricted.

Richard reviewed four computer-aided synthesis design programs that are currently in use, but concluded that they are essentially ideas generators. He contended that it should be possible to use data more effectively, illustrating his point with an example having limited information in the reaction database, more in the publication, but with far more information not there. More data will become available, particularly from ELNs, but researchers have to be willing to share and provide data in a format that can be exchanged. Automated capture of experimental conditions can help considerably, as experimenters tend to record only a fraction of what they actually observe. How we make effective use of the flood of data that will be produced will be a real big data challenge: should we keep it all or could we throw some away? Ultimately, the challenge is to make the most effective use of the existing data to predict reaction outcomes.

Big, broad and blighted data

Jeremy Frey

Chemical data is diverse and heterogeneous, which breadth differentiates it from the big data produced by, for example, the Large Hadron Collider. However, chemical data is sometimes not what we think it is, so might as a result be 'blighted'. These characteristics arise in part because chemical data comes from a lot of sources of different sizes: the distribution has a long tail, as it does when chemical data is analysed by country of origin. There is a huge amount of heterogeneous information now available from a lot of countries around the world. Moreover, since 1980, chemical information publications have given way to chemical informatics.

The use of social networking has increased the amount of user-generated content, but in a form that is potentially unprocessable. Such content might even include information about failed reactions, albeit emerging by unconventional routes. The community could also become involved in carrying out work, although questions of control inevitably arise.

Echoing Richard's message, Jeremy advanced the need to automate data capture, emphasising the importance of metadata, which people are known to be reluctant to assign. Metadata has to be captured at source; there are risks with adding it later. Insight and further information are essential for climbing the Data-Information-Knowledge-Wisdom pyramid, as Jeremy illustrated by reference to black bananas.

Semantic web technologies offer hope, with the caveat that human understanding of machine-machine interactions is important; otherwise we will not trust the findings. Typically, the further information we need will represent context, as Jeremy illustrated by reference to keto-enol tautomerism and to water. Work is going on with chemical schema and ontologies, including building links to other disciplines.

Jeremy then introduced this theme of *reducing uncertainty* (which we might relate to increasing understanding), illustrating his point with images of the Amazon and Okavanga deltas. The analogy is with data streams coming together to form a data river that then spreads out into a delta. However, the Okavanga spreads into a desert, rather than an ocean, which would not be useful in the case of data. Exploiting the integrated data relies not only on knowing its provenance, but also being confident about its correctness. If an assertion is subsequently shown to be untrue, it is almost impossible to remove that defect, because the information necessary to do so is not there. The challenge is to get people to do the work to ensure their data is reusable.

Digital disruption in the laboratory: joined-up science?

John Trigg

After teasing us with a "spot the difference" between pictures of labs old and new - his answer being that it is no longer necessary to wear a tie when working in a lab - John embarked on a comprehensive overview of the transformation wrought by the evolution of digital technologies. The result has been a fundamental change in the way we communicate, arising from the implementation of digital technologies, causing disruption. We no longer rely on a third party; we can do things ourselves. We have to manage IP differently; to adopt different business models, doing more with less; and to adjust our scientific method to be conscious of data curation, provenance, integrity, and preservation.

Against a background of knowledge and expertise being dispersed geographically, and chemistry becoming more complex and less certain, John used the Snowden and Stanbridge landscape of management diagram to reason that the current drive is towards process engineering (rules and order), because social complexity (un-order and heuristics) is difficult to manage.

John believes that the nature of laboratory work will change, creating a need for more education (for understanding) as opposed to training (for doing). Currently we have too much of the latter, raising the question: what happens if/when no longer need cognitive input? The Internet of Things is increasing the number of devices with machine-machine protocols, giving us unprecedented opportunities to exploit new technologies, provided we increase our understanding and do not rely on black-box technology without cognitive input.

John concluded with Charles Darwin's quote to the effect that the species that survives is the one most responsive to change.

Big data chemistry

Jonathan Goodman

Jonathan began by asking whether chemistry has big data, inviting comparison with astronomy, which can generate more data than all the CCTV cameras in the UK. He cited the Wikipedia view that big data is characterised by being difficult to process with traditional techniques, noting that chemistry has few reactions that we really understand and many more that we would like to understand.

