Increasing the Scope for Polymorph Prediction using e-Science

H. Nowell,[a] B. Butchart,[b] D. S. Coombes,[c] S. L. Price,[a] W. Emmerich,[b] C. R. A. Catlow[c]

a. Dept. of Chemistry, University College London, 20 Gordon Street, London, WC1H 0AJ, UK
b. Dept. of Computing Science, University College London, Gower Street, London, WC1E 6BT, UK
c. Davy Faraday Research Laboratory, The Royal Institution of Great Britain, 21 Albemarle Street, London, W1S 4BS, UK

Abstract

This poster demonstrates how e-Science can accelerate and facilitate scientific discovery in developing computational methods of predicting the possible crystal structures of organic molecules. We will show how re-engineering an existing scientific process using state-of-the-art distributed computing technologies will not only improve the speed of the scientific process but also make it possible to extend the methodology to pharmaceuticals so the computations can aid drug development. The polymorph prediction of a moderately challenging molecule, and subsequent morphology calculations, are used to illustrate the new technologies.

Introduction

e-Science is being used to extend a crystal structure prediction process, previously limited to small organic molecules, to molecules more typical of commercially important molecular materials. The new distributed computing application incorporates several distinct types of calculation needed to predict the possible crystal structures of an organic molecule.

Crystal structure prediction is potentially of great value for the pharmaceutical industry[1] to warn of the possible appearance of new crystalline forms of a pharmaceutical during scale-up, production or storage, and in patent protection. A pharmaceutical can only be marketed in the licensed crystal form, as different crystal structures (polymorphs) of a drug are likely to have different properties such as bioavailability, solubility and morphology (shape). Pharmaceutical molecules are generally flexible; the complexity of polymorph prediction[2] is thus increased because of the need not only to search through the huge range of possible crystal packings but also to consider the range of energetically plausible conformers (different shapes of the molecule). The international ‘Blind Tests’ of crystal structure prediction [3,4], organised by the Cambridge Crystallographic Data Centre, provide a measure of the progress in crystal structure prediction technology and methodology. In the recent test[5] there were no successful predictions for the flexible molecule.

Crystal Structure Prediction Method

Our method for predicting polymorphs involves a number of programs that have traditionally been run sequentially with manual editing of input and output files. For a flexible molecule it is necessary to perform thorough conformational analysis to produce a series of energetically plausible conformers. Using this manual method, a polymorph prediction study typically takes several months of work for each flexible molecule studied.

Each conformer is treated as a rigid probe in independent searches. For a molecule with a wide range of plausible conformers (with a potential energy surface that is flat or has a number of minima) the number of searches that is required increases from one (where the molecule is rigid) to many; enough such that every low energy grid point is used for a search.

Until now we have not had the capability of establishing what range of conformers, both in terms of energy and structural variation, need to be considered for a successful search. Recent experience with the ‘Blind Test’ has shown that even conformers that deviate considerably from the gas phase molecular structure must be considered to maximise the chances of locating observed polymorphs. This molecule was found to have 6 flexible torsion angles and so considering ~50,000 crystal structures and eight conformers was insufficient to locate even a qualitatively correct crystal structure. It is therefore necessary to perform as many searches using as wide a range of different conformers as reasonably possible with the time and resources available. This would be made far more effective if the scientist can be guided by the initial searches to decide which conformers to use for subsequent searches.

The search is implemented in the program MOLPAK[6]. Crystallographic relations are used in the search for crystal structures. At present 13 space groups, represented by 29 of the most common packing types (as identified from the Cambridge Crystallographic Database[7]) may be searched routinely. Up to 200 densely packed crystal structures are found for each packing type and each is input to DMAREL[8] for lattice energy minimisation and calculation of properties.

Many of the valid crystal structures will represent the same minimum, so the post search analysis must begin with removal of equivalent structures. The remaining unique structures are sorted in terms of energy (lattice energy plus a measure of the intramolecular energy, $\Delta E_{\mathrm{intra}}$, of the specific conformer) and property calculations are performed to determine which structures are more likely to be observed experimentally.

One property that affects the manufacture of organic materials is the crystal shape (morphology). The shape of a drug crystal can influence its dissolution rate, which is dependent on surface area, and hence the effective dose. Thus morphology predictions play a role in product development. Morphology predictions can also be used to indicate whether hypothetical crystal structures, found in the crystal structure prediction process, are likely to be observable polymorphs with advantageous properties[9].

In this work we calculate the attachment energy, i.e. the energy released when a stoichiometric layer of material is placed onto a surface, to predict the morphology of sets of observed and hypothetical crystal structures. This model assumes that the growth rate of a face is proportional to the absolute value of the attachment energy. The morphology can be visualised using a Wulff plot in which the distance from the origin to the (h,k,l) face, R_{hkl} is proportional to the magnitude of the attachment energy (which is negative), i.e. R_{hkl} |Eatt(h,k,l)|. Hence, we can compare both the growth rates of the slowest growing faces and the relative volume growth rates, by calculating the volume enclosed within the Wulff shape.

