42

Combinatorial chemistry and the Grid

●Au: Please spell out the initials for all the chapter authors.

Jeremy G. Frey, M. Bradley, ●J.W. Essex, M.B. Hursthouse, S.M. Lewis, M.M Luck, L.A.V.M. Moreau, D.C. DeRoure, M. Surridge, and A.H. Welsh

University of Southampton, Southampton, United Kingdom

42.1 INTRODUCTION

In line with the usual chemistry seminar speaker who cannot resist changing the advertised title of a talk as the first, action of the talk, we will first, if not actually extend the title, indicate the vast scope of combinatorial chemistry. 'Combinatorial Chemistry' includes not only the synthesis of new molecules and materials, but also the associated purification, formulation, 'parallel experiments' and 'high-throughput screening' covering all areas of chemical discovery. This chapter will demonstrate the potential relationship of all these areas with the Grid.

In fact, observed from a distance all these aspects of combinatorial chemistry may look rather similar, all of them involve applying the same or very similar processes in parallel to a range of different materials. The three aspects often occur in conjunction with each other, for example, the generation of a library of compounds, which are then screened for some specific feature to find the most promising drug or material. However, there are

Grid Computing – Making the Global Infrastructure a Reality, edited by F. Berman, G. Fox and T. Hey.

many differences in detail and the approaches of the researchers involved in the work and these will have consequences in the way the researchers will use (or be persuaded of the utility of?) the Grid.

42.2 WHAT IS COMBINATORIAL CHEMISTRY?

Combinatorial chemistry often consists of methods of parallel synthesis that enable a large number of combinations of molecular units to be assembled rapidly. The first applications were on the production of materials for the semiconductor industry by IBM back in the 1970s but the area has come into prominence over the last 5 to 10 years because of its application to lead optimisation in the pharmaceutical industry. One early application in this area was the assembly of different combinations of the amino acids to give small peptide sequences. The collection produced is often referred to as a library. The synthetic techniques and methods have now broadened to include many different molecular motifs to generate a wide variety of molecular systems and materials.

42.3 'SPLIT & MIX' APPROACH TO COMBINATORIAL CHEMISTRY

The procedure is illustrated with three different chemical units (represented in Figure 42.1 by a circle, square, and triangle). These units have two reactive areas so that they can be coupled one to another forming, for example, a chain. The molecules are usually 'grown' out from a solid support, typically a polymer bead that is used to 'carry' the results of the reactions through the system. This makes it easy to separate the product from the reactants (not linked to the bead).

At each stage the reactions are carried out in parallel. After the first stage we have three different types of bead, each with only one of the different units on them.

The results of these reactions are then combined together – not something a chemist would usually do having gone to great effort to make separate pure compounds – but each bead only has one type of compound on it, so it is not so hard to separate them if required. The mixture of beads is then split into three containers, and the same reactions as in the first stage are carried out again. This results in beads that now have every combination of two of the construction units. After n synthetic stages, 3^n different compounds have been generated (Figure 42.2) for only $3 \times n$ reactions, thus giving a significant increase in synthetic efficiency.

Other parallel approaches can produce thin films made up of ranges of compositions of two or three different materials. This method reflects the very early use of the combinatorial approach in the production of materials used in the electronics industry (see Figure 42.3).

In methods now being applied to molecular and materials synthesis, computer control of the synthetic process can ensure that the synthetic sequence is reproducible and recorded. The synthesis history can be recorded along with the molecule, for example, by being coded into the beads, to use the method described above, for example, by using an

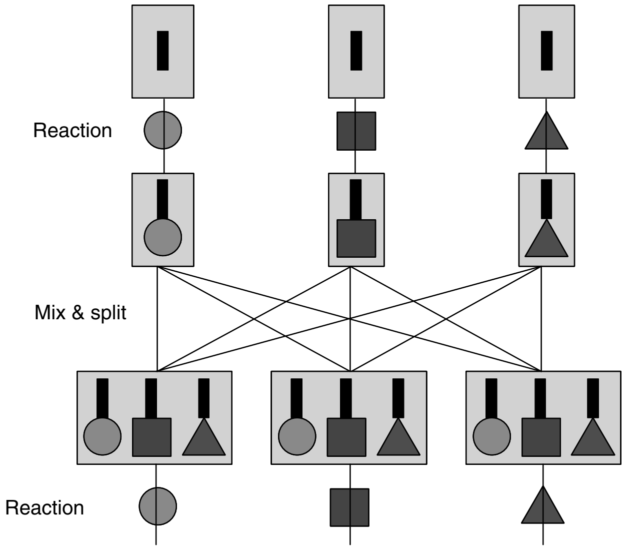

Figure 42.1 The split and mix approach to combinatorial synthesis. The black bar represents the microscopic bead and linker used to anchor the growing molecules. In this example, three molecular units, represented by the circle, the square and the triangle that can be linked in any order are used in the synthesis. In the first step these units are coupled to the bead, the reaction products separated and then mixed up together and split back to three separate reaction vessels. The next coupling stage (essentially the same chemistry as the first step) is then undertaken. The figure shows the outcome of repeating this process a number of times.

●Au: We have rephrased this sentence. Please clarify if it retains the original intended meaning.

RF tag or even using a set of fluorescent molecular tags added in parallel with each synthetic step – such tags are much more readily detected than the structure of the molecule produced from the tiny amounts of a single bead●. In cases in which materials are formed on a substrate surface or in reaction vessels arranged on a regular Grid, the synthetic sequence is known (i.e. it can be controlled and recorded) simply from the physical location of the selected molecule (i.e. where on a 2D plate it is located, or the particular well selected) [1].