He depicted a machine for making molecules as a box, which in its Mark III incarnation would take sunlight plus raw ingredients such as O_2, N_2, and CO_2 and deploy a library of processes to make the best molecules with specific properties. He asked us to identify the simplest molecule that we could not make: his answer was tri-t-butyl isopropyl methane.

Jonathan then suggested that we need different ways of looking at molecules, while acknowledging that there would be some resistance to change. To build his Mark III machine, we will also need new models for 'understanding' chemical data. Not all of the data that we would need is openly available, yet if life depends on it, we will want to know that a structure is correct, giving the debate about the structure of maitotoxin as an example. Even when we have a lot of data, do we understand it?

How long might it take from an invention to a derived product being available from Tesco, for example? Jonathan gave several examples: Teflon to non-stick pans; lasers to CD players; and (somewhat tongue-in-cheek) molecule-making machines to ready meals.

Discussion

Panel comprising the first session speakers

Asked whether chemists would be happy publishing failed reactions, Richard Whitby replied that we need a culture change: we know there are errors, so should not be afraid to disclose them. John Leonard queried what a failed reaction might be, as a 1% yield might still be regarded as a success.

In response to a question about the scientific method, Jeremy Frey said that it should be taught in schools, adding that ethics has to be dealt with early, a point reinforced by John Trigg.

The observation that, under US Patent law, machines cannot make inventions elicited a range of responses. Jonathan Goodman saw it as a challenge for lawyers; Jeremy Frey argued that encouragements for creativity deserve rewards; John Trigg pointed out that inventions are kicked off ideas, which prompted Jeremy to point out that cognitive computing (such as IBM's Watson) could suggest new ideas.

Asked why chemistry lags behind in depositing data to a repository, as required for open access, Jonathan Goodman observed that it was intrinsically difficult, then Donna Blackmond said that funding agencies require a plan. Jeremy Frey asserted that intelligently accessible data is the lifeblood of future progress. Depositing data needs to to become good, standard practice.

Approaches to Managing Big Data and Maximising Opportunities

Managing and searching large chemical structure data resources

Mark Forster

Mark's perspective was that chemical data is not necessarily big data, but computations could be big. As part of its portfolio, Syngenta is interested in developing new pesticides, which, with *in vivo* testing, can go from hypothesis to bioactivity testing in a few weeks, so the model for compound discovery is quite dissimilar to pharmaceuticals.

Mark's first example of the large structure datasets that Syngenta manages was ChEMBL: there are 28,000 compounds in the pesticide literature that are not in ChEMBL, but are now being added as a result of a collaboration between Syngenta and the EBI ChEMBL group.

Syngenta are investigating new search processes to find candidate compounds, surveying both corporate and vendor compounds: for example, Openbabel Open Source Chemistry Toolbox and Chemfp which uses fingerprints. They also use a script to find 'new' compounds using the eMolecules public chemical data set, extracting those not seen previously and performing property calculations and analyses. Pesticide physical property scoring produces HFI similarity scores: H(erbicide), F(ungicide), I(nsecticide) likeness. They also calculate a compound's novelty relative to Syngenta corporate compounds. Mark illustrated the use of Knime to filter and visualise structure, with the data flow set by search criteria. He showed a scatter plot in which size represented H score and colour the Novelty. Mark noted that they use InChI keys for linking data, observing that one can sometimes can put an InChI into Google and find a web page about the compound.

Data-rich organic chemistry: enabling and innovating the study of chemical reactions

Donna Blackmond

Donna's presentation was based on a two-day NSF-sponsored workshop held in Washington, DC, in September 2014. She brought copies of the reports for us to take away. The motivation for the workshop came mainly from the pharmaceutical industry, their collective interest being in: enabling technologies for capturing process information; precompetitive collaborations; and promoting the use of tools for reaction monitoring.

One aim of the workshop was to find new ways to fund academic research and to train the next generation of 'workers'. There were five talks about models for collaboration. The Caltech model is that of a central catalysis facility where different research groups can access both sophisticated instrumentation and expertise. The collaboration with Merck, for example, operates by Merck offering postdocs short assignments in the company, working in ways that will not harm competitive aspects.

The workshop also covered recent progress with pre-competitive collaboration models. The Pfizer approach relies on data being transportable, which requires compatibility of software, enabling the integration of data into a searchable architecture. For Merck, data-rich tools should be able to be to run without headaches, so that they can make use of the data.

