R.J. DAVEY

Figure 1 **Hey, brown sugar: Cacciuto *et al.*[3] offer new insight into the seeding of crystallization.**

that it can exist in two forms, or enantiomers, that are non-superimposable mirror images of each other. In 1866, Desiré Gernez wrote to his former colleague Pasteur, describing the result of an interesting experiment[4]. Following on from Pasteur's work, Gernez had discovered that the addition of seed crystals of pure enantiomer to a racemic solution of the tartrate — one containing equal amounts of the two enantiomers — yielded, not a racemic solid, but crystals of the same chirality as the seed. Separations based on this observation have become known as 'resolutions by entrainment'[4] and are part of the armoury of the modern-day chemical-process developer.

Not surprisingly, it was chemical engineers, interested in designing continuous crystallization processes, for whom seeding (or secondary nucleation, as they termed it) became a central issue. In 1934, Ting and McCabe[5] showed that solutions of magnesium sulphate could be nucleated more reproducibly at moderate supersaturations in the presence of seeds. Today, commercial crystallization processes operate at suspension densities of perhaps 20%, ensuring that seeding levels are always high.

Some clever experiments[6] in the 1970s on the seeded nucleation of enantiomers of sodium chlorate revealed that, as long as the supersaturations were not too high, all the crystals were enantiomerically identical to the seed. The experiments also showed that the new crystals originated from the seeds through their contact with the crystallizing vessel. We now know that secondary nucleation, and hence seeding, can often be more effective because of mechanical and liquid-shear damage at the seed surface[7]. Such damage would remove potential nuclei from the seed, allowing them to become free-growing crystals. This seems to be a tantalizing reflection of what Cacciuto *et al.*[3] have now shown.

Time, I think, for some new experiments.

Roger J. Davey is at the Molecular Materials Centre, Department of Chemical Engineering, UMIST, Manchester M60 1QD, UK.
e-mail: roger.davey@umist.ac.uk

1. Ostwald, F. W. *Z. Phys. Chem.* **22**, 289–302 (1897).
2. Garside, J. & Davey, R. J. *From Molecules to Crystallizers — An Introduction to Crystallization* Ch. 3 (Oxford Univ. Press, 2000).
3. Cacciuto, A., Auer, A. & Frenkel, D. *Nature* **428**, 404–406 (2004).
4. Jacques, J., Collet, A. & Wilen, S. H. *Enantiomers, Racemates, and Resolutions* 223 (Wiley, Chichester, 1994).
5. Ting, H. H. & McCabe,W. L. *Ind. Eng. Chem.* **26**, 1201–1207 (1934).
6. Denk, E. G. & Botsaris, G. D. *J. Cryst. Growth* **13/14**, 493–499 (1972).
7. Mullin, J. W. *Crystallization* 3rd edn, Chs 5, 6 (Butterworth–Heinemann, Burlington, MA, 1992).

RNA interference

Human genes hit the big screen

Andrew Fraser

Genetic screens are powerful tools for identifying the genes involved in specific biological processes. At last, RNA interference makes large-scale screens possible in mammalian cells.

One of the most intuitive ways to learn how a complicated machine works is to take it apart piece by piece — a directed 'learning by breaking'. For biologists, teasing apart the machinery underlying the form and function of an organism can be done, most simply, by removing genes one at a time and looking at the effect. One experimental method for turning genes off is known as RNA interference (RNAi; Box 1, overleaf); this has shot to prominence because it allows almost any gene of known sequence to be shut down with apparently magical ease[1].

In two of the biologist's favourite model organisms, nematode worms and fruitflies, RNAi has been used to turn off almost every one of their genes[2,3]. Such genome-wide RNAi surveys of gene function have remained out of reach in mammals — until now, that is, for on pages 427 and 431 of this issue Paddison *et al.*[4] and Berns *et al.*[5] report the generation of tools to allow RNAi mass-screening of mammalian genes. This at last makes it possible to carry out genetic screens in mammalian cells in culture.