In conjunction with parallel synthesis comes parallel screening of, for example, potential drug molecules. Each of the members of the library is tested against a target and those with the best response are selected for further study. When a significant response is found, then the structure of that particular molecule (i.e. the exact sequence XYZ or YXZ for example) is then determined and used as the basis for further investigation to produce a potential drug molecule.

It will be apparent that in the split and mix approach a library containing 10 000 or 100 000 or more different compounds can be readily generated. In the combinatorial synthesis of thin film materials, if control over the composition can be achieved, then the number of distinct ‘patches’ deposited could easily form a Grid of 100 × 100 members. A simple measurement on such a Grid could be the electrical resistance of each patch,

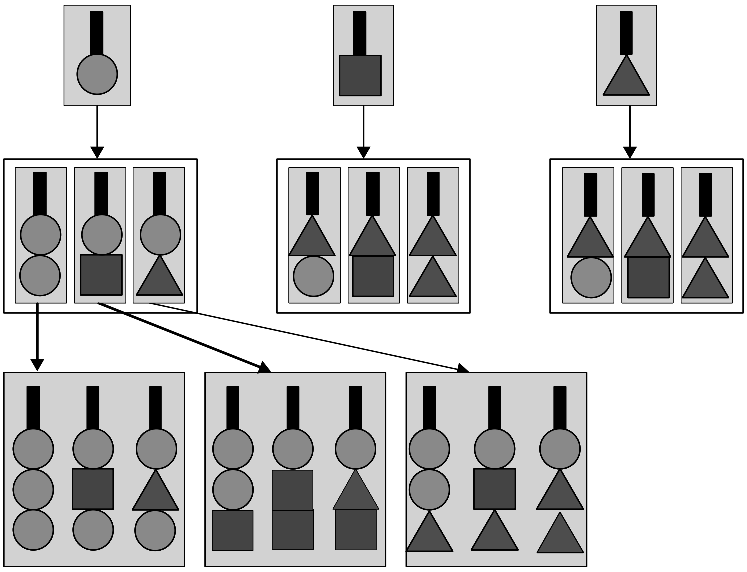

Figure 42.2 A partial enumeration of the different species produced after three parallel synthetic steps of a split & mix combinatorial synthesis. The same representation of the molecular units as in Figure 42.1 is used here. If each parallel synthetic step involves more units (e.g. for peptide synthesis, it could be a selection of all the naturally occurring amino acids) and the process is continued through more stages, a library containing a very large number of different chemical species can be readily generated. In this bead-based example, each microscopic bead would still have only one type of molecule attached.

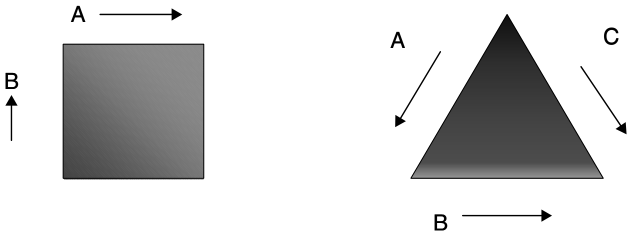

Figure 42.3 A representation of thin films produced by depositing variable proportions of two or three different elements or compounds (A, B & C) using controlled vapour deposition sources. The composition of the film will vary across the target area; in the figure the blending of different colours represents this variation. Control of the vapour deposition means that the proportions of the materials deposited at each point can be predicted simply by knowing the position on the plate. Thus, tying the stochiometry (composition) of the material to the measured properties – measured by a parallel or high throughput serial system – is readily achieved.

already a substantial amount of information, but nothing compared to the amount of the data and information to be handled if the Infrared or Raman vibrational spectrum (each spectrum is an xy plot), or the X-ray crystallographic information, and so on is recorded for each area (•see Figure 42.4. In the most efficient application of the parallel screening measurements of such a variable composition thin film, the measurements are all made in parallel and the processing of the information becomes an image processing computation.

•Au: There was no reference to Figure 42.4 in the text. Please clarify if the citation provided by us is appropriate.

Almost all the large chemical and pharmaceutical companies are involved in combinatorial chemistry. There are also many companies specifically dedicated to using combinatorial chemistry to generate lead compounds or materials or to optimise catalysts or process conditions. The business case behind this is the lower cost of generating a large number of compounds to test. The competition is from the highly selective synthesis driven by careful reasoning. In the latter case, chemical understanding is used to predict which species should be made and then only these are produced. This is very effective when the understanding is good but much less so when we do not fully understand the processes involved. Clearly a combination of the two approaches, which one may characterise as 'directed combinatorial chemistry' is possible. In our project, we suggest that the greater use of *statistical experimental design* techniques can make a significant impact on the parallel synthesis experimentation.

The general community is reasonably aware of the huge advances made in understanding the genetic code, genes and associated proteins. They have some comprehension of

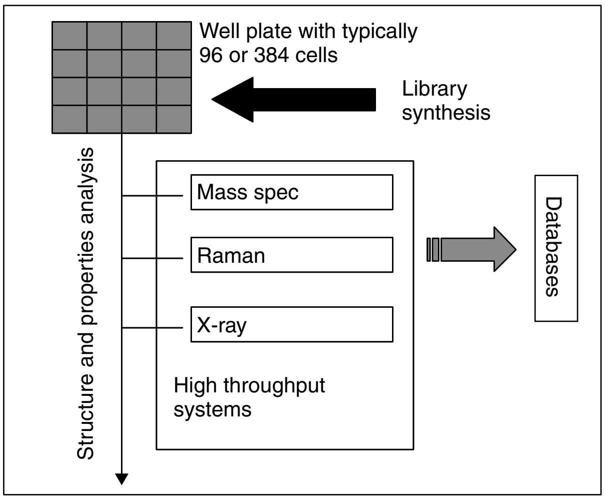
Well plate with typically 96 or 384 cells
Library synthesis
Structure and properties analysis
Mass spec
Raman
X-ray
High throughput systems
Databases

Figure 42.4 High throughput measurements are made on the combinatorial library, often while held in the same well plates used in the robotic driven synthesis. The original plates had 96 wells, now 384 is common with 1556 also being used. Very large quantities of data can be generated in this manner and will be held in associated databases. Electronic laboratory notebook systems correlate the resulting data libraries with the conditions and synthesis. Holding all this information distributed on the Grid ensures that the virtual record of all the data and metadata is available to any authorised users without geographical restriction.