Donna then talked about the need for transformative solutions: obtaining quality in a way that can accelerate development with fewer people. Pfizer are developing the concept of the Lab of the Future, which would require new skills and education as well as training.

Among the challenges is the development of a common data framework, which the Allotrope Foundation is working towards, developing standards and aiming to: improve integrity, reduce waste, realize the full value of the data, bridge the gap between ideas and execution.

The IQ Consortium aims to share ideas without hurting commercialisation, so the Allotrope Foundation and IQ are dealing with IP issues and seeking agreement about what collaborators are looking for. Donna envisaged the possibility of a new heyday of physical organic chemistry with new tools, asking what might have been if the pioneers of the discipline had had our tools. Achieving the aims requires education to enhance critical skills in data-rich science.

The workshop identified four ideas that would be important for the future, of which were given priority: developing new educational models and the development of a Caltech-like centre for data-rich experimentation. Donna expects these topics to arise at the next CCR Meeting, in May 2015, in a session entitled "Disruption in Biotechnology and Process Chemistry".

Use of data standards and metadata in information exchange

Rachel Uphill

Rachel began by identifying categories of big data, such as: gene expression profiles; interactions; reactions in our bodies; and citations. Pharmaceutical companies have a lot of data, which is increasingly complex and of higher dimensionality. They also get data from other sources, albeit often as PDFs, which might contain structures, but do not really use it.

Metadata is the answer. To integrate substance, result, experiment, and project data, we have to rely on metadata; there is no point in storing big data and not doing anything with it. However, questions do arise about the integrity of data, sometimes

epitomized as Garbage In Garbage Out (GIGO). Without the right data, and the right metadata, we are not going to get correct answers.

For GSK, the need is to be able to use the data, adding standards and then embedding the tools according to an information blueprint. This process involves stewardship and governance, to find, understand, and use the data, and to integrate with the information blueprint. There is a range of requirements and measures to give trust in the data and to enable its use. Master Data Management (MDM) provides one reference point: one view of the information.

With regard to standards, that is where the Allotrope Foundation comes in. Data held in Allotrope format does not lose context, so we can look back at its provenance. Rachel advised against creating a new standard without looking first at what is already out there; Allotrope is looking at the gaps, aiming for an open document standard and open metadata repository. Allotrope is also integrating the regulatory aspects, which lead to more requests for information.

Rachel concluded with an example from GSK, joining datasets together in different ways, integrating from external as well as internal sources.

100 million compounds, 100K protein structures, 2 million reactions, 4 million journal articles, 20 million patents and 15 billion substructures: Is 20TB really big data?

Noel O'Boyle

Citing the Wikipedia article, Noel suggested that any dataset could be considered big data if we lack the means to process it, giving examples of large numbers of 'things', adding some wry comments.

He went on to talk about searches for matched pairs [2] and matched series [>=3] in the ChEMBL dataset, which identified 391,000 matched series. In contrast substructure searches are relatively slow, especially when compared with a typical Google search, which can be very fast, owing to its look-ahead feature. Ideally, a sketch search should be underway, using a similar feature. A fingerprinting screen would be fast, although it would produce false positives, and could be followed by slow matching. However, in some situations, such as a structure containing benzene, a fingerprinting screen is not very effective. As the worst-case behaviour arises from slow matching, Noel went on described attempts to speed that up.

The approach is to pre-process the database, matching the rarer atoms first, which Noel showed to be significantly faster, with or without fingerprinting. An alternative approach is to pre-process all substructures, using NextMove's SmallWorld technology, which takes a lot of time and requires a lot of space. Maximum common subgraph techniques are computationally expensive but can be implemented efficiently using SmallWorld.

NextMove also have text mining technologies, which extract chemical names from text, and can find ~90% of the structures (131,000) in all the open access papers from PubChem. Noel ended with his view that many classic cheminformatics problems can be handled with today's techniques.

Dealing with the wealth of open source data

John Holliday

John began with overviews of the Sheffield research areas and the open source data available circa 1999/2000. In comparison, he showed a "scatter plot" of the open resources now available, comprising drug databases and compound databases, offering breadth as well as depth.

Sheffield will be using new as well as old techniques to investigate approaches such as hyperstructures, virtual screening, and data fusion; they are using CASREACT for reaction schemes. They are also exploring cross-database integration issues, for example, multiple formats, with various databases distributed in various formats; consistency (e.g., gaps) is a problem that they cannot do much about. If a test has not been done, they cannot use the data. However, they can report the issue back to EBI, for example, asserting that a particular assay is wrong.

