Cryptic crystallography

GAUTAM R. DESIRAJU is at the School of Chemistry, University of Hyderabad, Hyderabad 500 046, India.
e-mail: desiraju@uohyd.ernet.in

A method for predicting crystal structures from just molecular formulae has eluded scientists for more than 50 years. The problem is currently being addressed by two very different approaches. But which one is more likely to succeed?

A molecule can be defined, in strictly geometrical terms, as a group of atoms in which the distance from any one atom to at least one other in the group is much smaller than any distance between atoms in different groups. Such a definition was given nearly 50 years ago by the great Russian crystallographer Alexander I. Kitaigorodskii[1,2]. Interactions between atoms within a molecule are therefore much stronger, by almost two orders of magnitude, than those between atoms from different molecules. This is why we have a far less precise idea about the ways in which molecules assemble in condensed media than we have about the ways in which atoms bond within molecules. Trying to establish *a priori* exactly how molecules come together in crystals is in fact a very difficult problem, and is known by the name 'crystal structure prediction' (CSP). CSP is a problem of formidable proportions because the solution requires a complete understanding of the mechanism for crystallization — the ultimate goal of solid-state supramolecular chemistry.

The problem is stated easily enough. Given the structural formula of a small organic molecule with fewer than, say, 20 non-hydrogen atoms (A), and with no more than two or three conformationally flexible carbon–carbon or carbon–A bonds, is it possible to predict its crystal structure with the kind of accuracy that is obtained with an X-ray diffractometer? In other words, we are looking to predict the size of the crystal unit cell, correct to about two decimal places, the space group symmetry, and the positions of all the atoms in the asymmetric unit. Research groups from all over the world have taken up this challenge and much of their effort has been streamlined in the form of two blind tests conducted by the Cambridge Crystallographic Data Centre (CCDC) in 1999 and 2001. The crystal structures of three molecules were determined, and the results held in confidence by an independent referee. Some 18 groups were invited to submit up to three ranked predictions for each molecule within four months. To keep things 'easy', the crystal structures were restricted to the ten most common space groups, to have only a single molecule in the asymmetric unit, and to be ordered and unsolvated. Yet, out of a total of nearly 75 predictions made in the two tests combined, the number of correct predictions was fewer than ten. No research group predicted all three structures correctly in either of the two tests[3,4].

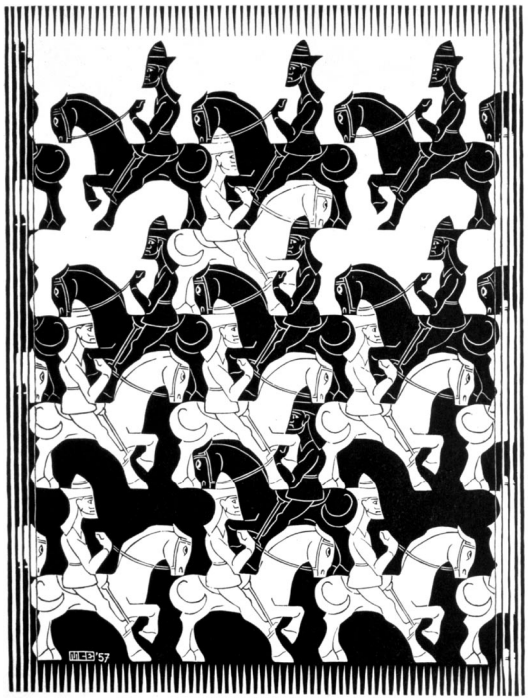

Figure 1 **Two-dimensional closest packing as seen in M. C. Escher's 'Horsemen'. For molecules to come together in this way, all portions of their surfaces must be equally 'sticky'. The presence of a few 'stickier' regions leads to distortions from ideal close packing, and eventually to a breakdown of the geometrical model.**

Back in 1988, John Maddox (former editor of *Nature*) pointed to the need for CSP in materials research[5]. He wrote that it was "one of the continuing scandals" that a general method for the prediction of crystal structures from molecular structures was not yet available. This provocative piece was much quoted in articles devoted to the emerging subject of crystal engineering — the dream of designing organic solids with specific and desired properties[6]. Later, in 1996, Philip Ball wrote that "a large part of the scandal remains"[7]. Maddox was hopeful at the time about predicting structures for extended inorganic solids,