There is a range of effective strategies for RNAi in mammalian cells (reviewed in ref. 6), and they differ principally in the method for getting the double-stranded RNA (dsRNA) that specifically interferes with the target gene into the cells. In one method, rather than synthesize the dsRNA chemically before introducing it into the cells, the interfering dsRNA is made directly by the cells themselves. A vector directing the transcription of precise short hairpin RNAs — shRNAs — by RNA polymerase III is introduced into the cells; these transcribed shRNAs are processed by the cell to give the small interfering dsRNAs (siRNAs) that turn off the target gene. shRNA-expressing vectors allow for sustained RNAi in a wide range of cell lines (including embryonic stem cells, the subject of much current research). A complete library of shRNA-expressing vectors designed to target each and every gene in a mammalian genome would thus allow genome-wide RNAi-based genetic screens in cells in culture. Put simply, for any process that we are interested in (cell division, response to DNA damage and so on), with such an shRNA library we could screen every gene in the human genome and ask if it is involved.

Both groups[4,5] have converged on the same basic shRNA library approach, each generating a retrovirus-based library capable of targeting around a third of human genes; the genes were chosen for their potential roles in disease. Different shRNAs often interfere to differing extents with a target gene, so at least three shRNAs have been cloned for most genes. This multiple coverage not only provides an internal control, but may also allow comparison of both strong and weak 'knock-downs' of a specific gene in an analogous way to a classical genetic approach[7].

Berns *et al.*[5] used their library to search for genes that affect the function of *p53*, a tumour-suppressor gene that kills or 'arrests' cells with damaged DNA. They screened around 8,000 human genes to find those required for a *p53*-dependent arrest of cell proliferation and identified six genes, including *p53* itself. Further assays confirm that these genes — which include a histone acetyl transferase and a histone deacetylase, two key regulators of gene expression — do indeed play a role in *p53*-induced cell-cycle arrest and senescence. This ability to survey the gene functions of a full third of the human genome so rapidly is breathtaking, and the success of the subsequent assays underscores the quality of this approach.

The retrovirus-based vectors used by both groups are excellent for many cell-based screens. But they cannot be used for the stable expression of shRNAs in all cell types — this requires moving the shRNA-encoding inserts to different vectors. The shRNA library described by Paddison *et al.*[4] incorporates an elegant system for shuttling the inserts into any destination vector simply using bacterial mating. Their sequence-verified shRNA library targets almost 10,000 human genes; the shRNAs have also been chosen to allow targeting of the mouse orthologues (equivalents) of those human genes, if possible, and over 5,000 mouse

Box 1 RNA interference: a primer

RNA interference silences a target gene through the specific destruction of that gene's messenger RNA, the intermediary molecule between DNA and protein. Double-stranded RNA (dsRNA) is central to the technique: when dsRNA with identical sequences to a specific mRNA is introduced into cells, the mRNA is recognized and degraded by a multiprotein body called the RNA-induced silencing complex. Destruction of the target mRNA leads to a drop in the levels of its encoded protein, and thus to inhibition of the target gene.

In worms and flies, dsRNAs of hundreds of nucleotides can be used to target a gene. However, in mammalian cells long dsRNAs induce a potent anti-viral response, shutting down the synthesis of all proteins. So more sophisticated strategies are required, and small interfering RNAs (siRNAs) are used instead. These siRNAs are about 21 nucleotides long, and are efficiently used by the RNA-induced silencing complex but are too short to activate a full-blown anti-viral dsRNA response.

siRNAs can either be made *in vitro* and subsequently introduced into cells, or they can be made directly in cells through the expression of short hairpin RNAs (shRNAs). shRNAs fold back on themselves, creating a region of dsRNA and a loop. This hairpin is processed enzymatically to remove the loop and generate a mature siRNA. Expression of shRNAs can be used to induce RNAi in transgenic mice as well as in cell lines, so the technique can be applied to investigate gene function in whole animals. **A.F.**

genes can be targeted using their clones. This library of easily transferable shRNAs is a beautifully designed resource, and should permit an impressive range of analyses in diverse cell types.