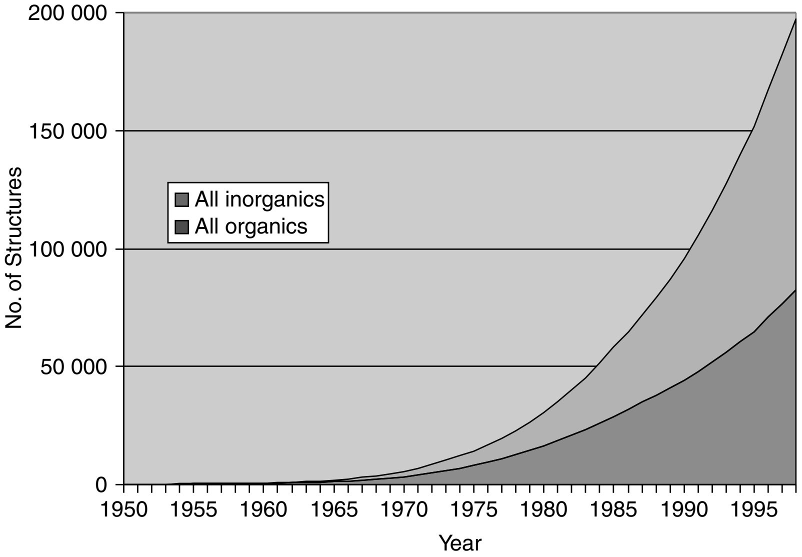

Figure 42.5 The number of X-ray crystal structures of small molecules in the Cambridge Crystallographic Data Centre database (which is one of the main examples of this type of structural databases) as a function of the year. Inorganic and organic represents two major ways in which chemists classify molecules. The rapid increase in numbers started before the high throughput techniques were available. The numbers can be expected to show an even more rapid rise in the near future. This will soon influence the way in which these types of databases are held, maintained and distributed, something with which the gene databases have already had to contend.

the incredibly rapid growth in the quantities of data on genetic sequences and thus by implication some knowledge of new proteins. The size and growth rates of the genetic databases are already almost legendary. In contrast, in the more mature subject of chemistry, the growth in the numbers of what we may call nonprotein, more typical 'small' molecules (not that they have to be that small) and materials has not had the same general impact. Nonetheless the issue is dramatic, in some ways more so, as much more detailed information can be obtained and held about these small molecules.

To give an example of the rapid rise in this 'Chemical' information Figure 42.5 shows the rise in the numbers of fully resolved X-ray structures held on the Cambridge Crystallographic Database (CCDC). The impact of combinatorial synthesis and high throughput crystallography has only just started to make an impact and so we expect even faster rise in the next few years.

42.4 CHEMICAL MARKUP LANGUAGE (cML)

In starting to set up the Comb-e-Chem project, we realized that it is essential to develop mechanisms for exchanging information. This is of course a common feature of all the e-Science projects, but the visual aspects of chemistry do lead to some extra difficulties. Many chemists have been attracted to the graphical interfaces available on computers (indeed this is one of the main reasons why many in the Chemistry community used Macs). The drag-and-drop, point-and-shoot techniques are easy and intuitive to use but present much more of a problem to automate than the simple command line program interface. Fortunately, these two streams of ideas are not impossible to integrate, but it does require a fundamental rethink on how to implement the distributed systems while still retaining (or perhaps providing) the usability required by a bench chemist.

One way in which we will ensure that the output of one machine or program can be fed in to the next program in the sequence is to ensure that all the output is wrapped with appropriate XML. In this we have some advantages as chemists, as Chemical Markup Language (cML) was one of the first (if not the first) of the XML systems to be developed (www.xml-cml.org) by Peter Murray-Rust [2].

Figure 42.6 illustrates this for a common situation in which information needs to be passed between a Quantum Mechanical (QM) calculation that has evaluated molecular properties [3] (e.g. in the author's particular laser research the molecular hyperpolarisibility) and a simulation programme to calculate the properties of a bulk system or interface (surface second harmonic generation to compare with experiments). It equally applies to the exchange between equipment and analysis. A typical chemical application would involve, for example, a search of structure databases for the details of small molecules,

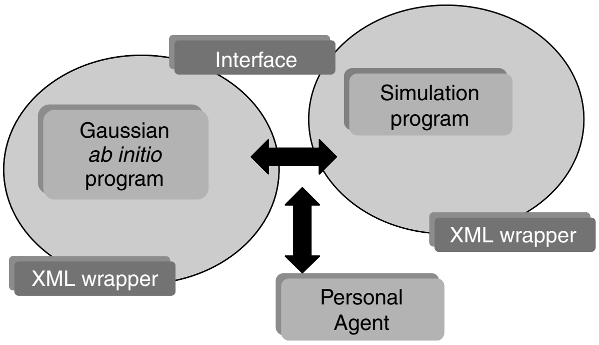

Figure 42.6 Showing the use of XML wrappers to facilitate the interaction between two typical chemistry calculation programs. The program on the left could be calculating a molecular property using an *ab initio* quantum mechanical package. The property could, for example, be the electric field surrounding the molecule, something that has a significant impact on the forces between molecules. The program on the right would be used to simulate a collection of these molecules employing classical mechanics and using the results of the molecular property calculations. The XML (perhaps cML and other schemas) ensures that a transparent, reusable and flexible workflow can be implemented. The resulting workflow system can then be applied to all the elements of a combinatorial library automatically. The problem with this approach is that additional information is frequently required as the sequence of connected programs is traversed. Currently, the expert user adds much of this information ('on the fly') but an Agent may be able to access the required information from other sources on the Grid further improving the automation.

followed by a simulation of the molecular properties of this molecule, then matching these results by further calculations against a protein binding target selected from the protein database and finally visualisation of the resulting matches. Currently, the transfer of data between the programs is accomplished by a combination of macros and Perl scripts each crafted for an individual case with little opportunity for intelligent reuse of scripts. This highlights the use of several large distributed databases and significant cluster computational resources. Proper analysis of this process and the implementation of a workflow will enable much better automation of the whole research process [4].