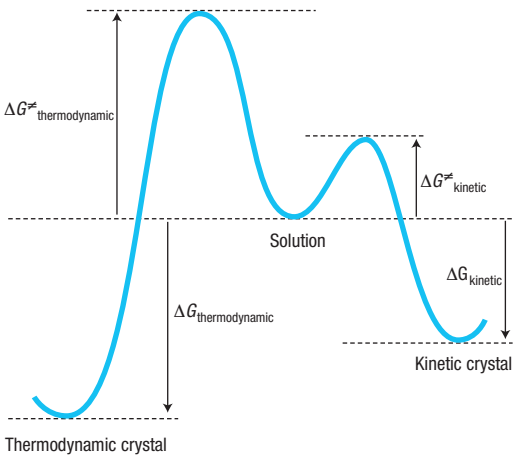

Figure 2 **The Curtin–Hammett principle: the most stable or the fastest? The route to the kinetically favoured crystal would be the fastest, because the activation energy (*G*) barrier to that state is lower ($\Delta G^{\neq}_{\text{kinetic}}$). The thermodynamically favoured crystal would take longer to form because the activation barrier is much higher ($\Delta G^{\neq}_{\text{thermodynamic}}$), but it would be more stable because the final energy state is the lowest. If the same crystalline form is both kinetically and thermodynamically favoured, polymorphism is highly unlikely.**

but for organic molecular solids one can say that, 15 years on, the situation has not changed greatly[8]. Because we still lack a general and accurate method for CSP, there is no known example of a designed crystalline molecular material in wide technological use. Perhaps it was not appreciated how difficult CSP really is.

The original, and by far the more popular, approach to CSP is based on geometry and goes back to Kitaigorodskii[1,2]. Interactions between molecules are assumed to be very weak and lacking in directionality; it is further assumed that all interactions taper off at longer distances in roughly the same way. In this isotropic model, crystal structures are governed by close packing. The structure that makes the most economical use of space is the best one, and molecules crystallize so that the bumps in the surface of one fit into the hollows in the surface of the other (Fig. 1). Using this model as a starting point, computational techniques can generate several hypothetical crystal structures that approximately satisfy these close-packing conditions[9]. These calculations generate many putative structures, maybe up to 20, within 2 kJ mol^{-1} of the lowest energy structure, known as the 'global minimum'[10]. The number of structures generated is large because the intermolecular interactions are weak.

Of course, most molecules do not live in these idealized isotropic conditions, and so the proponents of this school of thought introduce various modifications into their calculations to take into account the electrostatic and directional effects that are unquestionably present in organic crystals. Hydrogen bonding, being the prime example of such effects, is very important in crystal structures. Molecules that can hydrogen bond always do so, and this needs to be taken into account in CSP. Some of the resulting computations are quite sophisticated, but in the end, the close-packing approach has two, seemingly insurmountable, problems.

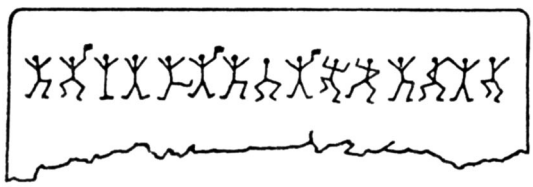

Figure 3 **Cryptology and crystallization: hidden but definite. Crystallography needs its own Sherlock Holmes to crack the secrets of molecules as they assemble into crystals.**

"Dancing Men" sourced from *The Complete Adventures of Sherlock Holmes* by Sir Arthur Conan Doyle. Published by Martin Secker and Warburg Ltd, London, 1981.

First, the fluctuations and variations in the interactions are too numerous to handle properly, and depend on molecular structural features that may be quite remote from the site of the interactions themselves[11]. Second, the global minimum structure is not the observed structure because kinetic effects dominate. In other words, one cannot be sure of getting the best structure; usually it is the one that can be obtained the most quickly.