To increase the speed of RNAi screening, both groups[4,5] borrow a sequence identifier (bar-code) system, developed in studies on yeast, for the quantitative analysis of pools of genes[8]. Each shRNA construct has a unique bar-code — Berns *et al.* use the shRNA sequence itself, whereas Paddison *et al.* have an independent bar-code, which they report as being of far greater effectiveness. The abundance of each shRNA construct in a pool of constructs can be assessed by monitoring the relative levels of each bar-code using a microarray. Thus any screen for genes that confer a growth advantage (or defect) can be carried out by the simultaneous screening of large pools of shRNA-expressing vectors, greatly increasing the throughput. Bar-coding is still in its infancy but has great potential for analysing RNAi selection screens.

There are still some uncertainties surrounding mammalian cell RNAi, especially regarding both specificity and efficiency of targeting. According to one report[9], a sequence identity of as few as 11–12 nucleotides between an interfering RNA and a messenger RNA may be sufficient for interference to occur. If it is, cross-reactivity is a substantial problem: far from targeting one gene, many expressed shRNAs may target several genes simultaneously. Similar analyses[10] came to the opposite conclusion, however, so it remains to be seen whether this is a general problem. Even if cross-reactivity does occur, there are straightforward controls for specificity: most simply, if two independent shRNAs targeting the same gene give similar effects, it is probably safe to conclude that this is specific to the targeted gene, and not due to some 'off-target' cross-reaction. This is precisely the approach adopted by Berns *et al.* and the presence in each of the libraries reported here of multiple shRNAs against each gene should make these internal controls relatively easy.

As regards RNAi targeting efficiency, it is clear that — as in worms or flies — different genes in mammalian cells are turned off with differing efficiencies. For example, Paddison *et al.* screened their library to identify components of the proteasome, a cellular machinery that degrades many unwanted proteins and that is implicated in certain diseases. Although genes encoding some subunits (those for the 19S base, for example) were apparently easily identified, others (such as those of the 19S lid or 20S core) were harder to hit. Like any screening tool, RNAi is unlikely ever to be perfect. As the rules for predicting effective shRNAs continue to improve, however, the false-negative rate will drop, and the libraries will improve.

Despite these notes of caution, we will no doubt see an explosion in RNAi screening of mammalian cells over the coming months. As with any genetic screen, the power of each RNAi screen depends on the appropriate choice of functional read-out, and that will require development of a variety of cell-based assays (such as the assay for proteasomal function reported by Paddison *et al.*). As no single laboratory can specialize in every aspect of gene function, the general availability of these shRNA libraries as communal resources is a major step forward, harnessing the screening expertise of the entire mammalian-cell research community. Pulling together the data from these varied RNAi screens in a common, central database will take our understanding of mammalian gene function a further giant stride forward. ■

Andrew Fraser is at the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,

100 YEARS AGO

It is not surprising to find that at last a 'motor' pocket book has appeared; in fact, it is a wonder such a work has not appeared sooner… Our author has a breezy style of expression which adds largely to the pleasure of reading the book. Take, for instance, his treatment of that all-important worry of the motorist, the 'police'. Mr O'Gorman says, "to pass unchallenged at a speed in excess of the legal limit — a thing which is daily accomplished by carts, hansoms, and even by the London omnibuses on almost every run when the gradients favour them… remember that by sitting upright with a calm face (on a quiet car) you produce no impression of speed except on turning a corner. If you turn a corner without being able to see down the road you are entering at over 20 miles per hour you deserve to be punished. If, however, you stoop forward… jamb your hat over your eyes, screw up your face, stare intently and anxiously, do a great deal of steering with visible swinging of your body, blow your horn in such a manner as to say 'Get out of my way' frequently, instead of pressing it slowly and peaceably, you will invariably be arrested."

From *Nature* 24 March 1904.