●Au: Please clarify if this reference is to Figure 42.7. If not, please provide the place of citation for Figure 42.7.

The example given in ● Figure 42.6, however, illustrates another issue; more information may be required by the second program than is available as output from the first. Extra knowledge (often experience) needs to be added. The Quantum program provides, for example, a molecular structure, but the simulation program requires a force field (describing the interactions between molecules). This could be simply a choice of one of the standard force fields available in the packages (but a choice nevertheless that must be made) or may be derived from additional calculations from the QM results. This is where the interaction between the 'Agent' and the workflow appears [5, 6].

●Au: We have changed the words 'In the limit' to 'Ultimately'. Please clarify if it is fine.

42.5 STATISTICS & DESIGN OF EXPERIMENTS

●Ultimately, the concept of combinatorial chemistry would lead to all the combinations forming a library to be made, or all the variations in conditions being applied to a screen. However, even with the developments in parallel methods the time required to carry out these steps will be prohibitive. Indeed, the raw materials required to accomplish all the synthesis can also rapidly become prohibitive. This is an example in which direction should be imposed on the basic combinatorial structure. The application of modern statistical approaches to 'design of experiments' can make a significant contribution to this process.

Our initial approach to this design process is to the screening of catalysts. In such experiments, the aim is to optimise the catalyst structure and the conditions of the reaction; these may involve temperatures, pressure, concentration, reaction time, solvent – even if only a few 'levels' (high, middle, low) are set for each parameter, this provides a huge parameter space to search even for one molecule and thus a vast space to screen for a library. Thus, despite the speed advantage of the parallel approach and even given the

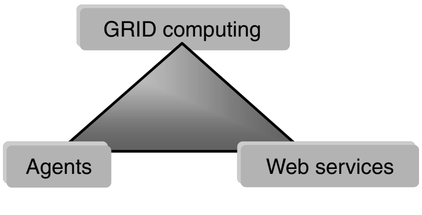

Figure 42.7 The Agent & Web services triangle view of the Grid world. This view encompasses most of the functionality needed for Comb-e-Chem while building on existing industrial based e-business ideas.

ability to store and process the resulting data, methods of trimming the exponentially large set of experiments is required.

The significant point of this underlying idea is that the interaction of the combinatorial experiments and the data/knowledge on the Grid should take place from the inception of the experiments and not just at the end of the experiment with the results. Furthermore, the interaction of the design and analysis should continue while the experiments are in progress. This links our ideas with some of those from RealityGrid in which the issue of experimental steering of computations is being addressed; in a sense the reverse of our desire for computational steering of experiments. This example shows how the combinatorial approach, perhaps suitably trimmed, can be used for process optimisation as well as for the identification of lead compounds.

42.6 STATISTICAL MODELS

The presence of a large amount of related data such as that obtained from the analysis of a combinatorial library suggests that it would be productive to build simplified statistical models to predict complex properties rapidly. A few extensive detailed calculations on some members of the library will be used to define the statistical approach, building models using, for example, appropriate regression algorithms or neural nets or genetic algorithms, that can then be applied rapidly to the very large datasets.

42.7 THE MULTIMEDIA NATURE OF CHEMISTRY INFORMATION

Chemistry is a multimedia subject – 3D structures are key to our understanding of the way in which molecules interact with each other. The historic presentation of results originally as text and then on a flat sheet of paper is too limiting for current research. 3D projectors are now available; dynamic images and movies are now required to portray adequately the chemist's view of the molecular world. This dramatically changes expectations of what a journal will provide and what is meant by 'publication'; much of this seems to be driven by the available technology–toys for the chemist. While there may be some justification for this view by early adopters, in reality the technology now available is only just beginning to provide for chemists the ability to disseminate the models they previously only held in the 'minds eye'.

●Au: There is no reference to Figure 42.8 in the text. Please clarify if this citation has been correctly placed.

Chemistry is becoming an information science [7], but exactly what information should be published? And by whom? The traditional summary of the research with all the important details (but these are not the same for all consumers of the information) will continue to provide a productive means of dissemination of chemical ideas. The databases and journal papers link to reference data provided by the authors and probably held at the journal site or a subject specific authority (●see Figure 42.8). Further links back to the original data take you to the author's laboratory records. The extent type of access available to such data will be dependent on the authors as will be the responsibility of archiving these data. There is thus inevitably a growing partnership between the traditional authorities in

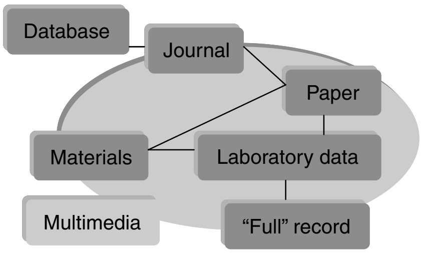

Figure 42.8 Publication @source: e-dissemination rather than simply e-publication of papers on a Web site. The databases and journal papers link to reference data provided by the authors and probably held at the journal site or a subject specific authority. Further links back to the original data take you to the author's laboratory records. The extent and type of access available to such data will be dependent on the authors as will be the responsibility of archiving these data.

publication and the people behind the source of the published information, in the actual publication process.