With unstructured data, the question arises whether one can be confident that the data is right. There are now more data types and chemical mime types, such as XML formats, including CML: essentially there are too many formats from too many different sources. Translation is feasible, but can get some loss of data in the process. Looking ahead, we might evolve standard format(s) by virtue of the way we use the data. John thought the situation could settle down with time, as everyone starts to use the same formats.

John then considered data management issues, such as: how to back up databases holding many terabytes; and the need for the right metadata, with the right metadata vocabulary, so they have to enforce the use of metadata. The overall need is for a proper management system. At Sheffield all databases are on MySQL, with a "nice" new front-end.

Using car design in 60s and 90s as his illustration of "soul", John argued for more human input into the design of decision support systems: include some "soul". We have to make sure our output is communicating to everyone. Sheffield uses benchmarking: they now have several benchmark sets for screening databases, where they used to use one dataset. People are now becoming more data aware, but have to pick and choose what is best for them. Although we are all becoming data scientists, programming skills are going down, despite an increase in the use of tools.

Discussion

Tom Hawkins noted that the examples used had been around the paradigm that a molecule has a single structure and a reaction a single result, then asked what was available in cheminformatics to support polymers and mixtures of products. Mark Forster replied that we can register a chemical without a structure; Rachel Uphill that we can register different structures with relationships between them.

Asked how today's tools would scale and what might happen in the future, Noel O'Boyle replied that NextMove tools all scale well and many are parallelisable. Rachel Uphill added that tools now operate on the databases: they are no longer downloaded.

A question about the lack of incentive and credit for publishing good, reproducible, data elicited several responses. Donna Blackmond noted that the "incentive to publish" system is working pretty well; John Holliday pointed out that there is a lot of data available now, so making the data fit the hypothesis might become an issue. Jeremy Frey then queried the possibility of safe "exchanges", to which Donna responded that there is so much incentive now, although there is concern about other groups knowing; however, she and her fellow moderators are trusted. Jeremy asked whether we could create blueprints, Donna replying that the NSF case studies were going to lead to some form of blueprint. Rachel Uphill proposed a hosting centre, so that each company doesn't have to go through the process each time.

Keynote: *Activities at the Royal Society of Chemistry to gather, extract and analyze big datasets in chemistry*

Tony Williams

Using a colourful slide, Tony illustrated big data in terms of the number of things going onto the web in 60 seconds, then showed a count of 95,736,025 substances in the CAS Registry at the time of capture in the afternoon of 22nd April 2015. Tony then traversed more chemistry-related numbers: the prophetic compounds in patents; the compounds in PubChem, including "similar" compounds that require manual curation; proliferation of InChIs, enables Google to access over 50 million records; and the chemicals held by ChemSpider. However, the reality is that relative to brontobytes (10^27), the numbers are not "big data".

The RSC has taken up open access and also open data, although it leads to some problematic conversations: funders tell you to make data available, but not how to do it; do you put data in a repository that might not be supported tomorrow? There is not as much open chemistry data as there should be. Some teams will want open access but be reluctant to release their own data, saying that it is "really important".

Turning to the scientific literature, Tony cited ContentMine, which claims to "liberate 100 million facts", but queried whether they were really facts or actually assertions about a particular measurement. The RSC has published more than 36,000 articles in 2015, posing several "how many" questions. However, much information is lost, particularly relationships, as publications are only a summary of work. Trying to find work from years ago is a problematic area, especially as much of the data in our hands still lies in PDF files. Data in publications should be available; it should not be locked up. Tony posed the question: how much data might be lost to pruning? Nobody will *rush* to publish to the Journal of Failed Reactions, so how much data is thrown away? How much data resides in ELNs? It would be great to sit at an ELN and make requests. How many compounds are made that are never reported? Tony thought he had probably published less than 5% of the work he did; the rest is mostly lost. There are data management systems in most institutions, so it should be feasible to share more data.

Tony then talked about his experiences with computer-assisted structure elucidation (CASE) and in associating structures with NMR spectra and selecting the highest ranked. One of the challenges of data analysis is access to raw data. For example, if 3 NMR peaks lie close together, they cannot be resolved from an image; you need the data in a CSV file. We publish into document formats, from which we have to extract data, whereas they could conform to a community norm. Currently, there isn't even a reference standard. Tony argues that we can solve these problems. In ChemSpider, supplementary spectra information is in JCAMP format: analytical data should be produced in standard rather than proprietary formats.