All chemical reactions are subject to the dictates of thermodynamics and kinetics — how far and how fast. Most chemical reactions that require the making and breaking of covalent bonds are kinetically controlled. On the other hand, supramolecular reactions that take place in solution are thermodynamically controlled and take place under conditions close to equilibrium[12]. But crystallization is different from other supramolecular reactions in that it is kinetically controlled. Like other kinetic processes, the outcome of crystallization depends a great deal on experimental variables such as temperature, solvent, rates of heating and cooling, impurities and shock. A particular crystal form may not be the global minimum in terms of free energy, but it may be the most common outcome because it is kinetically dictated by the reaction conditions.

The appearance of these 'local minimum' structures during crystallization makes a mockery of methods that look only for the global minimum. In the 1950s, David Y. Curtin and Louis P. Hammett showed that kinetic products may dominate, or even be formed exclusively in chemical reactions[13]. The Curtin–Hammett principle states that the distribution of products in a reaction that has many pathways need bear no relation to the relative stability of those products. Exactly the same rule holds for crystallization. Given a collection of molecules that can come together in many ways, the favoured route has little to do with the stability of the final ensemble, but rather with how fast this route can be travelled (Fig. 2). But if kinetics and not thermodynamics dominates, then a full dynamic treatment is required in CSP. This is beyond our most fanciful dreams. State-of-the-art dynamic simulations are now in the millisecond range, whereas crystal growth may take hours or even days. The maximum cluster size handled by the calculations may be around 10,000 atoms, of which 50% are surface atoms, and so the objects being modelled will have little in common with real crystals. In short, the events that take place during crystallization cannot be modelled computationally.

If *ab initio* prediction of kinetically controlled crystal forms is impossible, what is the solution? It is a basic rule of cryptology that even the most difficult of codes can be broken if a sufficient number of examples is available. Sherlock Holmes was able to decipher the sinister meaning of the messages in the *The Adventure of the Dancing Men*, because he started with the correct assumption that the figures occurring the most often in the bizarre messages corresponded to the letters 'S' and 'E', which occur the most frequently in the English language (Fig. 3). The same fingerprinting approach may be adopted in CSP.

Consider for example, molecule **I**, which was used in the 2001 CCDC blind test (Fig. 4). This molecule can, in principle, assemble into crystals in one of two ways through what is called the dimer synthon, **II**, or

the catemer synthon, **III**. The term 'synthon' or 'supramolecular synthon' refers to small, stable structural units that are readily accessed during crystallization[14]. The dimer is more stable than the catemer, but a crystal that contains the catemer grows more quickly. This is because the catemer can propagate by generating hydrogen bonds, which are strong and directional. In contrast, the dimer can associate with other dimers only through weak van der Waals interactions.

If we now check the Cambridge Structural Database, a depository of more than 250,000 accurately determined small molecule crystal structures (http://www.ccdc.cam.ac.uk/prods/csd/csd.html), we will find that a majority of molecules that are similar to molecule **I** give catemers rather than dimers. Coupling this information with computations that are generated using the close-packing approach (and that predict a crystal structure based on a dimer) results in a new ranking of the structures from which the experimental catemer structure hopefully emerges as the winner[15]. In the search for kinetically controlled crystal structures, such a method is on a surer footing than mere computations because it relies in part on experimental data. If a particular kind of structure has been formed often enough in the past, it is more likely that it will reappear.

But many questions remain. What is meant by molecular similarity? Is the identification of similar molecules a subjective matter? What parts of a molecule are the most critical in terms of defining the crystal structure that is ultimately formed? How large a database of crystal structures is required before most molecular recognition situations are covered? And most critically, what happens when more than one crystal form occurs experimentally?

The question of polymorphism — when more than one crystalline form exists — takes the difficult problem of CSP to an even higher level. Crystallization is a highly specific event, so how do alternative pathways become viable[16]? We know they do because for some categories of molecules, polymorphism is not rare. About 10% of organic substances may yield polymorphs easily, with another 20–30% possibly occurring if more exotic experimental conditions are used (such as high and low temperatures and pressures, hydrothermal methods, laser irradiation, annealing, supercritical liquids and unusual solvents). A plausible scenario is one in which either or both of the kinetic and thermodynamic forms are obtained during recrystallization. Many commercial substances, such as drugs or dyestuffs, produce many polymorphs, making this issue of critical importance to industry.