One of the most frustrating things is reading a paper and finding that the data you would like to use in your own analysis is in a figure so that you have to resort to scanning the image to obtain the numbers. Even if the paper is available as a pdf your problems are not much simpler. In many cases, the numeric data is already provided separately by a link to a database or other similar service (i.e. the crystallographic information provided by the •SIF data file). In a recent case of the 'publication' of the rice genome, the usual automatic access to this information to subscribers to the journal (i.e. relatively public access) was restricted to some extent by the agreement to place the sequence only on a company controlled Website.

•Au: Please spell out this abbreviation at the first instance.

In many cases, if the information required is not of the standard type anticipated by the author then the only way to request the information is to contact the author and hope they can still provide this in a computer readable form (assuming it was ever in this form?). We seek to formalise this process by extending the nature of publication to include these links back to information held in the originating laboratories. In principle, this should lead right back to the original records (spectra, laboratory notebooks •as is shown in Figure 42.9).

•Au: There is no reference to Figure 42.9 in the text. Please clarify if this citation has been correctly placed.

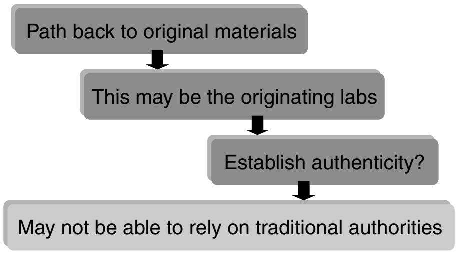
Path back to original materials
This may be the originating labs
Establish authenticity?
May not be able to rely on traditional authorities

Figure 42.9 What constitutes a trusted authority when publication @ source becomes increasingly important. Will adequate archives be kept? Will versioning be reliably supported? Can access be guaranteed?

It may be argued that for publicly funded research we have a responsibility to make all this information available. The immediate impact that many people may imagine on this is that it will make the detection of fraud much easier, but this is in fact a relatively minor issue. The main advantage will be the much greater use and reuse of the original data and the consequent checking of the data and different approaches to the analysis. The scientific process requires that as much as possible of the investigations are repeated and this applies just as much to the analysis as the initial data capture in the experiments.

42.8 THE PERVASIVE GRID AND METADATA

In my view if information (more valuable than just the data) is destined for the Grid, then this should be considered from the inception of the data. The relevant provenance, environment, and so on that traditionally characterise the metadata should not be added as an afterthought, captured separately in a laboratory notebook, but should be part of the data: it is this that differentiates information from mere data. If this background is not there from the beginning, how can we hope to propagate it efficiently over the Grid? This leads to the concept that the computing scaffold should be pervasive as well as forming a background transparent Grid infrastructure. The smart laboratory is another, and a highly challenging, environment in which to deploy pervasive computing [8].

Not only do we need to capture the data, the environment in which the data was generated but also the processes by which it was generated and analysed. In a research environment, the process capture is particularly interesting as the processes, for example, involved in analysing the data is not fully known in advance. Some basic aspects of the workflow will be predefined but it is usual to apply different methods (maybe different models or different computational methods) and the results compared (and compared with theory). We need to improve the tools to facilitate this process to capture and to represent it visually and to ensure that they can be used by a distributed research group/community to discuss the results and analysis. With a proper record of the process having been captured, the chosen route can then be quickly implemented for subsequent data.

The need to capture the experimental environment and the subsequent analysis process followed by integration with the existing knowledge shows the importance of the metadata. However, splitting the information into data and metadata may be counterproductive. The split between these two parts of the information is not well defined and depends on the perceived use of the information by the creator; which may, of course, be quite different from what actually happens. It is all too easy for the loss of the apparently useless metadata to render the data devoid of any real value.

More problematic is the subsequent need for information about the experiment not collected when the experiment was undertaken. This will always be a potential problem, as we cannot foresee all the parameters that may be significant. However, the smart environment will go a long way to help. The information will be automatically recorded and available if needed (and hidden if not to avoid confusion?). What is significant about the way we see the knowledge Grid is that this ‘metadata’ will always remain accessible from the foreground data even as the information propagates over the Grid. This is another part of the provenance of the information and something for which the current Web is not generally a good example.

Quoting one of the authors (Dave De Roure) '"Comb-e-Chem" is a *"real-time and pervasive semantic Grid", and as such provides a challenging environment in which to test many of the current ideas of the Human Computer Interface*'.

42.9 VIRTUAL DATA

A combinatorial library could be thought of as a 'Library Collection' with the material and information on that material all ideally cross-referenced. In a traditional system, if information is requested from the library collection then it can be provided if it is present in the collection. If the requested item it is not held in the collection, then a search can be made to find it elsewhere. This situation is paralleled with a molecular combinatorial library and applies not only to the physical material but also to the information held on the structure and properties of the molecules or materials in the library.

The power of the Grid-based approach to the handling of the combinatorial data is that we can go further than this 'static' approach. The combination of the laboratory equipment, the resulting information, together with the calculation resources of the Grid allows for a much more interesting system to be created. In the system outlined as an example described in Figure 42.10, an appropriate model can calculate the requested data. These models are themselves validated by comparison with the measured properties of the actual physical members of the library.

Depending on time or resource constraints, different types of models or different levels of implementation of a model can be chosen, ranging from resource hungry high-level QM calculation [9], through extensive simulations, to an empirically based approach; we thus have, in effect, a virtual entry in the database. Ultimately, this process has a close connection with the ideas of virtual screening of combinatorial libraries.