Using a Grid Infrastructure

Our approach to distributing the polymorph prediction process on grid resources is to wrap the MOLPAK and DMAREL Fortran codes as a set of loosely coupled Web Services. We have developed a prototype workflow enactment engine to orchestrate interactions between instances of these services using the Business Process Execution Language (BPEL). Isolating the computational workflow in a separate component gives users more flexibility to optimise the grid infrastructure to meet their needs and enables them to steer the scientific process in an interactive manner. Through recursive composition, BPEL also offers a convenient way to spread the activities in a workflow across multiple clusters facilitating load balancing and improving availability.

The BPEL engine is integrated with a server running the SUN Grid Engine (SGE) that acts a resource broker between the BPEL process and its partner services. Fig. 1 shows the architecture of these components.

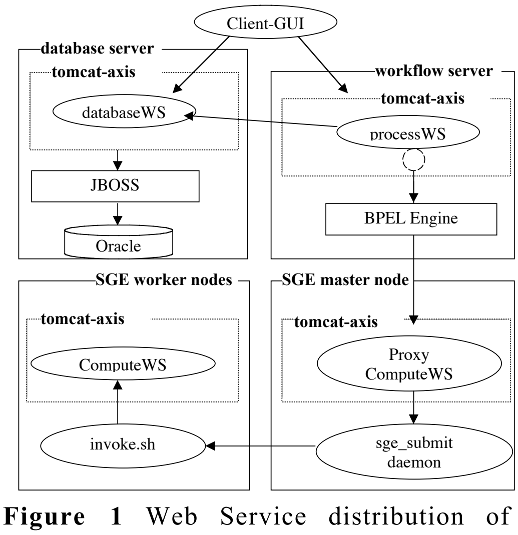

Figure 1 Web Service distribution of polymorph prediction process on a grid.

The client creates an initial BPEL process definition, which it sends along with the initial parameters of the polymorph search to the process Web Service running on the workflow server. The workflow server redirects the message to the BPEL engine, which subsequently starts to orchestrate proxy compute services hosted on the SGE_master node and the database server. The database server provides access to search metadata and stores results. The implementation interacts with Enterprise Java Beans deployed in the JBOSS Application Server with Oracle providing backend persistence.

The SGE_master node hosts Web Services that act as proxies for the compute services on the cluster worker nodes, which themselves are not directly accessible to the BPEL engine. Instead of running the Fortran binaries the proxy services persist the incoming SOAP message and make calls to the SGE submit_daemon, causing SGE to invoke a shell script on one of its worker nodes. Taking the persisted SOAP message as an argument, this script starts up Tomcat (with Axis) then invokes the Web Service on this local server instance. It is this instance that executes the FORTRAN binary eventually returning any output message to the proxy. XSLT scripts and purpose built parsers are employed to convert chemical data represented in CML in SOAP messages to and from FORTRAN input/output files. While it may seem wasteful to start and stop a new Tomcat instance for each invocation this is necessary to ensure the work done by the web services is accounted for properly by the SGE resource manager.

We have learnt of number of valuable lessons developing and using this infrastructure. Initially, we failed to recognise the importance of a high quality job scheduler to performance and reliability of the system. Since no existing middleware for scheduling invocations to Web Services was available, we built our own primitive component for obtaining references to available service instances. We discovered that attempting to control access to resources with standard Java synchronization primitives broke under high loads. More seriously, there was no way to prevent other users of the cluster running jobs on the same nodes, a situation that generally resulted in exhaustion of memory resources. Only when we redeployed onto a production environment, with Web Service invocation integrated with the SGE scheduler, did we make any progress with high loads. With the scheduling issue resolved, problems with other components emerged. While we benefited greatly from reuse of distributed open-source technologies such as Apache Tomcat, XML processors for handling CML, JBOSS and the Axis SOAP engine, we seriously underestimated the difficulty configuring these components to work together. We also encountered defects and memory leaks in some components when subjected to high loads.

In our first prototype design[10] we made extensive use of the GT3.0 OGSI implementation. We found that GT3 performed particularly badly under heavy load with client side message handlers intermittently mangling SOAP messages and the container rapidly exceeding memory limits when new service instances were created. The integration with SGE scheduling rendered the stateful nature of OGSI services redundant and we realized that only thread safe, stateless service implementations were likely to succeed under high load conditions. We are now phasing out OGSI and replacing these codes with standard Web Services deployed on Axis.

Test Case

A typical example of a moderately flexible molecule is the well-characterised nootropic drug, piracetam ($C_6H_{10}N_2O_2$), Fig. 2. Its conformation may be described in terms of three torsion angles. There are three reported polymorphs of piracetam[11,12,13] (denoted forms I, II and III), whose packing motifs are shown in Fig. 3, they contain essentially two molecular conformations; neither of which correspond to the predicted gas phase molecular conformer.

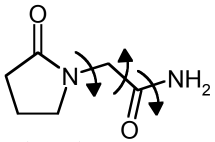

Figure 2 The piracetam molecule, arrows indicate flexible torsion angles.