50 YEARS AGO

Another statement claimed by Prof. Dingle to be fallacious is connected with an underlying assumption in experimental science; this assumption is that the repetition of an experiment will reproduce the original results. But experimental science is not based on an assumption; "it is an adventure in which you accept whatever you find, and although you may be guided in a particular case by an expectation, the experiment may reveal something totally different". An instance of this is found in the case of Schwabe, who counted sunspots with the object of finding an intra-Mercurial planet, and instead of doing so he found the eleven-year solar period... it would be futile to believe that the achievements of experimental science would necessarily lose all significance if it were discovered that some assumption proved baseless. In the realm of psychology it is accepted that no experiment when repeated produces the original result, and even in physics it has been held for a long time that no experiment is repeatable, the entropy of the universe never being twice the same.

From *Nature* 27 March 1954.

Hinxton, Cambridge CB10 1SA, UK.
e-mail: agf@sanger.ac.uk

1. Fire, A. *et al. Nature* **391,** 806–811 (1998).
2. Kamath, R. S. *et al. Nature* **421,** 231–237 (2003).
3. Kiger, A. *et al. J. Biol.* **2,** 27 (2003).
4. Paddison, P. J. *et al. Nature* **428,** 427–431 (2004).
5. Berns, K. *et al. Nature* **428,** 431–437 (2004).
6. Dykxhoorn, D. M., Novina, C. D. & Sharp, P. A. *Nature Rev. Mol. Cell Biol.* **4,** 457–467 (2003).
7. Hemann, M. T. *et al. Nature Genet.* **33,** 396–400 (2003).
8. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. *Nature Genet.* **14,** 450–456 (1996).
9. Jackson, A. L. *et al. Nature Biotechnol.* **21,** 635–637 (2003).
10. Chi, J. T. *et al. Proc. Natl Acad. Sci. USA* **100,** 6343–6346 (2003).

Learning theory

Past performance and future results

Carlo Tomasi

Learning from experience is hard, and predicting how well what we have learned will serve us in the future is even harder. The most useful lessons turn out to be those that are insensitive to small changes in our experience.

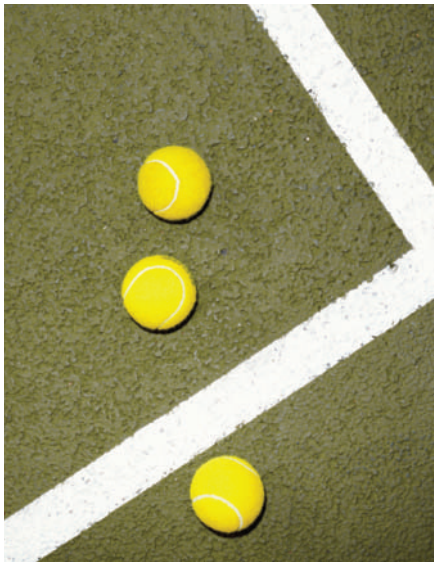
A hallmark of intelligent learning is that we can apply what we have learned to new situations. In the mathematical theory of learning, this ability is called generalization. On page 419 of this issue[1], Poggio *et al.* formulate an elegant condition for a learning system to generalize well.

As an illustration, consider practising how to hit a tennis ball. We see the trajectory of the incoming ball, and we react with complex motions of our bodies. Sometimes we hit the ball with the racket's sweet spot and send it where we want; sometimes we do less well. In the theory of supervised learning, an input–output pair exemplified by a trajectory and the corresponding reaction is called a training sample. A learning algorithm observes many training samples and computes a function that maps inputs to outputs. The learned function generalizes well if it does about as well on new inputs as on the old ones: if this is true, our performance during tennis practice is a reliable indication of how well we will play during the game.