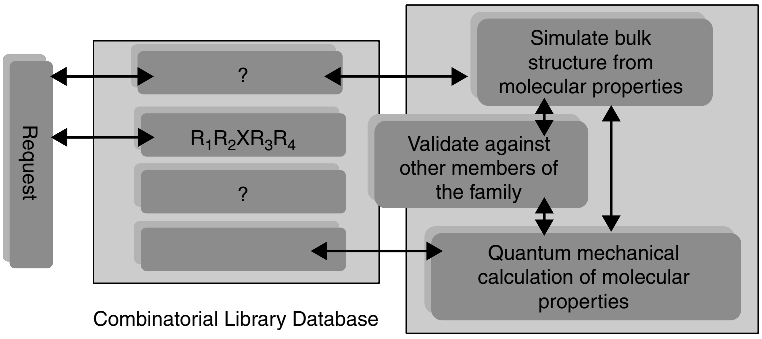
Request
?
$R_1R_2XR_3R_4$
?
Combinatorial Library Database
Simulate bulk structure from molecular properties
Validate against other members of the family
Quantum mechanical calculation of molecular properties

Figure 42.10 An example of virtual data interactions in the context of a combinatorial family. In this example, the combinatorial library is formed from the parallel synthesis of molecules with the form $R_1R_2XR_3R_4$ where X is a common building block and R_1 R_2 R_3 R_4 represent the variety of related groups that are added to this core to give the library. For example, they may be hydrocarbon chains of different length; CH_3-, CH_3CH_2-, $CH_3(CH_2)_n$-.

As in our model, the Grid extends down to the laboratory such that this virtual data idea can be extended to not only calculations but also to new experimental data acquisition or even to automated synthesis. That is, the direction of the synthesis or analysis in the automated laboratory would be controlled via a database request.

42.10 MULTIMEDIA COLLABORATION

A key issue for chemists making use of the Grid will be the support it can provide for distributed collaboration. This includes video, multimedia as well as the traditional need we have for visualisation. We have already demonstrated the need for significant, real-time, video interaction in the area of running a high throughput single crystal X-ray crystallography service. A demonstration Grid-aware system allowing users to interact with the UK Engineering and Physical Sciences Research Council (EPSRC) X-ray crystallography service bases at Southampton has highlighted a number of QoS and security issues that a Grid system must encompass if it is to provide an adequate infrastructure for this type of collaborative interactions. For example, the demands made on a firewall transmitting the video stream are very significant [10].

42.11 A GRID OR INTRA-GRIDS

It is possible that we may be able to enable query access to 'hidden databases' inside companies. In this model, certain types of general queries (e.g. checking structural motifs) could be asked without revealing the full nature of the compounds in the database. This could be useful, as we believe that this hidden information store exceeds considerably the publicly available information. Even with such a 'diode' placed in the dataflow to ensure reasonable isolation of company data, it seems likely that initially there will be intra-Grids in which large multinational companies use the Grid model we are proposing but restrict it to within the company (cf. intranets), so we will not initially have one Grid but many disjoint Grids.

There is also need for information to flow securely out of a company in support of equipment and other items that need to be serviced or run collaboratively with the manufacturers. The idea and indeed implementation of remote equipment diagnostics has been around for many years. In the computer industry, the remote support of PC is not uncommon – the supporting company can remotely control and 'fix' your PC (or router or other similar device). This has also been invaluable for training. •VNC provides this functionality in an open environment; one of our current projects is to marry this approach with the Web services model to provide more selective and secure interaction of this type. VNC was produced by Olivetti-AT&T labs (AT&T acquired the Olivetti Research Laboratory in 1999) and made open source [11].

•Au: Please clarify whether "VNC" can be spelt out as "Virtual Network Computing"

Discussions with some of the equipment manufacturers who have used remote diagnostics (usually via a dial-up) indicate that the interaction is insufficient even when the support team may have some physical presence at the site of the equipment and certainly is often inadequate to help the less well-trained users to fix the equipment. What is needed in addition to the direct connection to the machines is a connection with the users; this usually takes place by a concurrent telephone call.

An X-ray crystallography demonstration project based around the UK EPSRC National Crystallography Service (NCS), funded by the DTI e-Science core programme (M.S. Surridge & M.B. Hursthouse) has demonstrated how the use of audio and video over the Grid (despite the limitations of bandwidth, firewalls etc.) adds considerably to the quality of the interaction between users, experts, technicians and equipment. A full account of the security and associated issues uncovered by this demonstrator project will be the subject of a separate paper.

Taking the interactive approach to an experimental service used in the X-ray demonstrator project and linking them with the concerns of remote diagnostics, suggests that extensive remote monitoring of equipment will be possible over the Grid. This should allow pre-emptive maintenance, as frequent monitoring will be feasible. However, the remote diagnostics will often need to be augmented with the person-to-person multimedia links running synchronised with the control of and acquisition from equipment, for instruction and for any cases in which physical intervention is required by the user.

However, this discussion has taken the perspective of the researcher (or perhaps research manager) in the laboratory (or in the 'user' company). It must be married with the view from the other side (or of course as we are using the Grid, the other sides) of the interaction, the manufacturer support service. As already indicated, we believe the multimedia link directly benefits both sides of this interaction, and is already frequently undertaken by using a second separate communication channel (the phone). The use of parallel channels within the Grid is desirable as it allows for more efficient synchronisation of the information between the people and the equipment (or computers).

The Grid model allows for more. In the analysis of the combinatorial chemistry, structural and functional information models and calculations are used to link the data to extract information and knowledge (i.e. it is not just simply a mapping connection of data, though that is a significant activity, but the active development/extraction of new information).

Similarly, the diagnostic activates require comparison with a model that may be a physical model, that is, another copy of the equipment or, as the author suspects, because of the growing frequency of purpose designed individual systems, a physical device together with a computational model of the system. The integration and synchronisation of the computational model (which indeed may be viewed by some as essentially producing an Agent to help with the diagnostics) with the external machines is yet another area where the active properties of the Grid will be important.