Mandates do not offer data deposition solutions: we have to build them. The RSC will offer embargoing, collaborative sharing, and links to ELNs. There are standards, although students are not taught about them. There are also ontologies I use, so we should not create new ones.

Tony illustrated the issue of data quality with several examples, such as: detecting corrupted JCAMP files that have got flipped; an allegedly "high quality dataset" giving Mn++ as the symbol for a selenium oxide cadmium salt; a database with only 34 correct structures out of 149; and several other telling examples. His final example was that of domoic acid, for which C&E News had taken the (wrong) structure from Wikipedia rather than from SciFinder, because Wikipedia was free!

Tony then gave the Open PHACTS project as an example of ODOSOS (Open Data, Open Source, Open Standards) before moving on to open source validation with

CVSP (the Chemical Validation and Standardisation Platform), asking whether publishers could use it before submission if all rules were available. He noted that when he started his work, 8% of structures on Wikipedia were wrong: checking and correcting took 3 years.

After mentioning the RSC Archive, Tony gave an example of a reaction description, comprising a diagram and a method: the description would be useful, but we still do not know the context: that's in the publication.

With regard to modelling "big data", Tony talked about melting point models, showing a relatively narrow distribution. He went on discuss building a database of NMR spectra, noting the problems that can occur with names, especially those with brackets. Overall, there are issues with textual descriptions, such as erroneous and incomplete information.

In conclusion, Tony remarked that we are sitting on big data: what it takes is to apply the techniques and standards.

Discussion

Asked why, in the context of open data and open publishing, OpenArchive was not popular in the chemistry community, Tony replied that there was so much value in chemicals, so a reluctance to put data up.

With regard to data quality, it was suggested that a button to report errors would be very useful and should be encouraged. Tony said that ChemSpider publishes changes, but no one wants to take the feed. Responding to a question about publishing data, Tony said the need was to publish data without extracting it.

He was then asked about big data links to other sources and about appropriate curation of keys in other people's databases. Tony said that was what Open PHACTS was about: linking with biological data. It had a very specific focus; one would have to deal with appropriate organisations to deal with other areas.

Posters

Multiparameter optimization of pharmaceuticals: What 'BigData' can tell us about small groups that make a big difference

Al Dossetter

Matched Molecular Pair Analysis (MMPA) shows promise for assessing the pharmacology of new biological targets but the process requires many matched pairs to achieve statistical significance and thus new design rules. Combining and analyzing data from many pharmaceutical companies is both cheaper and faster than making and screening large numbers of compounds. To enable the contributing companies to share knowledge without exposing their intellectual property or critical data, datasets are encoded using only *changes* in structure and property. This poster illustrated the process that created and analyses the 'big data' involved, illustrated with examples of rules found within ChEMBL toxicity data.

Physical chemists' attitudes towards management of laboratory data

Isobel Hogg

This poster presented research into the data management needs of physical chemists, who often face difficulties owing to the quantity of data that they generate. The research also discovered issues with recording the narrative that goes alongside experiments and provides context to the data recorded in the lab, leading to an investigation of how physical chemists currently manage their data and the extent to

which their needs would be served by electronic laboratory notebooks (ELNs). The conclusion was that improved note taking and data organisation within a comprehensive system could improve working practices but data sharing would not be a strong driver for the adoption of an ELN.

Batch correction without QCs

Martin Rusilowicz

Quality Control (QC) samples are often used to assess and correct the variation between batch acquisitions of Liquid Chromatography – Mass Spectrometry (LC-MS) spectra for metabolomics studies. This poster showed how the use of QC samples could lead to certain problems. As an alternative, “background correction” methods use all the experimental data to estimate the variation over time, rather than relying on the QC samples alone. The poster reported comparisons of non-QC correction methods with standard QC correction.

Towards statistical descriptions of crystalline compounds

Philip Adler

This poster presented research demonstrating the use of statistical methods to address relationships between molecular and crystallographic structure. It illustrated example problem domains, and discussed issues with the methodology, both in terms of difficulty with statistical methods, and problems with gathering data in a standardised fashion from the published literature. In particular, sparse and uneven coverage of the chemical space by the literature, especially with respect to ‘failed’ experiments, has proven to be a large hindrance.