Given an appropriate measure for the 'cost' of a poor hit, the algorithm could choose the least expensive function over the set of training samples, an approach to learning called empirical risk minimization. A classical result[2] in learning theory shows that the functions learned through empirical risk minimization generalize well only if the 'hypothesis space' from which they are chosen is simple enough. That there may be trouble in a poor choice of hypotheses is a familiar concept in most scientific disciplines. For instance, a high-degree polynomial fitted to a set of data points can swing wildly between them, and these swings decrease our confidence in the ability of the polynomial to make correct predictions about function values between available data points. For similar reasons, we have come to trust Kepler's simple description of the elliptical motion of heavenly bodies more than the elaborate system of deferents, epicycles and equants of Ptolemy's *Almagest*, no matter how well the latter fit the observations.

The classical definition of a 'simple enough' hypothesis space is brilliant but technically involved. For instance, the set of linear functions defined on the plane has a complexity (or Vapnik–Chervonenkis dimension[2]) of three because this is the greatest number of points that can be arranged on the plane so that suitable linear functions assume any desired combination of signs (positive or negative) when evaluated at the points. This definition is a mouthful already for this simple case. Although this approach has generated powerful learning algorithms[2], the complexity of hypothesis spaces for many realistic scenarios quickly becomes too hard to measure with this yardstick. In addition, not all learning problems can be formulated through empirical risk minimization, so classical results might not apply.

Poggio *et al.*[1] propose an elegant solution to these difficulties that builds on earlier intuitions[3–5] and shifts attention away from the hypothesis space. Instead, they require the learning algorithm to be stable if it is to produce functions that generalize well. In a nutshell, an algorithm is stable if the removal of any one training sample from any large set of samples results almost always in a small change in the learned function. *Post facto*, this makes intuitive sense: if removing one sample has little consequence (stability), then adding a new one should cause little surprise (generalization). For example, we expect that adding or removing an observation in Kepler's catalogue will usually not perturb his laws of planetary motion substantially.



FREDERIKE HELWIG/GETTY IMAGES

In or out: success rests on learning algorithms that are stable against slight changes in input conditions[1].

The simplicity and generality of the stability criterion promises practical utility. For example, neuronal synapses in the brain may have to adapt (learn) with little or no memory of past training samples. In these cases, empirical risk minimization does not help, because computing the empirical risk requires access to all past inputs and outputs. In contrast, stability is a natural criterion to use in this context, because it implies predictable behaviour. In addition, stability could conceivably lead to a so-called online algorithm — that is, one that improves its output as new data become available.

Of course, stability is not the whole story, just as being able to predict our tennis performance does not mean that we will play well. If after practice we play as well as the best game contemplated in our hypothesis space, then our learning algorithm is said to be consistent. Poggio *et al.*[1] show that stability is equivalent to consistency for empirical risk minimization, whereas for other learning approaches stability only ensures good generalization. Even so, stability can become a practically important learning tool, as long as some key challenges are met. Specifically, Poggio *et al.*[1] define stability in asymptotic form, by requiring certain limits to vanish as the size of the training set becomes large. In addition, they require this to be the case for all possible probabilistic distributions of the training samples. True applicability to real situations will depend on how well these results can be rephrased for finite set sizes. In other words, can useful measures of stability and generalization be estimated from finite training samples? And is it feasible to develop statistical confidence tests for them? A new, exciting research direction has been opened. ■

Carlo Tomasi is in the Department of Computer Science, Duke University, Durham, North Carolina 27708, USA.
e-mail: tomasi@cs.duke.edu

1. Poggio, T., Rifkin, R., Mukherjee, S. & Niyogi, P. *Nature* **428,** 419–422 (2004).
2. Vapnik, V. N. *Statistical Learning Theory* (Wiley, New York, 1998).
3. Devroye, L. & Wagner, T. *IEEE Trans. Information Theory* **25,** 601–604 (1979).
4. Bousquet, O. & Elisseeff, A. J. *Machine Learning Res.* **2,** 499–526 (2002).
5. Kutin, S. & Niyogi, P. in *Proc. 18th Conf. Uncertainty in Artificial Intelligence, Edmonton, Canada,* 275–282 (Morgan Kaufmann, San Francisco, 2002).