Preliminary calculations indicate that it is unlikely for the known polymorphs to be located during a search using a gas phase optimised molecular structure; it is likely to be necessary to constrain some torsion angles to obtain a rigid conformer that can be used in a search to locate known structures. The predicted crystal morphologies for the three known crystal forms of piracetam are shown in Fig. 4.

The potential for the grid to allow simultaneous exploration using a number of conformers not only increases the coverage of search space, but also allows the scientist to develop a strategy for considering the conformers; a conformer with a relatively large ΔE_{intra} (therefore considered unlikely at first glance) may in fact have geometry that allows the formation of a specific intermolecular interaction that gives rise to a very favourable lattice energy. The scientist may therefore, on the basis of these intermediate results, choose to perform further searches with similar conformers.

The ability to perform more searches, quickly, enables a more thorough investigation of the search space and therefore increases the chance of establishing whether a different conformation could result in a more thermodynamically stable crystal structure than the known crystal. This type of prediction could have averted the crisis in the supply of the anti-HIV drug ritonavir[1].

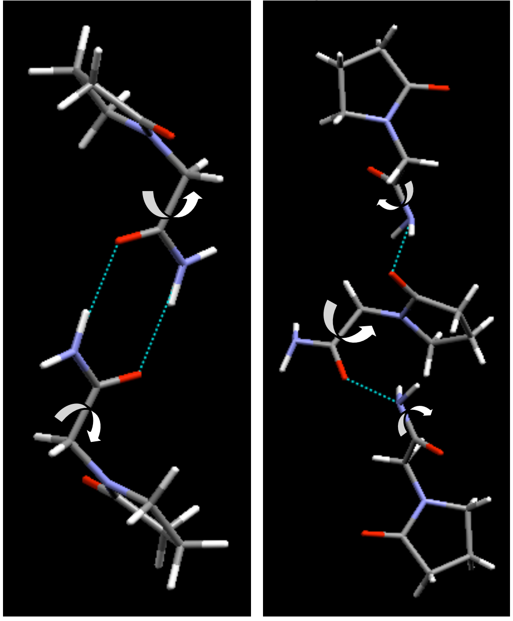

Figure 3 Above left is the hydrogen bonded dimer motif observed in piracetam forms II and III. No such dimers are observed in form I; hydrogen bonded chains, above right, are observed. The conformation of piracetam in form I is different to that in forms II and III; the white arrows indicate one torsion angle that differs between the conformers.

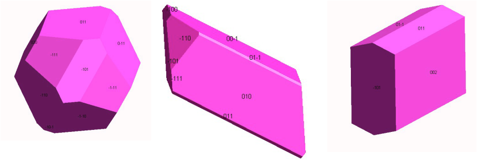

Figure 4 Left to right; predicted morphologies (Wulff plots) for piracetam forms I, II and III.

A typical preliminary search of the most common packing types takes less than 10 minutes, whereas a full search can be run in under two hours on about 200 commodity CPUs. The application may also increase the degree of flexibility that is deemed reasonable to approach for the purpose of polymorph prediction and may therefore widen the range of molecules that may be considered in a computational polymorph study.

Concluding Remarks

New technologies are being applied to an existing process for polymorph prediction. It is hoped that faster searches will make it more feasible to attempt crystal structure prediction for pharmaceuticals, where numerous searches using different low energy conformers are required. Indeed, responsive searching (the choice of which conformer to use in the next search based on previous search results) may become routine in a way that has previously not been possible.

Future work will focus on extending the existing workflow to handle new services, including morphology calculations on thermodynamically plausible crystal structures and integration with a data portal that has been established at CCLRC in Daresbury. This is the first step for building a database of computed and known crystal structures and properties for eventual data-mining to develop the science of polymorph prediction, using data generated by a Basic Technology Program project on "The Control and Prediction of the Organic-Solid-State".

Acknowledgements

This work was funded by the project "E-Science Technologies in the Simulation of Complex Materials".

References

1. S.L. Price, *Adv. Drug Delivery Rev.*, **56**, 301 (2004).
2. C. Ouvrard & S.L. Price, *Cryst. Growth Des.* submitted (2004).
3. J. P. M. Lommerse *et al.*, *Acta Crystallogr..*, **B58**, 697 (2000).
4. W. D. S. Motherwell *et al., Acta Crystallogr.*, **B58**, 647 (2002).
5. G. M. Day *et al*, in preparation.
6. J. R. Holden *et al.*, *J. Comput. Chem.*, **14**, 422 (1993).
7. F. H. Allen & O. Kennard, *Chem. Des. Autom. News*, **8**, 31 (1993).
8. D. J. Willock *et al.*, *J. Comput. Chem.*, **16**, 628 (1995).
9. D. S. Coombes *et al.*, *Cryst. Growth Des.*, in preparation.
10. B. Butchart *et al.*, OGSA First impressions - A case study, In. Proc All Hands Meeting 2003,.
11. G. Admiraal *et al.*, *Acta Crystallogr.*, **B38**, 2600 (1982).
12. D. Louër *et al.*, *Acta Crystallogr.*, **B51**, 182 (1995).
13. R. Céolin *et al.* , *J. Solid State Chem.*, **122**, 186 (1996).