The manner in which users, equipment, experts and servicing will be linked over the Grid will depend on which resources are most in demand and which are most limited in supply. The X-ray crystallography demonstrator is built around a very high-end diffractometer – there will not usually be many of these, coupled with experts in the collection and analysis of data from difficult crystal samples – people in even shorter supply. The connection model is thus that of Figure 42.11(a). An alternative model in which the equipment is relatively cheaper and easier to use, but nevertheless generated data that may require the help of an expert to understand fully, is shown in Figure 42.11(b). In both cases the Grid is enabling scarce resources to be shared while ensuring that all

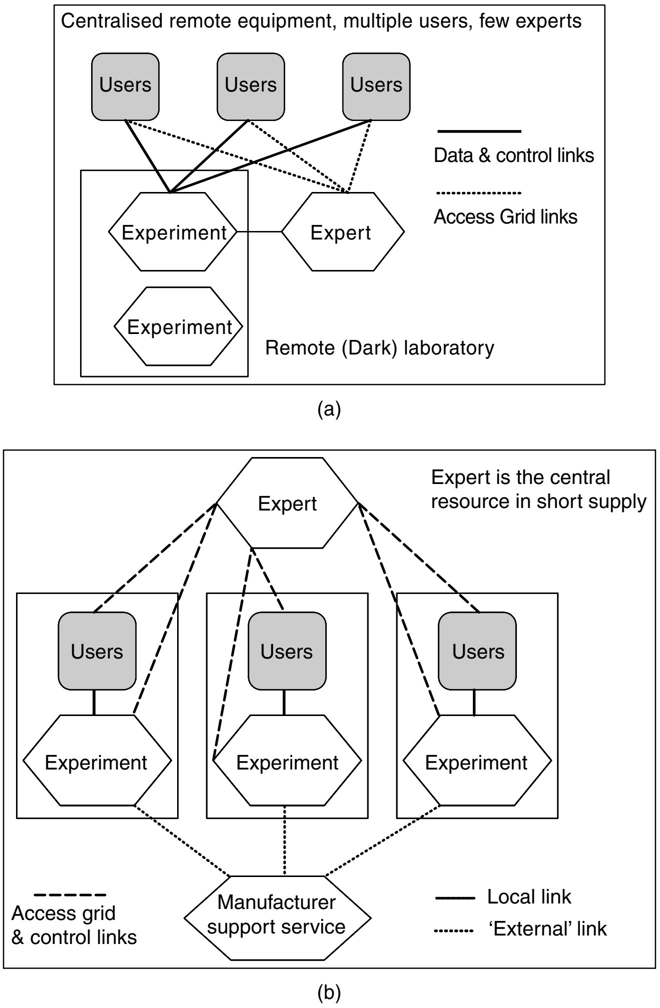

Figure 42.11 (a) shows a connection map for equipment providing a central service; (b) shows the situation for more commonly available equipment needing centralised support.

the 'stakeholders' in an experiment have the possibility of a presence in all stages of the procedures.

42.12 E-SCIENCE AND E-BUSINESS

I hope that this discussion has highlighted some of the specific aspects of chemistry research and, in particular, the way in which the application of combinatorial chemistry ideas together with the needs of collaborative research give rise to demands of the computational infrastructure that can be answered by the Grid. I hope it is clear that chemistry can use and catalyse many of the generic aspects of the knowledge Grid. The requirements of the research community are not that different from those carrying out e-Business. In particular, there is a common need for security (especially where patent-sensitive information is involved), authentication and provenance to ensure that the information can be trusted or at least investigated.

42.13 CONCLUSIONS

Chemistry in general and combinatorial chemistry in particular, will continue to make great demands on computational and network resources both for calculations and for knowledge management. The Grid will make an important impact in both these areas. The pervasive possibilities of the modern computing environment are ideal for extending the idea of a computational Grid down in the laboratory. The ability to automate both the experiments and the data analysis provides new possibilities and requirements for knowledge management. The exponentially growing quantities of data that combinatorial chemistry, in particular, is already delivering demonstrates the need for a Grid-based approach to handling the information generated [12]. The desirability of efficient use of resources (human, computational and equipment) in handling the resulting data is reflected in the need to employ statistical techniques both in the analysis of the large datasets and in the design of the experiments. The Grid will allow the desired close connection between the design, control and analysis of experiments (both physical and computational) to be implemented efficiently.

●Au: There is no reference to Figure 42.12 in the text. Please clarify if this citation has been correctly placed.

APPENDIX 1: THE COMB-e-CHEM e-SCIENCE PILOT PROJECT

Comb-e-Chem (www.combechem.org) is an interdisciplinary pilot project involving researchers in chemistry, mathematics and computer science (●see Figure 42.12) and is funded by the UK Engineering and Physical Science Research Council (www.epsrc.ac.uk) under the Office of Science and Technology e-Science initiative (www.research-councils.ac.uk/escience/).

A major aspect of the crystal structure measurement and modelling involved in this project will be to develop the e-Science techniques to improve our understanding of how molecular structure influences the crystal and material properties. The same molecule can crystallise in a number of different forms each with different physical and chemical properties; the same compound can frequently form many different materials – to take a very important example in food – chocolate can crystallise in six different forms but only one of them has the lustre and snap of good quality chocolate – the other forms are cheap and nasty. The same thing can happen with drugs, the wrong formulation can result in none of the drug being absorbed by the body. For one AIDS drug, the appearance of a new less soluble polymorph required $40 M to reformulate.