So CSP is not only a great scientific challenge, but has many implications in the chemical industry in areas such as catalysis, pharmaceuticals and separation. A reliable and general method for the CSP of polymorphic forms of a drug would transform the multibillion-dollar pharmaceutical industry. Perhaps, and given the great fundamental and practical importance of CSP, a good case can be made for a much bigger coordinated international effort to solve this problem, bringing CSP into the league of protein folding and the Human Genome project. In the context of CSP of polymorphs, pertinent questions are: (1) Is it possible to determine from the molecular structure whether a compound will be polymorphic? (2) If so, how many polymorphs will it have? (3) Under what conditions will these polymorphs be obtained? (4) Can one predict from the molecular structure whether a compound will crystallize with the solvent of crystallization, especially water? (5) Under what conditions will these pseudopolymorphs, or hydrates, be obtained?

These are very hard questions, but then the whole issue of CSP is full of extreme difficulty: the interactions between molecules are numerous, weak and variable; the supramolecular behaviour of a functional group acutely depends on the nature and even the locations of other functionalities in the molecule; there are many closely packed, low-energy structures; one can never really tell in advance whether polymorphism will occur or not; kinetics competes with thermodynamics; and one has no way of knowing from the molecular structure whether the kinetic structure would also be favoured thermodynamically. Research problems in CSP and crystal engineering are like the Himalayan Mountains. One tackles them because they are there, and one doesn't give up. But then, one comes face to face not only with Everest, but also with killers like Makalu, Lhotse and Annapurna — and in those magnificent ranges, peaks less than 20,000 feet are not even named.

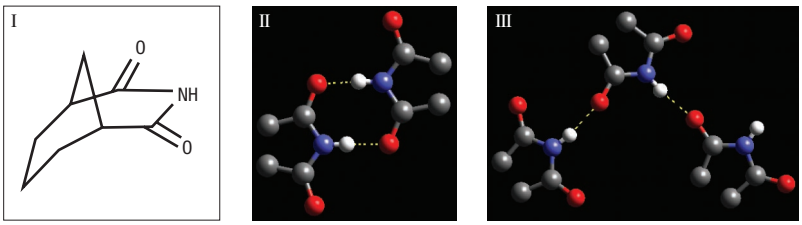

Figure 4 **A simple imide molecule in the 2001 CCDC blind test (I), and two alternative supramolecular synthons (II and III) that it can form. Synthon III was found in the experimental structure.**

References

1. Kitaigorodskii, A. I. *Organic Crystal Chemistry* (Izd. Akad. Nauk SSSR, Moscow, 1955).
2. Kitaigorodskii, A. I. *Molecular Crystals and Molecules* (Academic, New York, 1973).
3. Lommerse, J. P. M. *et al. Acta Crystallogr. B* **56**, 697–714 (2000).
4. Motherwell, W. D. S. *et al. Acta Crystallogr. B* **58**, 647–661 (2002).
5. Maddox, J. *Nature* **335**, 201 (1988).
6. Desiraju, G. R. *Crystal Engineering. The Design of Organic Solids* (Elsevier, Amsterdam, 1989).
7. Ball, P. *Nature* **381**, 648–650 (1996).
8. Beyer, T., Lewis, T. & Price, S. L. *Cryst. Eng. Comm.* **44**, 1–35 (2001).
9. Gavezzotti, A. *Synlett.* 201–214 (2002).
10. Karfunkel, H. R. & Gdanitz, R. J. *J. Comput. Chem.* **13**, 1171–1183 (1992).
11. Desiraju, G. R. *Nature* **412**, 397–400 (2001).
12. Rowan, S. J., Cantrill, S. J., Cousins, G. R. L., Sanders, J. K. M. & Stoddart, J. F. *Angew. Chem. Int. Edn Engl.* **41**, 898–952 (2002).
13. Curtin, D. Y. *Rec. Chem. Prog.* **15**, 111–128 (1954).
14. Desiraju, G. R. *Angew. Chem. Int. Edn Engl.* **34**, 2311–2327 (1995).
15. Sarma, J. A. R. P. & Desiraju, G. R. *Cryst. Growth Des.* **2**, 93–100 (2002).
16. Desiraju, G. R. *Science* **278**, 404–405 (1997).