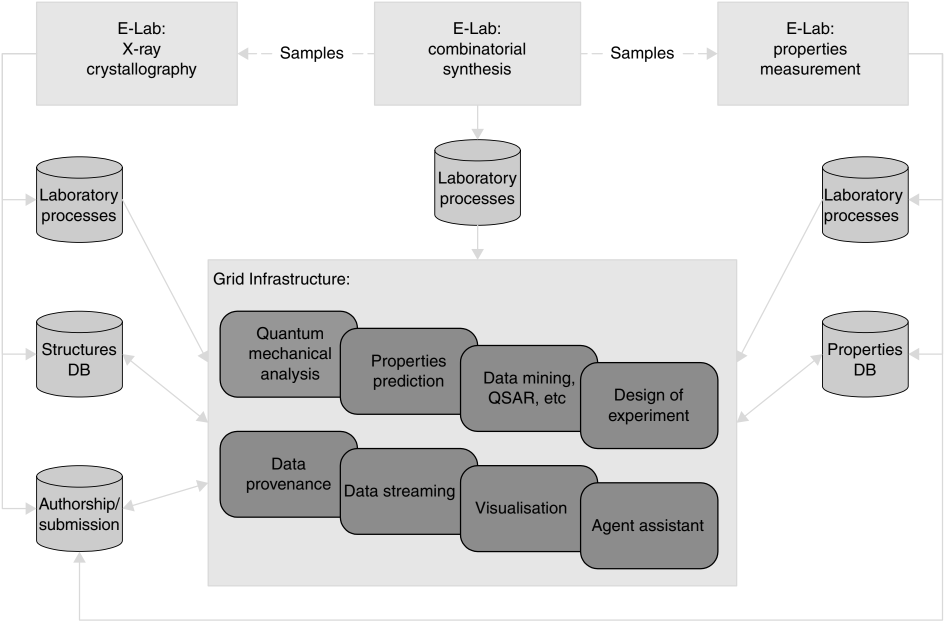

Figure 42.12 The Comb-*e*-Chem project development.

One of us (JGF) as Principle Investigator has the task of coordinating the other researchers who are in Chemistry (www.chemistry.soton.ac.uk) at the University of Southampton: Jeremy Frey, Mike Hursthouse, Jon Essex and Chris Frampton (formally at Roche); Chemistry at the University of Bristol (www.bris.ac.uk/Depts/Chemistry/Bristol_Chemistry.html), Guy Orpen; in Statistics at Southampton (www.maths.soton.ac.uk/stats/), Sue Lewis and Alan Welsh; in Electronics & Computer Science at Southampton, Dave De Roure, Mike Luck, and Luc Moreau; and Mike Surridge at IT-Innovation (www.it-innovation.soton.ac.uk). The project is in its early stages but already we have a very active team of researchers (a list can be found at www.combechem.org). We appreciate the significant help and assistance from our industrial collaborators, in particular, IBM UK at Hursley (www.hursley.ibm.com), The Cambridge Crystallographic Data Centre (CCDC www.ccdc.cam.ac.uk), Astra-Zeneka, and Pfizer).

REFERENCES

●Au: Please provide the volume number for this reference.

1. Scheemeyer, L. F. and van Dover, R. B. (2001) The combinatorial approach to materials chemistry, Chapter 10, in Keinan, E. and Schechter, I. (eds) *Chemistry for the 21st Century*. ISBN 2-527-30235-2, Weinheim, Germany: Wiley-VCH, pp. 151–174.
2. Murray-Rust, P. (1997) Chemical Markup language. *World Wide Web Journal*, ●135–147.
3. Crawford, T. D., Wesolowski, S. S., Valeev, E. F., King, R. A., Leininger, M. L. and Schaefer III, H. F. (2001) The past present and future of quantum Chemistry, Chapter 13, in Keinan, E.

and Schechter, I. (eds) *Chemistry for the 21st Century*. ISBN 2-527-30235-2, Weinheim, Germany: Wiley-VCH, pp. 219–246.

4. Leymann, F. and Roller, D. (1997) Workflow-based applications. *IBM Systems Journal*, **36** (1), 102–123.
5. Wooldridge, M. and Jennings, N. R. (1995) Intelligent agents: theory and practice. *The Knowledge Engineering Review*, **10**(2), 115–152.
6. Jennings, N. R. (2001) An agent-based approach for building complex software systems. *Communications of the ACM*, **44**(4), 35–41.
7. Lehn, J. M. 2001 Some reflections on chemistry, Chapter 1, in Keinan, E. and Schechter, I. (eds) *Chemistry for the 21st Century*. ISBN 2-527-30235-2, Weinheim, Germany: Wiley-VCH, pp. 1–7.
8. De Roure, D., Jennings, N., Shadbolt, N. (2001) *Research Agenda for the Semantic Grid: A Future e-Science Infrastructure*, Technical report UKeS-2002-02, •UK National e-Science Centre, November, 2001..
9. Q. Alchemy, M. L. Cohen, (2001), Chapter 14, in Keinan, E. and Schechter, I. (eds) *Chemistry for the 21st Century*. ISBN 2-527-30235-2., Weinheim, Germany: Wiley-VCH, pp. 247–270.
10. Buckingham Shum, S., De Roure, D., Eisenstadt, M., Shadbolt, N. and Tate, A. (2002) CoAKTinG: collaborative advanced knowledge technologies in the grid. *Proceedings of the Second Workshop on Advanced Collaborative Environments, Eleventh IEEE Int. Symposium on High Performance Distributed Computing (HPDC-11)*, Edinburgh, Scotland, July 24–26, 2002.
11. Richardson, T., Stafford-Fraser, Q., Wood, K. R. and Hopper, A. (1998) Virtual network computing. *IEEE Internet Computing*, **2**(1), 33–38.
12. Maes, P. (1994) Agents that reduce work and information overload. *Communications of the ACM*, **37**(7:31), 40.

•Au: Please provide the place of publication for this reference.